Statistica Sinica **27** (2017), 1485-1594
doi:http://dx.doi.org/10.5705/ss.2014.067

DISSECTING MULTIPLE IMPUTATION FROM A MULTI-PHASE INFERENCE PERSPECTIVE: WHAT HAPPENS WHEN GOD'S, IMPUTER'S AND ANALYST'S MODELS ARE UNCONGENIAL?

Xianchao Xie and Xiao-Li Meng

Harvard University

Abstract: Real-life data are almost never really *real.* By the time the data arrive at an investigator's desk or disk, the raw data, however defined, have most likely gone through at least one "cleaning" process, such as standardization, re-calibration, imputation, or de-sensitization. Dealing with such a reality scientifically requires a more holistic multi-phase perspective than is permitted by the usual framework of "God's model versus my model." This article provides an in-depth look, from this broader perspective, into multiple-imputation (MI) inference (Rubin (1987)) under uncongeniality (Meng (1994)). We present a general estimating-equation decomposition theorem, resulting in an analytic (asymptotic) description of MI inference as an integration of the knowledge of the imputer and the analyst, and establish a characterization of self-efficiency (Meng (1994)) for regulating estimation procedures. These results help to reveal *how* the quality of and relationship between the imputer's model and analyst's procedure affect MI inference, including how a seemingly perfect procedure under the "God-versus-me" paradigm is actually inadmissible when God's, imputer's, and analyst's models are uncongenial to each other. Our theoretical investigation also leads to useful procedures that are as trivially implementable as Rubin's combining rules, yet with confidence coverage guaranteed to be minimally the nominal level, under any degree of uncongeniality. We reveal that the relationship is very complex between the validity of approaches taken for individual phases and the validity of the final multi-phase inference, and indeed that it is a nontrivial matter to quantify or even qualify the meaning of *validity* itself in such settings. These results and many open problems are presented to raise the general awareness that the multi-phase inference paradigm is an uncongenial forest populated by thorns, as well as some fruits, many of which are still low-hanging.

Key words and phrases: Confidence validity, data cleaning, estimating equation decomposition, incomplete data, multi-phase inference, pre-processing, self-efficiency, strong efficiency, uncongeniality.

1. Multi-phase Inference: An Expanded Paradigm

With the dramatic increases in the size, diversity, and complexity of data available for scientific discoveries, medical advances, education reforms and

evidence-based policy making, to name a few, the entire enterprise of scientific quantitative inquiry has been presented with unprecedented challenges and opportunities. In particular, many current quantitative inquiries are not made by one or even two teams (i.e., data collectors and analysts), but rather by multiple teams/parties entering the process over several phases, such as data collection, processing, curation, and analysis. Due to constraints such as resource limitations and confidentiality, each team may not have adequate knowledge of or are unable to utilize the assumptions made by, and resources available to, those coming before or after its phase. Even in cases where several different phases involve a single team, the complexity of the data and different purposes of these phases often encourage or even force the team to adopt incompatible setups and assumptions across different phases.

This was the case for example in the area of estimating survival distributions using data from AIDS surveillance systems, which suffered from substantial reporting delay (see Tu, Meng, and Pagano (1993); Bouman, Dukic, and Meng (2005)). In theory, an encompassing model can be set up to estimate the survival distribution and reporting delay distribution simultaneously. But practically, the complexity of the data and the need to perform standard Cox regression for the survival distributions, compelled us to adopt a two-phase strategy. We first adopted a multiple imputation approach via a Bayesian modelling to impute the delayed cases, and then we performed the Cox regression using the observed *and* imputed cases. Neither the models nor the inference perspectives in these two phases were constrained to be the same, even though there was only one team involved, as documented in Tu, Meng, and Pagano (1993).

Therefore, by *multi-phase inference,* we mean an inference process where the ultimate conclusions are a result of several phases carried out *in a sequential order, each phase with its own goal, assumptions, and methods, not necessarily compatible across different phases.* The emphasis here is on the sequential nature of how the inferential conclusions are built upon a sequence of not necessarily (theoretically) compatible phases. As such inference processes become increasingly common in this age of Big Data, they compel us to rethink the traditional paradigms for statistical analysis and data preservation (see Blocker and Meng (2013)). From this expanded perspective, this article takes a critical look at Rubin's (1987) multiple imputation (MI) inference under uncongeniality, as formulated in Meng (1994). MI inference explicitly acknowledges the impact of an intermediate phase, namely the imputation phase, and hence it demonstrates nicely the necessity, intricacy, and opportunities of theoretical investigation of multi-phase inference.

1.1. Why is an expanded paradigm necessary?

Much of the statistical literature assumes data are generated by a "God's model", and then postulates one or multiple model classes for the purpose of inferring aspects of God's model and its implications. Here by "God's model", we mean the true status of nature, not a posited model class. In contrast, by *multiple model classes* we mean multiple sets of assumptions (hence, not necessarily parametric models), which are mathematically compatible within each set but incompatible across different sets—if they are not mathematically incompatible then we are merely postulating a larger model class. Whereas many model classes may be entertained, the commonly accepted paradigm explicitly identifies "my model", the ones used by an analyst, to approximate the (unknown) God's model. However, reality is far more complicated, compelling us to distinguish between *analyst's data* from *God's data*, the realizations from God's model that the analyst's data were collected to *approximate*. Any attempt to mathematically define such concepts is doomed to fail, but it is important to distinguish the two forms of data because the *approximation* process introduces an additional inference phase or even phases.

For example, in physical and biological sciences, the analyst's data typically are results of a series of pre-processing steps to deal with limitations or irregularities in recording God's data (e.g., discarding "outliers", re-calibration), yet typically the analyst at best has only partial information about this process. For social and behavioral sciences, many variables are "constructed variables", typically from a deterministic algorithm converting a set of questionnaire responses to an index, say, that whether a subject has depression. The algorithm is often a pitch black box because the analyst is unaware of what variables were used to produce the index. For large-scale public-use data files, virtually all data sets contain imputations because of non-responses or other forms of missing data, which means someone has "fixed the holes" in the data before releasing.

In all these examples, the key issue is that during the journey from God's data to the analyst's data, a set of assumptions have been introduced deliberately or accidentally. There is no "assumption-free" pre-processing; any attempt to make the data "better" or "more usable" implies that a judgement has been made. Under the God-versus-me paradigm, this intermediate "data cleaning" process has to be considered either as part of God's model, or of the analyst's model class, or of both, by somehow separating aspects of the process. Regardless of how we conceptualize, we are in an extremely muddy situation. If aspects of this intermediate process are considered to be part of God's model, then the analyst's inference is not merely about God's model but also about someone else's assumptions about it. If we relegate the pre-processing to the analyst's model class, then the analyst will need good information about the process.

Whereas understanding the entire data forming mechanism is a critical on-going emphasis of our profession, the reality is that for the vast majority of real-life data sets, especially large-scale ones, it is simply impossible to trace how the data were collected or pre-processed. Indeed, many such processes are nowhere documented, and some are even protected by confidentiality constraints.

Such "data cleaning" pre-processes motivate the *multi-phase inference* paradigm. The key distinction between the multi-phase paradigm and the God-versus-me paradigm is not that the former involves multiple model classes or even multiple investigating teams. Rather, in the multi-phase paradigm, we explicitly acknowledge the *sequential nature* of the phases and that the assumptions made at different phases are permitted to be *different* or even *contradictory*. The key aims here are (I) to understand the consequences of permitting such discrepancy and contradictions, and (II) to develop methods with theoretical validity and even optimality for the ultimate inference results in such complex but realistic settings.

1.2. Multiple imputation under uncongenality

A great example of multi-phase inference is the MI inference (Rubin (1987)) under *uncongeniality* (Meng (1994)). In a nutshell, uncongeniality means that the imputation model class and the analyst's model class are incompatible. There are many reasons for such incompatibility, including different aims of imputation (where one wants to use as many variables as possible even if causal directions are incorrectly specified) and of analysis (where one may be interested only in a subset of variables with specified causal directions). In this paper we assume that the imputer's model class is (approximately) valid, which means that it has properly taken into account the missing data mechanism (MDM), regardless of whether it is ignorable such as missing at random (MAR) or non-ignorable (Rubin (1976)). Indeed, the more sophisticated handling by the imputer is often a source of uncongenality because subsequent users of the imputed data do not possess the necessary knowledge or resource to properly handle the missing data themselves.

Rubin's MI inference, a general approach for addressing serious defects of single imputation, was originally justified from the Bayesian perspective (Rubin (1987)) under the (implicit) assumption of congeniality. Specifically, let $Z_{\rm com}$ denote the complete data, $Z_{\rm obs}$ and $Z_{\rm mis}$ respectively the observed and missing data, and θ the analyst's estimand (e.g., a population quantity). The imputer uses a Bayesian method (or its equivalent) to fill in the missing values by drawing independently m times from the predictive distribution $p^I(\tilde{Z}_{\rm mis}|Z_{\rm obs})$. This produces m complete data sets $\tilde{Z}^{(\ell)}_{\rm com} = (Z_{\rm obs}, \tilde{Z}^{(\ell)}_{\rm mis}), \ell = 1, \ldots, m$ available for any potential analysis. The superscript I signifies that the imputation is done under

the *Imputer's model class*. In the analyzing phase, the analyst applies a chosen complete-data procedure $\mathcal{P}_c = [\hat{\theta}^A(Z_{\rm com}), U^A(Z_{\rm com})]$, where $\hat{\theta}^A(Z_{\rm com})$ is the analyst's point estimator of his/her estimand θ and $U^A(Z_{\rm com})$ is its associated variance estimator, to each $\tilde{Z}_{\rm com}^{(\ell)}$ to produce $\hat{\theta}^{(\ell)} \equiv \hat{\theta}^A(\tilde{Z}_{\rm com}^{(\ell)})$ and $U^{(\ell)} \equiv U^A(\tilde{Z}_{\rm com}^{(\ell)})$, for $\ell = 1, \ldots, m$.

Rubin's MI inference refers to the use of Rubin's (1987) rules that estimate θ by the average of the m individual estimators and its variance as the sum of two terms:

$$\bar{\theta}_m = \frac{1}{m}\sum_{\ell=1}^{m} \hat{\theta}^{(\ell)} \quad \text{and} \quad T_m = \bar{U}_m + \left(1 + \frac{1}{m}\right) B_m \ , \tag{1.1}$$

where $\bar{U}_m$ and B_m are the estimated within- and between-imputation variances

$$\bar{U}_m = \frac{1}{m}\sum_{\ell=1}^{m} U^{(\ell)} \quad \text{and} \quad B_m = \frac{1}{m-1}\sum_{\ell=1}^{m} (\hat{\theta}^{(\ell)} - \bar{\theta}_m)(\hat{\theta}^{(\ell)} - \bar{\theta}_m)^\top.$$

The factor $(1 + 1/m)$ in (1.1) is due to the finite number of imputations. For our theoretical investigation later, we will focus on $m = \infty$, in order to study the performance of Rubin's MI inference in the absence of Monte Carlo errors. In a nutshell, Rubin's MI is simply a size m Monte Carlo simulation from $p^I(\tilde{Z}_{\rm mis}|Z_{\rm obs})$, with the ultimate *estimands*

$$\bar{\theta}_\infty = \lim_{m\to\infty} \bar{\theta}_m, \quad \text{and} \quad T_\infty = \bar{U}_\infty + B_\infty : \tag{1.2}$$

$$\bar{U}_\infty = \lim_{m\to\infty} \bar{U}_m \ = \ E^I[U^A(\tilde{Z}_{\rm com})|Z_{\rm obs}]; \tag{1.3}$$

$$B_\infty = \lim_{m\to\infty} B_m \ = \ V^I[\hat{\theta}^A(\tilde{Z}_{\rm com})|Z_{\rm obs}]. \tag{1.4}$$

Justifying Rubin's combining rules is easy under congeniality as formulated by Meng (1994):

(I) The complete-data analysis procedure can be embedded into a Bayesian model class with

$$\hat{\theta}^A(\tilde{Z}_{\rm com}) = E^A(\theta|\tilde{Z}_{\rm com}) \text{ and } U^A(\tilde{Z}_{\rm com}) = V^A(\theta|\tilde{Z}_{\rm com}), \text{ for all } \tilde{Z}_{\rm com}, \tag{1.5}$$

where the superscript A indexes expectation with respect to the analyst's model class;

(II) The imputation model class and the analysis (embedding) model class are the same for the purposes of predicting missing data:

$$P^I(\tilde{Z}_{\rm mis}|Z_{\rm obs}) = P^A(\tilde{Z}_{\rm mis}|Z_{\rm obs}), \quad \text{for all } \tilde{Z}_{\rm mis} \text{ (but the given } Z_{\rm obs}). \tag{1.6}$$

Under (I) and (II), the MI inference $\mathcal{P}_\infty = [\bar{\theta}_\infty, T_\infty]$ is the same as the posterior mean and variance of θ under the analyst's (embedded) model class given Z_{obs}. That is, $[\bar{\theta}_\infty, T_\infty] \equiv \mathcal{P}_\infty = \mathcal{P}_{\text{obs}} \equiv [E^A(\theta|Z_{\text{obs}}), V^A(\theta|Z_{\text{obs}})]$, a fact that can be verified by using iterative expectations. That is, under congeniality, Rubin's MI can be viewed as performing Monte Carlo integration for the analyst to obtain $\mathcal{P}_{\text{obs}}$, without any knowledge of it, using only the complete-data procedure $\mathcal{P}_c$.

When the imputation model class and the (embedded) analyst's model class differ, the behavior of Rubin's rules becomes very complicated, capable of producing inconsistent variance estimators, a matter that has received recurrent criticisms (Fay (1991, 1992); Kott (1995); Nielsen (2003)). To address such criticisms, Meng (1994) identified the concept uncongeniality as the key ingredient for studying the complex behavior of Rubin's MI inference. It is worth emphasizing that the uncongenality as defined by (I) and (II) is a form of *estimation uncongenality* or more generally *inferential uncongenality*, because (I) is determined by a particular estimation procedure. For example, suppose both imputer and analyst adopt the same $N(\theta, 1)$ model (and the imputer adopts a constant prior on θ), yet the analyst decides to estimate θ by both the sample mean and the sample median, the latter being an attempt to robustify. Then the imputer's model is congenial to analyst's sample-mean procedure, but not to the sample-median procedure, because for the latter the embedding (sampling) model class is not $N(\theta, 1)$, but rather (say) Laplace $L(\theta, 1)$.

Meng (1994) obtained some initial theory under this inferential uncongeniality, including conditions for Rubin's MI inference to be confidence valid, i.e., the interval estimator has at least the nominal coverage. In particular, this theory indicates that the bias in Rubin's variance estimator is caused by a lack of orthogonality in an ANOVA-type decomposition under uncongeniality, confirming and explaining counterexamples in several previous studies (Fay (1992); Kott (1995)). Consequently, several proposals (e.g., Robins and Wang (2000); Kim et al. (2006)) were made to correct the bias, assuming that the imputer provides information *beyond the imputed values*.

Here we revisit these issues in light of a series of theoretical results discovered through our study under the multi-phase inference paradigm, and the results here extend Meng's (1994) both in scope (e.g., covering multi-dimensions) and in depth (e.g., showing how MI estimators integrate the imputer's and analyst's knowledge). Section 2 illustrates the complexities of multi-phase inference and summarizes our major findings. Section 3 presents a general decomposition of an estimator resulting from a decomposable estimating equation. The result then is applied in Section 4 to Rubin's MI point estimator to arrive at a matrix-weighted representation, a result which in turn is applied in Section 5 for variance calculations and for deriving an exceedingly simlple standard-error combining rule,

as well as a variance doubling rule, for confidence validity under uncongeniality. Revealing a hidden implication of Rubin's variance combining rule that makes the assumption of self-efficiency important, Section 6 presents a general result about when an estimating equation is self-efficient. Armed with these results, Section 7 then characterizes Rubin's variance combining rules under nested models. Section 8 explores issues such as measuring the degree of uncongeniality and possible cancellation of errors from different phases, issues that are unique to the multi-phase paradigm and need to be addressed before its fruitful foundation can be established in general.

1.3. But what is *Validity* in multi-phase inference and for *whom*?

We have already invoked the phrase "valid" several times, but as a reviewer rightly pointed out, its meaning requires careful qualification and quantification. Since our ultimate goal is to infer aspects of God's model, *validity* apparently must mean that our inferential conclusions should be consistent with God's specifications if we had an unlimited amount of data. But such *consistency* requirement is only one part of *statistical validity*, which also regulates uncertainty assessments in our inferential statements. These uncertainty assessments, such as confidence coverage and hypothesis testing errors, give rise to more ambiguity in the multi-phase setting than in the familiar God-vs-me paradigm, as revealed in the MI setting.

Whereas there are many inferential perspectives (Bayesian, frequentist, likelihood, fiducial, etc; see Liu and Meng (2016)), there is essentially only one scientific way to evaluate and compare statistical procedures — show how they work when applied repeatedly, in reality, via simulation or in thought experiments. Therefore, assessing uncertainty or more generally quality of any statistical procedure is inherently a frequentist endeavor, even for Bayesians (see Rubin (1984)). The key question then is over what replications we should evaluate our procedures.

Under the familiar God-vs-me paradigm, a well accepted replication scheme is for *me* to imagine that God uses the same process $G(D)$ that generated *my* data, denoted by D_0, to produce many more identically distributed *data sets*, either independently, or conditionally (on some ancillary feature of D_0 for example) independently, of D_0. Denoting these hypothetical date sets by $\{D_i, i \in \mathcal{I}\}$ and the procedure under evaluation by $\mathcal{P}_{ro}$, I can compute whatever operating characteristics of $\mathcal{P}_{ro}$ that are of interest (e.g., variance, coverage, Type I errors) by statistically summarizing $\{\mathcal{P}_{ro}(D_i), i \in \mathcal{I}\}$; or theoretically, I can calculate the property of $\mathcal{P}_{ro}(D)$ over the God's model/process $G(D)$ I perceive. How to perceive a relevant God process $G(D)$ for my particular data set D_0 is nontrivial and indeed is at the heart of any statistical inference, as argued in Liu and Meng

(2016). Nevertheless, there is only one God process I need to contemplate and only *I*, wearing the hat as an analyst, need to contemplate.

In a multi-phase setting, a complication arises because either there will be multiple "God" processes I need to contemplate, or there will be multiple "I"s doing the contemplation, or both. For example, in the MI setting, in addition to *me* as an analyst, there will be the imputer (which could just be *me* but wearing a different hat). From *me* the analyst's perspective, my interest is not that different from *me* as in the God-vs-me paradigm — I want to ensure my inference is valid in the same sense as before, with the complication that I now need to include the imputation process as a part of the God's model; or more precisely to treat the imputer as a Demigod, and add the Demigod process to the God process to form a Super God process. Using the notation of Section 1.2, we can express this Super God's process as $G(Z_{\text{obs}}) \prod_{\ell=1}^{m} P^I(\tilde{Z}_{\text{mis}}^{(\ell)}|Z_{\text{obs}})$, yielding *my* data $D_0 = \{Z_{\text{obs}}, \tilde{Z}_{\text{mis}}^{(1)}, \cdots, \tilde{Z}_{\text{mis}}^{(m)}\}$.

But from the imputer's perspective, especially those who are responsible for producing public-use data files for many analysts, the validity is no longer about a particular *me*, but to ensure that the imputation quality is such that as many subsequent analysts will be able to reach the validity they care about without having to worry (too much) about how the imputation was done. This is no different from the perspective of a data collector for public-use data files. Indeed, for large-scale public-use data files, such as those put out by the US Census Bureau, it is beyond virtually any analysis team's capacity to assess the data quality, be observed or imputed. This effectively means that subsequent analysts would have to invoke the aforementioned "Super God" perspective, however subconsciously or involuntarily, in their contemplations of relevant replications for assessing uncertainty and validity. Although they typically have little interest in inferring any aspect of the Demigod model, they need to treat the imputation process as a "nuisance process", affecting the quality of the inference they care about.

Therefore, a key consideration of validity from the imputer's perspective must be to ensure that the imputation process is as small a nuisance as possible to as many subsequent analyses as possible. This means that to assume the imputation model as precisely consistent with God's specific model for producing Z_{obs} *as the imputer believes* can actually be very harmful. This is because the more precise a specification (e.g., setting a regression coefficient to zero), the fewer analysts would include it in their contemplations. Therefore, a more restrictive imputation model will typically do more harm to more subsequent analyses, especially considering the analysts have essentially zero chance to correct the problem or even to suspect that there is a problem.

It is therefore well understood from the early days of the debate on MI inference, as documented in Meng (1994), that the imputation model class should be as saturated as practically feasible, when it is compared with the analysts' (embedding) model classes. The theoretical results in this paper further support this general advice, which is applicable even when *the same team* carries out both the imputation and analysis phases. This is because the need for separating the two phases in the first place typically means that the team has faced practical or theoretical constraints that have compelled it to take care of the missing data problem before performing any desired analyses. Consequently, it is in the team's interest to not unduly tie its own hands by using an overly restrictive imputation model, so it can perform more subsequent analyses at the analysis phase without having to regret or even redo the imputation.

However, there has been virtually no study of how to quantify or even qualify the types of analysis to be conducted on any public-data file, nor is it clear how to meaningfully conduct such studies given the on-going methodological evolution after the release of any specific data set. Therefore, since a primary goal of this paper is to investigate theoretically the consequences of having uncongenial models via the perspective of multi-phase inference, we must put the imputation model and analysis procedure on an equal footing in order to study how their relationships and interactions would influence the final MI inference. We do so by investigating how a single class of imputation models interacts with a single class of analysis models, as an effective building block for understanding the interactions between multiple classes of imputation and analysis models. Consequently, the notion of validity in this article is with respect to the original God's process that creates what we observe, i.e., $G(Z_{\text{obs}})$, where G encompasses the entire process of creating Z_{obs}, including God's missing-data mechanism. The issue of imputation uncertainty due to a finite m disappears in our theoretical results, because we assume $m = \infty$ to separate the complication due to uncongenality from Monte-Carlo errors because of a finite m.

Furthermore, the central controversy about MI has been the possibility of an invalid inference, under God's $G(Z_{\text{obs}})$, when both the imputation and analysis model classes are *correctly* specified. We therefore restrict ourselves to such a setting as well, by assuming both classes contain God's model as a special case. The cases where one or both model classes are misspecified are of greater practical interest, just as in reality essentially all model classes are misspecified. Nevertheless, theoretical insights are typically developed by first studying what can go wrong under controlled environments. As we reveal below, even within our restrictive environments, the findings are substantially more intricate than previously anticipated. Such intricacies seem to be the rule rather than exception in multi-phase inference (see Blocker and Meng (2013)), and we hope they can

entice those with adventurous spirit to explore with us this essentially virgin forest of multi-phase inference.

2. Summarizing and Illustrating Key Findings

Uncongenial models complicate our life. To help readers to decide if they want explore such a complicated life style, here we use three simple but informative examples to illustrate general guidelines, with precise theoretical conditions and statements given in Section 7.

- The validity of both the imputer's and analyst's model/procedure *does not* guarantee the validity of the resulting MI inference, especially when the analyst's procedure is completely unregulated (e.g., when it is not self-efficient);
- When the imputer's model is more saturated than the model underlying the analyst's procedure and the analyst's procedure is self-efficient, Rubin's rules are confidence proper (i.e., with the correct coverage) and possess good robustness properties;
- When the imputer's model is less saturated than the analyst's (embedding) model and the analyst's procedure is self-efficient, Rubin's rules achieve super efficiency when the fractions of missing information for all parameters are the same; otherwise confidence validity (i.e., with at least the nominal coverage) is not guaranteed;
- In general, uncongenality should be regarded as the rule rather than the exception, and a simple confidence valid procedure to combat any degree of uncongenality is to double Rubin's MI variance estimate. For univariate estimand, a less conservative but still confidence valid approach is to apply Rubin's additive combining rule in terms the standard errors instead of variances — to form the MI standard error as the sum of the with-imputation standard error and the between-imputation standard error.

A key reason that the validity of individual models does not necessarily guarantee the overall validity is that "a valid model" here really means "a valid model class". They all share the completely specified God's model (including MDM) as a special case, but can differ in any other aspects, including having different parameter spaces. Consequently, we use a superscript to denote whether a model parameter θ, as a generic notation, comes from the analyst's model (θ^A) or from the imputer's model (θ^I). Although from this point on we adopt the common practice of not distinguishing between "models" (completely known) and "model classes" (which contains unknowns), our examples illustrate the complications caused by the flexibility of a valid *model class*. These analytical tractable examples demonstrate well the intricate nature of dealing with even the simplest

two-phase inference with *two uncongenial model classes* (imputation and analysis), and with God's model being nested within each. Real-life situations are far more complex than these stylistic examples, as the regression example given in the on-line supplement demonstrates (also see Section 8.3), even though the regression example itself is on the simplistic side.

2.1. Example 1: Illustrating uncongeniality and self-efficiency

In the first example, we assume that the complete data are N independent normal observations with mean θ_0, which is *arbitrary* but fixed. However, the first n ($< N$) observed ones have variance 1 and the remaining $N - n$ missing ones have variance σ_0^2. The imputer's model correctly captures this unequal variance setting, but it is unknown to the analyst. Nevertheless, the analyst's complete-data procedure is still valid. Specifically, we assume

- *God's Model*: $Z_{\text{obs}} = (Y_1, \ldots, Y_n)$ with $Y_i \overset{i.i.d.}{\sim} N(\theta_0, 1)$ for $i = 1, \ldots, n$ and $Z_{\text{mis}} = (Y_{n+1}, \ldots, Y_N)$ with $Y_i \overset{i.i.d.}{\sim} N(\theta_0, \sigma_0^2)$ for $i = n+1, \ldots, N$.
- *Imputer's Model*: $Y_i \overset{i.i.d.}{\sim} N(\theta, 1)$ for $i = 1, \ldots, n$ and $Y_i \overset{i.i.d.}{\sim} N(\theta, \sigma_0^2)$ for $i = n+1, \ldots, N$; prior $p(\theta) \propto 1$; imputed values are obtained as posterior predictive draws by sampling

$$\tilde{\theta}|Z_{\text{obs}} \sim N(\bar{Y}_n, n^{-1}) \quad \text{and then} \quad \tilde{Y}_i|\tilde{\theta} \overset{i.i.d.}{\sim} N(\tilde{\theta},\ \sigma_0^2), \text{ for } i = n+1, \ldots, N,$$

  where $\bar{Y}_n$ is the average of the observed sample: $Z_{\text{obs}} = (Y_1, \ldots, Y_n)$.
- *Analyst's Complete-data Procedure:*

$$\hat{\theta}^A_{\text{com}} = \bar{Y}_N, \quad \hat{V}^A_{\text{com}} = \hat{V}(\hat{\theta}^A_{\text{com}}) = \frac{1}{N} S_N^2,$$

  where $\bar{Y}_N$ and S_N^2 are the sample mean and sample variance of $Z_{\text{com}} = \{Y_1, \ldots, Y_N\}$.

To simplify the algebra, we replace $\hat{V}^A_{\text{com}}$ by its asymptotic equivalent

$$\bar{U}_\infty = \frac{1}{N}[(1-f) + f\sigma_0^2], \quad \text{where } f = \frac{N-n}{N}, \tag{2.1}$$

because $\hat{V}^A_{\text{com}} - \bar{U}_\infty = O_p(N^{-3/2})$, assuming the fraction of missing data $f = 1 - n/N$ is bounded away from 1 as $N \to \infty$. Thus, for our asymptotic comparisons where we ignore anything of $o_p(N^{-1})$ order, we can assume $\mathcal{P}_c = [\bar{Y}_N, \bar{U}_\infty]$ with $\bar{U}_\infty$ given by (2.1) even though the analyst is unaware of the heteroscedasticity in the observations.

Given the above setup, it is straightforward to verify that $\bar{\theta}_\infty = \bar{Y}_n$ and

$$T_\infty = \bar{U}_\infty + B_\infty = \frac{1}{N}\left[(1-f) + f\sigma_0^2\right] + \frac{f}{N}\left[\frac{f}{1-f} + \sigma_0^2\right]. \tag{2.2}$$

Clearly, T_∞ differs in general from the sampling variance of $\bar{\theta}_\infty$, $V_\infty \equiv V(\bar{Y}_n) = 1/n$, resulting in an *asymptotic bias*

$$\Delta_n \equiv n(T_\infty - V_\infty) = 2f(1-f)(\sigma_0^2 - 1). \tag{2.3}$$

Other than the trivial case when $f = 0$, we see that Δ_n is zero if and only if $\sigma_0^2 = 1$ (we have assumed $f < 1$), namely, if and only if $\mathcal{P}_c$ is congenial to the imputer's model.

When $\sigma_0^2 \neq 1$, the bias in T_∞ can be either positive or negative. It also illustrates something more subtle. Because $\bar{\theta}_\infty$ is the same as analyst's procedure applied to Z_{obs}, i.e., $\bar{\theta}_\infty = \bar{Y}_n = \hat{\theta}^A_{\text{obs}}$, if Rubin's rule $T_\infty = \bar{U}_\infty + B_\infty$ *were* to provide the correct variance for $\bar{\theta}_\infty = \bar{Y}_n$, it would imply that (asymptotically under God's model)

$$V(\hat{\theta}^A_{\text{obs}}) = V(\bar{Y}_n) = T_\infty = \bar{U}_\infty + B_\infty \geq \bar{U}_\infty = V(\bar{Y}_N) = V(\hat{\theta}^A_{\text{com}}). \tag{2.4}$$

One might take (2.4) for granted, since $\bar{Y}_N$ is based on more observations than $\bar{Y}_n$. But (2.4) is true in general only when $\{Y_1, \ldots, Y_N\}$ are *exchangeable*, which clearly is not the case under heteroscedasticity. Indeed, $V(\bar{Y}_n) = 1/n$ and $V(\bar{Y}_N) = (1 - f + f\sigma_0^2)/N$. Hence, the inequality (2.4) holds if and only if $\sigma_0^2 \leq 1 + (1-f)^{-1}$ (assuming $f > 0$). If σ_0^2 is too big, then the analyst would *obtain a more efficient estimator with fewer observations*.

This seemingly paradoxical phenomenon arises because the analyst's procedure, which gives all observations equal weight, is optimal only when all observations *deserve* to be weighted equally. Otherwise, by giving those observations with large variabilities more weights than they deserve, we actually can hurt ourselves with more observations because their large variabilities outweigh the gain in sample size; see Meng and Xie (2014) for a general discussion of such phenomenon. This implies that for Rubin's variance combining rule to hold, we minimally need to impose (2.4). Actually, as shown in Meng (1994), to avoid this "paradoxical phenomenon" we need to require the analyst's procedure to be *self-efficient*, i.e., $\hat{\theta}^A_{\text{com}}$ needs to be most efficient among the class $\{\lambda\hat{\theta}^A_{\text{obs}} + (1-\lambda)\hat{\theta}^A_{\text{com}} : \lambda \in \mathbf{R}\}$ with respect to mean-squared loss. This assumption is violated here when $\sigma_0^2 \neq 1$, because then

$$\lambda = \frac{(1-f)(\sigma_0^2 - 1)}{\sigma_0^2(1-f) + f} \tag{2.5}$$

will render $\lambda\hat{\theta}^A_{\text{obs}} + (1-\lambda)\hat{\theta}^A_{\text{com}} = \lambda\bar{Y}_n + (1-\lambda)\bar{Y}_N = \hat{\theta}_{\text{MLE}}$. Here $\hat{\theta}_{\text{MLE}}$ is the complete-data MLE of θ under the correct imputer's model, which is more efficient than the analyst's $\hat{\theta}^A(Z_{\text{com}}) = \bar{Y}_N$ under God's model whenever $\sigma_0^2 \neq 1$. (Recall σ_0^2 is known.)

Linking Rubin's variance combining rule to the self-efficient formulation allows us to investigate when T_∞ is a conservative estimator or anti-conservative estimator, a useful distinction for investigating confidence validity; see Section 7. We notice that the bias Δ_n of (2.3) and the λ of (2.5) have identical sign, confirming the corresponding general result of Meng (1994). In particular, here T_∞ is conservative if and only if $\sigma_0^2 > 1$. We illustrate in the next two examples that this conservative tendency is a general phenomenon when the imputer's model and analyst's (embedding) model form a nested relationship in either direction, although the two directions have different flavors of conservatism. (For the current example, $N(\theta, 1)$ and $N(\theta, \sigma_0^2)$ do not form a nested pair in either direction when $\sigma_0^2 \neq 1$.)

Before we proceed, we illustrate a remarkable and practical finding that will be proved in Section 5.2. Although T_∞ may underestimate V_∞, $2T_\infty$ turns out will never be below V_∞. Indeed, a shaper upper bound for V_∞ exists for a univariate estimand; the sum of the within-imputation standard error $\sqrt{\bar{U}_\infty}$ and between-imputation standard error $\sqrt{B_\infty}$ is never below the standard error of $\bar{\theta}_\infty$, $\sqrt{V_\infty}$. (This is sharper because $(\sqrt{\bar{U}_\infty} + \sqrt{B_\infty})^2 \leq 2(\bar{U}_\infty + B_\infty) = 2T_\infty$). Thus they provide us with very simple ways to obtain confidence valid inference regardless of the degree of uncongenality and such inferences are sharp in the multi-phase inference paradigm when we need to protect ourselves from any degree of incompatibility between phases (Section 5.2).

To verify this in the current example, we need only show that

$$\sqrt{\bar{U}_\infty} + \sqrt{B_\infty} \geq \sqrt{V_\infty} \tag{2.6}$$

holds for all values of σ_0^2, which governs the degree of uncongenality in this case. But it's easy to see from (2.2) that the left-hand side of (2.6), as a function of σ_0^2, achieves its minimum when $\sigma_0^2 = 0$. This minimum value is given by, using the fact $1 - f = n/N$,

$$\sqrt{\frac{1-f}{N}} + \frac{f}{\sqrt{N(1-f)}} = \frac{1}{\sqrt{N(1-f)}} = \frac{1}{\sqrt{n}}, \tag{2.7}$$

which is exactly the right-hand side of (2.6).

2.2. Example 2: Hidden robustness—When analyst assumes more

Here we assume that the complete data are two normal samples that actually can be treated as one. This fact is used by the analyst, but not by the imputer, who models each sample with its own mean. But both model classes contains God's model as a sub-model, as depicted in Figure 1.

Specifically, maintaining the notation Z_{com} and Z_{obs} for complete data and observed data, respectively (and hence here $Z_{\text{com}} = \{X_{\text{com}}, Y_{\text{com}}\}$ and $Z_{\text{obs}} = \{X_{\text{obs}}, Y_{\text{obs}}\}$), we have

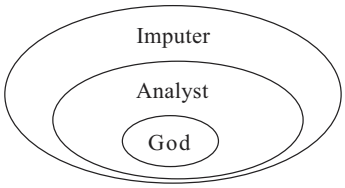

Figure 1. The scenario where the imputer assumes less.

- *God's Model:* $X_{\text{com}} = \{X_{\text{obs}}, X_{\text{mis}}\} \equiv \{(X_1, \ldots, X_{n_x}), (X_{n_x+1}, \ldots, X_{N_x})\}$ and $Y_{\text{com}} = \{Y_{\text{obs}}, Y_{\text{mis}}\} \equiv \{(Y_1, \ldots, Y_{n_y}),\ (Y_{n_y+1}, \ldots, Y_{N_y})\}$ with $X_i \overset{i.i.d.}{\sim} N(\theta_0, 1)$ and $Y_i \overset{i.i.d.}{\sim} N(\theta_0, 1)$; the X sample and Y sample are independent, denoted by $X \perp Y$.
- *Imputer's Model*: $X_i \overset{i.i.d.}{\sim} N(\theta_x, 1)$ and $Y_i \overset{i.i.d.}{\sim} N(\theta_y, 1)$, and $X \perp Y$; prior $p(\theta_x, \theta_y) \propto 1$; MIs are obtained as posterior predictive draws by sampling

$$\tilde{\theta}_x \sim N(\bar{X}_{n_x}, n_x^{-1}) \quad \text{and then} \quad \tilde{X}_i | \tilde{\theta}_x \overset{i.i.d.}{\sim} N(\tilde{\theta}_x, 1), \text{ for } i = n_x + 1, \ldots, N_x;$$

$$\tilde{\theta}_y \sim N(\bar{Y}_{n_y}, n_y^{-1}) \quad \text{and then} \quad \tilde{Y}_i | \tilde{\theta}_y \overset{i.i.d.}{\sim} N(\tilde{\theta}_y, 1), \text{ for } i = n_y + 1, \ldots, N_y.$$

- *Analyst's Complete-data Procedure:* Analyst's underlying model: $X_i \overset{i.i.d.}{\sim} N(\theta, 1)$ and $Y_i \overset{i.i.d.}{\sim} N(\theta, 1)$, and $X \perp Y$; the corresponding procedure for inferring θ then is

$$\hat{\theta}_{\text{com}} = w_x^{(c)} \bar{X}_{N_x} + w_y^{(c)} \bar{Y}_{N_y}, \quad V(\hat{\theta}_{\text{com}}) \equiv \bar{U}_\infty = \frac{1}{N}, \tag{2.8}$$

where $N = N_x + N_y$, $w_x^{(c)} = N_x/N$, and $w_y^{(c)} = N_y/N$.

Given this setting, it is straightforward to verify that Rubin's rules yield

$$\bar{\theta}_\infty = w_x^{(c)} \bar{X}_{n_x} + w_y^{(c)} \bar{Y}_{n_y} \quad \text{and} \quad T_\infty = \bar{U}_\infty + B_\infty = \frac{[w_x^{(c)}]^2}{n_x} + \frac{[w_y^{(c)}]^2}{n_y}. \tag{2.9}$$

But T_∞ is exactly the variance of $\bar{\theta}_\infty$ (conditioning on n_x and n_y) so Rubin's T_∞ remains consistent for $V(\bar{\theta}_\infty)$ even though the analyst's model and the imputer's model are uncongenial. The reason is that the analyst's complete-data procedure provides the best possible answer under either the analyst's model or the imputer's model, a condition that is sufficient (but not necessary) for the validity of $\mathcal{P}_\infty = [\bar{\theta}_\infty, T_\infty]$ (see Theorem 5 of Section 7).

What are the consequences of being uncongenial then? From the standard paradigm perspective, a negative consequence is that $\mathcal{P}_\infty = [\bar{\theta}_\infty, T_\infty]$ given by (2.9) is not optimal in general, since the MLE for θ under the analyst's model given Z_{obs} is

$$\hat{\theta}^A_{\text{obs}} = w_x^{(o)} \bar{X}_{n_x} + w_y^{(o)} \bar{Y}_{n_y}, \tag{2.10}$$

where $w_x^{(o)} = n_x/n$ and $w_y^{(o)} = n_y/n$. Clearly the variance of $\hat{\theta}^A_{\text{obs}}$ is $1/n$ (conditioning on the size n), smaller than T_∞ except when $w_x^{(o)} = w_x^{(c)}$ (and $w_y^{(o)} = w_y^{(c)}$), in which case $\bar{\theta}_\infty = \hat{\theta}^A_{\text{obs}}$. The difference between $\bar{\theta}_\infty$ of (2.9) and $\hat{\theta}^A_{\text{obs}}$ of (2.10) lies in their weights. Using the complete-data proportion $w_x^{(c)}$ is not as efficient as using the observed $w_x^{(o)}$ when the analyst's assumption of common mean holds. However, a positive consequence of not using this assumption is robustness. When the assumption fails, $\bar{\theta}_\infty$ is still a consistent estimator of the *overall mean* of the combined population, but $\hat{\theta}^A_{\text{obs}}$ is not. This is a standard bias-variance tradeoff, with the twist that the robustness was built into MI, enjoyed by the analyst but without necessarily knowing it.

Indeed, from the multi-phase inference perspective, this hidden benefit of $\bar{\theta}_\infty$ may outweigh the gain in efficiency by $\hat{\theta}^A_{\text{obs}}$. To see this, suppose an analyst wants to estimate the average income of Asian Americans in United States from a survey. Unknown to the analyst, both the income level and response probability depend substantially on US born versus foreign born, which is not provided, rendering consistent estimates of the average income impossible based on the observed responses alone. However, the nativity data are made available to the imputer, who therefore imputes the income levels for the two groups separately. Since (2.8) does not require nativity, the analyst can use Rubin's rules to reach the valid inference (2.9) without ever knowing the nativity. Clearly, the fact that there *could* be a more efficient $\hat{\theta}^A_{\text{obs}}$ if the analyst *were* given the nativity information is irrelevant in this multi-phase inference setting. Such a separation of information is particularly important for preserving data confidentiality (e.g., Reiter (2009a,b)).

In Section 7, we show that the phenomenon reported here is general. Under the scenario depicted by Figure 1, and the assumption that the analyst's procedure is self-efficient (plus some regularity conditions), Rubin's $\mathcal{P}_\infty = [\bar{\theta}_\infty, T_\infty]$ is consistent. Furthermore, $V(\bar{\theta}_\infty) \geq V(\hat{\theta}^A_{\text{obs}})$ because the efficiency due to the analyst's additional assumption is not used by the imputer, a necessary premium for the robustness built into the imputation. These results are obtained by an asymptotic decomposition of $\bar{\theta}_\infty$, as illustrated below.

Let $f = 1 - n/N$ and $w_x^{(m)} = (N_x - n_x)/(N - n)$ be the proportion of missing data that belong to the X sample, and $w_y^{(m)} = 1 - w_x^{(m)}$. Rubin's $\bar{\theta}_\infty$ then can be re-written as

$$\begin{aligned}\bar{\theta}_\infty &= (1-f)\left(w_x^{(o)}\bar{X}_{n_x} + w_y^{(o)}\bar{Y}_{n_y}\right) + f\left(w_x^{(m)}, w_y^{(m)}\right)\begin{pmatrix}\bar{X}_{n_x}\\ \bar{Y}_{n_y}\end{pmatrix}\\ &\equiv (1-F)\hat{\theta}^A_{\text{obs}} + FK\hat{\theta}^I_{\text{obs}}. \qquad (2.11)\end{aligned}$$

Therefore, $\bar{\theta}_\infty$ is a weighted sum of two estimators: $\hat{\theta}^A_{\text{obs}}$ of (2.10)—the analyst's observed-data estimator, and $K\hat{\theta}^I_{\text{obs}}$—the *projection* of the imputer's observed-data estimator, $\hat{\theta}^I_{\text{obs}} = (\bar{X}_{n_x}, \bar{Y}_{n_y})^\top$. The role played by the projection matrix

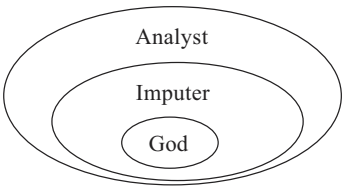

Figure 2. The scenario where the imputer assumes more.

$K = (w_x^{(m)}, w_y^{(m)})$ is crucial because θ^I and θ^A may live in different spaces. The weight is determined here by the *fraction of missing information* F, whose general expression will be given in Section 4.

The decomposition in (2.11) can also be derived from the viewpoint of estimating equations (EEs). In this example, Rubin's MI estimator $\bar{\theta}_\infty$ can be derived from EE

$$\begin{aligned} & h_n(Z_{\text{obs}};\theta) = u_n(Z_{\text{obs}};\theta) + v_n(Z_{\text{obs}};\theta) = 0, \\ \text{where} \quad & u_n(Z_{\text{obs}};\theta) = n_x\bar{X}_{n_x} + n_y\bar{Y}_{n_y} - n\theta\ , \\ \text{and} \quad & v_n(Z_{\text{obs}};\theta) = (N_x - n_x)\bar{X}_{n_x} + (N_y - n_y)\bar{Y}_{n_y} - (N-n)\theta. \end{aligned} \tag{2.12}$$

Here $u_n(Z_{\text{obs}};\theta) = 0$ and $v_n(Z_{\text{obs}};\theta) = 0$ can be viewed as the analyst's and imputer's EEs because they respectively yield $\hat{\theta}^A_{\text{obs}}$ and $K\hat{\theta}^I_{\text{obs}}$, as in (2.11). In Section 4, we show that this (asymptotical) correspondence between estimator and EE decompositions is indeed general.

2.3. Example 3: Super efficiency—When the imputer assumes more

This example adopts the same God's model as in Example 2, but with the imputer's model and the analyst's (embedding) model being switched, as depicted in Figure 2. Adopting the same notation as in Example 2, we have

- *God's model:* $X_{\text{com}} = \{X_{\text{obs}}, X_{\text{mis}}\} \equiv \{(X_1, \ldots, X_{n_x}), (X_{n_x+1}, \ldots, X_{N_x})\}$ and $Y_{\text{com}} = \{Y_{\text{obs}}, Y_{\text{mis}}\} \equiv \{(Y_1, \ldots, Y_{n_y}),\ (Y_{n_y+1}, \ldots, Y_{N_y})\}$ with $X_i \overset{i.i.d.}{\sim} N(\theta_0, 1)$ and $Y_i \overset{i.i.d.}{\sim} N(\theta_0, 1)$; and the two samples are independent.
- *Imputer's Model*: $X_i \overset{i.i.d.}{\sim} N(\theta, 1)$ and independently $Y_i \overset{i.i.d.}{\sim} N(\theta, 1)$; prior $p(\theta) \propto 1$; multiple imputations are obtained as the posterior predictive draws by sampling

$$\tilde{\theta}|Z_{\text{obs}} \sim N(\hat{\theta}^I_{\text{obs}}, n^{-1}), \text{and then } \tilde{X}_i|\tilde{\theta} \overset{i.i.d}{\sim} N(\tilde{\theta}, 1), \quad \tilde{Y}_i|\tilde{\theta} \overset{i.i.d}{\sim} N(\tilde{\theta}, 1), \tag{2.13}$$

where, as in Example 2, $n = n_x + n_y$, and

$$\hat{\theta}^I_{\text{obs}} = w_x^{(o)}\bar{X}_{n_x} + w_y^{(o)}\bar{Y}_{n_y}. \tag{2.14}$$

- *Analyst's Complete-data Procedure:* Analyst's underlying model: $X_i \overset{i.i.d.}{\sim} N(\theta_x, 1)$ and independently $Y_i \overset{i.i.d.}{\sim} N(\theta_y, 1)$. The resulting complete-data procedure is

$$\hat{\theta}^A_{\text{com}} \equiv \begin{pmatrix} \hat{\theta}^A_{x,com} \\ \hat{\theta}^A_{y,com} \end{pmatrix} = \begin{pmatrix} \bar{X}_{N_x} \\ \bar{Y}_{N_y} \end{pmatrix}, \quad V(\hat{\theta}^A_{\text{com}}) \equiv \bar{U}_\infty = \begin{pmatrix} \frac{1}{N_x} & 0 \\ 0 & \frac{1}{N_y} \end{pmatrix}. \tag{2.15}$$

Let $f_x = 1 - n_x/N_x$ and $f_y = 1 - n_y/N_y$. Adopting other notations in Example 2, we have

$$\bar{\theta}_\infty = \begin{pmatrix} (1-f_x)\bar{X}_{n_x} + f_x\hat{\theta}^I_{\text{obs}} \\ (1-f_y)\bar{Y}_{n_y} + f_y\hat{\theta}^I_{\text{obs}} \end{pmatrix}, \quad T_\infty = \begin{pmatrix} \frac{1-f_x^2}{n_x} + \frac{f_x^2}{n} & \frac{f_x f_y}{n} \\ \frac{f_x f_y}{n} & \frac{1-f_y^2}{n_y} + \frac{f_y^2}{n} \end{pmatrix}. \tag{2.16}$$

It can also be verified directly that the sampling variance of $\bar{\theta}_\infty$ is given by

$$V_\infty = \begin{pmatrix} \frac{(1-f_x)^2}{n_x} + \frac{f_x(2-f_x)}{n} & \frac{f_x+f_y-f_x f_y}{n} \\ \frac{f_x+f_y-f_x f_y}{n} & \frac{(1-f_y)^2}{n_y} + \frac{f_y(2-f_y)}{n} \end{pmatrix}. \tag{2.17}$$

Consequently, unless $f_x = f_y = 0$, Rubin's T_∞ has a non-vanishing asymptotic bias

$$\Delta_n = n(T_\infty - V_\infty) = \begin{pmatrix} 2f_x(1-f_x)\left(\frac{1}{n_x} - \frac{1}{n}\right) & 2f_x f_y - f_x - f_y \\ 2f_x f_y - f_x - f_y & 2f_y(1-f_y)\left(\frac{1}{n_y} - \frac{1}{n}\right) \end{pmatrix}. \tag{2.18}$$

Here we see that the two diagonal entries in (2.18) are always non-negative, implying that for estimating *individual components* of $\theta^A = (\theta_x, \theta_y)^\top$, Rubin's T_∞ is conservative and hence the actual converge is no less than the nominal coverage. However, the matrix Δ_n is not non-negative definite in general; e.g., when $f_x = 0$, $\det(\Delta_n) = -f_y^2 < 0$, as long as the Y sample is not fully observed.

From the standard paradigm perspective, this conservatism means that Rubin's procedure is not optimal. However, from the perspective of multi-phase inference, this conservatism is inevitable when the analyst complete-data procedure is derived without the benefit of the imputer's extra information. In Example 2, adopting $\mathcal{P}_\infty = [\bar{\theta}_\infty, T_\infty]$ makes it possible for the analyst to benefit from the robustness built into the MIs while only using a complete-data procedure. In the same vein, for the current example, adopting $\mathcal{P}_\infty = [\bar{\theta}_\infty, T_\infty]$ permits the analyst to benefit from the extra efficiency built into the imputations without being aware of it.

To see this clearly, we write $\bar{\theta}_\infty = (\hat{\theta}_{x,\infty}, \hat{\theta}_{y,\infty})^\top$, and let $T_{x,\infty}$ and $T_{y,\infty}$ be the diagonal elements of T_∞ of (2.16). Consider the X component. Evidently, without MIs, the analyst's estimator for θ_x will be $\bar{X}_{n_x}$, with variance $V(\bar{X}_{n_x}) =$

n_x^{-1} (conditioning on n_x). Although $T_{x,\infty}$ overestimates $V(\hat{\theta}_{x,\infty})$, it is easy to verify from the expression of T_∞ in (2.16) that

$$V(\bar{X}_{n_x}) - T_{x,\infty} = f_x^2 \left(\frac{1}{n_x} - \frac{1}{n} \right) = \frac{f_x^2 n_y}{n n_x} \geq 0, \tag{2.19}$$

which is zero if and only if either $f_x = 0$, no missing X's, or $n_y = 0$, no Y sample to help. The inequality (2.19) implies that any (asymptotic) confidence interval estimator for θ_x based on the analyst's observed-data procedure is *inadmissible*. For example, although $(\bar{X}_{n_x} - 1.96 n_x^{-1/2},\ \bar{X}_{n_x} + 1.96 n_x^{-1/2})$ has the correct 95% coverage, it is at least as wide as $(\hat{\theta}_{x,\infty} - 1.96 T_{x,\infty}^{1/2},\ \hat{\theta}_{x,\infty} + 1.96 T_{x,\infty}^{1/2})$ because of (2.19), yet the latter has at least 95% coverage because $T_{x,\infty} \geq V(\hat{\theta}_{x,\infty})$. This seemingly paradoxical phenomenon is because $\hat{\theta}_{x,\infty}$ is more efficient than $\bar{X}_{n_x}$, and the over-estimation by $T_{x,\infty}$ for $V(\hat{\theta}_{x,\infty})$ still does not exceed the added variance in $\bar{X}_{n_x}$ compared with that of $\hat{\theta}_{x,\infty}$, i.e., $V(\bar{X}_{n_x}) - V(\hat{\theta}_{x,\infty})$.

Since the conservativeness result holds for individual components, shouldn't it also hold for any linear combination of them? After all, we can always re-parameterize. To see why this is not the case, we write, analogous to (2.11),

$$\bar{\theta}_\infty = \begin{pmatrix} (1-f_x)\bar{X}_{n_x} + f_x \hat{\theta}^I_{\text{obs}} \\ (1-f_y)\bar{Y}_{n_y} + f_y \hat{\theta}^I_{\text{obs}} \end{pmatrix} \equiv (I - F)\hat{\theta}^A_{\text{obs}} + FK\hat{\theta}^I_{\text{obs}}, \tag{2.20}$$

where $\hat{\theta}^I_{\text{obs}}$ is given by (2.14), and

$$F = \begin{pmatrix} f_x & 0 \\ 0 & f_y \end{pmatrix},\ K = \begin{pmatrix} 1 \\ 1 \end{pmatrix},\ \hat{\theta}^A_{\text{obs}} = \begin{pmatrix} \bar{X}_{n_x} \\ \bar{Y}_{n_y} \end{pmatrix}. \tag{2.21}$$

Again we find that $\bar{\theta}_\infty$ is a *matrix weighted* sum of $\hat{\theta}^A_{\text{obs}}$, the analyst's observed-data estimator and $K\hat{\theta}^I_{\text{obs}}$, a projection of the imputer's observed-data estimator. As in Section 2.2, the decomposition (2.20) has its corresponding EEs for $\bar{\theta}_\infty$, $\hat{\theta}^A_{\text{obs}}$ and $K\hat{\theta}^I_{\text{obs}}$ as

$$\begin{aligned} h_n(Z_{\text{obs}}; \theta) &= u_n(Z_{\text{obs}}; \theta) + v_n(Z_{\text{obs}}; \theta), \\ u_n(Z_{\text{obs}}; \theta) &= \begin{pmatrix} n_x \bar{X}_{n_x} - n_x \theta \\ n_y \bar{Y}_{n_y} - n_y \theta \end{pmatrix}, \\ v_n(Z_{\text{obs}}; \theta) &= \begin{pmatrix} (N_x - n_x)\hat{\theta}^I_{\text{obs}} - (N_x - n_x)\theta \\ (N_y - n_y)\hat{\theta}^I_{\text{obs}} - (N_y - n_y)\theta \end{pmatrix}. \end{aligned} \tag{2.22}$$

Now consider a re-parametrization $\phi = C\theta$ under the analyst's model, where C is a 2×2 matrix. From (2.20), unless F is proportional to the identity matrix, the matrices C and F (and hence $(I - F)$) generally *do not commute*. Consequently, in general,

$$\bar{\phi}_\infty = C\bar{\theta}_\infty \neq (I - F)C\hat{\theta}^A_{\text{obs}} + F[CK\hat{\theta}^I_{\text{obs}}] = (I - F)\hat{\phi}^A_{\text{obs}} + F[CK\hat{\theta}^I_{\text{obs}}].$$

That is, the matrix-weighted decomposition is *not* invariant even to linear transformations.

This lack of invariance reflects the inherent complexity of missing-data mechanism, which can easily affect different parameters differently. To illustrate, suppose the analyst's estimand is $\phi = \theta_x + \theta_y$. From (2.16), the Rubin's MI estimator for ϕ is

$$\hat{\phi}_\infty = \hat{\theta}_{x,\infty} + \hat{\theta}_{y,\infty} = [(1 - f_x)\bar{X}_{n_x} + (1 - f_y)\bar{Y}_{n_y}] + (f_x + f_y)\hat{\theta}^I_{\text{obs}}. \tag{2.23}$$

Because the MLE under the analyst's observed-data model is $\hat{\phi}^A_{\text{obs}} = \bar{X}_{n_x} + \bar{Y}_{n_y}$ and under the imputer's model is $\hat{\phi}^I_{\text{obs}} = 2\hat{\theta}^I_{\text{obs}}$ with $\hat{\theta}^I_{\text{obs}}$ given by (2.14), the different weights for $\bar{X}_{n_x}$ and $\bar{Y}_{n_y}$ as seen in (2.23) make it impossible to express $\hat{\phi}_\infty$ of (2.23) as $(1 - F)\hat{\phi}^A_{\text{obs}} + F\hat{\phi}^I_{\text{obs}}$ for some scalar quantity F, unless $f_x = f_y$ (exactly or asymptotically). This impossibility is a consequence of the fact that, *under the analyst's complete-data* model, $p^A(Z_{\text{com}}|Z_{\text{obs}}; \theta_x, \theta_y)$ depends on both θ_x and θ_y, unless $f_x = f_y$, in which case, trivially, it depends on ϕ only.

Consequently, the bias in Rubin's variance estimator for ϕ is not guaranteed to be non-negative. For example, in the simple case where $f_x = 0$ and $N_x = N_y/2$ (i.e., the Y sample is twice as large as the X sample when both are fully observed), the asymptotic bias in Rubin's variance estimator for ϕ is given by $-f_y$. However, the variance estimator derived from using $2T_\infty$, $2\vec{c}^\top T_\infty \vec{c}$, will serve as an upper bound for $V(\hat{\phi}_\infty) = \vec{c}^\top V_\infty \vec{c}$, where $\vec{c} = (1, 1)^\top$. This is because $\tilde{\Delta}_\infty \equiv 2T_\infty - V_\infty$ is non-negative definite, since by (2.16)$-$(2.17) both diagonal elements of $\tilde{\Delta}_\infty$ are bounded below by n^{-1} for $n \geq \max\{n_x, n_y\}$, yet the absolute value of its off-diagonal element is bounded above by n^{-1} because $|f_x + f_y - 3f_x f_y| \leq 1$ for all $0 \leq f_x, f_y \leq 1$.

Compounded by the issue of uncongenality, which typically implies that the imputer's model and analyst's (embedding) model have different (nuisance) parameters, greater caution is therefore in order when generalizing univariate results to multivariate ones. In what follows, we exercise such caution by carefully laying out key assumptions and regularity conditions, although for greater statistical insights, we do not strive for the weakest possible technical conditions. Nor do we claim all the technical derivations or results are really new. Indeed, a few of the results below can be obtained via direct asymptotic calculations (e.g., Robins and Wang (2000)). We adopt the decomposing approach because it permits us to gain deeper, and newer, theoretical insights on how different phases contribute to the final results.

3. General Decomposition of Estimating Equations

The general decomposition result in this section is instrumental to our understanding of Rubin's MI inference as an integration of the knowledge of the

imputer and the analyst. As illustrated by (2.11) and (2.20), the key here is to express an estimator as a matrix-weighted combination of two relevant estimators. For example, a complete data procedure can be written as the sum of a (projected) marginal procedure based on the observed data and a (residual) conditional one based on the missing data given the observed. Such decompositions are essential for studying multi-phase inferences because they explicate the impacts from different phases.

Let Z be a generic notation for the data we have and

$$h_n(Z;\theta) = 0 \tag{3.1}$$

be the d-dimensional EE we adopt, where $\theta \in \Theta \subset \mathbf{R}^d$, and n is a deterministic index of the information in Z under our posited assumptions such that it is (probabilistically) meaningful to postulate limiting behavior (e.g., consistency) when n is permitted to grow indefinitely. We can view the EE in (3.1) as having been properly normalized by a positive sequence $\{a_n, n \geq 1\}$ such that its derivative with respect to θ is proportional to n asymptotically – see Definition 1 below for a precise statement. (Due to space limitation, we omit discussion of the vast related literature on EEs, other than mentioning a very readable overview article by Desmond (1997).)

Now suppose for $h_n(Z;\theta)$ we have the decomposition

$$h_n(Z;\theta) = u_n(Z;\theta) + v_n(Z;\theta)\ , \tag{3.2}$$

where $u_n(Z;\theta)$ and $v_n(Z;\theta)$ are *also EEs*, as in (2.12) and (2.22). It is logical to expect a certain relationship, at least asymptotically, among the three corresponding roots. To rigorously establish this relationship, we need notation and (standard) regularity conditions. Hereafter, E^G (similarly "V^G" and "Cov^G") denotes an expectation with respect to God's model. Since God's model does not involve any unknown parameter (to God), the "true value" θ_0 below refers to the value of our model parameter θ such that $f(Z|\theta_0)$ coincides with God's model. For simplicity, we adopt the L^2 norm for vectors and matrices, and the shorthand $\frac{\partial}{\partial\theta} f(\theta_0)$ for $\frac{\partial}{\partial\theta} f(\theta)|_{\theta=\theta_0}$.

Definition 1. Second-Order Regularity (SOR). We say an EE $g_n(Z;\theta) = 0$ satisfies Weak SOR if

(i) $g_n(Z;\theta)$ is twice differentiable with respect to θ for any given Z;

(ii) Asymptotically $g_n(Z;\theta)$ is unbiased, $E^G[g_n(Z;\theta_0)] = o(\sqrt{n})$, and it has the unique root $\hat{\theta}_g$, which is $\sqrt{n}$-consistent, $\sqrt{n}\|\hat{\theta}_g - \theta_0\| = O_p(1)$, where θ_0 is the true value of θ;

(iii) There exists a finite $J_g(\theta_0)$ satisfying (almost surely)

$$\lim_{n\to\infty} -\frac{1}{n}\frac{\partial}{\partial\theta}g_n(Z;\theta_0) = \lim_{n\to\infty} -\frac{1}{n}E^G\left[\frac{\partial}{\partial\theta}g_n(Z;\theta_0)\right] = J_g(\theta_0),$$

where $J_g(\theta_0)$ is continuous at θ_0 and its determinant satisfies $0 < |J_g(\theta_0)| < \infty$.

(iv) There exist some $\epsilon > 0$ and $M_n(Z)$ such that $E^G[M_n(Z)] < K$ for some $K < \infty$ and

$$\left\|\frac{\partial^2}{\partial\theta^2}g_n(Z;\theta)\right\| \le nM_n(Z) \quad \text{for any } \theta \text{ such that } \ ||\theta - \theta_0|| \le \epsilon.$$

We say g_n satisfies *SOR* if in addition to (i)−(iv), the following hold:

(v) There exists some $\epsilon > 0$ and $M < \infty$ such that $E^G\left[\sqrt{n}\|\hat{\theta}_g - \theta_0\|\right]^{2+\epsilon} < M$ for any n.

(vi) There exists some $\epsilon > 0$ and $M < \infty$ such that $E^G\left[\|g_n(Z;\theta)\|/\sqrt{n}\right]^{2+\epsilon} < M$ for any n.

The following lemma links the EE and its root, and serves as a building block for Theorem 1, which provides our key decomposition result. (All proofs are given in the on-line Appendix I.)

Lemma 1. *If an EE $g_n(Z;\theta) = 0$ satisfies Weak SOR, then we have for its root $\hat{\theta}_g$,*

$$R_n \triangleq \sqrt{n}\left[\left(\hat{\theta}_g - \theta_0\right) - \left[nJ_g(\theta_0)\right]^{-1} g_n(Z;\theta_0)\right] \xrightarrow{p} 0. \tag{3.3}$$

If it also satisfies SOR, then $R_n \xrightarrow{L^2} 0$.

Theorem 1. *Assume EEs $h_n(Z;\theta)$, $u_n(Z;\theta)$ and $v_n(Z;\theta)$ all satisfy Weak SOR and*

$$h_n(Z;\theta) = u_n(Z;\theta) + v_n(Z;\theta).$$

Then their corresponding roots $\hat{\theta}_h, \hat{\theta}_u$, and $\hat{\theta}_v$ obey the asymptotic relationship

$$D_n \triangleq \sqrt{n}\left[\hat{\theta}_h - \left((I - F)\hat{\theta}_u + F\hat{\theta}_v\right)\right] \xrightarrow{p} 0,$$

where $F = J_h(\theta_0)^{-1}J_v(\theta_0)$ is the "fraction of information" contained in $v_n(Z;\theta)$. If in addition all three EEs satisfy SOR, then $D_n \xrightarrow{L^2} 0$.

Here the fraction of information contained in v_n reduces to the conventional fraction of missing information in the context of EM algorithm when h_n is the complete-data score function and u_n is the observe-data score function (and hence $v_n = h_n - u_n$ captures the missing information).

There are circumstances where $u_n(Z;\theta)$ and $v_n(Z;\theta)$ are (asymptotically) orthogonal, resulting in $\hat{\theta}_u$ and $\hat{\theta}_v$ being uncorrelated. One important class is captured below.

Corollary 1. *Suppose that $u_n(Z;\theta)$ and $v_n(Z;\theta)$ in Theorem 1 have the form $u_n(Z;\theta) = u_n(Z_1;\theta)$ and $v_n(Z;\theta) = v_n(Z_1, Z_2;\theta)$, where*

$$E^G[v_n(Z_1, Z_2;\theta_0)|Z_1] = 0. \tag{3.4}$$

Then $Cov^G(\hat{\theta}_u, \hat{\theta}_v) = o\left(n^{-1}\right)$. In particular, $\hat{\theta}_v$ is asymptotically uncorrelated with any consistent estimator $\hat{\theta}_u(Z_1)$ such that $\sqrt{n}(\hat{\theta}_u(Z_1) - \theta_0)$ converges in L^2 to a mean-zero variable.

An important application has $\hat{\theta}^A_{\rm com}(Z_{\rm com})$ as the root of an EE defined by $S^A(Z_{\rm com};\theta)$:

$$h^A_n(Z_{\rm com};\theta) \equiv S^A(Z_{\rm com};\theta) = 0. \tag{3.5}$$

We can then write $h^A_n(Z_{\rm com};\theta)$ as the sum of the two EEs

$$u^A_n(Z_{\rm obs};\theta) = E^A\left[S^A(Z_{\rm com};\theta)|Z_{\rm obs};\theta\right] \quad \text{and} \tag{3.6}$$

$$v^A_n(Z_{\rm com};\theta) = S^A(Z_{\rm com};\theta) - E^A\left[S^A(Z_{\rm com};\theta)|Z_{\rm obs};\theta\right]\ . \tag{3.7}$$

The roots corresponding to $h^A_n = 0$, $u^A_n = 0$ and $v^A_n = 0$ are respectively denoted by $\hat{\theta}^A_{\rm com}(Z_{\rm com})$, $\hat{\theta}^A_{\rm obs}(Z_{\rm obs})$ and $\hat{\theta}^A_{\rm mis}(Z_{\rm com})$. Theorem 1 then allows us to write

$$\hat{\theta}^A_{\rm com}(Z_{\rm com}) = (I - F^A)\hat{\theta}^A_{\rm obs}(Z_{\rm obs}) + F^A\hat{\theta}^A_{\rm mis}(Z_{\rm com}) + \frac{R^A_n}{\sqrt{n}}, \tag{3.8}$$

where $R^A_n \xrightarrow{L^2} 0$, and F^A is given by Theorem 1 with all its ingredients specified by the two terms given in (3.6)−(3.7). Corollary 1 then tells us that asymptotically $\hat{\theta}^A_{\rm obs}(Z_{\rm obs})$ and $\hat{\theta}^A_{\rm mis}(Z_{\rm com})$ are uncorrelated because (3.4) follows from (3.6)−(3.7). Consequently

$$V^G(\hat{\theta^A}_{\rm com}) = (I - F^A)V^G(\hat{\theta}^A_{\rm obs})(I - F^A)^\top + F^A V^G(\hat{\theta}^A_{\rm mis})(F^A)^\top + o\left(n^{-1}\right), \tag{3.9}$$

a decomposition that plays an important role in Section 5. When $S^A(Z_{\rm com};\theta)$ is a complete-data score function, $u^A_n(Z_{\rm obs};\theta)$ of (3.6) is simply the observed-data score function $S^A(Z_{\rm obs};\theta)$ because of the Fisher identity $E^A[S^A(Z_{\rm com};\theta)|Z_{\rm obs};\theta] = S^A(Z_{\rm obs};\theta)$, the key identity underlying the EM algorithm — see Meng and van Dyk (1997). Clearly (3.4) then follows.

4. Application of the Key Decomposition to MI Inference

4.1. A theoretical setup and simplification

In order to apply the general decomposition result in Section 3 to study Rubin's MI inference, we need to adopt a theoretical simplification. Specifically, consider two ways of imputation.

Table 1. Four different multiple imputation estimators.

	Averaging Estimates	Averaging EEs
Posterior Predictive	$\bar{\theta}_\infty^{(11)}$	$\bar{\theta}_\infty^{(12)}$
Plug-in Predictive	$\bar{\theta}_\infty^{(21)}$	$\bar{\theta}_\infty^{(22)}$

(i) *Posterior Predictive*: Draw $\tilde{\theta}$ from the posterior distribution $p^I(\theta|Z_{\rm obs})$ and then draw the predictive value $\tilde{Z}_{\rm mis}$ of $Z_{\rm mis}$ from $p^I(\tilde{Z}_{\rm mis}|Z_{\rm obs};\tilde{\theta})$.

(ii) *Plug-in Predictive*: Draw the predictive value $\tilde{Z}_{\rm mis}$ of $Z_{\rm mis}$ from the "plug-in" distribution $p^I(\tilde{Z}_{\rm mis}|Z_{\rm obs};\hat{\theta}^I_{\rm obs})$, where $\hat{\theta}^I_{\rm obs}$ is the imputer's estimator of θ (e.g., MLE).

In general the proper way of performing MI is to use (i), whereas draws from (ii) generally lead to under-dispersion, as discussed in Kim (2011). However, for theoretical studies of the point-estimator $\bar{\theta}_\infty$ (but *not* for its variance estimator T_∞), (i) and (ii) are equivalent under the usual regularity conditions that ensure Bayesian and likelihood inferences are asymptotically equivalent. Regardless of how imputations are drawn, there are (at least) two ways to form $\bar{\theta}_\infty$.

(a) *Averaging Estimators:* Compute individual estimators $\hat{\theta}^{(\ell)}$ and then take the average:

$$\bar{\theta}_\infty = \lim_{m\to\infty} \frac{1}{m}\sum_{\ell=1}^{m} \hat{\theta}^{(\ell)} = E^I[\hat{\theta}^A(\tilde{Z}_{\rm com})].$$

(b) *Averaging EEs*: Form the average EE first and then compute $\bar{\theta}_\infty$ as its root:

$$0 = E^I[S^A(\tilde{Z}_{\rm com};\theta)] = \lim_{m\to\infty} \frac{1}{m}\sum_{\ell=1}^{m} S^A(\tilde{Z}^{(\ell)}_{\rm com};\theta).$$

Rubin's rules adopt (a), but (b) is easier for applying the decomposition result in Section 3.

In a nutshell, by crossing (i)−(ii) with (a)−(b), we have four ways of constructing $\bar{\theta}_\infty$, as summarized in Table 1. Rubin's (1987) original proposal is $\bar{\theta}_\infty^{(11)}$, but Wang and Robins (1998) and Robins and Wang (2000) considered the estimator from the following EE

$$E^I[S^A(\tilde{Z}_{\rm com};\theta)|Z_{\rm obs};\hat{\theta}^I_{\rm obs}] = 0\ , \tag{4.1}$$

which is $\bar{\theta}_\infty^{(22)}$ in Table 1. They argued that their estimator is asymptotically the same as $\bar{\theta}_\infty^{(11)}$. Indeed, all four estimators in Table 1 are asymptotically equivalent under suitable regularity conditions, especially the assumption that SOR is preserved when God's $Z_{\rm com}$ in $S^A(Z_{\rm com};\theta)=0$ is replaced by the "completed-data" $\tilde{Z}_{\rm com} = \{Z_{\rm obs}, \tilde{Z}_{\rm mis}\}$ (similar to the "Super God" perspective in Section 1.3)

from $p(Z_{\text{obs}}|\theta_0)p(\tilde{Z}_{\text{mis}}|Z_{\text{obs}};\hat{\theta}^I_{\text{obs}})$. This assumption (apparently new) reflects the intuition that in order for the MI inference to be valid, the imputed data must capture enough probabilistic properties of the actual unobserved data. Under these assumptions, the asymptotic equivalence of all estimators in Table 1 is proved in Appendix I; we thus can use a generic $\bar{\theta}_\infty$.

To apply Theorem 1, we write

$$h_n(Z_{\text{obs}};\theta) = u_n(Z_{\text{obs}};\theta) + v_n(Z_{\text{obs}};\theta), \tag{4.2}$$

where

$$h_n(Z_{\text{obs}};\theta) = E^I\left[S^A(Z_{\text{obs}},\tilde{Z}_{\text{mis}};\theta)\,\middle|\,Z_{\text{obs}};\hat{\theta}^I_{\text{obs}}\right], \tag{4.3}$$

$$u_n(Z_{\text{obs}};\theta) = E^A\left[S^A(Z_{\text{obs}},Z_{\text{mis}};\theta)\,|Z_{\text{obs}};\theta\right], \tag{4.4}$$

$$\begin{aligned} v_n(Z_{\text{obs}};\theta) = {} & E^I\left[S^A(Z_{\text{obs}},\tilde{Z}_{\text{mis}};\theta)|Z_{\text{obs}};\hat{\theta}^I_{\text{obs}}\right] \\ & -E^A\left[S^A(Z_{\text{obs}},Z_{\text{mis}};\theta)|Z_{\text{obs}};\theta\right], \end{aligned} \tag{4.5}$$

assuming all expectations are well defined (but see Section 7.2 for a counterexample). As with (3.6), solving $u_n(Z;\theta)=0$ is interpreted as the analyst's observed-data procedure, the projection of the analyst's complete-data procedure onto the observed-data space *under the analyst's (embedding) model.* The $v_n(Z;\theta)$ term, in contrast to the v^A_n term defined in (3.7), is more complicated because it depends on both the analyst's model and the imputer's model. However, as Theorem 3 below will show, its root can be interpreted as a projection of the imputer's observed-data estimator $\hat{\theta}^I_{\text{obs}}$ onto the analyst's parameter space. This is particularly clear from the regression example presented in Section 8.3, and more fully in the on-line Appendix II.

4.2. Key results on integrating the imputer's and analyst's knowledge

With the setup in Section 4.1, we now present our key result.

Theorem 2. *Let $h_n(Z_{\text{obs}};\theta)$, $u_n(Z_{\text{obs}};\theta)$, and $v_n(Z_{\text{obs}};\theta)$ be the EEs defined in (4.2)–(4.5) and $\bar{\theta}_\infty$, $\hat{\theta}^A_{\text{obs}}$, and $\hat{\theta}^H_{\text{obs}}$, be their corresponding roots. Assuming that the three EEs satisfy SOR, we have*

$$\sqrt{n}\left[\bar{\theta}_\infty - \left((I-F)\hat{\theta}^A_{\text{obs}} + F\hat{\theta}^H_{\text{obs}}\right)\right] \xrightarrow{L^2} 0, \tag{4.6}$$

where F is the "fraction of missing information" given by

$$F = I_{\text{com}}(\theta_0)^{-1} I_{\text{mis}}(\theta_0) = \left[I_{\text{obs}}(\theta_0) + I_{\text{mis}}(\theta_0)\right]^{-1} I_{\text{mis}}(\theta_0), \tag{4.7}$$

with $I_{\text{com}}(\theta_0) = J_h(\theta_0)$, $I_{\text{mis}}(\theta_0) = J_v(\theta_0)$ and $I_{\text{obs}}(\theta_0) = J_u(\theta_0)$, where the J matrix is provided in condition (iii) *of Definition* 1.

Unlike in (3.8), the decomposition in Theorem 2 is not orthogonal. Indeed, here both $\hat{\theta}^H_{\text{obs}}$ and $\hat{\theta}^A_{\text{obs}}$ are functions of Z_{obs} only and they are inherently dependent, where the superscript H stands for *hybrid* because $\hat{\theta}^H_{\text{obs}}$ is based on both the imputer's model and the analyst's (embedding) model. When the same parametric model is used by the analyst and the imputer and both adopt the MLE approach, it is easy to see that $\hat{\theta}^I_{\text{obs}}$ is also a root of $v_n(Z_{\text{obs}};\theta) = 0$. Consequently, given our assumption that the root is unique, $\hat{\theta}^H_{\text{obs}} = \hat{\theta}^I_{\text{obs}} = \hat{\theta}^A_{\text{obs}}$ and hence the decomposition reduces to the congenial case where $\bar{\theta}_\infty = \hat{\theta}^A_{\text{obs}}$. This is also expected from the perspective of the EM algorithm, because under congeniality, performing MI with an infinite number of imputations (and with the "plug-in" predictive imputation) is the same as carrying out the final EM iteration.

Theorem 2 would be sufficient if our goal were merely to derive the asymptotic variance of $\bar{\theta}_\infty$. (In fact the asymptotic variance can be derived directly as in Wang and Robins (1998) and Robins and Wang (2000)). However, to study Rubin's T_∞ of (1.2), we need to understand how the imputer's knowledge enters into $\hat{\theta}^H_{\text{obs}}$. This turns out to be a challenging task, with several subtleties. For example, there is not necessarily a direct link between θ^A and θ^I; both the imputer and the analyst are free to adopt whatever serves their purposes, especially when they do not share information. Nevertheless, θ^A and θ^I can be linked indirectly through moments of the data, or more generally common distributional summaries (e.g., percentiles) or even the entire predictive distribution under each model. For example, suppose the analyst's interest is a population mean, expressible via $E^A(Z|\theta^A) \equiv \mu^A(\theta^A)$. If under the imputer's model, $E^I(Z|\theta^I) \equiv \mu^I(\theta^I)$ is also well defined, then $\mu^A(\theta^A) = \mu^I(\theta^I)$ serves as a natural map between θ^A and θ^I.

Clearly any single map is not always one-to-one, and in general there are infinitely many such maps by equating $E^A[g(Z)|\theta^A] = E^I[g(Z)|\theta^I]$ and varying g. How to construct a suitable set of maps will depend on the objective of the study. Here we raise the issue of mapping to help to understand the following theorem, which will link $\hat{\theta}^H_{\text{obs}}$, an estimator of θ^A, to $\hat{\theta}^I_{\text{obs}}$, an estimator of θ^I. The link below is linear because we have assumed both θ^A and θ^I are of continuous type living in Euclidian spaces, and we are concerned only with asymptotical results for which the usual linearization via Taylor expansion takes place. Generalizations to other type of parameters, such as discrete ones arising in model selection, are open problems.

Theorem 3. *Under regularity conditions stated in the proof (see Appendix I), we have*

$$\sqrt{n}\left[(\hat{\theta}^H_{\text{obs}} - \theta^A_0) - K(\hat{\theta}^I_{\text{obs}} - \theta^I_0)\right] \xrightarrow{L^2} 0,$$

where K is a projection matrix given by

$$K = [I^A_{\text{mis}}(\theta^A_0)]^{-1} \times \lim_{n\to\infty} \frac{1}{n} E^G \left[v^A_n(Z_{\text{com}}; \theta^A_0) \left\{ \frac{\partial}{\partial \theta^I} \log p^I(Z_{\text{mis}}|Z_{\text{obs}}; \theta^I_0) \right\}^\top \right],$$

with $I^A_{\text{mis}}(\theta^A_0) = J_{v^A}(\theta^A_0)$ and $v^A_n(Z_{\text{com}}; \theta^A)$ is given by (3.7).

4.3. Results for nested cases and a subtlety

When θ^A and θ^I measure the same population quantity, one may expect the "link matrix" K in Theorem 3 to be the identity matrix. This turns out to be false in general, owing to a subtle issue in defining the meaning of "measuring the same population quantity"—do we mean the exact value used in God's data-generating distribution or more generally the "population parameter", for which the actual value used by God is only one realization?

To see why this distinction matters, consider an example where God's model is $N(0, 1)$ and the imputer's model is $N(\mu, 1)$. However, the analyst's procedure is

$$S^A(Z_{\text{com}}; \mu) = \sum_i (Z_i^3 - \mu) = 0. \tag{4.8}$$

Hence $\hat{\mu}^H_{\text{obs}} = \bar{Z}^3_{\text{obs}} + 3\bar{Z}_{\text{obs}}$ and $\hat{\mu}^I_{\text{obs}} = \bar{Z}_{\text{obs}}$, implying $\sqrt{n}[(\hat{\mu}^H_{\text{obs}} - \mu_0) - 3(\hat{\mu}^I_{\text{obs}} - \mu_0)] \xrightarrow{p} 0$ *when the true value* $\mu_0 = 0$. Here, the projection matrix K is 3 instead of 1.

The problem here is that the analyst's procedure is valid only when $\mu_0 = 0$. Had God used a slightly different μ_0, it would fail to lead to a consistent estimator. The real estimand in (4.8) should be $\theta \equiv \theta(\mu) = \mu^3 + 3\mu = E(Z^3)$, even though $\theta(\mu) = \mu$ when $\mu = 0$. For θ, it is easy to verify that $\hat{\theta}^H_{\text{obs}} = \bar{Z}^3_{\text{obs}} + 3\bar{Z}_{\text{obs}} = \hat{\theta}^I_{\text{obs}}$, and therefore trivially $\sqrt{n}\left[(\hat{\theta}^H_{\text{obs}} - \theta_0) - (\hat{\theta}^I_{\text{obs}} - \theta_0)\right] \xrightarrow{p} 0$ regardless of the actual *value* of θ_0, restoring the projection matrix K for θ to be 1.

This seemingly contrived example reveals an important issue for studying multi-phase inference — it may be insufficient to assume that models at all phases are *correctly specified*, by which we typically mean that they are all consistent with the God's model that generates our current data. For features of the models we declare to be the same, we may need to assume that "sameness" holds even when God's model is somewhat perturbed from the one that generates the our data. Such assumptions are implicit in many asymptotic studies under the standard paradigm (e.g., regularity conditions imposed in an ϵ-neighbor of the true value), but for multi-phase inferences, they may play more critical roles than serving merely as "technical regularities".

For many applications, it is possible to construct a class of distributions indexed by a vector parameter $\breve{\theta} \in \breve{\Theta}$ such that both the analyst's (embedding)

model and the imputer's model are subclasses of $\mathcal{F} = \{f_{\breve{\theta}}(Z) : \breve{\theta} \in \breve{\Theta}\}$ in the sense that both θ^A and θ^I are sub-vectors of $\breve{\theta}$. Thus the analyst's parameter space and the imputer's one can be divided into a trichotomy: (i) the overlapping part $\theta^{I\cap A}$, which may be empty; (ii) the leftover part for the imputer: $\theta^{I\setminus A}$; and (iii) the leftover part for the analyst: $\theta^{A\setminus I}$. By imposing essentially a "smooth transition" condition on the imputation model with respect to the overlapping $\theta^{I\cap A}$, we can ensure a broader "sameness" for estimating $\theta^{I\cap A}$, and hence avoid the complications demonstrated by the example above.

Definition 2. Imputation Validity Under Perturbation. Let $\tilde{v}(Z_{\text{obs}}; \theta^{I\cap A}, \theta^{A\setminus I}, \theta^{I\setminus A})$ be the projection of the analyst's EE (3.7) under the imputer's model,

$$\tilde{v}(Z_{\text{obs}}; \theta^{I\cap A}, \theta^{A\setminus I}, \theta^{I\setminus A}) \equiv E^I \left[v_n^A(Z_{\text{com}}; \{\theta^{I\cap A}, \theta^{A\setminus I}\}) \mid Z_{\text{obs}}; \{\theta^{I\cap A}, \theta^{I\setminus A}\} \right]. \tag{4.9}$$

We say the imputation model $p^I(\tilde{Y}_{\text{mis}}|Z_{\text{obs}}; \theta^I)$ is valid under perturbation with respect to the analyst's complete-data (second-order regular) EE $S^A(Z_{\text{com}}; \theta^A) = 0$ if there exists an $\epsilon > 0$ such that for all $\theta^{I\cap A}$ satisfying $||\theta^{I\cap A} - \theta_0^{I\cap A}|| \leq \varepsilon$, we have

$$\tilde{v}(Z_{\text{obs}}; \theta^{I\cap A}, \theta_0^{A\setminus I}, \theta_0^{I\setminus A}) = o_p(n^{1/2}) \quad \text{and} \quad \frac{\partial \tilde{v}(Z_{\text{obs}}; \theta^{I\cap A}, \theta_0^{A\setminus I}, \theta_0^{I\setminus A})}{\partial \theta^{I\cap A}} = o_p(n),$$

where the subscript "0" on any parameter indicates its true value.

We can now simplify the matrix K in Theorem 3 when the parameter spaces are nested.

Corollary 2. *Assuming the imputation model is valid under perturbation, then under the regularity conditions of Theorem* 3*, we have*

(1) *If θ^I is a sub-parameter of θ^A, then $K = [I_I, \mathbf{0}]^\top$, where I_I is the identity matrix with dimension corresponding to θ^I.*
(2) *If θ^A is a sub-parameter of θ^I, then $K = [I_A, B]$, where I_A is the identity matrix with dimension corresponding to θ^A and*

$$B = [I_{\text{mis}}^A(\theta_0^A)]^{-1} \times \lim_{n\to\infty} \frac{1}{n} E^G \left[v_n^A(Z_{\text{com}}; \theta_0^A) \left\{ \frac{\partial}{\partial \theta^{I\setminus A}} \log p^I(Z_{\text{mis}}|Z_{\text{obs}}; \theta_0^I) \right\}^\top \right].$$

5. Biases in and Bounds on MI Variance Estimators

In a nutshell, the developments in Section 3 and Section 4 — especially (3.8), Theorem 2, and Theorem 3 — give us two key asymptotic decompositions:

$$\hat{\theta}_{\text{com}}^A - \theta_0^A = (I - F^A)(\hat{\theta}_{\text{obs}}^A - \theta_0^A) + F^A(\hat{\theta}_{\text{mis}}^A - \theta_0^A) \tag{5.1}$$

and

$$\bar{\theta}_\infty - \theta_0^A = (I - F)(\hat{\theta}_{\text{obs}}^A - \theta_0^A) + FK(\hat{\theta}_{\text{obs}}^I - \theta_0^I), \tag{5.2}$$

with all the quantities as previously defined. In particular, $\hat{\theta}_{\text{mis}}^A$ is asymptotically uncorrelated with either $\hat{\theta}_{\text{obs}}^A$ or $\hat{\theta}_{\text{obs}}^I$. These two decompositions apparently employ different measures of fraction of missing information, i.e., the F^A used in (3.8) and the F of (4.7). They differ in the definition of $I_{\text{mis}}(\theta_0) = J_v(\theta_0)$ (see (4.7)) because F uses the v_n defined by (4.5) but F^A uses the v_n^A of (3.7). However, a closer inspection of (4.5) and (3.7) reveals that the difference, for calculating I_{mis}, is only in the two (limiting) expectations of the same partial derivative,

$$\lim_{n\to\infty} \frac{1}{n} E^G \left[\frac{\partial}{\partial\theta} S^A(Z_{\text{com}}; \theta_0) \right] \text{ versus } \lim_{n\to\infty} \frac{1}{n} E^G \left[E^I \left(\frac{\partial}{\partial\theta} S^A(Z_{\text{com}}; \theta_0) \Big| Z_{\text{obs}}; \hat{\theta}_{\text{obs}}^I \right) \right]. \tag{5.3}$$

But under mild regularity conditions, the $\hat{\theta}_{\text{obs}}^I$ in the second expectation can be replaced by its limiting value θ_0 without affecting the limiting expectation. It follows that the limits in (5.3) are actually the same and therefore we will use F and F^A interchangeably.

Decomposition (5.2) allows us to derive the asymptotic variance of $\bar{\theta}_\infty$, and the two decompositions also permit us to understand and measure the bias in Rubin's T_∞, as shown below.

5.1. Bias in Rubin's variance estimator under uncongeniality

To apply (5.2), we need a stronger condition than SOR to ensure that the asymptotic variance of $\bar{\theta}_\infty$ is the variance of its asymptotic distribution.

Definition 3. Strong SOR. We say the estimating equation (3.1) satisfies Strong SOR if in addition to the SOR conditions given in Definition 1, we have

(vii) The estimator $\hat{\theta}_g$ is asymptotically normally distributed as

$$\sqrt{n}(\hat{\theta}_g - \theta_0) \xrightarrow{L^2} N(0, V_g),$$

where

$$V_g = \lim_{n\to\infty} J_g^{-1}(\theta_0) \times \left[\frac{V^G(g_n(Y; \theta_0))}{n} \right] \times \left[J_g^{-1}(\theta_0) \right]^\top.$$

This assumption guarantees the following result, first obtained by Robins and Wang (2000).

Corollary 3. *Under Strong SOR, the asymptotic variance of $\bar{\theta}_\infty$ is given by*

$$\begin{aligned} V_\infty = {} & (I - F)V_{\text{obs}}^A(I - F^\top) + FKV_{\text{obs}}^I K^\top F^\top + (I - F)C_{\text{obs}}^{A,I} K^\top F^\top \\ & + FK(C_{\text{obs}}^{A,I})^\top (I - F^\top) + o\left(n^{-1}\right) \end{aligned}$$

with V^A_{obs}, V^I_{obs} and $C^{A,I}_{\text{obs}}$ being the asymptotic variances of $\hat\theta^A_{\text{obs}}$ and of $\hat\theta^I_{\text{obs}}$, and their covariance.

The potential bias of Rubin's $T_\infty = \bar U_\infty + B_\infty$ is then best understood by expressing $\bar\theta_\infty = \hat\theta^A_{\text{com}} + (\bar\theta_\infty - \hat\theta^A_{\text{com}})$. It follows that

$$V_\infty \equiv V^G(\bar\theta_\infty) = V^G(\hat\theta^A_{\text{com}}) + V^G(\bar\theta_\infty - \hat\theta^A_{\text{com}}) - D_\infty - D_\infty^\top, \tag{5.4}$$

where $D_\infty = Cov^G(\hat\theta^A_{\text{com}}, \hat\theta^A_{\text{com}} - \bar\theta_\infty)$. Under the congeniality assumption (I)−(II) of Section 1.2, D_∞ vanishes, $\bar U_\infty$ and B_∞ are, respectively, consistent estimators of $V^G(\hat\theta^A_{\text{com}})$ and $V^G(\bar\theta_\infty - \hat\theta^A_{\text{com}})$, and hence Rubin's $T_\infty = \bar U_\infty + B_\infty$ is consistent for $V^G(\bar\theta_\infty)$.

Under uncongeniality, the consistency of either $\bar U_\infty$ or B_∞ is unaffected, as discussed below. However, the cross term D_∞ is no longer negligible as $n \to \infty$ because the orthogonality underlying "Total Variance = Within Variance + Between Variance" is destroyed by uncongeniality, as emphasized by Kott (1995). We can therefore adopt D_∞, or better, the correlation (matrix)

$$C_{un} = Corr^G(\hat\theta^A_{\text{com}}, \hat\theta^A_{\text{com}} - \bar\theta_\infty) \tag{5.5}$$

as a measure of uncongeniality with respect to estimand θ; see Section 8.2.

Intuitively, the consistency of $\bar U_\infty$ remains under uncongeniality because $\bar U_\infty$ is the average of the analyst's complete-data variance estimator, denoted by $U^A(Z_{\text{com}})$, calculated on each imputed data set. Therefore, as long as $U^A(Z_{\text{com}})$ is consistent for $V^G(\hat\theta^A(Z_{\text{com}}))$ in the sense that $U^A(Z_{\text{com}}) = V^G(\hat\theta^A(Z_{\text{com}})) + o_p(N^{-1})$, and that the imputer's model is correctly specified (which we always assume, but see Section 8.4), the consistency of $\bar U_\infty$ follows under mild regularity conditions. In addition, the decomposition (3.8) implies that

$$\bar U_\infty = (I - F)V^A_{\text{obs}}(I - F)^\top + FV^A_{\text{mis}}F^\top + o_p\left(n^{-1}\right), \tag{5.6}$$

where V^A_{obs} and V^A_{mis} denote respectively $V^G(\hat\theta^A_{\text{obs}})$ and $V^G(\hat\theta^A_{\text{mis}})$.

Connecting B_∞ with $V^G(\bar\theta_\infty - \hat\theta^A_{\text{com}})$ is a bit more involved because this is where we cannot use the approximation based on the plug-in predictive imputation, which would lead to under-dispersion in the imputation value. But for its intended estimand, $V^G(\bar\theta_\infty - \hat\theta^A_{\text{com}})$, (5.1) and (5.2) together imply that (asymptotically)

$$\bar\theta_\infty - \hat\theta^A_{\text{com}} = FK\hat\theta^I_{\text{obs}} - F\hat\theta^A_{\text{mis}} - F[K\theta^I_0 - \theta^A_0]. \tag{5.7}$$

Because $\hat\theta^I_{\text{obs}}$ and $\hat\theta^A_{\text{mis}}$ are asymptotically orthogonal, a consequence of Corollary 1, we have

$$V^G(\bar\theta_\infty - \hat\theta^A_{\text{com}}) = FKV^I_{\text{obs}}K^\top F^\top + FV^A_{\text{mis}}F^\top + o\left(n^{-1}\right), \tag{5.8}$$

where $V^I_{\text{obs}} = V^G(\hat{\theta}^I_{\text{obs}})$. The consistency of B_∞ is then established by proving that asymptotically the two terms on the righthand side of (5.8) correspond to the two steps in the construction of B_∞: the posterior draw of θ based on the imputation model and the imputed missing values given the draw θ; see Appendix I for details.

Given the consistency of both $\bar{U}_\infty$ and B_∞ under uncongeniality, assessing the bias in T_∞ amounts to calculating the D_∞ term in (5.4). Using the fact that $\hat{\theta}^A_{\text{mis}}$ is asymptotically uncorrelated with both $\hat{\theta}^I_{\text{obs}}$ and $\hat{\theta}^A_{\text{obs}}$ (again because of Corollary 1), we have

$$D_\infty = FV^A_{\text{mis}}F^\top - (I-F)C^{A,I}_{\text{obs}}K^\top F^\top, \tag{5.9}$$

where $C^{A,I}_{\text{obs}} = Cov^G(\hat{\theta}^A_{\text{obs}}, \hat{\theta}^I_{\text{obs}})$. The bias in T_∞ for estimating V_∞ is then given by

$$T_\infty - V_\infty = D_\infty + D_\infty^\top + o_p\left(n^{-1}\right). \tag{5.10}$$

5.2. A standard error combining rule under multi-phase paradigm

Evidently, the bias given in (5.10) is essentially impossible to eliminate in practice without further knowledge/assumptions about the nature of the uncongeniality. Requiring the analyst to investigate such issues would largely defeat the main purpose of MI, that is, to permit the analyst to reach valid inferences without having to deal with the missing data problem. The central question then is that, given $\{\bar{\theta}_\infty, U_\infty, B_\infty\}$ *only*, is it still possible to produce a statistical procedure that enjoys some validity regardless of the degree of uncongeniality? The answer is no if we insist on validity as defined by consistent variance estimators, as demonstrated previously. However, as discussed in Meng (1994) and Rubin (1996), in the context of constructing confidence intervals, confidence validity permits the actual coverage to exceed the nominal level (Neyman (1937)), and hence a conservative variance is accordingly acceptable.

Accepting such confidence validity, we can obtain a simple procedure that is valid for θ of any dimension—we simply double Rubin's variance estimator T_∞, which has a similar flavor as doubling the variance in Copas and Eguchi (2005) to deal with model mis-specification. To see this, we note that (5.6) and (5.8), respectively, imply

$$(I-F)V^A_{\text{obs}}(I-F)^\top \le \bar{U}_\infty \quad \text{and} \quad FKV^I_{\text{obs}}K^\top F^\top \le B_\infty, \tag{5.11}$$

where $\le$ is defined such that $A \le B$ for two squared matrices if $B - A$ is nonnegative definite. Using (5.11) and the simple fact that $V(X+Y) \le 2[V(X) + V(Y)]$, we have

$$V_\infty = V^G(\bar{\theta}_\infty) = V^G\left((I-F)\hat{\theta}^A_{\text{obs}} + FK\hat{\theta}^I_{\text{obs}}\right) \le 2(\bar{U}_\infty + B_\infty) = 2T_\infty. \tag{5.12}$$

To illustrate, consider again Example 1, where

$$\bar{U}_\infty(\sigma_0^2) = \frac{1}{N}\left[(1-f) + f\sigma_0^2\right] \quad \text{and} \quad B_\infty(\sigma_0^2) = \frac{f}{N}\left[\frac{f}{1-f} + \sigma_0^2\right]. \tag{5.13}$$

From (2.3), T_∞ over-estimates (under-estimates) $V_\infty = n^{-1}$ if and only if $\sigma_0^2 > 1$ ($\sigma_0^2 < 1$). Because both $\bar{U}_\infty(\sigma_0^2)$ and $B_\infty(\sigma_0^2)$ are monotone increasing functions of σ_0^2, whenever $T_\infty(\sigma_0^2) = \bar{U}_\infty(\sigma_0^2) + B_\infty(\sigma_0^2)$ overestimates, the amount of over-estimation is unbounded as σ_0^2 increases. However, whenever $T_\infty(\sigma_0^2)$ underestimates, $T_\infty(\sigma_0^2) \le V_\infty \le 2T_\infty(\sigma_0^2)$, and hence the normal confidence interval using $2T_\infty$ with nominal coverage 95% will have coverage between 95% and 99.5%, the latter corresponding to $1.96\sqrt{2} = 2.77$ standard deviations. This is a general result for doubling T_∞ *whenever it under-estimates.*

In the case of scalar estimands, we can reduce the conservativeness in (5.12). Specifically, suppose our estimand is $\phi = c\theta^A$,where c is a row vector (hence θ^A can be of any dimension) . Using the (scalar) inequality $V(X+Y) \le [\sqrt{V(X)} + \sqrt{V(Y)}]^2$, we have

$$V_\infty^\phi \equiv V(c\bar{\theta}_\infty) \le \left[\sqrt{c\bar{U}_\infty c^\top} + \sqrt{cB_\infty c^\top}\right]^2 \equiv \left[\sqrt{\bar{U}_\infty^\phi} + \sqrt{B_\infty^\phi}\right]^2, \tag{5.14}$$

which gives a tighter bound than (5.12).

These results lead to extremely simple procedures for MI under uncongeniality with a finite number of imputations, m. The "doubling-variance" rule (5.12) simply replaces T_m of (1.1) by $2T_m$. For the "Combining-Standard-Errors" rule for *univariate* T_m, we let

$$\tilde{T}_m = \left(\sqrt{\bar{U}_m} + \sqrt{B_m}\right)^2 + \frac{1}{m}B_m \tag{5.15}$$

and substitute $\tilde{T}_m$ for T_m in MI inference. Because the orthogonal ANOVA decomposition $V(\bar{\theta}_m) = V(\bar{\theta}_m - \bar{\theta}_\infty) + V(\bar{\theta}_\infty)$ holds regardless of uncongeniality (due to the fact that $E^I(\bar{\theta}_m|Z_{\text{obs}}) = \bar{\theta}_\infty$), we need only to add the B_m/m term to the finite-m counterpart of (5.14) in order to take into account the Monte Carlo error. However, the right-hand side of (5.15) is bounded above by $(\sqrt{\bar{U}_m} + \sqrt{(1+m^{-1})\,B_m})^2$, hence it is acceptable to take square root of the two terms in Rubin's original variance rule (1.1) when forming the standard error combining rule. But (5.15) is shaper, especially when the fraction of missing information is low.

Although doubling variance or adding up standard errors may be viewed as extremely conservative under the "God-versus-me" paradigm, we emphasize that under the multi-phase inference paradigm it is likely to be a necessary premium to insure us against the unknown degrees of uncongeniality. Of course, more

research is needed to investigate the general properties of these bounds, especially their effect on estimating fractions of missing information.

6. Self-efficiency and Strong Efficiency

Example 1 demonstrates the need to regulate the analyst's complete-data procedure in order to avoid the seemingly counter-intuitive phenomena that more data actually lead to less efficient estimators. Further demonstrations are given in the on-line Appendix III, as well as in Meng and Xie (2014). Such phenomena have been discussed in the literature, e.g., Chernoff (1983); Meng (2001, 2005), but their negative consequences are particularly pronounced for multi-phase inferences. Actually, we need to regulate the procedure further to prevent other counter-intuitive phenomena from happening, the idea of which is captured by the notion of self-efficiency (Meng (1994)) defined in terms Mean Squared Error (MSE), recast below for θ of arbitrary dimension.

Definition 4. Self-Efficiency. Let $Z_{\rm com}$ be a data set and $Z_{\rm obs}$ be a subset of $Z_{\rm com}$ created by a selection mechanism. An estimation procedure $\hat{\theta}(\cdot)$ for θ is said to be self-efficient (with respect to the selection mechanism) if, for any $\lambda \in (-\infty, \infty)$, $\hat{\theta}_{\rm com}$ dominates $\lambda\hat{\theta}_{\rm obs} + (1-\lambda)\hat{\theta}_{\rm com}$ in terms of MSE.

This concept of comparing two estimators can be generalized to the following notion.

Definition 5. Strong Efficiency. Suppose two estimators $\hat{\theta}_u$ and $\hat{\theta}_v$ of the same θ are $\sqrt{n}$-consistent in L^2, and their covariance is well defined. We say that $\hat{\theta}_u$ is (asymptotically) strongly more efficient than $\hat{\theta}_v$, denoted as $\hat{\theta}_u \succ \hat{\theta}_v$, if the orthogonality relationship $\hat{\theta}_u \perp (\hat{\theta}_v - \hat{\theta}_u)$ holds asymptotically,

$$Cov^G(\hat{\theta}_u, \hat{\theta}_v - \hat{\theta}_u) = o\left(n^{-1}\right). \tag{6.1}$$

When $\hat{\theta}_u$ is an estimator of a sub-vector of θ, $\hat{\theta}_u \succ \hat{\theta}_v$ is to be understood as $(\hat{\theta}_u, \theta_0^{v\backslash u})^\top \succ \hat{\theta}_v$, where $\theta_0^{v\backslash u}$ is the true value of the part of θ that is estimated by $\hat{\theta}_v$, but not by $\hat{\theta}_u$.

Given this definition, self-efficiency amounts to requiring $\hat{\theta}_{\rm com}$ be strongly more efficient than $\hat{\theta}_{\rm obs}$, when both of them are $\sqrt{n}$-consistent in L^2 . Strong efficiency is also closely related to Rao-Blackwellization. If $\hat{\theta}_u$ is a Rao-Blackwellized $\hat{\theta}_v$, $E(\hat{\theta}_v|S) = \hat{\theta}_u$, where S is a (correctly specified) sufficient statistic, then $\hat{\theta}_u$ must be strongly more efficient than $\hat{\theta}_v$, because $E^G(\hat{\theta}_u\hat{\theta}_v) = E^G[E(\hat{\theta}_v|S)E(\hat{\theta}_v|S)]$ implies $Cov^G(\hat{\theta}_u, \hat{\theta}_v - \hat{\theta}_u) = 0$. Conversely, letting $D = \hat{\theta}_v - \hat{\theta}_u$, we see that if the delta method is applicable to linearize $E^G(D|\hat{\theta}_u)$ as a function of $\hat{\theta}_u$, then (6.1) implies $V[E^G(D|\hat{\theta}_u)] \equiv Cov[D, E^G(D|\hat{\theta}_u)] = o(n^{-1})$. Consequently, $E^G(\hat{\theta}_v|\hat{\theta}_u) = c + \hat{\theta}_u + o_p(n^{-1/2})$, for some constance c, which must be zero when

both $\hat{\theta}_u$ and $\hat{\theta}_v$ are $\sqrt{n}$-consistent in L^2. Therefore, $\hat{\theta}_u$ can be viewed as a Rao-Blackwellization of $\hat{\theta}_v$ with $\hat{\theta}_u$ itself serving as the "sufficient statistic". It is in this sense we say $\hat{\theta}_u$ is *strongly more efficient* than $\hat{\theta}_v$. We remark that strong efficiency is not a total ordering and can even be non-transitive. Nevertheless, as we will show in Section 7, it is a useful concept for Rubin's MI inference under uncongeniality.

One subtlety in formulating self-efficiency lies in the definition of $\hat{\theta}_{\text{obs}}$. A principled method such as MLE or Bayesian analysis is applicable for all data patterns. However, for an arbitrary *complete-data procedure*, it is generally unclear what the *corresponding* procedure would be when the data are incomplete. In fact, mathematically speaking, the notation $\hat{\theta}(\cdot)$ in Definition 4 is ambiguous because of the varying dimension of its argument while moving from Z_{com} to Z_{obs}.

When an analyst adopts an EE $S^A(Z_{\text{com}};\theta)=0$, the corresponding observed-data EE can be defined as its projection

$$S^A(Z_{\text{obs}};\theta) \equiv E^A[S^A(Z_{\text{com}};\theta)|Z_{\text{obs}};\theta] = 0. \tag{6.2}$$

Caution is needed in defining this projection when there is a nuisance parameter ξ, in which case (6.2) may not be a meaningful EE for θ because the projection $E^A[S^A(Z_{\text{com}};\theta)|Z_{\text{obs}};\theta,\xi]$ can depend on ξ. This can cause problems; see Section 7.2.

As expected, a fully efficient *procedure* such as an MLE or a Bayes estimator is self-efficient, under the assumption that no new information is introduced to the missing-data mechanism that can improve the complete-data estimators (Rubin (1976)). Self-efficiency is a weaker requirement, as we will show shortly. Nevertheless, as Example 1 demonstrates, self-efficiency does exclude certain seemingly ideal estimators.

The orthogonality between $\hat{\theta}_{\text{com}}$ and $\hat{\theta}_{\text{obs}}-\hat{\theta}_{\text{com}}$ as rendered by self-efficiency plays a critical role in establishing some general theoretical results in Section 7. The following result is particularly useful in building insights regarding the behavior of T_∞.

Lemma 2. *Assume the analyst's procedure is self-efficient. Then*

$$2FV^A_{\text{mis}}F^\top = (I-F)V^A_{\text{obs}}F^\top + FV^A_{\text{obs}}(I-F)^\top + o\left(n^{-1}\right). \tag{6.3}$$

Consequently, (5.9) *is simplified to*

$$D_\infty = (I-F)\left[V^A_{\text{obs}} - C^{A,I}_{\text{obs}}K^\top\right]F^\top. \tag{6.4}$$

Thus, T_∞ *is* (*asymptotically*) *unbiased if and only if* (*assuming* F *and* $I-F$ *are of full rank*)

$$V^A_{\text{obs}} = C^{A,I}_{\text{obs}}K^\top + o_p\left(n^{-1}\right). \tag{6.5}$$

The proof of this lemma relies on the following characterization of EEs that produce self-efficient estimators. For simplicity below, we write $E^G(a) = E^G[a(Y;\theta_0)]$.

Theorem 4. *Suppose a complete-data estimator $\hat{\theta}_{\text{com}}$ is given by $h_{\text{com}}(Z_{\text{com}};\theta) = 0$, and the observed-data estimator $\hat{\theta}_{\text{obs}}$ is given by $h_{\text{obs}}(Z_{\text{obs}};\theta) = 0$, and that both $h_{\text{com}}(Z_{\text{com}};\theta)$ and $h_{\text{obs}}(Z_{\text{obs}};\theta)$ satisfy SOR. Then the corresponding estimating procedure is self-efficient (asymptotically) if and only if*

$$\left[E^G\left(\frac{\partial h_{\text{obs}}}{\partial\theta}\right)\right]^{-1} E^G\left(h_{\text{obs}}h_{\text{com}}^\top\right) = \left[E^G\left(\frac{\partial h_{\text{com}}}{\partial\theta}\right)\right]^{-1} E^G\left(h_{\text{com}}h_{\text{com}}^\top\right) + o(1), \tag{6.6}$$

where all h functions and their derivatives are evaluated at $\theta = \theta_0$.

The following is a class of EEs that satisfies (6.6), yet it does not necessarily lead to fully efficient estimators. Let $Z_{\text{com}} = (Y_1, \ldots, Y_N)$ be a sequence of i.i.d. random variables and $Z_{\text{obs}} = (Y_1, \ldots, Y_n)$. Choose h such that the EEs

$$h_{\text{obs}}(Z_{\text{obs}};\theta) = \sum_{i=1}^{n} h(Y_i;\theta) \quad \text{and} \quad h_{\text{com}}(Z_{\text{com}};\theta) = \sum_{i=1}^{N} h(Y_i;\theta) \tag{6.7}$$

satisfy SOR. An additional property the EEs in (6.7) enjoy is that $E^G[h_{\text{obs}}(h_{\text{com}} - h_{\text{obs}})^\top] = 0$. The following corollary gives a characterization of self-efficient EEs in such cases.

Corollary 4. *If in addition to the conditions stated in Theorem 4, we have that*

$$\left[E^G\left(\frac{\partial h_{\text{obs}}}{\partial\theta}\right)\right]^{-1} E^G\left[h_{\text{obs}}(h_{\text{com}} - h_{\text{obs}})^\top\right] = o(1)\ ,$$

then the corresponding estimating procedure is self-efficient (asymptotically) if and only if

$$\left[E^G\left(\frac{\partial h_{\text{obs}}}{\partial\theta}\right)\right]^{-1} E^G\left(h_{\text{obs}}h_{\text{obs}}^\top\right) = \left[E^G\left(\frac{\partial h_{\text{com}}}{\partial\theta}\right)\right]^{-1} E^G\left(h_{\text{com}}h_{\text{com}}^\top\right) + o(1). \tag{6.8}$$

When we adopt MLEs, the above equality can be shown by noticing that

$$\left[E\left(\frac{\partial h_{\text{obs}}}{\partial\theta}\right)\right]^{-1} E\left(h_{\text{obs}}h_{\text{obs}}^\top\right) = I = \left[E\left(\frac{\partial h_{\text{com}}}{\partial\theta}\right)\right]^{-1} E\left(h_{\text{com}}h_{\text{com}}^\top\right), \tag{6.9}$$

where I is the identity matrix, because of the familiar second Bartlett identity for likelihood. In (6.9) the expectation operator E is with respect to the same class of models that underlies the likelihood, and hence it does not have the superscript G on it. Therefore, we can view (6.8) and, more generally, (6.6), as

generalizations of the second Bartlett identity for EEs, when we replace their E^G operator by E. The value of Theorem 4 is that it provides a practical way to construct, identify, and verify self-efficient estimation procedures.

7. General Behavior of Rubin's Variance Combining Rule

In this section we first establish a necessary and sufficient condition for the consistency of Rubin's variance estimator. We then investigate two contrasting scenarios: (i) the analyst's estimator dominates the imputer's in terms of strong efficiency, $\hat{\theta}^A_{\text{obs}} \succ \hat{\theta}^I_{\text{obs}}$, (ii) the imputer's estimator dominates the analyst's, $\hat{\theta}^I_{\text{obs}} \succ \hat{\theta}^A_{\text{obs}}$. Given these scenarios and assuming self-efficiency of the analyst's procedure, we identify circumstances where Rubin's variance estimator is consistent or conservative. The validity of Theorem 5 relies on that of the previous theorems and corollaries, whose regularity conditions are needed but not repeated below for brevity.

Theorem 5. *The MI variance estimator T_∞ is consistent for V_∞ if and only if $\hat{\theta}^A_{\text{com}} \succ \bar{\theta}_\infty$.*

The proof of this result requires the regularity conditions for Theorem 2 and (5.10), but the intuitive argument is rather immediate. By definition, $\hat{\theta}^A_{\text{com}}$ is (asymptotically) strongly more efficient than $\bar{\theta}_\infty$ if and only if we can write (asymptotically)

$$V_\infty = V^G(\bar{\theta}_\infty) = V^G(\hat{\theta}^A_{\text{com}}) + V^G(\bar{\theta}_\infty - \hat{\theta}^A_{\text{com}}). \tag{7.1}$$

But as discussed in Section 5.1, the two terms on the right-hand side of (7.1) are consistently estimated, respectively, by $\bar{U}_\infty$ and by B_∞, hence the consistency of T_∞.

Whereas technically verifying $\hat{\theta}^A_{\text{com}} \succ \bar{\theta}_\infty$ typically is no easier than directly verifying the consistency of T_∞, Theorem 5 leads to an important insight. It establishes that if the analyst uses fully efficient *complete-data* estimators, such as MLE, then T_∞ will be consistent as long as the imputer's model does not bring in "secret information" unused in forming the analyst's complete-data estimator (e.g., such as the imputer's information on equal means in Example 3). This result therefore provides a concrete two-part practical guideline: (A) the analyst should adopt a fully efficient estimator (under the analyst's model) as much as is possible, and (B) the imputer should employ an imputation model that is as saturated as feasible. Point (B) has been well emphasized throughout the MI literature (e.g., Rubin (1987, 1996), and Meng (1994)), but point (A) has received much less emphasis.

This imbalance seems to be due to the desire to allow users the complete freedom to apply their favorite methods to the imputed data sets. Just as in

other cases of robustness and efficiency trade-off, there is a price to be paid for this "applicability robustness", namely, the potential bias in T_∞. However, in the context of variance estimation, it is often acceptable or even preferable to have a certain degree of over-estimation as a way to mitigate the negative effect from the usual tendency of under-assessing the overall uncertainty (e.g., due to model uncertainty). Somewhat remarkably, as demonstrated below, as long as the analyst's procedure is self-efficient, T_∞ has a tendency to overestimate rather than to underestimate. The rationale for invoking self-efficiency has been discussed previously, but its relevance can also be seen in Theorem 5, which obviously is applicable in the special case of congeniality. However, under congeniality, $\bar{\theta}_\infty = \hat{\theta}^A_{\text{obs}}$, and hence $\hat{\theta}^A_{\text{com}} \succ \bar{\theta}_\infty$ is the same as requiring the analyst's procedure to be self-efficient.

7.1. The scenario where the analyst assumes more than the imputer

Moving beyond the congenial case, we first discuss the circumstance in which the analyst makes more assumptions than the imputer so that $\hat{\theta}^A_{\text{obs}} \succ \hat{\theta}^I_{\text{obs}}$. This occurs, for example, when the analyst's model is nested within the imputer's, and both parties adopt the MLE approach. In such cases, θ^A is a sub-parameter of θ^I, and hence $\hat{\theta}^A_{\text{obs}} \succ \hat{\theta}^I_{\text{obs}}$ is understood as having $\{\hat{\theta}^A_{\text{obs}}, \theta_0^{I\setminus A}\}$ dominate $\hat{\theta}^I_{\text{obs}}$, as discussed before. Theorem 6 below states that if in addition the analyst's procedure is self-efficient, then Rubin's T_∞ is consistent and hence confidence proper. At the same time, because the additional (correct) information assumed by the analyst is unknown/unknowable to the imputer, the MI estimator $\bar{\theta}_\infty$ cannot be more efficient than the analyst's estimator based directly on the observed data, $\hat{\theta}^A_{\text{obs}}$ (e.g., the observed-data MLE). Again, the validity of Lemma 3 and Theorem 6 requires the (unlisted) conditions underlying the previous results.

Lemma 3. $\hat{\theta}^A_{\text{obs}} \succ \hat{\theta}^I_{\text{obs}}$ *implies* $\hat{\theta}^A_{\text{obs}} \succ \hat{\theta}^H_{\text{obs}}$.

Theorem 6. *Assuming that* $\hat{\theta}^A_{\text{obs}} \succ \hat{\theta}^I_{\text{obs}}$, *we have*

(i) *Rubin's variance estimator* T_∞ *is consistent if the analyst's procedure is self-efficient.*

(ii) *The MI estimator* $\bar{\theta}_\infty$ *cannot be more efficient than* $\hat{\theta}^A_{\text{obs}}$, $V_\infty \geq V^A_{\text{obs}}$.

This result provides a general theoretical guarantee of the "hidden robustness" phenomenon illustrated in Section 2.2. Because strong efficiency is about orthogonality, the geometric insight for (i) of Theorem 6 is as follows. First, Theorem 1 allows us to place the relevant estimators in a Euclidean space as in Figure 3. Second, respectively, Corollary 1, Lemma 3 and the self-efficiency of

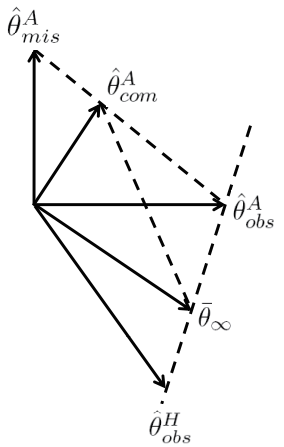

Figure 3. A geometric interpretation for Theorem 6.

$\hat{\theta}^A_{\text{com}}$ imply (1), (2) and (3) below, leading to

$$\left.\begin{array}{l}(1)(\hat{\theta}^A_{\text{obs}}-\hat{\theta}^H_{\text{obs}})\perp\hat{\theta}^A_{\text{mis}}\\(2)(\hat{\theta}^A_{\text{obs}}-\hat{\theta}^H_{\text{obs}})\perp\hat{\theta}^A_{\text{obs}}\end{array}\right\}\Rightarrow(\hat{\theta}^A_{\text{obs}}-\hat{\theta}^H_{\text{obs}})\perp\hat{\theta}^A_{\text{com}}$$

$$\left.\begin{array}{l}\\(3)\ \hat{\theta}^A_{\text{com}}\succ\hat{\theta}^A_{\text{obs}}\end{array}\right\}\Rightarrow(\hat{\theta}^A_{\text{obs}}-\hat{\theta}^A_{\text{com}})\perp\hat{\theta}^A_{\text{com}}$$

$$\Rightarrow \text{Span}\{\hat{\theta}^A_{\text{obs}}-\hat{\theta}^H_{\text{obs}},\hat{\theta}^A_{\text{obs}}-\hat{\theta}^A_{\text{com}}\}\perp\hat{\theta}^A_{\text{com}}.$$

Thus, $\hat{\theta}^A_{\text{com}}$ is perpendicular to the plane formed by the three dashed lines in Figure 3. It is then clear that $\hat{\theta}^A_{\text{com}}\succ\bar{\theta}_\infty$, implying that T_∞ is consistent because of Theorem 5.

7.2. The scenario where the imputer assumes more than the analyst

Next we assume $\hat{\theta}^I_{\text{obs}}\succ\hat{\theta}^A_{\text{obs}}$. This holds, for example, when both analyst and imputer adopt MLEs, but with the imputer's model being a sub-model of the analyst's. Unlike in Theorem 6, which guarantees that T_∞ is consistent, here we can only guarantee that it is conservative, and even this less precise result is obtained under the restrictive assumption that the fraction of missing information is the same for all parameters.

Theorem 7. *Assuming that* $\hat{\theta}^I_{\text{obs}}\succ\hat{\theta}^A_{\text{obs}}$, *we have the following asymptotical results.*

(i) *The MI estimator* $\bar{\theta}_\infty$ *is no less efficient than the analyst's estimator* $\hat{\theta}^A_{\text{obs}}$, $V_\infty\le V^A_{\text{obs}}$.

(ii) *If we also assume the analyst's procedure is self-efficient, then* $T_\infty\le V^A_{\text{obs}}$.

(iii) *In addition to these conditions, if the fraction of missing information matrix* F *defined in* (4.7) *is proportional to an identity matrix, then* T_∞ *is conservative,* $V_\infty \le T_\infty$.

Theorem 7 provides a theoretical backbone for the "super efficiency" phenomenon illustrated in Example 3 . There is, however, a subtlety in applying Theorem 7 to that example, because the F there is generally not proportional to the identity matrix I_2 when we consider $\theta = (\theta_x, \theta_y)^\top$ as the analyst's parameter. However, we can apply Theorem 7 to θ_x and θ_y separately because the $u_n(Z_{\text{obs}}; \theta)$ of (4.4) can be decomposed into two "self-contained" projections for estimating θ_x and θ_y individually:

$$E^A\left[S(Z_{\text{com}};\theta)\middle|\, Z_{\text{obs}};\theta\right] = \begin{pmatrix} E^A\left[S(Z_{\text{com}};\theta_x)\middle|\, Z_{\text{obs}};\theta_x\right] \\ \\ E^A\left[S(Z_{\text{com}};\theta_y)\middle|\, Z_{\text{obs}};\theta_y\right] \end{pmatrix} = \begin{pmatrix} E^A\left[\bar{X}_{n_x} - \theta_x\middle|\, X_{\text{obs}};\theta_x\right] \\ \\ E^A\left[\bar{Y}_{n_y} - \theta_y\middle|\, Y_{\text{obs}};\theta_y\right] \end{pmatrix}. \tag{7.2}$$

However, for $\phi = \theta_x + \theta_y$, the complete-data EE is

$$S^A(X_{\text{com}}, Y_{\text{com}}; \phi) = \bar{X}_{N_x} + \bar{Y}_{N_y} - \phi = 0, \tag{7.3}$$

for which the "self-contained" projection $E^A[S^A(X_{\text{com}}, Y_{\text{com}}; \phi)|X_{\text{com}}, Y_{\text{com}}, \phi]$ does not exist under the analyst's model, because when $f_x \neq f_y$, the required conditional expectation will depend on individual values of θ_x and θ_y. Hence, Theorem 7 is inapplicable.

Technically, one may be tempted to get around this problem by re-defining the projection in (4.4)−(4.5), replacing the E^A operator by a projection that is consistent with God's model. Recall for Example 3, God's model assumes $\theta_x = \theta_y$ and hence both of them equal to $\phi/2$. Using such a projection, E, we have

$$E[S^A(X_{\text{com}}, Y_{\text{com}}; \phi)|X_{\text{obs}}, Y_{\text{obs}}; \phi] = \frac{n_x\bar{X}_{n_x} + (N_x - n_x)\phi/2}{N_x} + \frac{n_y\bar{Y}_{n_y} + (N_y - n_y)\phi/2}{N_y} - \phi,$$

which, upon being set to zero, leads to the observed-data estimator

$$\hat{\phi}_{\text{obs}} = \frac{2[(1-f_x)\bar{X}_{n_x} + (1-f_y)\bar{Y}_{n_y}]}{(1-f_x) + (1-f_y)}. \tag{7.4}$$

This re-definition of (4.4)−(4.5) however does not solve but only postpones our problem. This is because the "self-efficiency" condition $\hat{\phi}_{\text{com}} \succ \hat{\phi}_{\text{obs}}$ no longer holds whenever $f_x \neq f_y$. We must put "self-efficiency" in quotes here because

the construction of $\hat{\phi}_{\text{obs}}$ violates the spirit of the original formulation of self-efficiency, since the derivation of (7.4) has used the imputer's knowledge $\theta_x = \theta_y$, unavailable to the analyst. This also explains that, although $\hat{\phi}_{\text{com}}$ is the MLE of ϕ under the analyst's model, it does *not* dominate $\hat{\phi}_{\text{obs}}$ precisely because $\hat{\phi}_{\text{com}}$ is not the MLE under the additional assumption that $\theta_x = \theta_y$.

The proportionality assumption in (iii) of Theorem 7 holds when the missing data are missing completely at random (MCAR; see Rubin (1976)) in the regression example given in Appendix II. Exploring such connections in general is one of many open problems that are worth investigating. Even though this assumption appears to be rather restrictive, it is only a sufficient instead of necessary condition.

8. Subtleties and Open Problems

The results reported so far have helped us to decipher the complex behavior of Rubin's MI inference under uncongeniality, providing a valuable exploration of the multi-phase inference paradigm. Nevertheless, our expedition into the uncongenial forest has encountered several "subtlety traps." We share some stories here in hoping to entice readers to join our adventure.

8.1. Example 4: Complications with multi-phase inference

In discussing Theorem 5, we made a point that self-efficiency is a sufficient and necessary condition for T_∞ to be consistent under congeniality. However, neither self-efficiency nor congeniality is a necessary condition for T_∞ to be consistent in general, because the lack of self-efficiency can be somehow compensated by uncongeniality. To demonstrate this "two wrongs make a right" complexity of multi-phase inference, we follow the setting of Example 1 but replace $N(\theta, \sigma^2)$ by Laplace $L(\theta, \tau) : p(y|\theta, \tau) = (1/2\tau) \exp(-|y - \theta|/\tau)$.

- *God's Model*: $Z_{\text{obs}} = (Y_1, \ldots, Y_n)$ with $Y_i \overset{i.i.d.}{\sim} L(\theta_0, 1)$ for $i = 1, \ldots, n$, and $Z_{\text{mis}} = (Y_{n+1}, \ldots, Y_N)$ with $Y_i \overset{i.i.d.}{\sim} L(\theta_0, \tau_0)$ for $i = n + 1, \ldots, N$.
- *Imputer's Model*: $Y_i \overset{i.i.d.}{\sim} L(\theta, 1)$ for $i = 1, \ldots, n$ and $Y_i \overset{i.i.d.}{\sim} L(\theta, \tau_0)$ for $i = n+1, \ldots, N$; prior $p(\theta) \propto 1$; MI draws are obtained by first sampling $\tilde{\theta}$ from $p(\theta|Z_{\text{obs}}) \propto \exp(-\sum_{i=1}^{n} |y_i - \theta|)$, and then sampling $\tilde{Y}_i \overset{i.i.d.}{\sim} L(\tilde{\theta}, \tau_0)$ for $i = n + 1, \ldots, N$. For asymptotic calculations, we use $p(\theta|Z_{\text{obs}}) \approx N(Y_{(n/2)}, n^{-1})$, where $Y_{(n/2)}$ is the median of $\{Y_1, \ldots, Y_n\}$, which is the MLE of θ under $L(\theta, \tau)$.
- *Analyst's Complete-data Procedure:* $\hat{\theta}^A_{\text{com}} = \bar{Y}_N$ and $\hat{V}^A_{\text{com}} = \hat{V}(\hat{\theta}^A_{\text{com}}) = N^{-1} S^2_N$.

As in Example 1, the analyst's procedure is not self-efficient whenever $\tau_0 \neq 1$. Nor could it be congenial to the imputer's model because the MLE of θ is the sample median, which is not asymptotically equivalent to the sample mean. We thus have $\hat{\theta}^A_{\text{obs}} = \bar{Y}_n$, $\hat{\theta}^I_{\text{obs}} = \hat{\theta}^H_{\text{obs}} = Y_{(n/2)}$, and $\bar{\theta}_\infty = (1-f)\hat{\theta}^A_{\text{obs}} + fK\hat{\theta}^I_{\text{obs}}$ with $f = (N-n)/N$ and $K = 1$, verifying Theorems 2 and 3.

Because here $\hat{\theta}^I_{\text{obs}}$ is the MLE under the correct model, we have $\hat{\theta}^I_{\text{obs}} \succ \hat{\theta}^A_{\text{obs}}$, in the realm of Theorem 7. To verify its (i), we note the variance of $L(\theta, \tau)$ is $2\tau^2$, and hence $V^A_{\text{obs}} = V^G(\bar{Y}_n) = 2/n$. For $V_\infty \equiv V^G(\bar{\theta}_\infty)$, the expression of $\bar{\theta}_\infty$ above implies that asymptotically

$$V_\infty = \frac{1}{n}[1 + (1-f)^2] \leq \frac{2}{n} = V^A_{\text{obs}}. \tag{8.1}$$

Formula (8.1) verifies (i) of Theorem 7, and it also shows that the efficiency of $\bar{\theta}_\infty$ is an *increasing* function of the fraction of missing data f for fixed n. This is no surprise because as f increases, the percentage (f) of the imputer's $\hat{\theta}^I_{\text{obs}} = Y_{(n/2)}$ in $\bar{\theta}_\infty$ increases while the percentage $(1-f)$ of the doubly more variable $\hat{\theta}^A_{\text{obs}} = \bar{Y}_n$ decreases, effectively reducing the impact of a defect in $\hat{\theta}_{\text{com}}$.

Because the first two-moment calculations of $N(\theta, \sigma^2)$ and $L(\theta, \tau)$ are identical once we equate $\sigma^2 = 2\tau^2$, we can obtain the current T_∞ from (2.2), after substituting 1 by 2 for the variance of any observed Y_i's and σ_0^2 by $2\tau_0^2$ for any unobserved Y_i's; that is,

$$T_\infty = \bar{U}_\infty + B_\infty = \frac{1}{N}\left[2(1-f) + 2f\tau_0^2\right] + \frac{f}{N}\left[\frac{f}{1-f} + 2\tau_0^2\right]. \tag{8.2}$$

From (8.1)−(8.2), simple algebra yields

$$n(V^A_{\text{obs}} - T_\infty) = f^2 + 4f(1-f)(1-\tau_0^2); \tag{8.3}$$

$$n(T_\infty - V_\infty) = 2f(1-f)(2\tau_0^2 - 1). \tag{8.4}$$

Therefore, without further conditions on τ_0^2, the conclusion is that neither (ii) nor (iii) of Theorem 7 can hold. However, (ii) of Theorem 7 also assumes that the analyst's procedure is self-efficient, which means $\tau_0^2 = 1$ in the current case. But when $\tau_0^2 = 1$, (8.3) and (8.4) are non-negative, verifying both (ii) and (iii) of Theorem 7. Also, as expected and following a similar inequality as in (2.7), the inequality (2.6) is verified directly, demonstrating again its applicability.

Less expected is that when $\tau_0^2 = 1/2$, $T_\infty = V_\infty$ even though $\tau_0^2 = 1/2$ corresponds to neither self-efficiency nor congeniality. We do not have an insight why $\tau_0^2 = 1/2$ is special, other than the observation that $\tau_0^2 = 1/2$ leads to the posterior predictive variance $1 + n^{-1}$ (asymptotically), the same as that under congeniality for the normal setting in Example 1. It seems to suggest an

"effective" congeniality of some sort, though currently we are unable to ascertain what it is.

8.2. Measuring uncongeniality and the use of partial knowledge

Since in practical situations, the analysis model happens to be congenial to the imputer's model is an event with very small (zero?) probability, the central question of interest is not much about detecting whether uncongenality has occurred (it has!), but rather to what degree. The Example 4 above indicates that measuring uncongeniality is both an important and challenging task, even for a univariate estimand (which we assume here).

For Example 4, the uncongeniality index (for estimating θ) defined in (5.5) is given by (due to (8.2) and (8.4))

$$C_{un} = \frac{\eta}{\sqrt{(2-f)/f+\eta}\sqrt{1/(1-f)+\eta}}, \quad \text{where} \quad \eta = 2\tau_0^2 - 1. \tag{8.5}$$

We see that as τ_0^2 varies from 0 to ∞, η varies from -1 to ∞ , and C_{un} moves monotonically from $-1/\sqrt{2}$ to 1. Here we can use C_{un} to index the degree of the bias in Rubin's variance combining rule, as well as our standard error combining rule (5.15) (with $m = \infty$). In general, (5.4)$-$(5.5) and the consistency of $\bar{U}_\infty$ and B_∞ imply that asymptotically we have

$$V_\infty = \bar{U}_\infty + B_\infty - 2C_{un}\sqrt{\bar{U}_\infty B_\infty} = T_\infty - 2C_{un}\sqrt{\bar{U}_\infty B_\infty}. \tag{8.6}$$

Therefore, as long as $C_{un} \geq 0$, Rubin's T_∞ will overestimate V_∞, with $C_{un} = 1$ indexing the most extreme overestimation, for then V_∞ reaches its lower bound given by

$$(\sqrt{\bar{U}_\infty} - \sqrt{B_\infty})^2 \leq V_\infty \leq (\sqrt{\bar{U}_\infty} + \sqrt{B_\infty})^2. \tag{8.7}$$

Similarly, when $C_{un} < 0$, T_∞ underestimates with $C_{un} = -1$ representing the extreme bias, although in the case of (8.5), the upper bound in (8.7) is unreachable because $C_{un} \geq -1/\sqrt{2}$.

This example indicates the possibility for the analyst to derive a lower (or upper) bound on C_{un} by examining extreme cases of uncongeniality without full knowledge of the imputer's model. Here bounds are derived using only the trivial knowledge that $0 \leq \tau_0^2 < \infty$ (in addition to the form of C_{un}). Such partial knowledge can help to reduce the confidence over-coverage. Knowing $C_{un} > C_{\min}$ permits us to replace our standard error combining rule by

$$\tilde{S}_\infty = \left[\bar{U}_\infty + B_\infty - 2C_{\min}\sqrt{\bar{U}_\infty B_\infty}\right]^{1/2} = R\left(\sqrt{\bar{U}_\infty} + \sqrt{B_\infty}\right), \tag{8.8}$$

where

$$R = \left[1 - \frac{1 + C_{\min}}{2}\gamma\right]^{1/2} \quad \text{with } \gamma = \frac{4\sqrt{\bar{U}_\infty B_\infty}}{\left(\sqrt{\bar{U}_\infty} + \sqrt{B_\infty}\right)^2} \tag{8.9}$$

is a deflation factor on the over-estimation (for the worst possible scenario) by the standard error combining rule. Because $\gamma \in [0, 1], R \in [0, 1]$. For our current example, we can show that for a given f, γ reaches its maximal possible value 1 when $\tau_0^2 \to \infty$, but its minimal possible value is bounded away from zero, as it is achieved when $\tau_0^2 = 0$, yielding

$$\gamma_{\min} = 4\sqrt{2}\left[\sqrt{2\frac{1-f}{f}} + \sqrt{\frac{f}{1-f}}\right]^{-2}.$$

Consequently, recalling $C_{\min} = -1/\sqrt{2}$, we see for any given f

$$\left[\frac{1+\sqrt{2}}{2\sqrt{2}}\right]^{1/2} = \left[\frac{1-C_{\min}}{2}\right]^{1/2} \le R \le \left[1 - \frac{1 + C_{\min}}{2}\gamma_{\min}\right]^{1/2} = \left[1 - \frac{\sqrt{2}-1}{2\sqrt{2}}\gamma_{\min}\right]^{1/2}. \tag{8.10}$$

Therefore, when, say, $f = 1/3$, $\gamma_{\min} = 0.772$, leading to $0.924 \le R \le 0.942$.

This calculation demonstrates a stability of an apparently very conservative rule, seeing how close R is to 1. This is not to suggest extrapolation (from a toy example) but rather exploration. We need to explore general methods for estimating the degree of uncongeniality, how such estimates can be used with (8.8), how the finite number of imputation enters the picture, etc. Another practical issue needs be addressed is how conservativeness in variance estimation affects our estimates of fraction of missing information, which is a key part in deriving approximate distributions for confidence intervals and hypothesis testing; see, for example, Li et al. (1991); Li, Raghunathan, and Rubin (1991), Meng and Rubin (1992), Barnard and Rubin (1999).

8.3. Difficulties with non-nested cases

Our theoretical results, up to Section 6, do not require any assumption about the relationship between the imputer's model and analysis model. However, we have not been able to obtain theoretical results that can render statistical acumen, such as those given in Section 7, without making assumptions about this relationship. This is particularly frustrating for us, because strong efficiency is a strong condition, albeit in cases of nested models, it is often satisfied, as in Examples 2-3. (But Example 4 shows that model nesting is not a necessary condition for strong efficiency dominance.) Here we present a simple case of non-nested models to illustrate the difficulties.

Consider a regression setting $Y = X_1\theta_1 + X_2\theta_2 + \epsilon$ with $\epsilon \sim N(\mathbf{0}, I_N)$, where the covariates X are fully observed but the response Y is missing at random (MAR), i.e., the missing-data mechanism can depend on X but not on Y itself. Without loss of generality, we write

$$\left.\begin{array}{ccc} Y_1 : X_{1,1} & X_{1,2} \\ \vdots \quad & \vdots \quad & \vdots \\ Y_n : X_{n,1} & X_{n,2} \end{array}\right\} X_{\text{obs}}; \qquad \left.\begin{array}{ccc} ? : X_{n+1,1} & X_{n+1,2} \\ \vdots \quad & \vdots \quad & \vdots \\ ? : \ X_{N,1} & X_{N,2} \end{array}\right\} X_{\text{mis}},$$

where "?" indicates that the value is missing. That is, only the first n of $\{Y_i, i = 1, \ldots, N\}$ are observed.

Assume that God's model here is the null model, with both θ_1 and θ_2 are zero, whereas the analyst sets $\theta_2 = 0$ and $\theta^A = \theta_1$, but the imputer adopts $\theta_1 = 0$ and $\theta^I = \theta_2$. The estimators $\hat{\theta}^A_{\text{obs}}$ and $\hat{\theta}^I_{\text{obs}}$ are taken to be the least square estimator under the corresponding model. Because we assume MAR, the plug-in predictive imputation model is simply the linear regression model $Y = X^I_{\text{mis}}\hat{\theta}^I_{\text{obs}} + \epsilon$, and hence the averaged/projected estimating equation is

$$\left(\sum_{i=1}^{n} x_{i,1}y_i + \sum_{i=n+1}^{N} x_{i,1}x_{i,2}\hat{\theta}^I_{\text{obs}}\right) - \left(\sum_{i=1}^{N} x^2_{i,1}\right)\theta^A = 0 \ ,$$

which can be decomposed as

$$\left[\left(\sum_{i=1}^{n} x_{i,1}y_i\right) - \left(\sum_{i=1}^{n} x^2_{i,1}\right)\theta^A\right] + \left[\left(\sum_{i=n+1}^{N} x_{i,1}x_{i,2}\hat{\theta}^I_{\text{obs}}\right) - \left(\sum_{i=n+1}^{N} x^2_{i,1}\right)\theta^A\right] = 0.$$

The corresponding decomposition of $\bar{\theta}_\infty$ is $\bar{\theta}_\infty = (1-f)\hat{\theta}^A_{\text{obs}} + fK\hat{\theta}^I_{\text{obs}}$ where

$$f = \frac{\sum_{i=n+1}^{N} x^2_{i,1}}{\sum_{i=1}^{N} x^2_{i,1}} \quad \text{and} \quad K = \frac{\sum_{i=n+1}^{N} x_{i,1}x_{i,2}}{\sum_{i=n+1}^{N} x^2_{i,1}},$$

which verifies Theorems 2 and 3. Clearly f is the fraction of missing information under the analyst's model because it is the same as $1 - [V(\hat{\theta}^A_{\text{com}})/V(\hat{\theta}^A_{\text{obs}})]$, the relative loss of precision (the reciprocal of variance) due to missing Y's. And K is the project coefficient from θ^I to θ^A because it is the regression coefficient estimator, $\hat{\beta}^{(2,1)}_{\text{mis}}$, from regressing X_2 on X_1 using these units with missing $Y's$. To evaluate the bias of T_∞, simple algebra yields

$$T_\infty - V_\infty = 2f(1-f)\left[\left(\sum_{i=1}^{n} x^2_{i,1}\right)^{-1} - K\frac{\sum_{i=1}^{n} x_{i,1}x_{i,2}}{\sum_{i=1}^{n} x^2_{i,1}}\left(\sum_{i=1}^{n} x^2_{i,2}\right)^{-1}\right] \quad (8.11)$$

$$= 2f(1-f)[V^G(\hat{\theta}^A_{\text{obs}}) - \hat{\beta}^{(2,1)}_{\text{mis}}\hat{\beta}^{(2,1)}_{\text{obs}}V^G(\hat{\theta}^I_{\text{obs}})], \quad (8.12)$$

where $\hat{\beta}_{\text{obs}}^{(2,1)}$ is the counterpart of $\hat{\beta}_{\text{mis}}^{(2,1)}$ but it is based on those units where Y is observed. This verifies (5.10), but it is less clear how to use it to generate practical advice for either the imputer or analyst. But it does imply that in the case X_1 and X_2 are uncorrelated either among the sub-population where Y's is observed ($\hat{\beta}_{\text{obs}}^{(2,1)} = 0$) or where it is missing ($\hat{\beta}_{\text{mis}}^{(2,1)} = 0$), Rubin's T_∞ will be conservative. This suggests that the confidence validity results in Section 7 can be generalized to the non-nested cases under additional conditions, but currently it is not clear what these conditions are other than the tautological ones such as directly assuming $V^G(\hat{\theta}_{\text{obs}}^A) \geq \hat{\beta}_{\text{obs}}^{(2,1)}\hat{\beta}_{\text{mis}}^{(2,1)}V^G(\hat{\theta}_{\text{obs}}^I)$. In contrast, as shown in Appendix II using a general regression setting, for nested models, we will have either $T_\infty - V_\infty = 0$ (when the covariates used in analysis model form a subset of those used in the imputation model) or $T_\infty - V_\infty \geq 0$ (when the covariates used in imputer's model is a subset of those used in the analysis model and when the fraction of missing information is the same for all components of Y), verifying the general results in Section 7.

8.4. Ultimate challenges

An even harder problem is to obtain general results when the imputer and/or the analyst have misspecified their models. Our intuition says that in general the resulting MI inference would be invalid. However, as discussed in Section 2.2, one case in which our results continue to hold in the presence of model misspecification lies in a slight twist of Example 2. Specifically, if God's model in Example 2 is changed to have different population means in the two groups, even though the analyst's model then becomes mis-specified, all the derivations there remain valid as long as the analyst's parameter of interest is the overall population mean. On the other hand, defects in the imputation model may cause great damage, as they can affect any subsequence analysis, as emphasized in Section 1.3. But how do we quantify or even formulate this intuition? Furthermore, what conditions do we need in order to capture the type of scenarios suggested by Example 4 when $\sigma_0^2 = 1/2$, two or more mis-specifications may somehow cancel each other?

Finally, all such questions are relevant for general multi-phase inferences. Defects incurred in earlier phases may cause more damages, just as problems caused by the data collection phase are usually harder to deal with than problems in the analysis phase, especially when some of those phases are "irreversible". Indeed, the question of "what to keep" has been much discussed and debated in the rapidly growing literature on data curation (e.g., Borgman (2010); Edwards et al. (2011)). Currently there is little participation of statisticians in such discussions. We venture that the lack of statisticians' participation is partially due to the fact

that our current "God-versus-me" paradigm is inadequate to address many critical problems in that emerging literature. We believe the multi-phase inference paradigm would allow us to be a part of the dialogues that can directly affect "scientific standards and reliability" because "we really are thinking about the same problem from different perspectives!" (Borgman, personal communication). Indeed, Blocker and Meng (2013) were able to make some headway in formulating a class of preprocessing theory under the multi-phase perspective, but clearly that was a small step. In general, as argued in Meng (2014), multi-phase inference, is one of the three large classes of problems with increasing frequencies and urgency in this age of Big Data (the other two being *multi-resolution inference* and *multi-source inference*) and, as such, much more need to be done and can be done. So please join us!

Acknowledgement

The main theoretical results in this article formed a part of Xie's Ph.D. thesis under the supervision of Meng. The discussion of multi-phase inference in Section 1 also formed a basis for the entry on "Multi-Party Inference and Uncongeniality" by Meng in *International Encyclopedia of Statistical Science* (2011, Springer), for which he thanks the Editor Miodrag Lovric for the invitation. (But here we have adopted the more encompassing term, *multi-phase* inference). Both authors thank colleagues, especially Alex Blocker, Christian Borgman, Jan Hannig and Donald Rubin, for helpful exchanges, Steven Finch and Kin Wai Chan for careful editing and proofreading, reviewers for very encouraging and constructive comments, and NSF (US) for partial financial support.

Appendix I: Technical Proofs

Proof of Lemma 1. Whereas we believe this lemma is a standard result and similar arguments have been invoked in many papers (e.g., Copas and Eguchi (2005)), we are unable to find a direct reference and hence provide a proof for completeness. To simplify notation, we assume θ to be scalar. By Taylor expanding $0 = g_n(Z, \hat{\theta}_g)$ around θ_0, we have

$$0 = g_n(Z;\theta_0) + g_n'(Z,\theta_0)(\hat{\theta}_g - \theta_0) + \frac{1}{2} g_n''(Z,\theta^*)(\hat{\theta}_g - \theta_0)^2,$$

with θ^* lying between θ_0 and $\hat{\theta}_g$. Dividing both sides by $g_n'(Z,\theta_0)$ (assuming it is not zero) yields

$$(\hat{\theta}_g - \theta_0) + \frac{g_n(Z;\theta_0)}{g_n'(Z,\theta_0)} = -\frac{1}{2n} \frac{g_n''(Z,\theta^*)/n}{g_n'(Z,\theta_0)/n} \times \left(\sqrt{n}(\hat{\theta}_g - \theta_0)\right)^2.$$

The regularity conditions (ii), (iii), and (iv) of Definition 1 together then imply

$$\sqrt{n}\left(\hat{\theta}_g - \theta_0 + \frac{g_n(Z;\theta_0)}{g_n'(Z,\theta_0)}\right) = O_p\left(\frac{1}{\sqrt{n}}\right). \tag{A.1}$$

Noting $\sqrt{n}(\hat{\theta}_g - \theta_0) = O_p(1)$, we know

$$\frac{\sqrt{n}g_n(Z;\theta_0)}{g_n'(Z,\theta_0)} = O_p(1),$$

which, together with (iii) in Definition 1, implies

$$\frac{\sqrt{n}g_n(Z;\theta_0)}{g_n'(Z,\theta_0)}\left(-\frac{g_n'(Z,\theta_0)}{nJ_g(\theta_0)} - 1\right) \xrightarrow{p} 0. \tag{A.2}$$

Adding (A.1) and (A.2), we then have

$$R_n = \sqrt{n}\left[(\hat{\theta}_g - \theta_0) - \frac{g_n(Z;\theta_0)}{nJ_g(\theta_0)}\right] \xrightarrow{p} 0.$$

To prove convergence in L^2, since conditions (v) and (vi) imply that $[\sqrt{n}(\hat{\theta}_g - \theta_0)]^2$ and $[g_n(Z;\theta_0)/\sqrt{n}J_g(\theta_0)]^2$ are uniformly integrable, so is R_n^2 . Hence $R_n^2 \xrightarrow{p} 0$ implies $R_n \xrightarrow{L^2} 0$.

Proof of Theorem 1. For simplicity, assume θ to be scalar. From Lemma 1 and our assumption that the three EEs satisfy SOR, we know

$$R_n^h = \sqrt{n}\left[\left(\hat{\theta}_h - \theta_0\right) - \frac{h_n(Z;\theta_0)}{nJ_h(\theta_0)}\right] \xrightarrow{p} 0;$$
$$R_n^u = \sqrt{n}\left[\left(\hat{\theta}_u - \theta_0\right) - \frac{u_n(Z;\theta_0)}{nJ_u(\theta_0)}\right] \xrightarrow{p} 0;$$
$$R_n^v = \sqrt{n}\left[\left(\hat{\theta}_v - \theta_0\right) - \frac{v_n(Z;\theta_0)}{nJ_v(\theta_0)}\right] \xrightarrow{p} 0;$$

$$h_n(Z;\theta) = u_n(Z;\theta) + v_n(Z;\theta) \quad \text{and} \quad J_h(\theta_0) = J_u(\theta_0) + J_v(\theta_0).$$

With some simple manipulations, we obtain

$$\sqrt{n}\left[(\hat{\theta}_h - \theta_0) - \left((I-F)(\hat{\theta}_u - \theta_0) + F(\hat{\theta}_v - \theta_0)\right)\right]$$
$$= \left[R_n^h - ((I-F)R_n^u + FR_n^v)\right] \xrightarrow{p} 0.$$

If the three EEs also satisfy condition (v) and (vi), the remainder terms R_n^h, R_n^u, and R_n^v are then known to converge to zero in L^2, which implies

$$\sqrt{n}\left[\hat{\theta}_h - \left((I-F)\hat{\theta}_u + F\hat{\theta}_v\right)\right] \xrightarrow{L^2} 0.$$

Proof of Corollary 1. From the properties of conditional expectations and the assumption $E^G(v_n(Z_1, Z_2; \theta_0)|Z_1) = 0$, we have

$$\begin{aligned} Cov^G(u_n(Z;\theta_0), v_n(Z;\theta_0)) &= E^G(u_n(Z;\theta_0)v_n(Z;\theta_0)) \\ &= E^G(E^G(u_n(Z_1;\theta_0))v_n(Z_1, Z_2;\theta_0)|Z_1) \\ &= E^G(u_n(Z_1;\theta_0)E^G(v_n(Z_1, Z_2;\theta_0)|Z_1)) = 0. \end{aligned}$$

From SOR and Lemma 1, we have

$$\begin{aligned} \hat{\theta}_u &= \theta_0 + \frac{u_n(Z;\theta_0)}{nJ_u(\theta_0)} + \frac{R_n^u}{\sqrt{n}}, \\ \hat{\theta}_v &= \theta_0 + \frac{v_n(Z;\theta_0)}{nJ_v(\theta_0)} + \frac{R_n^v}{\sqrt{n}}, \end{aligned}$$

with $R_n^u \xrightarrow{L^2} 0$ and $R_n^v \xrightarrow{L^2} 0$, implying $V^G(u_n(Z;\theta_0)) = O(n)$, $V^G(v_n(Z;\theta_0)) = O(n)$ and

$$\begin{aligned} Cov^G(\hat{\theta}_u, \hat{\theta}_v) &= Cov^G\left(\frac{u_n(Z;\theta_0)}{nJ_u(\theta_0)} + \frac{R_n^u}{\sqrt{n}}, \frac{v_n(Z;\theta_0)}{nJ_v(\theta_0)} + \frac{R_n^v}{\sqrt{n}}\right) \\ &= Cov^G\left(\frac{u_n(Z;\theta_0)}{nJ_u(\theta_0)}, \frac{R_n^v}{\sqrt{n}}\right) + Cov^G\left(\frac{R_n^u}{\sqrt{n}}, \frac{v_n(Z;\theta_0)}{nJ_v(\theta_0)}\right) \\ &\quad + Cov^G\left(\frac{R_n^u}{\sqrt{n}}, \frac{R_n^v}{\sqrt{n}}\right). \end{aligned}$$

Applying the Cauchy-Schwarz inequality to each of the terms above, it is then obvious $Cov(\hat{\theta}_u, \hat{\theta}_v) = o(1/n)$. As for the last assertion of the Corollary, in the proof above, we used regularity conditions to ensure $R_n^u \xrightarrow{L^2} 0$. Thus, when $\hat{\theta}_u - \theta_0 = R_n^u/\sqrt{n}$ with $R_n^u \xrightarrow{L^2} 0$, the proof will be valid as well.

Proof of Theorem 2. It is proved by applying Theorem 1 to $h_n(Z_{\text{obs}};\theta) = u_n(Z_{\text{obs}};\theta) + v_n(Z_{\text{obs}};\theta)$.

Proof of Theorem 3. By a Taylor expansion, we have

$$\begin{aligned} &\int v_n^A(Z_{\text{com}};\hat{\theta}_{\text{obs}}^H)p^I(Z_{\text{mis}}|Z_{\text{obs}};\hat{\theta}_{\text{obs}}^I)dZ_{\text{mis}} \\ &= \int v_n^A(Z_{\text{com}};\theta_0^A)p^I(Z_{\text{mis}}|Z_{\text{obs}};\theta_0^I)dZ_{\text{mis}} \\ &\quad + \int \frac{\partial}{\partial\theta^A}v_n^A(Z_{\text{com}};\theta_0^A)p^I(Z_{\text{mis}}|Z_{\text{obs}};\theta_0^I)dZ_{\text{mis}} \times (\hat{\theta}_{\text{obs}}^H - \theta_0^A) \\ &\quad + \int v_n^A(Z_{\text{com}};\theta_0^A)\left[\frac{\partial}{\partial\theta^I}p^I(Z_{\text{mis}}|Z_{\text{obs}};\theta_0^I)\right]^\top dZ_{\text{mis}} \times (\hat{\theta}_{\text{obs}}^I - \theta_0^I) + R_n. \end{aligned}$$

But by the definition of $\hat{\theta}_{\text{obs}}^H$ and $v_n^A(Z_{\text{com}};\theta^A)$, we have

$$E^I\left[v_n^A(Z_{\text{com}};\hat{\theta}^H_{\text{obs}})|Z_{\text{obs}};\hat{\theta}^I_{\text{obs}}\right]=\int v_n^A(Z_{\text{com}};\hat{\theta}^H_{\text{obs}})p^I(Z_{\text{mis}}|Z_{\text{obs}};\hat{\theta}^I_{\text{obs}})dZ_{\text{mis}}=0,$$

$$\int v_n^A(Z_{\text{com}};\theta_0^A)p^I(Z_{\text{mis}}|Z_{\text{obs}};\theta_0^I)dZ_{\text{mis}}=0.$$

Consequently, we have

$$\int \frac{\partial}{\partial\theta^A}v_n^A(Z_{\text{com}};\theta_0^A)p^I(Z_{\text{mis}}|Z_{\text{obs}};\theta_0^I)dZ_{\text{mis}}\times(\hat{\theta}^H_{\text{obs}}-\theta_0^A)$$
$$+\int v_n^A(Z_{\text{com}};\theta_0^A)\left[\frac{\partial}{\partial\theta^I}p^I(Z_{\text{mis}}|Z_{\text{obs}};\theta_0^I)\right]^\top dZ_{\text{mis}}\times(\hat{\theta}^I_{\text{obs}}-\theta_0^I)+R_n=o_p(n^{1/2}).$$

Assuming regularity conditions that the limit

$$\lim_{n\to\infty}\frac{1}{n}\int v_n^A(Z_{\text{com}};\theta_0^A)\left[\frac{\partial}{\partial\theta^I}p^I(Z_{\text{mis}}|Z_{\text{obs}};\theta_0^I)\right]^\top dZ_{\text{mis}}$$

exists, and the remaining term R_n satisfies $R_n=o_p(n^{1/2})$, we have

$$\sqrt{n}\left[(\hat{\theta}^H_{\text{obs}}-\theta_0^A)-K(\hat{\theta}^I_{\text{obs}}-\theta_0^I)\right]\xrightarrow{p}0.$$

Finally, assuming that the square of the difference is uniformly integrable, we have the convergence in L^2.

Proof of Corollary 2. From the validity under perturbation assumption, we know

$$\frac{\partial}{\partial\theta^{I\cap A}}\int v_n^A(Z_{\text{com}};\theta^{I\cap A},\theta_0^{A\setminus I})p^I(Z_{\text{mis}}|Z_{\text{obs}};\theta^{I\cap A},\theta_0^{I\setminus A})dZ_{\text{mis}}=o_p(n)$$

for any $\theta^{I\cap A}$ such that $||\theta^{I\cap A}-\theta_0^{I\cap A}||\le\varepsilon$. Assuming the exchangeability of differentiation and integration, we know

$$\int\frac{\partial}{\partial\theta^{I\cap A}}v_n^A(Z_{\text{com}};\theta^{I\cap A},\theta_0^{A\setminus I})p^I(Z_{\text{mis}}|Z_{\text{obs}};\theta^{I\cap A},\theta_0^{I\setminus A})dZ_{\text{mis}}$$
$$+\int v_n^A(Z_{\text{com}};\theta^{I\cap A},\theta_0^{A\setminus I})\frac{\partial}{\partial\theta^{I\cap A}}p^I(Z_{\text{mis}}|Z_{\text{obs}};\theta^{I\cap A},\theta_0^{I\setminus A})^\top dZ_{\text{mis}}=o_p(n),$$

from which one can verify that the matrix K has the given form for the two nested cases.

Proof of Corollary 3. From the results in Theorem 2 and Theorem 3, it is immediate that

$$\bar{\theta}_\infty-\theta_0=(I-F)(\hat{\theta}^A_{\text{obs}}-\theta_0)+FK(\hat{\theta}^I_{\text{obs}}-\theta_0)+\frac{R_n}{\sqrt{n}},$$

with $R_n\xrightarrow{L^2}0$. Therefore,

$$V^G(\bar{\theta}_\infty) = V^G\left((I-F)\hat{\theta}^A_{\text{obs}} + FK\hat{\theta}^I_{\text{obs}} + \frac{R_n}{\sqrt{n}}\right).$$

Applying the Cauchy-Schwarz inequality, it is easy to show

$$V^G(\bar{\theta}_\infty) = V^G\left((I-F)\hat{\theta}^A_{\text{obs}} + FK\hat{\theta}^I_{\text{obs}}\right) + o\left(\frac{1}{n}\right).$$

Proof of Lemma 2. Let

$$h_{\text{com}} = S^A(Z_{\text{com}};\theta^A) \quad \text{and} \quad h_{\text{obs}} = S^A(Z_{\text{obs}};\theta^A) = E^A[S^A(Z_{\text{com}};\theta^A)|Z_{\text{obs}};\theta^A].$$

Because $E^G[h_{\text{obs}}(h_{\text{com}} - h_{\text{obs}})^\top] = 0$, we know that condition of Corollary 4 (proved below) is satisfied. Consequently, the self-efficiency condition implies that

$$\left(E^G\frac{\partial h_{\text{obs}}}{\partial\theta}\right)^{-1} E^G(h_{\text{obs}}h^\top_{\text{obs}}) = \left(E^G\frac{\partial h_{\text{com}}}{\partial\theta}\right)^{-1} E^G(h_{\text{com}}h^\top_{\text{com}}) + o(1),$$

which implies

$$(nJ_{\text{obs}})^{-1}E^G(h_{\text{obs}}h^\top_{\text{obs}}) = (nJ_{\text{com}})^{-1}E^G(h_{\text{com}}h^\top_{\text{com}}) + o(1).$$

This, together with the fact that $E^G(h_{\text{obs}}h^\top_{\text{mis}}) = E^G(h_{\text{obs}}(h_{\text{com}} - h_{\text{obs}})^\top) = 0$, leads to

$$(J_{\text{mis}})(J_{\text{obs}})^{-1}E^G(h_{\text{obs}}h^\top_{\text{obs}}) = E^G(h_{\text{mis}}h^\top_{\text{mis}}) + o(n).$$

Multiplying both sides by some common factors, we get

$$J^{-1}_{\text{com}}J_{\text{mis}}J^{-1}_{\text{obs}}E^G(h_{\text{obs}}h^\top_{\text{obs}})(J^\top_{\text{com}})^{-1} = J^{-1}_{\text{com}}E^G(h_{\text{mis}}h^\top_{\text{mis}})(J^\top_{\text{com}})^{-1} + o(n).$$

Now from the Strong SOR condition and some algebraic manipulations, we find

$$FV^A_{\text{obs}}(I-F)^\top = FV^A_{\text{mis}}F^\top + o(n^{-1}). \tag{A.3}$$

The symmetry of the right-hand side of (A.3) then is sufficient to establish (6.3).

Proof of Theorem 4. For simplicity, we assume θ to be a scalar, but the argument is general. From Lemma 1 and Strong SOR, we know

$$\begin{aligned}
Cov^G(\hat{\theta}_{\text{com}}, \hat{\theta}_{\text{obs}} - \hat{\theta}_{\text{com}}) &= E^G(\hat{\theta}_{\text{com}}\hat{\theta}_{\text{obs}}) - E^G(\hat{\theta}^2_{\text{com}}) \\
&= E^G\left[\left(\frac{h_{\text{com}}}{nJ_{\text{com}}} + \frac{R^{\text{com}}_n}{\sqrt{n}}\right)\left(\frac{h_{\text{obs}}}{nJ_{\text{obs}}} + \frac{R^{\text{obs}}_n}{\sqrt{n}}\right)\right] - E^G\left(\frac{h_{\text{com}}}{nJ_{\text{com}}} + \frac{R^{\text{com}}_n}{\sqrt{n}}\right)^2 \\
&= \frac{E^G(h_{\text{com}}h_{\text{obs}})}{n^2J_{\text{com}}J_{\text{obs}}} - \frac{E^Gh^2_{\text{com}}}{n^2J^2_{\text{com}}} + o(n^{-1}) \\
&= \frac{E^G(h_{\text{com}}h_{\text{obs}})}{(E^G\frac{\partial h_{\text{com}}}{\partial\theta})(E^G\frac{\partial h_{\text{obs}}}{\partial\theta})} - \frac{E^Gh^2_{\text{com}}}{(E^G\frac{\partial h_{\text{com}}}{\partial\theta})^2} + o(n^{-1}).
\end{aligned}$$

The result then follows from the definition of strong efficiency (6.1).

Proof of Corollary 4. This is a direct consequence of Theorem 4 and the fact that

$$-\lim_{n\to\infty}\frac{1}{n}E^G\left(\frac{\partial h_{\text{obs}}}{\partial\theta}\right)=J_{\text{obs}}.$$

Proof of Theorem 5. From the discussion in Section 5.1, if we can establish that

$$\bar{U}_\infty = V(\hat{\theta}^A)+o_p(n^{-1})\ , \tag{A.4}$$

$$B_\infty = V(\bar{\theta}_\infty-\hat{\theta}^A)+o_p(n^{-1})\ , \tag{A.5}$$

then the result follows directly from (5.10). Establishing (A.4) is straightforward, as presented in Section 5.1, Establishing (A.5) is not, because we cannot use the approximation based on the plug-in predictive imputation, which would lead to under-dispersion in the imputation value. But for its intended estimand, $V^G(\bar{\theta}_\infty-\hat{\theta}^A_{\text{com}})$, (5.1) and (5.2) together imply (asymptotically)

$$\bar{\theta}_\infty-\hat{\theta}^A_{\text{com}}=FK\hat{\theta}^I_{\text{obs}}-F\hat{\theta}^A_{\text{mis}}.$$

Because $\hat{\theta}^I_{\text{obs}}$ and $\hat{\theta}^A_{\text{mis}}$ are asymptotically orthogonal (due to Corollary 1), we then have

$$V^G(\bar{\theta}_\infty-\hat{\theta}^A_{\text{com}})=FKV^I_{\text{obs}}K^\top(F)^\top+FV^A_{\text{mis}}(F)^\top+o\left(n^{-1}\right), \tag{A.6}$$

where $V^I_{\text{obs}}=V^G(\hat{\theta}^I_{\text{obs}})$. Applying (3.8), but with Z_{com} replaced by $\tilde{Z}_{\text{com}}=(Z_{\text{obs}},\tilde{Z}_{\text{mis}})$, where $\tilde{Z}_{\text{mis}}$ is a draw from the imputer's posterior predictive distribution $p^I(Z_{\text{mis}}|Z_{\text{obs}})$, we obtain

$$\begin{aligned}B_\infty &= V^I\left[\hat{\theta}^A(\tilde{Z}_{\text{com}})|Z_{\text{obs}}\right]=V^I\left[(I-F)\hat{\theta}^A_{\text{obs}}+F\hat{\theta}^A_{\text{mis}}(Z_{\text{obs}},\tilde{Z}_{\text{mis}})|Z_{\text{obs}}\right]+o_p(n^{-1})\\ &= FV^I\left[\hat{\theta}^A_{\text{mis}}(Z_{\text{obs}},\tilde{Z}_{\text{mis}})|Z_{\text{obs}}\right](F)^\top+o_p(n^{-1})\\ &= FV^I\left[e(Z_{\text{obs}};\tilde{\theta})|Z_{\text{obs}}\right](F)^\top+FE^I\left[\sigma(Z_{\text{obs}};\tilde{\theta})|Z_{\text{obs}}\right](F)^\top+o_p(n^{-1}),\end{aligned} \tag{A.7}$$

where

$$\sigma(Z_{\text{obs}};\tilde{\theta})=V^I\left[\hat{\theta}^A_{\text{mis}}(Z_{\text{obs}},\tilde{Z}_{\text{mis}})|Z_{\text{obs}};\tilde{\theta}\right], \tag{A.8}$$

$$e(Z_{\text{obs}};\tilde{\theta})=E^I\left[\hat{\theta}^A_{\text{mis}}(Z_{\text{obs}},\tilde{Z}_{\text{mis}})|Z_{\text{obs}};\tilde{\theta}\right]. \tag{A.9}$$

The identity in (A.7) is due to $V(X)=E[V(X|Y)]+V[E(X|Y)]$.

For the second term in (A.7), because we assume the imputer's model is correctly specified, asymptotically we have

$$E^I\left[\sigma(Z_{\text{obs}};\tilde{\theta})|Z_{\text{obs}}\right]=\sigma(Z_{\text{obs}};\theta_0)+o_p(n^{-1}). \tag{A.10}$$

Then

$$\begin{aligned} V_{\text{mis}}^A \equiv & V^G\left[\hat{\theta}_{\text{mis}}^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}})\right] \\ = & V^G\left[E^I\left(\hat{\theta}_{\text{mis}}^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}})|Z_{\text{obs}}; \theta_0\right)\right] + E^G\left[V^I\left(\hat{\theta}_{\text{mis}}^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}})|Z_{\text{obs}}; \theta_0\right)\right] \\ = & V^G[e(Z_{\text{obs}}; \theta_0)] + E^G[\sigma(Z_{\text{obs}}; \theta_0)]. \end{aligned} \tag{A.11}$$

Here we can switch the superscript "G" to "I" in the inner layer conditional expectations because the imputer's model is correctly specified and hence $p^G(Z_{\text{mis}}|Z_{\text{obs}}) = p^I(Z_{\text{mis}}|Z_{\text{obs}}; \theta_0)$. But when both the imputer's and the analyst's models are correctly specified, the estimating equation for $\hat{\theta}_{\text{mis}}^A(\tilde{Z}_{\text{com}})$, $v_n^A(Z_{\text{com}}; \theta)$ of (3.7) is conditionally unbiased (conditioning on Z_{obs} and $\theta = \theta_0$),

$$E^I\left[v_n^A(\tilde{Z}_{\text{com}}; \theta_0)|Z_{\text{obs}}; \theta_0\right] = 0.$$

By Corollary 1, the corresponding root $\hat{\theta}_{\text{mis}}^A(\tilde{Z}_{\text{com}})$ is asymptotically uncorrelated with its conditional expectation $e(Z_{\text{obs}}; \theta_0)$. Because $Cov(X, E(X|Y)) = V(E(X|Y))$, the first term on the right-hand side of (A.11) is asymptotically negligible compared to the second term, and hence the second term in (A.7) is a consistent estimator of the second term on the right-hand side of (5.8).

For the first term on the righthand side of (A.7), we *cannot* replace $e(Z_{\text{obs}}; \tilde{\theta})$ by $e(Z_{\text{obs}}; \theta_0)$ or any $e(Z_{\text{obs}}; \hat{\theta}_{\text{obs}})$ because then the variance due to (conditional) uncertainty in the parameter estimation would be incorrectly set to zero. However, using essentially the same argument (see below) as in establishing the asymptotic equivalence of the estimators in Table 1, we can prove that $e(Z_{\text{obs}}; \tilde{\theta})$ is asymptotically the same as the root of the *imputed* $v_n^A(Z_{\text{com}}; \theta)$ (see (3.7)) under the plug-in predictive imputation assuming $\theta = \tilde{\theta}$, that is, $e(Z_{\text{obs}}; \tilde{\theta})$ can be viewed as the root of

$$\begin{aligned} & E^I(v_n^A(Z_{\text{com}}; \theta)|Z_{\text{obs}}; \tilde{\theta}) \\ & = E^I\left[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta)|Z_{\text{obs}}; \tilde{\theta}\right] - E^A\left[S^A(Z_{\text{obs}}, Z_{\text{mis}}; \theta)|Z_{\text{obs}}; \theta\right] = 0. \end{aligned} \tag{A.12}$$

Comparing (A.12) with (4.5) reveals that they differ only in the plug-in value: $\hat{\theta}_{\text{obs}}^I$ verse $\tilde{\theta}$. Thus essentially the same argument used for proving Theorem 3 can be applied to establish that

$$e(Z_{\text{obs}}; \tilde{\theta}) = \theta_0^A + K(\tilde{\theta} - \theta_0^I) + o_p(n^{-1/2}). \tag{A.13}$$

It follows then that

$$V^I\left[e(Z_{\text{obs}}; \tilde{\theta})|Z_{\text{obs}}\right] = KV^I(\tilde{\theta}|Z_{\text{obs}})K^\top + o_p(n^{-1}).$$

Under the usual regularity conditions that guarantee the asymptotic equivalence between the posterior variance $V^I(\tilde{\theta}|Z_{\text{obs}})$ and the sampling variance $V^G(\hat{\theta}^A_{\text{obs}})$, we can then conclude that the first term on the right-hand side of (5.8) is a consistent estimator for the first term on the right-hand side of (A.7), and hence (A.5) is established.

Proof of Lemma 3. The fact that $\hat{\theta}^A_{\text{obs}} \succ \hat{\theta}^I_{\text{obs}}$ implies θ^A is a sub-parameter of θ^I. Theorem 3 and Corollary 2 then imply

$$\hat{\theta}^H_{\text{obs}} - \theta^A_0 = K(\hat{\theta}^I_{\text{obs}} - \theta^I_0) + R_n,$$

where $\sqrt{n}R_n \xrightarrow{L^2} 0$ and $K = [I_A, B]$. We then know from $\hat{\theta}^A_{\text{obs}} \succ \hat{\theta}^I_{\text{obs}}$ that $\hat{\theta}^A_{\text{obs}} \succ \hat{\theta}^H_{\text{obs}}$.

Proof of Theorem 6. From our assumption $\hat{\theta}^A_{\text{obs}} \succ \hat{\theta}^I_{\text{obs}}$, we can re-arrange the parameter space of the imputer's model as $(\theta^A, \theta^{I\setminus A})$, where θ^A is a parameter to both the analyst and the imputer, but $\theta^{I\setminus A}$ is a parameter to the imputer only. Therefore we have

$$C^{A,I}_{\text{obs}} \equiv Cov^G(\hat{\theta}^A_{\text{obs}}, \hat{\theta}^I_{\text{obs}}) = \Big[V^G(\hat{\theta}^A_{\text{obs}}), 0\Big] + o(n^{-1}).$$

From Corollary 2, we know $K = [I, B]$, which implies $C^{A,I}_{\text{obs}}K^\top = V^G(\hat{\theta}^A_{\text{obs}}) + o(n^{-1})$, and hence the consistency of T_∞, as a consequence of Lemma 2.

For (ii), notice from Corollary 3 and the above discussion that

$$\begin{aligned}
V_\infty &= (I-F)V^A_{\text{obs}}(I-F^\top) + FKV^I_{\text{obs}}K^\top F^\top + (I-F)C^{A,I}_{\text{obs}}K^\top F^\top \\
&\quad + FK(C^{A,I}_{\text{obs}})^\top(I-F^\top) + o(n^{-1}) \\
&= (I-F)V^A_{\text{obs}}(I-F^\top) + FKV^I_{\text{obs}}K^\top F^\top + (I-F)V^A_{\text{obs}}F^\top \\
&\quad + FV^A_{\text{obs}}(I-F^\top) + o(n^{-1}), \\
V^A_{\text{obs}} &= (I-F)V^A_{\text{obs}}(I-F^\top) + FV^A_{\text{obs}}F^\top + (I-F)V^A_{\text{obs}}F^\top + FV^A_{\text{obs}}(I-F^\top).
\end{aligned}$$

It is then sufficient to prove $KV^I_{\text{obs}}K^\top \geq V^A_{\text{obs}}$, which is obvious because

$$V^I_{\text{obs}} \geq \begin{pmatrix} V^A_{\text{obs}} & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad V^A_{\text{obs}} = K \begin{pmatrix} V^A_{\text{obs}} & 0 \\ 0 & 0 \end{pmatrix} K^\top.$$

Proof of Theorem 7. From Theorem 2 and Theorem 3, we know

$$\bar{\theta}_\infty - \theta_0 = (I-F)(\hat{\theta}^A_{\text{obs}} - \theta_0) + F[I,0]^\top(\hat{\theta}^I_{\text{obs}} - \theta_0).$$

The result (i) then simply is a consequence of the assumption that $\hat{\theta}^I_{\text{obs}} \succ \hat{\theta}^A_{\text{obs}}$. For (ii), from (5.6), (5.8), and Lemma 2, we have

$$T_\infty = (I-F)V^A_{\text{obs}}(I-F)^\top + FKV^I_{\text{obs}}K^\top F^\top + FV^A_{\text{obs}}(I-F)^\top + (I-F)V^A_{\text{obs}}F^\top.$$

Since $V^A_{\text{obs}} = [(I-F)+F]V^A_{\text{obs}}[(I-F)+F]^\top$, it is then sufficient to prove $V^A_{\text{obs}} \geq KV^I_{\text{obs}}K^\top$, which follows from the fact that $K = [I,0]^\top$ and $\hat{\theta}^I \succ \hat{\theta}^A$.

To prove (iii), first consider the case where θ^A is a scalar: let ϕ denote this scalar parameter. Re-arrange the imputer's estimator as $\hat{\theta}^I_{\text{obs}} = (\hat{\phi}^I_{\text{obs}}, \hat{\eta}^I_{\text{obs}})$ ($\hat{\phi}^I_{\text{obs}}$ could be ϕ_0 if the imputer knows the true value of ϕ). Since $\hat{\theta}^I_{\text{obs}} \succ \hat{\theta}^A_{\text{obs}}$, the analyst's procedure can be expanded to a procedure that estimates ϕ and η together by appending the score function on η. From Theorem 3,

$$\hat{\phi}^H_{\text{obs}} - \phi_0 = K(\hat{\theta}^I_{\text{obs}} - \theta^I_0) + R_n = (\hat{\phi}^I_{\text{obs}} - \phi_0) + R_n,$$

where $\sqrt{n}R_n \xrightarrow{L^2} 0$ and $K = [I,0]^\top$. Therefore, we know

$$Cov^G(\hat{\theta}^A_{\text{obs}}, \hat{\theta}^I_{\text{obs}})K^\top = V^G(\hat{\phi}^I_{\text{obs}}) + o(n^{-1}).$$

The bias in T_∞ can be re-written as

$$T_\infty - V_\infty = 2F(1-F)\left(V^G(\hat{\phi}^A_{\text{obs}}) - V^G(\hat{\phi}^I_{\text{obs}})\right) + o(n^{-1}).$$

From the assumptions $\hat{\theta}^A_{\text{com}} \succ \hat{\theta}^A_{\text{obs}}$ and $\hat{\theta}^I_{\text{obs}} \succ \hat{\theta}^A_{\text{obs}}, 0 \leq F \leq 1$ and $V^G(\hat{\phi}^A_{\text{obs}}) \geq V^G(\hat{\phi}^I_{\text{obs}})$; hence $T_\infty - V_\infty \geq 0$. The same proof applies when θ^A is a vector and $F \propto I$.

Proof of Equivalence of Four Estimators in Table 1. Here we assume EE $S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta)(=0)$ satisfies SOR, and the ratio n/N is bounded away from zero as the *complete-data* size $N \to \infty$. To prove $\bar{\theta}^{(21)}_\infty$ and $\bar{\theta}^{(22)}_\infty$ are asymptotically the same, we first recall $\bar{\theta}^{(21)}_\infty = E^I[\hat{\theta}^A(\tilde{Z}_{\text{com}})|Z_{\text{obs}}; \hat{\theta}^I_{\text{obs}}]$, where $\hat{\theta}^A(\tilde{Z}_{\text{com}})$ is a root of $S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta) = 0$. Then, following the proof in Lemma 1,

$$\sqrt{n}\left(\left(\hat{\theta}^A(\tilde{Z}_{\text{com}}) - \theta_0\right) - \frac{S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)}{NJ_S(\theta_0)}\right) \xrightarrow{L^2} 0.$$

From the property of convergence in L^2, we know

$$\sqrt{n}\left(\left(E^I[\hat{\theta}^A(\tilde{Z}_{\text{com}})|Z_{\text{obs}}; \hat{\theta}^I_{\text{obs}}] - \theta_0\right) - \frac{E^I[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)|Z_{\text{obs}}; \hat{\theta}^I_{\text{obs}}]}{NJ_S(\theta_0)}\right) \xrightarrow{L^2} 0\ ,$$

which implies

$$\sqrt{n}\left(\left(\bar{\theta}^{(21)}_\infty - \theta_0\right) - \frac{E^I[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)|Z_{\text{obs}}; \hat{\theta}^I_{\text{obs}}]}{NJ_S(\theta_0)}\right) \xrightarrow{L^2} 0. \tag{A.14}$$

But $E^I[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)|Z_{\text{obs}}; \hat{\theta}^I_{\text{obs}}]$ is the averaged EE used by the plug-in predictive imputation (see Table 1). Therefore, again by Lemma 1, we have

$$\sqrt{n}\left(\left(\bar{\theta}^{(22)}_{\infty} - \theta_0\right) - \frac{E^I[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)|Z_{\text{obs}}; \hat{\theta}^I_{\text{obs}}]}{NJ_S(\theta_0)}\right) \xrightarrow{L^2} 0. \tag{A.15}$$

Consequently, $\sqrt{n}(\bar{\theta}^{(21)}_{\infty} - \bar{\theta}^{(22)}_{\infty}) \xrightarrow{L^2} 0$. The assertion $\sqrt{n}(\bar{\theta}^{(11)}_{\infty} - \bar{\theta}^{(12)}_{\infty}) \xrightarrow{L^2} 0$ can be established by a similar argument. In the asymptotic results above, we have replaced $\sqrt{N}$ by $\sqrt{n}$ because the ratio n/N is bounded away from zero.

To establish $\sqrt{n}(\bar{\theta}^{(12)}_{\infty} - \bar{\theta}^{(22)}_{\infty}) \xrightarrow{L^2} 0$, we again notice that,

$$\sqrt{n}\left(\left(\bar{\theta}^{(12)}_{\infty} - \theta_0\right) - \frac{E^I[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)|Z_{\text{obs}}]}{NJ_S(\theta_0)}\right) \xrightarrow{L^2} 0. \tag{A.16}$$

Results (A.15) and (A.16) imply that we need to show

$$\sqrt{n}\left(\frac{E^I[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)|Z_{\text{obs}}; \hat{\theta}^I_{\text{obs}}]}{NJ_S(\theta_0)} - \frac{E^I[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)|Z_{\text{obs}}]}{NJ_S(\theta_0)}\right) \xrightarrow{L^2} 0. \tag{A.17}$$

First notice that

$$\begin{aligned}
&E^I[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)|Z_{\text{obs}}; \hat{\theta}^I_{\text{obs}}] - E^I[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)|Z_{\text{obs}}] \\
&= E^I[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)|Z_{\text{obs}}; \hat{\theta}^I_{\text{obs}}] - E^I[E^I[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)|Z_{\text{obs}}; \theta]|Z_{\text{obs}}] \\
&= -E^I\left[d_n(Z_{\text{obs}}; \theta) \mid Z_{\text{obs}}\right],
\end{aligned} \tag{A.18}$$

where the last expectation is with respect to the imputer's posterior distribution $p^I(\theta|Z_{\text{obs}})$, and

$$d_n(Z_{\text{obs}}; \theta) = E^I[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)|Z_{\text{obs}}; \theta] - E^I[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)|Z_{\text{obs}}; \hat{\theta}^I_{\text{obs}}].$$

Let

$$f_n(Z_{\text{obs}}; \theta) = \frac{\partial E^I[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)|Z_{\text{obs}}; \hat{\theta}^I_{\text{obs}}]}{\partial \theta}(\theta - \hat{\theta}^I_{\text{obs}}). \tag{A.19}$$

Then a Taylor expansion yields

$$d_n(Z_{\text{obs}}; \theta) - f_n(Z_{\text{obs}}; \theta) = \frac{1}{2}\frac{\partial^2 E^I[S^A(Z_{\text{obs}}, \tilde{Z}_{\text{mis}}; \theta_0)|Z_{\text{obs}}; \theta^*]}{\partial \theta^2}(\theta - \hat{\theta}^I_{\text{obs}})^2, \tag{A.20}$$

where θ^* is a value between θ and $\hat{\theta}^I_{\text{obs}}$. The right-hand side of (A.20) is of order $O_p(n)O_p(1/n) = O_p(1)$, and thus we can conclude, assuming both d_n and f_n are uniformly integrable, that

$$n^{-1/2}[d_n(Z_{\text{obs}}; \theta) - f_n(Z_{\text{obs}}; \theta)] \xrightarrow{L^2} 0. \tag{A.21}$$

Under the usual regularity assumptions that ensure asymptotic equivalence between Bayesian estimator and MLE, $E^I(\theta|Z_{\rm obs}) - \hat{\theta}^I_{\rm obs} = O_p\left(n^{-1}\right)$. This together with the assumption that

$$\frac{\partial}{\partial\theta} E^I[S^A(Z_{\rm obs}, \tilde{Z}_{\rm mis}; \theta_0)|Z_{\rm obs}; \hat{\theta}^I_{\rm obs}] = O_p(n)$$

implies that $n^{-\frac{1}{2}} E^I[f_n(Z_{\rm obs}; \theta)|Z_{\rm obs}] \stackrel{L^2}{\to} 0$. Consequently, $n^{-\frac{1}{2}} E^I[d_n(Z_{\rm obs}; \theta)|Z_{\rm obs}] \stackrel{L^2}{\to} 0$, beacause of (A.21). But this implies (A.17) because of (A.18).

Appendix II: Illustrating Theoretical Results via a Regression Setting

Suppose in a regression model $Y = X\theta + \epsilon$ with $\epsilon \sim N(\mathbf{0}, I_N)$, the covariates X are fully observed but the responses Y are only missing at random (MAR). The notation is the same as in the regression example of Section 8.3, but here we have p covariates, instead of just two. We set a model by setting some of the coefficients θ to be zero or, equivalently, we can use a set of covariates to denote the nonzero part. Therefore, we use I and A to also denote, respectively, the set of covariates used by the imputer and the analyst. The estimators $\hat{\theta}^A_{\rm obs}$ and $\hat{\theta}^I_{\rm obs}$ are taken to be the least square estimator in the corresponding model. Under MAR, the plug-in predictive imputation model is simply the linear regression model $Y = X^I_{\rm mis}\hat{\theta}^I_{\rm obs} + \epsilon$, with the corresponding averaged/projected EE

$$\left(X^A_{\rm com}\right)^\top \left[\begin{pmatrix} Y_{\rm obs} \\ X^I_{\rm mis}\hat{\theta}^I_{\rm obs} \end{pmatrix} - X^A_{\rm com}\theta^A\right] = 0.$$

Denote $M^{(c)}_{U,V} = [X^U_{\rm com}]^\top X^V_{\rm com}$, e.g., $M^{(c)}_{A,A} = [X^A_{\rm com}]^\top X^A_{\rm com}$, and similarly, $M^{(o)}_{U,V}$ and $M^{(m)}_{U,V}$ for the observed-data and missing-data counterparts. Straightforward algebra then yields

$$\bar{\theta}_\infty = [M^{(c)}_{A,A}]^{-1} M^{(o)}_{A,A}\hat{\theta}^A_{\rm obs} + [M^{(c)}_{A,A}]^{-1} M^{(m)}_{A,A}\hat{\theta}^H_{\rm obs}\ , \tag{A.22}$$

where

$$\hat{\theta}^H_{\rm obs} = [M^{(m)}_{A,A}]^{-1} M^{(m)}_{A,I}\hat{\theta}^I_{\rm obs} \equiv K_N\hat{\theta}^I_{\rm obs}. \tag{A.23}$$

Let

$$F = [M^{(c)}_{A,A}]^{-1} M^{(m)}_{A,A} \tag{A.24}$$

be the fraction of missing information, then we have from (A.22) that

$$\bar{\theta}_\infty = (I - F)\hat{\theta}^A_{\rm obs} + F\hat{\theta}^H_{\rm obs}\ ,$$

confirming the result in Theorem 2. Assuming $K = \lim_{N\to\infty} K_N$ exists, and noting $\theta^A_0 = K_N\theta^I_0$, we know from (A.23) that

$$\sqrt{n}\left[(\hat{\theta}^H_{\rm obs} - \theta^A_0) - K(\hat{\theta}^I_{\rm obs} - \theta^I_0)\right] = (K_N - K)[\sqrt{n}(\hat{\theta}^I_{\rm obs} - \theta^I_0)] = o_p\left(1\right),$$

which verifies the general result in Theorem 3.

Next we verify the form of K in the two nested cases. On the one hand, when the analyst's model and imputer's model are nested as $A \supseteq I$, we see that

$$\hat{\theta}^H_{\text{obs}} = [M^{(m)}_{A,A}]^{-1} M^{(m)}_{A,I} \hat{\theta}^I_{\text{obs}} = ({X^A_{\text{mis}}}^\top X^A_{\text{mis}})^{-1} \left[{X^A_{\text{mis}}}^\top \left(X^I_{\text{mis}}, X^{A\setminus I}_{\text{mis}} \right) \right] \begin{pmatrix} \hat{\theta}^I_{\text{obs}} \\ 0 \end{pmatrix}$$
$$= \begin{pmatrix} \hat{\theta}^I_{\text{obs}} \\ 0 \end{pmatrix} = \begin{pmatrix} I_I \\ 0 \end{pmatrix} \hat{\theta}^I_{\text{obs}} ,$$

implying that $\hat{\theta}^H_{\text{obs}}$ is simply the imputer's estimator appended by the true value, verifying (1) of Corollary 2. This result is anticipated because the EE for $\hat{\theta}^H_{\text{obs}}$ is in effect defined with infinitely many imputed "data points". Thus the estimator will be the one used by the imputer, including true values of the part of θ that the imputer assumed.

On the other hand, when $I \supseteq A$, we have

$$\hat{\theta}^H_{\text{obs}} = [M^{(m)}_{A,A}]^{-1} M^{(m)}_{A,I} \hat{\theta}^I_{\text{obs}} = ({X^A_{\text{mis}}}^\top X^A_{\text{mis}})^{-1} \left[{X^A_{\text{mis}}}^\top \left(X^A_{\text{mis}}, X^{I\setminus A}_{\text{mis}} \right) \right] \hat{\theta}^I_{\text{obs}} = K \hat{\theta}^I_{\text{obs}},$$

where $K = [I_A, B]$ with $B = [M^{(m)}_{A,A}]^{-1} M^{(m)}_{A,I\setminus A}$, verifying (2) of Corollary 2. Since $\theta^{I\setminus A}_0 = \mathbf{0}$, we have $\hat{\theta}^H_{\text{obs}} - \theta^A_0 = K(\hat{\theta}^I_{\text{obs}} - \theta^I_0)$, i.e., $\hat{\theta}^H_{\text{obs}}$ is a projection of $\hat{\theta}^I_{\text{obs}}$ onto the analyst's parameter space.

To verify Corollary 3, straightforward calculation yields

$$V_\infty = [M^{(c)}_{A,A}]^{-1} (J_1 + J_2 + J_3 + J_3^\top) [M^{(c)}_{A,A}]^{-1}, \tag{A.25}$$

where $J_1 = M^{(o)}_{A,A}$, $J_2 = M^{(m)}_{A,I} [M^{(m)}_{I,I}]^{-1} M^{(m)}_{I,A}$ and $J_3 = M^{(o)}_{A,I} [M^{(o)}_{I,I}]^{-1} M^{(m)}_{I,A}$. The four individual terms in (A.25) can be verified to correspond to those in Corollary 3. For example,

$$\begin{aligned} &[M^{(c)}_{A,A}]^{-1} J_1 [M^{(c)}_{A,A}]^{-1} \\ &\quad = ({X^A_{\text{com}}}^\top X^A_{\text{com}})^{-1} ({X^A_{\text{obs}}}^\top X^A_{\text{obs}}) ({X^A_{\text{obs}}}^\top X^A_{\text{obs}})^{-1} ({X^A_{\text{obs}}}^\top X^A_{\text{obs}}) ({X^A_{\text{com}}}^\top X^A_{\text{com}})^{-1} \\ &\quad = (I - F) V^A_{\text{obs}} (I - F)^\top. \end{aligned}$$

Similarly to verify Lemma 2, and particularly (6.5), we note that

$$V^A_{\text{obs}} = [M^{(o)}_{A,A}]^{-1}, \quad C^{A,I}_{\text{obs}} = [M^{(o)}_{A,A}]^{-1} M^{(o)}_{A,I} [M^{(o)}_{I,I}]^{-1} \tag{A.26}$$

and the exact expression of K depends on how the analyst's and imputer's models are nested with each other. For example, to illustrate both Theorem 5 and Theorem 6, we assume the analyst's set of covariates is a subset of the imputer's set (and that both sets contain the actual covariates used by God). Then the

analyst's $\hat{\theta}^A_{\text{com}}$ is the MLE even under the imputer's model, hence is strongly more efficient than $\bar{\theta}_\infty$, implying T_∞ is consistent. We can reach the same conclusion from Theorem 6 by noting that the analyst's procedure is self-efficient since it is the MLE and $\hat{\theta}^A_{\text{obs}} \succ \hat{\theta}^I_{\text{obs}}$ because $\hat{\theta}^A_{\text{obs}}$ is the MLE under a submodel of the one that underlies $\hat{\theta}^I_{\text{obs}}$.

We can also establish $T_\infty - V_\infty = 0$ directly by verifying (6.5). Given (A.26), because $K = [M^{(m)}_{A,A}]^{-1} M^{(m)}_{A,I}$, verifying (6.5) is the same as proving $M^{(o)}_{A,I}[M^{(o)}_{I,I}]^{-1} M^{(m)}_{I,A}[M^{(m)}_{A,A}]^{-1} = I$, which is a simple consequence of the identity

$$\begin{aligned}&(X^A_{\text{obs}}, X^{I\setminus A}_{\text{obs}})^\top X^I_{\text{obs}} (X^I_{\text{obs}}{}^\top X^I_{\text{obs}})^{-1} (X^I_{\text{mis}}{}^\top X^A_{\text{mis}}) \\ &\quad = X^I_{\text{mis}}{}^\top X^A_{\text{mis}} = (X^A_{\text{mis}}, X^{I\setminus A}_{\text{mis}})^\top X^A_{\text{mis}}.\end{aligned}$$

Finally, this regression setting illustrates the importance of having the proportionality assumption on F in (iii) of Theorem 7, and indicates how it holds when the missing data are missing completely at random (MCAR). Specifically, consider cases where $I \subseteq A$. Re-write θ as $\theta^A = (\theta^{A\setminus I}, \theta^I)$. Then, noticing that $[M^{(s)}_{A,A}]^{-1} M^{(s)}_{A,I} = [I_I, \mathbf{0}]^\top$ for $s = c, o, m$, we see from (A.26) that $C^{A,I}_{\text{obs}} = [I_I, \mathbf{0}]^\top [M^{(o)}_{I,I}]^{-1}$, and from (1) of Corollary 2 that $K = [I_I, \mathbf{0}]^\top$. Consequently,

$$C^{A,I}_{\text{obs}} K^\top = [I_I, \mathbf{0}]^\top [M^{(o)}_{I,I}]^{-1} [I_I, \mathbf{0}] = \begin{pmatrix} V^I_{\text{obs}} & 0 \\ 0 & 0 \end{pmatrix},$$

which is exactly the variance of imputer's estimator $(\hat{\theta}^I_{\text{obs}}, \theta_0^{A/I})$ for the analyst's parameter θ^A.

It follows then that we can rewrite D_∞ of (6.4) as

$$D_\infty = (I - F)\left[V^A_{\text{obs}} - K V^I_{\text{obs}} K^\top\right] F^\top. \tag{A.27}$$

Since Theorem 7 assumes $\hat{\theta}^I_{\text{obs}} \succ \hat{\theta}^A_{\text{obs}}$, we know $V^A_{\text{obs}} - K V^I_{\text{obs}} K^\top \geq 0$ in the sense of being a non-negative definite matrix. However, this does not imply $D_\infty + D_\infty^\top$ is non-negative definite because the matrix F generally does not commute with $V^A_{\text{obs}} - K V^I_{\text{obs}} K^\top$. But when F is proportional to an identity matrix, then (5.10) and (A.27) together imply that

$$T_\infty - V_\infty = 2F(I - F)\left[V^A_{\text{obs}} - K V^I_{\text{obs}} K^\top\right] + o_p\left(n^{-1}\right), \tag{A.28}$$

which is (asymptotically) non-negative definite, and hence the conclusion of (iii) of Theorem 7.

Evidently the expression (A.27) is not restricted to the regression setting, but the regression setting provides an indication of when it is possible for the

the assumption $F \propto I$ to hold in general. Because $F = [M_{A,A}^{(c)}]^{-1} M_{A,A}^{(m)}$, as in (A.24), if the missing data are missing completely at random (MCAR; see Rubin (1976)), then $\lim_{n\to\infty} F = fI$, where f is the limit of the fraction of missing data $(N-n)/N$. This result comes as no surprise, because if MCAR holds it is intuitive that the loss of information should be the same for all parameters if the original complete data have an i.i.d. structure, as in the regression setting (on the joint space of $\{X_i, Y_i\}$). It would be useful to explore the general complete-data structures under which MCAR implies $F \propto I$.

Appendix III: Larger Does Not Guarantee Better

Let $Y = X\theta + \varepsilon$, where $\varepsilon_i \overset{ind}{\sim} N(0, X_i^\eta)$ for $i = 1, \ldots, n$ and $\eta > 0$ (obviously we require X_i^η is always positive). Consider the ordinary least squares (OLS) estimator

$$\hat{\theta}_n^{LS} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}, \tag{A.29}$$

whose variance is given by

$$V_n^{LS} = \frac{\sum_{i=1}^{n} X_i^{2+\eta}}{\left(\sum_{i=1}^{n} X_i^2\right)^2}. \tag{A.30}$$

Because V_n^{LS} is not necessarily a monotone decreasing sequence in n when the X_i's are not all identical, $\hat{\theta}_n^{LS}$ can be more efficient with less data, by throwing away the part of the data with large values of X_i^η. Consider the case where $X_i = 1/(101-i)$, $i = 1, \ldots, 100$ and $\eta = 2$. Throwing away the last 36 data points will lead to a far more efficient $\hat{\theta}_{64}^{LS}$ than using all the data, $\hat{\theta}_{100}^{LS}$, since:

$$V_{64}^{LS} = 0.0214 < V_{100}^{LS} = 0.4049. \tag{A.31}$$

As with Example 1, the reason for this phenomenon is quite simple. Whereas the least-squared estimator enjoys robustness, it is consistent (and unbiased) even when $\eta \neq 0$, we pay a price in efficiency for this robustness. The equally weighted least square estimator can be terribly inefficient because it gives those data points with large X_i's—and hence large variances—much more weight than they deserve. The MLE of θ corrects this problem by properly re-weighting, leading to (assuming η is known)

$$\hat{\theta}_n^{MLE} = \frac{\sum_{i=1}^{n} X_i^{1-\eta} Y_i}{\sum_{i=1}^{n} X_i^{2-\eta}}, \tag{A.32}$$

whose variance is now monotone decreasing in n because (recall $X_i^{2-\eta} > 0$)

$$V_n^{MLE} = \Big[\sum_{i=1}^{n} X_i^{2-\eta}\Big]^{-2}. \tag{A.33}$$

With MLE, we see

$$V_{64}^{MLE} = 0.0156 > V_{100}^{MLE} = 0.01. \tag{A.34}$$

Comparing (A.31) with (A.34), we see that OLS has only 2.5% efficiency relative to MLE when we use the complete data ($n = 100$), but about 73% efficiency when we use the incomplete data ($n = 64$). It is such unbalanced loss of efficiency that causes the seemingly paradoxical phenomenon of producing a less efficient estimator with more data. Further illustrations with this example, such as how the relative efficiency of the OLS changes with the sampling mechanism, are given in Meng and Xie (2014).

References

Barnard, J. and Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* **86**, 948-955.

Blocker, A. and Meng, X.-L. (2013). The potential and perils of preprocessing: Building new foundations. *Bernoulli* **19**, 1176-1211.

Borgman, C. L. (2010). Research Data: Who will share what, with whom, when, and why? China-North America Library Conference, Beijing.

Bouman, P., Dukic, V. and Meng, X. L. (2005). A Bayesian multiresolution hazard model with application to an AIDS reporting delay study. *Statist. Sinica* **15**, 325-357.

Chernoff, H. (1983). When it seems desirable to ignore data. In *A Festschrift for Erich L. Lehmann: In Honor of His Sixty-Fifth Birthday.* (Edied by P. J. Bickel, K. Doksum and J. L. Hodges). Wadsworth Inc, California.

Copas, J. B. and Eguchi, S. (2005). Local model uncertainty and incomplete data bias (with discussion). *J. Roy. Statist. Soc. Ser. B* **67**, 459-513.

Desmond, A. F. (1997). Optimal estimating functions, quasi-likelihood and statistical modelling, *J. Statist. Plann. Inference* **60**, 77-121.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C. and Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies Sci.* **41**, 667-690.

Fay, R. E. (1991). A design-based perspective on missing data variance. *Proceedings of the* 1991 *Annual Research Conference,* 429-440, U.S. Bureau of the Census, Washington, DC.

Fay, R. E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 227-232, Alexandria, VA.

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

Kim, J. K., Brick, M. J., Fuller, W. A. and Kalton, G. (2006). On the bias of the multiple imputation variance estimator in survey sampling. *J. Roy. Statist. Soc. Ser. B* **68**, 509-521.

Kott, P. S. (1995). A paradox of multiple imputation. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 380-383, Alexandria, VA.

Li, K. H., Meng, X.-L., Raghunathan, T. E. and Rubin, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statist. Sinica* **1**, 65-92.

Li, K. H., Raghunathan, T. E. and Rubin, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *J. Amer. Statist. Assoc.* **86**, 1065-1073.

Liu, K. and Meng, X.-L. (2016). There is individualized treatment. Why not individualized inference? *Annual Rev. Statist. Its Appl.* **3**, 79-111.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statist. Sci.* **9**, 538-573.

Meng, X.-L. (2001). A congenial overview and investigation of multiple imputation inference under uncongeniality. In *Survey Nonresponse* (Edited by R. Groves, D. Dillman, J. Eltinge and R. Little), 343-356. Wiley, New York.

Meng, X.-L. (2005). Discussion: Computation, survey and inference. *Statist. Sci.* **20**, 21-28.

Meng, X.-L. (2014). A trio of inference problems that could win you a Nobel prize in statistics (If you help fund it). In *Past, Present, and Future of Statistical Science* (Edited by X. Lin, et. al), 537-562. CRC Press, Boca Raton.

Meng, X.-L. and Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79**, 103-111.

Meng, X.-L. and van Dyk, D. A. (1997). The EM algorithm - An old folk song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59**, 511-567.

Meng, X.-L. and Xie, X. (2014). I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb? *Econometric Reviews* **33**, 218-250.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. R. Soc. Lond. Ser. A* **76**, 333-380.

Nielsen, S. F. (2003). Proper and improper multiple imputation. *Internat. Statist. Rev.* **71**, 593-607.

Reiter, J. P. (2009a). Using multiple imputation to integrate and disseminate confidential microdata. *Internat. Statist. Rev.* **77**, 179-195.

Reiter, J. P. (2009b). Multiple imputation for disclosure limitation: Future research challenges. *Journal of Privacy and Confidentiality* **1**, 223-233.

Robins, J. M. and Wang N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113-124.

Rubin, D. B. (1976). Inference with missing data. *Biometrika* **63**, 581-592.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12**, 1151-1172.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *J. Amer. Statist. Assoc.* **91**, 473-489

Tu, X. M., Meng, X.-L. and Pagano, M. (1993). The AIDS epidemic: Estimating survival after AIDS diagnosis from surveillance data. *J. Amer. Statist. Assoc.* **88**, 26-36.

Wang, N. and Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85**, 935-948.

Department of Statistics, Harvard University, Cambridge, MA 02138-2901, USA.

E-mail: xie1981@gmail.com

Department of Statistics, Harvard University, Cambridge, MA 02138-2901, USA.

E-mail: meng@stat.harvard.edu

(Received February 2014; accepted July 2015)

DISCUSSION
GOD, DEVIL AND GURU IN THE LAND OF MULTIPLE IMPUTATION

Trivellore Raghunathan

University of Michigan

The multiple imputation approach for handling missing data is essentially derived from a Bayesian perspective and establishing general conditions for the validity from the repeated sampling perspective is an important, but a daunting, task. Rubin (1987) describes conditions for this validity in rather broad terms which has been subject to debate (See for example, Fay (1992); Wang and Robins (1998) and Kim et al. (2006)). The notion of uncongeniality was introduced by Meng (1994) as a framework for understanding and addressing the issues that arise when the models used by the imputer and the analyst are different, or when the analyst procedure is not fully efficient (for example, using the method of moments, instead of the maximum likelihood to estimate the parameters). This paper addresses further dissection of issues and formally establishes conditions for "validity" of multiple imputation inferences from the repeated sampling perspective. I want to commend Xianchao Xie and Xiao-Li Meng (XM, here after) for taking a highly complex topic and developing a principled way of approaching the statistical inferences when multiple imputation is used to handle

missing data. Also, I want thank the Statistica Sinica Editors for giving me the opportunity to contribute discussion to this important paper.

For this discussion, let us look at the realistic, but simplified, version of the situation: (1) The imputer and the analyst(s) operate rather independently using different model classes, each one assuming that the "God model" is captured within their model class; (2) the analyst may behave incoherently (that is, not using the optimal procedure within their own assumed model class); and (3) the Imputer might also behave incoherently by imposing assumptions and not using the optimal imputation procedure within his/her own model class.The key situations are described in Figures 1 and 2 (in XM) with concentric circles representing the model classes used by the Analyst and Imputer and a complement (not pictured) where the circles are not concentric but may even be disjoint. In this kind of possibly anarchistic situation, the question arises, what is the definition of validity? How should one conceptualize the repeated sampling thought experiment? In particular, is $\widehat{\theta}^A_{obs}$ always preferable to $\bar{\theta}_\infty$ (or the multiple imputation estimate $\bar{\theta}_M$ where M is the number of imputations)? Should we even compare V^A_{obs} to T_∞? What is the relevance of V_∞? Does the analyst even have the needed information to make the assessment?

To be concrete, suppose that the statistical experiment involves collecting information on two variables (X, Y) generated under the God model with the density function, $g_o(x, y)$. For now, assume that there are no missing values in a random sample of size n. An analyst only interested in the parameter of the marginal distribution of Y, posits the model class $p(y|\theta), \theta \in \Theta$ and assumes that the "God model" (in his/her world view), $p_o(y) = p(y|\theta_o)$ belongs to the model class. The analyst does not need the conditional distribution $q(x|y, \theta, \phi)$ or, equivalently, assume that this conditional distribution is totally arbitrary. This p-analyst develops a procedure to infer about θ and evaluates the procedure through a thought experiment that only involves repeated sampling from $p(y|\theta)$.

Similarly, another analyst only interested in the parameter of the marginal distribution of X, posits the model class, $r(x|\phi), \phi \in \Phi$, assumes that the God model (in his/her world view), $r_o(x) = r(x|\phi_o)$ belongs to the model class and leaves the conditional distribution $s(y|x, \theta, \phi)$ completely unspecified. The repeated sampling thought experiment involves only $r(x|\phi)$ for this r-analyst. The same is true for the analysts interested in the joint (g-analyst) or conditional distributions(q-analyst, s-analyst). Note that the p-analyst (or the r-analyst) is using a much richer model class compared to the g-analyst because the q-model class (or the s-model class) can be unspecified. All these analysts can operate

independently, dipping into the same well of data, without getting entangled with each other, having their own God models, their own procedures and their own way of thought experiment. This is a perfect "Hindu" setup with every aspect of the life process (marginal, conditional, joint) having its own God and the corresponding proper propitiation (procedures and thought experiments). By the way, if you don't like all this God business, then blame XM because they did it first!!

In this backdrop of complete data inference setup (all our statistical training is with this set up), the plot thickens: where there is God, there is also a Devil! The Devil keeps some values of (Y, X) intact, erases some values of X, some values of Y. The devils can have a Casper-like quality (MCAR), a benign quality (MAR) or really a malignant quality (MNAR). With the devilish actions all the peace, tranquility and independence are lost. If the Devil operates with the Casper-like quality then the analysts can still maintain independence but obtain less boons (efficiency). Let us assume that the Devil is of benign quality. Obviously the independence is lost. The thought experiment using p (or r) does not yield the desired results. The joint modeler (of (X, Y)) is the only analyst that can do something in this distraught landscape.

The Guru comes to the rescue (recognizes that the joint model is needed and can be used to deal with the inferential questions for the p or r-analysts) and uses the available information to multiply impute the missing values in X and Y, and provides several completed data sets with simple instructions for drawing inferences about θ or ϕ. In the process, however, posits a joint model $g(x, y|\theta, \phi)$ and assumes that the God model $g_o(x, y) = g(x, y|\theta_o, \phi_o)$ belongs to the class. The completed-data is not a complete data and so the p-analyst (or the r-analyst) cannot be independent because the information from the q-model (or the s-model) seeps into the completed data. The only relevant thought experiment for all the analysts is the repeated sampling from the joint model or the g-model.

Given that all the analysts (regardless of their parameters of interest) will have to work with a joint model, the dispute is a standard one between one analyst with the other, even in the complete data world: "My model" versus "Your model". Since all analysts need a joint model, the question is which is the best fitting model. The repeated sampling properties of the analyst statistics, $(\bar{\theta}_\infty, T_\infty)$, under the imputer model, if it is the best fitting model, seems to be ideal from the inferential perspective. Not sure $(\widehat{\theta}^A_{obs}, V^A_{obs})$ is even relevant in this case given the repeated sampling under the p-model is not meaningful at all.

The question is, does the analyst has any reasons to question the imputation model? Some diagnostics procedures are available for the analyst to check the imputations. See, for example, Aboyomi, Gelman and Levy (2008); Bondarenko and Raghunathan (2016). If the analyst has reason to question the imputer joint model ("your joint model") relative to his or her own joint model ("my joint model"), then the analyst can do his or her own multiple imputation inference under his/her model (or the maximum likelihood, fully Bayesian etc).

The practical situation, however, is more complex. Suppose that the imputer has the knowledge of a variable Z (For example, Y is the self-reported income and Z is income from an administrative data source, such as Tax records.) which can be used for imputation but cannot be released to the analyst. The imputer uses this additional information in the imputation process using a joint model $h(y, x, z|\lambda)$ and releases the multiply imputed data sets, $(Y^{(l)}, X^{(l)}), l = 1, 2, \ldots, M$. The analyst has no information to conceptualize the needed joint model (unless willing to make the assumption that Z is not related to the missing data mechanism or to (Y, X)). In this case, the best option for the analyst is to use $\bar{\theta}_{MI}$ and T_{MI} for inference purposes, since Figure 2 in XM is the likely scenario and the analyst has no information to model the conditional distribution of Z given (Y, X) (nor the joint distribution of (Y, X) given only the imputed data sets and a MAR mechanism, conditional on the observed values of (Y, X, Z) with Z unavailable to the analyst). In other words, the analyst has to make heroic assumptions in lieu of using the multiply imputed data sets created based on the joint distribution of (Y, X, Z).

The Example 4 illustrates this pitfall more clearly. The implicit model under which the Analyst procedure is optimal is $N(\theta, \tau^2)$. Any sensible analyst will question this judgement after a cursory inspection of the histogram of the observed and imputed values. Even in the case of a careless analyst, he or she is better off using the multiply imputed data sets rather than the observed data sample mean as the estimate. The sampling calculations under the poorly fitting models is of questionable (no?) value.

This dissection by XM also help us understand the importance of the imputer being a careful modeler of all available information and to be a trusted partner for the analysts who do not have enough information to be independent as they have been led to believe through their training in the complete-data inference system. Dealing with missing data requires a collaboration between the data producer (through careful design to collect needed information to compensate for missing data), imputer (through careful modeling and creation of imputed data sets), and

the analysts (with a penchant for using the best available procedure) to ensure that all available information are used to compensate for the missing data. Any system contrary to this collaborative efforts will only harm the analysts, in the long run. For me, the dissection by XM reinforces this point much more clearly and, perhaps, pitting the imputer against the analyst is a red-herring exercise.

References

Aboyomi, K., Gelman, A. and Levy, M. (2008). Diagnostics for multiple imputations. *Applied Statistics* **57**, 273-291.

Bondarenko, I. and Raghunathan, T. E. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputation and imputation models. *Statistics in Medicine* **35**, 3,007-3,020.

Fay, R. E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 227-232.

Kim, J. K., Michael Brick, J., Fuller, W. A. and Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 509-521.

Meng, X. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* **9**, 538-558.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.

Wang, N. and Robins, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85**, 935-948.

Survey Research Center, University of Michigan, Ann Arbor, Michigan, 48109 USA
E-mail: teraghu@umich.edu

(Received July 2016; accepted July 2016)

DISCUSSION

David Draper

University of California, Santa Cruz

This interesting and important paper encourages all of us to expand our standard horizons and consider what Xie and Meng (hereafter XM) call *multi-*

phase inference, in which

(a) different teams of analysts (or possibly even the same analysts at different points in time) may be involved in different phases of an analysis, viewed comprehensively from data collection and {data wrangling and curation} to data analysis (possibly consisting of multiple phases itself) and interpretation, but

(b) the statistical models used in some or all of the phases that involve modeling may be based on incompatible assumptions.

I can reinforce the need for multi-phase thinking in contemporary statistical work by relating some of my recent experience in data science at two large eCommerce companies, denoted (for reasons of confidentiality) by X and Y:

- In both companies, the end-product of much analysis and decision-making is a web site that can be visited by people wishing to buy or sell various items. This site is supported by a large amount of experimentation and modeling aimed at improving the user experience. Each company has between 10 and 100 groups/teams, working with various degrees of independence from each other, all tinkering with fundamental aspects of how the web site functions (an example at company X is a recommender system to help users either sharpen or broaden their searches for products similar to the one they're looking at now). It's frequently the case that the analytic output of one group forms the input to another group, and it's often true that there is sufficiently little communication between groups that the team receiving an analysis has little understanding of how it was arrived at. It may seem hard to believe that successful companies permit this level of inefficiency of communication and lack of multi-phase thinking, but they do.
- A specific example of failure to adopt a whole-systems perspective at company Y is as follows. There is a team that makes decisions on behalf of the entire company about how the data stream, generated by users of the web site, at its most granular level — time-stamped data about where in the tree of company Y's web pages the user left-clicked his or her mouse, and even spatio-temporal data tracking the location of the mouse arrow to the millisecond — is summarized for analysis by other teams in the company. I discovered that this data summary team had made a statistically unfortunate decision, namely that data that kept track of demand for a particular item in a given time period recorded a 0 for two completely different reasons:

a 0 would be entered into the data base that the rest of the company used either if no items were bought or if the item in question was not yet in the catalog of items offered to the users (!). When I inquired about what would be involved in fixing this self-inflicted problem, I was told that it would be politically unwise to pursue a solution, because the data-summary team was under a different Vice President in the corporate hierarchy than I was (!).

Alex Terenin and I have recently been thinking about a framework that includes XM's multiple imputation instance of multi-phase inference as a special case: viewing the output of one team's Bayesian analysis sequentially as the input to the next team can be referred to as *Bayesian model composition*, in the functional-analytic sense that team 1 operates on the available data D, yielding $f_1(D)$, which is then operated upon by team 2, yielding $f_2\left(f_1(D)\right)$, and so on. One question that immediately arises from this perspective is: How can team i craft its f_i in such a way that no important information is lost in the sequential analysis (when compared, for instance, with an ideal all-encompassing Bayesian analysis by a single meta-team)? In extremely simple situations we know that the usual "yesterday's posterior is today's prior" sequential use of Bayes's Theorem accomplishes this goal, but in complex settings it's not at all obvious how to build no-information-loss operators f_i. XM's work can be seen as a detailed attempt to wrestle with this question, in the context of trying to cope optimally with missing data.

- In Section 1.1 XM point out that "... the key issue is that during the journey from God's data to the analyst's data, a set of assumptions have been introduced deliberately or accidentally." I've recently run into a somewhat nonstandard example of this in the teaching of introductory statistics to undergraduates. One of my final-exam problems in fall 2015 began as follows:

  > In one of the largest and most famous public health experiments ever conducted, in 1954 a randomized controlled trial was run to see whether a vaccine developed by a doctor named Jonas Salk was effective in preventing paralytic polio. A total of 401,974 children, chosen to be representative of those who might be susceptible to the disease, were randomized to two groups: 200,745 children were injected with a harmless saline solution and the other 201,229 chil-

> dren were injected with Salk's vaccine. ... The results of the trial were as follows: 33 of the 201,229 children who got the vaccine later developed paralytic polio, whereas 115 of the other 200,745 children suffered this fate.

I had obtained the background information for this problem by the usual (and lazy) route of reading about the Salk trial in statistics textbooks. When I finally decided to do a bit of proper scholarship and dig into the literature on the Salk trial, I was amazed to find that the actual experiment was vastly messier than the textbook treatment: Meldrum (1998) tells the true story, in which 623,972 children were actually injected either with vaccine or placebo, "and more than a million others participated as 'observed' controls." Meldrum goes on as follows:

> *The statistical design used in this great experiment was singular, prompting criticism at the time and since. Eighty–four test areas in 11 [U.S.] states used the textbook model: in a randomised, blinded design all participating children in the first three grades of school (ages 6–9) received injections of either vaccine or placebo and were observed for evidence of the disease. But 127 test areas in 33 states used an "observed control" design: participating children in the second grade (ages 7–8) received injections of vaccine; no placebo was given, and children in all three grades were then observed for the duration of the polio "season."*

The sample sizes 200,745 and 201,229 appear nowhere in Meldrum's article! To paraphrase XM, in the journey from the actual trial to textbook summaries of it, a set of assumptions was introduced deliberately or accidentally, resulting in a substantial over-simplification of reality.

- The word *valid* with respect to statistical analyses is used frequently in this paper. For example, in Section 1.2 XM say "Meng (1994) obtained some initial theory under this inferential uncongeniality, including conditions for Rubin's MI inference to be *confidence valid*, i.e., the interval estimator has at least the claimed nominal coverage" (italics mine), and in Section 2 the authors offer a simple general recipe: "In general, uncongenality should be regarded as the rule rather than the exception, and a simple confidence-valid procedure to combat any degree of uncongenality is to double Rubin's MI variance estimate." While it's arguably true that failing to cover at

the advertised level is worse in confidence interval construction than creating intervals that are (much) wider than necessary to achieve the nominal coverage, I'm uncomfortable with relying only on confidence validity when what John Tukey used to refer to as *robustness of efficiency* — are the intervals indeed wider than they need to be while still hitting the coverage target? — is unaddressed: the phrase "at least the claimed nominal coverage" is equally satisfied at nominal 95% by intervals whose actual coverage is 95.01% and 99.999%, and the latter intervals will of course be substantially wider than the former.

This issue arises again in Section 5.2, where XM say "... in the context of constructing confidence intervals, confidence validity permits the actual coverage to exceed the nominal level (Neyman (1937)), and hence a [variance estimate that's biased high by an unknown amount] is accordingly acceptable." I have great respect for Mr. Neyman and his work — as it happens, he was my statistical grandfather, and (as a graduate student at Berkeley) I had the pleasure of many statistical discussions with him; I'm confident (pun intended) that Mr. Neyman would agree with me that inflated variance estimates are only useful for *a fortiori* arguments of the form "my '95%' confidence interval, based on a positively-biased variance estimate, with coverage at least nominal, doesn't include 0, so the effect I've identified is unlikely to be a statistical artifact." But what can we say if such an interval *does* include 0? XM of course understand this; they conclude Section 5.2 with the statement "... much more research is needed to investigate the general properties of these bounds ..."; hear, hear.

Having grumbled about inflated variance estimates, I'll now hypocritically congratulate XM for having made calculations leading to the simple rule "double T_∞ to yield a nominal 95% interval with actual coverage between 95% and 99.5%," and — if this were a *Royal Statistical Society* Read Paper — it would be my pleasure to either propose or second a vote of thanks.

References

Meldrum M (1998). "A calculated risk": the Salk polio vaccine field trials of 1954. *British Medical Journal* **317**, 1233–1236.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statist. Sci.* **9**, 538-573.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of

probability. *Philos. Trans. R. Soc. Lond. Ser. A* **76**, 333-380.

Baskin School of Engineering, University of California, Santa Cruz, CA 95064 USA
E-mail: draper@ucsc.edu

(Received October 2016; accepted October 2016)

DISCUSSION

David Banks and Victor Peña

Duke University

We congratulate Xianchao Xie and Xiao-Li Meng on a paper that fundamentally broadens the perspective of applied statistics. And we are deeply impressed that the authors are able to make such a significant expansion, which entails considerable mathematical complexity, and nonetheless find practical solutions that admit full and even elegant analysis. This is an important paper.

The research takes a new perspective that is relevant to many situations. Often there is true model that generates the data ("God's model"), but the data collection, cleaning and preparation process distort the data in important ways, systematically and/or stochastically. And then the statistician's analysis uses a model that is different from the one implied by the concatenation of the true model with the distortion. The Xie and Meng paper explores this situation in several imputation contexts, and finds analytic solutions and that convey insight into longstanding questions in the field (cf. Fay (1992); Kott (1995)).

Of course, as the authors point out, the need for end-to-end analysis arises ubiquitously, not just in the context of imputation. Multiphase inference could be used to understand the effects of many different processes that can be applied to "raw" data, such as are coarsening, rounding, censoring, and Winsorization. We would be interested in knowing if the authors have any thoughts about how their paradigm plays out in this broader problem space.

In some sense, a general solution strategy is straightforward. The statistician uses a nonparametric Bayesian model to represent her uncertainty about God's model, and an additional nonparametric Bayesian model to describe the distor-

tion process. Then the analyst finds the solution that maximizes her expected utility against that concatenated model for the multiphase data generation mechanism. If her uncertainty is honestly expressed, then her inference is honestly accurate. And if her prior knowledge is both honest and precise, then her solution will generally be accurate and precise as well. But if her beliefs are woefully mistaken, then her inference will often be sadly wrong. However, as the authors show in the context of imputation, multiphase applicationss are complicated and there can be counterintuitive surprises.

Of course, this general solution strategy can be difficult to implement. But there are circumstances in which the statistician has strong knowledge of the data preparation process (the imputation technique, or the number of decimal places to which the data are recorded, or the rules for handling outliers). For example, in Tu, Meng and Pagano (1993), the imputers were also the analysts, and thus the analyst had full information on the distortion. And regarding God's model, statisticians regularly address model uncertainty. Nonetheless, the Devil is in the details.

But we would like now to focus the discussion more tightly upon some research issues inspired by Example 1 in the paper. Suppose the true data generating mechanism is random sampling from the $N(\mu, \sigma^2)$ distribution, and assume there are two statisticians, Bob and Carol. For simplicity, let $\mu = 0$ and $\sigma^2 = 1$, but these values are unknown to Bob and Carol.

A sample of size N is drawn. Carol observes all of the data, but Bob sees only the first n values. But he also observes $N - n$ additional synthetic values that are generated by Carol based upon the full data set. For example, Carol might generate $N - n$ independent observations from a normal distribution with mean and variance equal to the sample mean and sample variance in the full data set. This situation could arise in practice if the last $N - n$ values were confidential.

The Xie and Meng paper gives results for estimating population means, providing sensible standard errors, and ensuring nominal coverage levels. In contrast, we consider hypothesis testing, because, if it is common for noncongeniality to strongly influence decision making, then the issue is urgent. As noted, multiphase inference arises in many cases, and we hope that statisticians have not been misled too often.

To explore this, we consider two examples. In the first, Carol provides an unbiased sample and Bob wants to test a null hypothesis. In the second, she induces a constant location bias (which is plausible in certain adversarial circumstances;

e.g., Carol may be trying to make her class's test scores seem higher), and Bob wants to estimate the population mean.

Unbiased Pre-Processing

Suppose that Carol's prior specification is $\sigma^2 \sim \text{IG}(a/2, a/2)$ and $\mu \,|\, \sigma^2 \sim N(0, \sigma^2\tau^2)$. Note that her prior expectation for μ is correct (recall that the true data generating mechanism is $N(0,1)$). After seeing the data, Carol's posterior predictive predictive distribution is a t-distribution with updated parameters. If a is large and τ^2 small, her posterior predictive will be close to the data-generating mechanism. On the other hand, if a is small and τ^2 is big, her synthetic datasets will be "unbiased" (in the sense that their marginal expectation will be correct) but will have thicker tails (and greater variance) than a $N(0,1)$ distribution, especially if the sample size is rather small.

If Bob wants to test the point null hypothesis $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$ and runs a two-sided t-test with the full data (real and synthetic), problems arise unless Carol's prior is strongly informative, especially if the fraction of unobserved individuals is not small—for any given dataset, Carol's posterior predictive distribution is always centered at a nonzero mean, so the point null hypothesis is technically wrong. Note that this difficulty cannot be circumvented by using a nonparametric test such as Wilcoxon. An easy way out is throwing away all synthetic data and performing a test with the real data, but this seems undesirable.

From a Bayesian perspective, Bob can construct a model that mimics Carol's preprocessing (which would involve modeling her imputation scheme and incorporating that belief into his analysis) and then make a decision based upon the posterior probability of the null hypothesis and his loss function. We haven't tested the practical utility of this Bayesian approach, although we believe that it would be interesting to study. The main message of our example is that even "good" preprocessing can invalidate inferences.

Biased Pre-Processing

Now suppose that Carol's prior specification is $\sigma^2 \sim \text{IG}(a/2, a/2)$ and $\mu \,|\, \sigma^2 \sim N(\delta, \sigma^2\tau^2)$, where δ could be nonzero. If $\delta \neq 0$, then Carol's prior induces a systematic location bias δ that would carry over to her posterior predictive distribution. In that circumstance, if Bob reports the sample mean using all the data he receives (real and synthetic), his estimate would be (marginally and conditionally) biased.

What could Bob do? From a Bayesian perspective, he could model his Carol's distortion of the data by putting a prior distribution on δ (which is similar in spirit to adversarial risk analysis; cf. Banks, Rios and Ríos Insua (2015)). This approach is useful if Bob has prior information about the true population mean, and could be supported by examining the difference in sample means between the n good observations and the $N - n$ synthetic observations. The practicality of that examination depends upon both the magnitude of n and δ. And, of course, it assumes that there are no "lurking" variables that induce differences between the observed and synthetic individuals.

From a frequentist perspective, if the number of observed values is sufficiently large, Bob can compare the means of the real and synthetic observations. If these means are very different, he can either discard the synthetic data or bias-correct them. Unfortunately, this approach wouldn't be applicable in the case of fully synthetic datasets, whereas the Bayesian approach can still be helpful if there is strong prior information (from other studies, for example) that the population mean should lie within a relatively narrow range.

In general, Bob can try to robustify his inferences by considering that the real and synthetic groups can have different means, and he could even consider nonparametric models to alleviate the effects of model misspecification (cf. Berger and Berliner (1986)). However, this conservative approach can lead to less precise inferences.

Some Questions and Conclusions

In summary, these are some of the future challenges that were brought to mind after reading the article (most of which were introduced in the examples):

- Should we model the process that has generated the data? If we don't, what are the implications? What are the conditions under which we can ignore the process? The answer to these questions will depend on the estimand, but how?
- What should a Bayesian do? If we truly want to reflect our uncertainty about the data-generating mechanism, we should arguably model the pre-processing/imputation steps. Our intuition suggests that "bad" subjective assessments about intermediate steps can have catastrophic consequences, whereas "good" subjective assessments can be very helpful, in that we could potentially correct for biases or mistakes that were made at some previous step.

- In some cases, inferences can (potentially) be made robust by using nonparametric approaches and "expanding" models (as in our second example, where the real and synthetic data had different means). In most cases, we would have to sacrifice some precision in the inferences. How can we quantify the precision one trades off for robustness?

We would also like to know if the authors have thought about applying multiphase inference for studying cases where the estimands are quantities that depend heavily on the tails of the distribution, or examples where sufficient statistics are hard to come by.

We end our discussion by congratulating the authors again. This paper provides a new paradigm for a large class of practical problems, it does so with mathematical power, deep insight, and a soupçon of graceful humor.

References

Banks, D., Rios, J. and Ríos Insua, D. (2015). *Adversarial Risk Analysis*. CRC Press, Boca Raton, FL.

Berger, J. and Berliner, M. (1986). Robust Bayes and empirical Bayes analysis with ϵ-contaminated priors. *Annals of Statistics* **14**, 461–486.

Fay, R. E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 227–232, Alexandria, VA.

Kott, P. S. (1995). A paradox of multiple imputation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 380–383, Alexandria, VA.

Tu, X. M., Meng, X.-L. and Pagano, M. (1993). The AIDS epidemic: Estimating survival after AIDS diagnosis from surveillance data. *Journal of the American Statistical Association* **88**, 26–36.

Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708 USA
E-mail: banks@stat.duke.edu
Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708 USA
E-mail: vp58@stat.duke.edu

(Received June 2016; accepted June 2016)

DISCUSSION

Yi-Hau Chen

Academia Sinica

1. Introduction

The authors make an important contribution to the discussion of the variance estimation of Rubin's multiple imputation (MI) inference (Rubin (1987)). In particular, assuming the imputer's model is correctly specified while the analyst's may not be, the "uncongeniality" considered in the paper, the authors identify sufficient conditions for the validity of the MI inference in terms of the relative efficiency between the imputer's and the analyst's observed-data estimators. Although it has been well known in practice that imputation should be based on a sufficiently saturated model, the results in Xie and Meng (2017), especially Theorems 6 and 7, do provide substantial new insights into how the MI inference works in general.

Using the notation in the paper, the two components of Rubin's MI variance estimator T_∞, $\bar{U}_\infty$ and B_∞, are respectively consistent estimators for the variances of $\widehat{\theta}^A_{com}$ and $\bar{\theta}_\infty - \widehat{\theta}^A_{com}$, where $\widehat{\theta}^A_{com}$ and $\bar{\theta}_\infty$ are, respectively, the analyst's complete-data and the MI estimates for the analyst's model parameter, regardless of whether the imputer's and the analysts's models are congenial or not. The sufficient and necessary condition for T_∞ consistently estimating the variance of $\bar{\theta}_\infty$ is thus

$$\mathrm{Cov}(\widehat{\theta}^{\mathrm{A}}_{\mathrm{com}}, \bar{\theta}_\infty - \widehat{\theta}^{\mathrm{A}}_{\mathrm{com}}) = \mathrm{o}(\mathrm{n}^{-1}), \tag{1.1}$$

the asymptotic orthogonality between $\widehat{\theta}^A_{com}$ and $\bar{\theta}_\infty - \widehat{\theta}^A_{com}$ (Theorem 5 of the paper). Section 6 of the paper introduces the notion of *strong efficiency* and *self efficiency* so that a sufficient condition for (1.1), and hence consistency of T_∞, to hold is that

$$\widehat{\theta}^A_{com} \text{ is self-consistent } (\widehat{\theta}^A_{com} \succ \widehat{\theta}^A_{obs}) \text{ and } \widehat{\theta}^A_{obs} \succ \widehat{\theta}^I_{obs} \tag{1.2}$$

where $a \succ b$ means "a is strongly more efficient than b".

My first comment is that, in practice, the condition (1.1) and hence consistency of T_∞ can be satisfied under more general settings than those dictated by (1.2). For example, when the analyst's inference is based on a weighted estimating equation where the weights are used to account for the mechanism of

sampling and/or missingness itself, Seaman et al. (2012) showed that, in the linear model with missing outcome data, Rubin's MI variance estimator for the analyst's estimator $\widehat{\theta}^{A}_{obs}$ obtained from the weighted estimating equation considered is consistent if the imputed outcomes are drawn from a linear model that incorporates an interaction term formed by the covariates in the analyst's model multiplied by the weight variable used. Such a result can be extended to the generalized linear model (GLM) framework considered for robust imputation discussed by Chen (2000). These results not only echo the practical and working knowledge that the imputation models should be as saturated as possible, but also indicate an explicit way to make the imputation model "saturated enough" to lead to valid MI inference. Moreover, although a fully efficient analyst's estimator such as MLE is a sufficient condition for the consistency of Rubin's MI variance estimator (Theorem 6 in Xie and Meng (2017)), the results in Seaman et al. (2012) and in the GLM framework of Chen (2000) suggest that the consistency can be reached for a general estimation-equation based analysis scheme, provided a corresponding imputation procedure ensuring valid MI inference has been designed and performed. This fact is especially encouraging given that where the missing data issue is particularly prominent, such as in longitudinal studies and complex surveys, it is rarely feasible to implement a fully efficient analysis but that some inefficient methods are usually more implementable.

The other point that may deserve further discussion is the issue of model selection for the analyst's model given that a correct (or at least approximately correct) imputation model has been employed to impute the missing data. This issue has been largely ignored in the literature. Although the authors have presented a very simple "doubling-variance" or "combining-standard-errors" procedure to ensure robust inference under incompatibility (uncongeniality) between imputer's and analysts' models, a more prudent analyst may wish to conduct a serious model comparison/selection procedure to choose the most suitable model among a pool of candidate analysis models. Shen and Chen (2013) considered information criterion-based methods for selection of the generalized estimating equation (GEE) analysis models with multiply imputed missing longitudinal data. In the setting considered in Shen and Chen (2013), although the analysis model of interest is the marginal mean model for the longitudinal outcomes, their imputation model for a missing outcome utilizes all the available information, including the observations for the past outcomes, in the hope of making the imputation as precise as possible. More in-depth studies of related issues are needed.

The points made in this discussion are meant only to highlight issues that

may warrant further considerations and investigations. The original contribution of the paper is really timely, important, and insightful, inspiring more innovative thinking in both the theory and practice of multiple imputation. I sincerely congratulate the authors on this excellent accomplishment.

References

Chen, Y.-H. (2000). A robust imputation method for surrogate outcome data. *Biometrika* **87**, 711–716.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley & Sons, New York.

Seaman, S. R., White, I. R., Copas, A. J. and Li, L. (2012). Combining multiple imputation and inverse-probability weighting. *Biometrics* **68**, 129–137.

Shen, C.-W. and Chen, Y.-H. (2013). Model selection of generalized estimating equations with multiply imputed longitudinal data. *Biometrical Journal* **55**, 899–911.

Xie, X. and Meng, X.-L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: what happens when God's, imputer's and analyst's models are uncongenial? *Statistica Sinica* **27**, 1485-1545.

Institute of Statistical Science, Academia Sinica, Nankang, Taipei, Taiwan
E-mail: yhchen@stat.sinica.edu.tw

(Received April 2016; accepted April 2016)

DISCUSSION

Anthony F. Desmond

University of Guelph

1. Introduction

When I was first invited to discuss this paper, I was stimulated by the challenge of looking at this novel area of multi-phase inference and was particularly interested in the role of estimating functions in missing data situations. At first glance, the paper looked quite congenial to me (congenial in the conventional sense of being stimulating and thought-provoking). At second reading, the brow became increasingly furrowed and I realized that this new area of multi-phase inference was going to be considerably more taxing or thorny than anticipated. I hasten to add that on subsequent readings the paper remains congenial to

me, again in the conventional sense of being thought-provoking and stimulating. However, as a newcomer to this area, I needed to understand ideas relatively new to me such as uncongeniality, self-efficiency etc. Also, old ideas, such as validity, which are (arguably, see below) relatively straightforward in single-phase inference, take on a more difficult aspect in the multi-phase paradigm as discussed in Section 1.3. I found this section particularly challenging and initially at least somewhat mystifying (perhaps even mystical!); the great varieties of actors here, Gods, demi-Gods, imputers, and analysts I found rather daunting.

2. Validity

Xie and Meng reference the classic paper by Neyman (1937) as justification for the notion of confidence validity, which implies conservative coverage properties for confidence intervals. This led me to look more closely at the early work of Neyman. Many texts such as Lehmann's (1959) classic, and also texts such as Bickel and Doksum (1977) and Casella and Berger (2002), do indeed define confidence coefficients in conservative terms, i.e. a $1 - \alpha$ confidence interval has *at least* coverage probability of $1 - \alpha$. In teaching I always felt this was just a means to allow for the difficulties with discrete distributions in attaining exact coverage probabilities for a given α; Bickel and Doksum, for example, explicitly mention this difficulty. However, it is interesting to trace the evolving attitudes of Neyman through the years via the original classic paper Neyman (1934) and subsequent papers Neyman (1935), Neyman (1937), Neyman (1941), and finally Neyman (1977).

In the 1934 paper (read before the Royal Statistical Society) Neyman presents another classic, both for its pioneering contributions to survey sampling and for the first development in English of Neyman's approach to interval estimation. This appears to be the paper in which the terms confidence interval and confidence coefficient are introduced into the English language for the first time, although earlier work in Poland, Pytkowski (1932), contains the equivalent terms, and Neyman had been using the equivalent terms in lectures in Poland for some time; see Reid (1982). Bowley, in the discussion, sees nothing new and refers to a 'confidence trick'. However, Neyman here, p. 562, does indeed use the conservative definition and defines the confidence coefficent as the lower bound for the confidence coverage. He suggests that he is merely providing alternative derivations to Fisher's fiducial intervals describing "its main lines in a way somewhat different to that followed by Fisher" and says: "Thus the new solution of the problems of estimation consists mainly in a rigorous justification of what has

been generally considered correct on more or less intuitive grounds."

On p. 586 equation (43) he again uses the conservative condition although later:

"On the contrary, erroneous judgments of the form (43) must happen, but it is known how often they will happen in the long run: their probability is *equal to* ϵ (my italics)." This is possibly just a *lapsus linguae*, to use a phrase that Neyman was fond of using in referring to Fisher's way of explaining the fiducial argument. The formal theory underlying the confidence intervals is delegated to Note I of appendix VI, pp. 589-593. Thus, one of the most influential ideas (for good or bad!) in modern statistics is relegated to a note in an appendix to an admittedly pioneering paper! Neyman's illustrative example here is the binomial , so conservatism is inevitable. There is a quasi-Bayesian flavour as admitted by the later Neyman (1977), in that an unknown prior probability distribution $\phi(\theta)$ is assumed for the collective character? (what we would now call a parameter in the population. Fisher's response is generally positive in his comments "on those applications of inductive logic which constituted so illuminating and refreshing an aspect of the evening's paper." He compliments Neyman on his generalization of the fiducial argument for "its perfect clarity". He then takes issue with three issues: (1) lack of uniqueness, (2) the uses of inequalities for discrete distributions, and (3) the difficulties in the multi-parameter case. It is (2) that is of most interest here as it appears to be an inspiration for Neyman (1935). Of course, fiducial intervals for discrete distributions posed a major difficulty for Fisher.

In the 1935 paper, Neyman revisits the problem of confidence intervals. He begins by mentioning Fisher's criticism that his (Neyman's) extension of his (Fisher's) work concerning the fiducial argument to the case of *discontinuous* (my italics) distributions is obtained at great expense, namely the replacement of equalities by inequalities. He then shows that exact equalities are not possible in general for discontinuous distributions. In particular for the binomial he suggests that the well-known Clopper-Pearson intervals involving inequalities are best possible. This is now known to be false. Recent work by Agresti and Coull (1998) and Brown, Cai and DasGupta (2001) and references therein to earlier work, show that the 'gold-standard' exact Clopper-Pearson intervals are extremely conservative and much better alternatives are available. There is an interesting discussion in the Brown et al paper as to whether one should insist on conservatism or whether being close to the nominal level is a preferable criterion, suggesting that modern statisticians are not entirely in agreement with the textbook definition. It is interesting that Neyman in 1935 continues to invoke a

prior distribution in his argument and that he appears to still regard his work as an extension of Fisher's approach to fiducial intervals.

In his classic 1937 paper, the one to which the authors refer, Neyman gives a treatment of confidence intervals, which gives a solution to the problem of confidence intervals without recourse to any a priori distribution and answers the question posed in the last sentence of his 1935 paper at least for continuous distributions. Equalities of confidence coverage and confidence coefficient are maintained throughout and in the general treatment and the examples only continuous distributions are used. In his review of previous attempts at interval estimation the fiducial argument is studiously avoided, although a footnote indicates that this review is incomplete! In later work, his notes of 1952, for example, he maintains the equality of coverage and confidence coefficient, although the conservatism for discrete distributions is touched on briefly. The later Neyman (1977), in a delightfully contentious paper in Synthese, returns to equality of coverage probabilities and again considers only continuous distributions.

My reading of Neyman (1934) is that the conservative definition of the confidence coefficient is mainly to allow for the discrete nature of the binomial example used there. My main point is that I doubt the 1934 paper could be used as a general justification for the notion of confidence validity. That is not to say that the concept of confidence validity is not a useful one. The arguments in Rubin (1996), for example, seem sound to me.

3. The Multiphase Paradigm

As I understand it, the multi-phase inference paradigm is quite a general one and the multiple imputation special case is illustrative. As an academic statistician, I tried to think of situations in which I, or colleagues of mine, might have been part of a phase (or phases) to which the multi-phase paradigm might have (retrospectively) brought some useful perspective. It is often said that applications of statistics in science and technology are piece-meal; but the multi-phase-paradigm may be an attractive way of adding a useful formalism to counter-act this. Like most academic statisticians, I have been involved in consulting projects with colleagues from the life sciences, engineering etc. The closest I may have been to a multi-phase situation was a collaborative project with General Motors Canada. This did involve much data pre-processing, but Xie and Meng make me wonder whether we were as useful to the clients, in retrospect, as we might have been. We had a large amount of data on worker behaviour at a plant in Oshawa, of a very messy nature which needed to be cleaned etc. I was not

personally involved in in the pre-processing phase. The data collection phase involved engineers and technicians at the plant with little statistical knowledge. The pre-processing of the data was done by colleagues in a group at the University of Guelph involving students, a colleague and research assistants. The ultimate motivation was to simulate the process and see how we might improve productivity and increase profits for the company. One of my tasks was to apply a plausible, but in some ways overly simple model, to a small part of this large data set, which resulted in Desmond and Desmond and Chapman (1993). Other tasks resulted in several technical reports, which the team at GM seemed to find helpful. Although aspects of the data-preprocessing and the difficulties with the original very large data set were discussed with me, I was not involved in the earlier phase. In retrospect, I would now possibly consider some statistical learning techniques for Big Data. However, this was in the early nineties and the Big Data revolution had not yet begun! The warnings on the perils of preprocessing by Blocker and Meng (2013), cited here, are definitely on my reading list.

In the Multiple Imputation case discussed in this paper, we have large public-use data-bases with statistically sophisticated imputers relative to inexperienced users (analysts), who may indeed be non-statisticians unfamiliar with missing data procedures and likely to use off-the-shelf complete-data packages. The importance of some encompassing Multiphase paradigm makes a good deal of sense here, and the multiple imputation approach, despite its critics , seems a sensible pragmatic approach to this situation. Also, arguably, Bayesian imputation and frequentist analysis are reasonable at the present time, although with the increasing acceptance and use of Bayesian methods, and more importantly, the increasing availability of Bayesian-based software, that situation may change in the future. In the case of multi-party use of large public data bases, it is unlikely that the analyst would be as sophisticated as a Meng or a Xie say! In other applications, however, where the analyst is a scientist with considerable subject-matter knowledge is the Bayesian as Imputer versus Analyst as frequentist dichotomy really necessary? Why not consider Bayesian validity at the analysis phase along with the effect of uncongeniality on inferences? For example in the Tu, Meng and Pagano (1993) example, cited in the paper, one could use the Bayesian approach to impute the delayed cases; but should this prevent a principled Bayesian from using a Bayesian version of the Cox or other analysis, which are available in survival analysis; one could even possibly use informative priors for relative risks and baseline hazard based on the first phase?

4. Use of Estimating Functions

The use of estimating functions in the context of multiple imputation and more generally multiphase inference is fascinating. The key decomposition result in Section 3 is very interesting. This seems to fit the multiple imputation particularly well. Extensions and applications in the more general multiphase scenario presents an area for future research. Some earlier work on the use of estimating for missing data is referenced in a recent encyclopedia article by Desmond (2016). Also, conditional expectations such as those of Xie and Meng, regarded as projections, in L_2 spaces are useful analytical tools in deriving optimality results in the search for good estimating functions. Small and McLeish (1994) is a good introduction to this approach. The EM algorithm, so ably summarized in Meng and van Dyk (1997), can be extended (generalized) to the estimating function situation. A particularly interesting example, when no likelihood is available, but second order assumptions are made, is Heyde and Morton (1996). They develop what they call the Projection-Solution approach, in which the E-step is replaced by a projection into a space of estimating functions determined by the second order assumptions; The M-step is then replaced by the solution step on solving the estimating equation obtained from the Projection stage. One wonders whether the Projection-Solution approach could be useful in multiphase inference? The projection idea is a very powerful one in statistics generally. Pythagoras for estimating functions, rather than estimators may be more fruitful?

5. Conservatism

Xie and Meng let $m = \infty$, for the number of imputations. My, admittedly limited, reading of the multiple imputation literature suggests m quite small, say 2 to 5, is adequate for validity and this has been advocated as an advantage of the multiple imputation approach in terms of data storage e.g. Rubin (1996) and elsewhere. Under uncongeniality how does small m affect the double the variance rule? Intuitively a more conservative rule might seem appropriate? Yet the transition from (5.14) to (5.15) for the standard error rule for a scalar estimand suggests this is not the case? As Xie and Meng mention, the 'double the variance' rule is reminiscent of the discussed paper by Copas and Eguchi (2005). Some of the discussants of that paper expressed reservations about this rule. One issue is that practitioners might use such a rule automatically, and of course, there are dangers in that, which presumably will be even more challenging in the multi-phase case. Others, e.g. Little, suggested that a lower bound for

uncertainty is not very useful. Copas and Eguchi are appropriately cautious (see their reply and page 484 of their paper). Those authors dealt with the twin issues of incomplete data and model misspecification but within a single-phase paradigm. It is interesting that this results in lower bounds for uncertainty, as opposed to the current paper, which gives conservative inferences. Of course, the former paper deals with model misspecification, whereas the current paper does not. In the multi-phase case, with incomplete data and uncongeniality, if we add in model misspecification, the statistician may have to throw his/her hands up and admit defeat! Sometimes no valid inference is possible, and a range of sensitivity analyses, possibly not very informative will be all that is feasible. In the simpler case, of single-phase, nonignorable missing data, such sensitivity analysis seems to be the only recourse. However, Xie and Meng have presented a substantial challenge to the statistics profession to deal with the unholy trinity of missingness, misspecification and uncongeniality in the multi-phase paradigm. They have made substantial advances in illustrating this paradigm in the multiple imputation case and presented many open problems for future research. They have given us much to think about. It has been a great pleasure to have had the opportunity to comment on this excellent paper, which I expect to re-read frequently for its thoughtful and challenging contributions to a new paradigm in statistical methodology.

References

Agresti, A. and Coull, B. A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions, *The American Statistician* **52**, 119-126.

Bickel, P. and Doksum, K. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* Holden Day, San Francisco.

Blocker, A. and Meng, X.-L. (2013). The potential and perils of preprocessing: Building new foundations. *Bernoulli* **19**, 1176-1211.

Brown, L. D., Cai, T. T. and DasGupta, A. (2001). Interval estimation for a binomial proportion (with discussion). *Statist. Sci.* **16**, 101-133.

Casella, G. and Berger, R. L. (2002). *Statistical Inference.* Duxbury Press, New York, NY.

Copas, J. B. and Eguchi, S. (2005). Local model uncertainty and incomplete data bias (with discussion). *Journal of the Royal Statistical Society* B **67**, 459-513.

Desmond, A. F. (2016). *Estimating Functions.* Wiley StatsRef: Statistics Reference Online, 1-11, Wiley, New York.

Desmond, A. F. and Chapman, G. R. (1993). Modelling task completion data with inverse Gaussian mixtures. *Journal of the Royal Statistical Society* C **42**, 603-613.

Heyde, C. C. and Morton, R. (1996). Quasi-likelihood and generalizing the EM algorithm. *Journal of the Royal Statistical Society* B **58**, 317-327.

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. John Wiley, New York.

Meng, X. L. and van Dyk, D. A. (1997). The EM Algorithm- An old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society* B **59**, 511-567.

Neyman, J. (1934). On the two different aspects of the representative method. *Journal of the Royal Statistical Society* B **97**, 558-625.

Neyman, J. (1935). On the problem of confidence intervals. *Annals of Mathematical Statistics* **6**, 111-116.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. R. Soc. Lond.* A **236**, 333-380.

Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika*, **32** 128-150.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese* **36**, 97-131.

Pytkowski, W. (1932). *The Dependence of the Income in Small Farms upon Their Area, the Outlay and the Capital Invested in Cows*. Bibljoteka Pulawska, Warsaw.

Reid, C. (1982). *Neyman from Life*. Springer, New York.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Association* **91**, 473-489.

Small, C. G. and McLeish, D. L. (1994). *Hilbert Space Methods in Probability and Statistical Inference*. Wiley, New York.

Tu, X. M., Meng, X.-L. and Pagano, M. (1993). The AIDS epidemic: Estimating survival after AIDS diagnosis from surveillance data. *Journal of American Statistical Association* **88**, 26-36.

Department of Mathematics and Statistics, University of Guelph, Ontario, N1G 2W1, Canada
E-mail: tdesmond@uoguelph.ca

(Received June 2016; accepted June 2016)

DISCUSSION

BY X. XIE AND X. L. MENG

Shu Yang and Jae Kwang Kim

North Carolina State University and Iowa State University

1. Introduction

We would like to first congratulate Drs. Xie and Meng on their excellent work on investigating the mystery of multiple imputation. Multiple imputation (MI) has been promoted as a general purpose estimation tool for missing data,

but there are debates over its statistical validity in many practical situations. This article will certainly serve an important building block to address these debates from a multi-phase inference perspective.

Multiple imputation was originally designed to handle missing data for public-released databases. The imputation process and subsequent analyses of the imputed datasets are separate. Therefore, this multi-phase inference features the possibility of uncongeniality. The authors focused on $m = \infty$ to avoid Monte Carlo error and introduced simple examples to highlight a number of key concepts. Specifically, we would like to discuss robustness, self-efficiency, confidence validity, and the links with the EM algorithm and fractional imputation.

2. Robustness

The authors demonstrated the hidden robustness when the analyst assumes more than the imputer through a simple example in Section 2.2. In the missing data literature, two lines of research have focused on different parts of distributions: multiple imputation models the data distribution; inverse probability weighting and doubly robust estimation (Bang and Robins (2005); Kang and Schafer (2007)) model the response probability. To gain robustness, researchers have investigated combining inverse probability weighting and multiple imputation to improve robustness of estimation (Seaman et al. (2012); Han (2015)). The authors' theory for MI can be used to cover these phenomenons.

We would like to point out that robustness is generally achievable in many imputation methods. To illustrate the idea, consider the bivariate data $(x_i, y_i), i = 1, \cdots, N$, with y_i being subject to missingness. Without loss of generality, assume the first n $y's$ are observed and the other $N - n$ $y's$ are missing. Let $m(x; \beta)$ be the "working" model for $E(Y \mid x)$ and take $\widehat{y}_i = m(x_i; \widehat{\beta})$ as the imputed value for y_i, where $\widehat{\beta}$ satisfies $\sum_{i=1}^{n}\{y_i - m(x_i; \widehat{\beta})\} = 0$. In this case, the regression imputation estimator $\widehat{\theta}_I = N^{-1}\{\sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \widehat{y}_i\}$ is algebraically equivalent to the two-phase regression estimator

$$\widehat{\theta}_{tp,reg} = N^{-1}\sum_{i=1}^{N}\widehat{y}_i + n^{-1}\sum_{i=1}^{n}(y_i - \widehat{y}_i)\,.$$

Under MCAR, using the argument in Kim and Rao (2012), $\widehat{\theta}_I$ is asymptotically unbiased regardless of the choice of $m(x_i; \beta)$. If the response probability $\widehat{\pi}_i$ is available, then we can include $\widehat{\pi}_i^{-1}$ in X so that $\sum_{i=1}^{n}\widehat{\pi}_i^{-1}(y_i - \widehat{y}_i) = 0$ holds. Then, the regression imputation estimator is algebraically equivalent to

$$\widehat{\theta}_{tp,reg} = N^{-1}\sum_{i=1}^{N}\widehat{y}_i + N^{-1}\sum_{i=1}^{n}\widehat{\pi}^{-1}\left(y_i - \widehat{y}_i\right),$$

which is also asymptotically unbiased regardless of the choice of $m(x_i;\beta)$. Thus, as long as the column space of X includes $\widehat{\pi}_i^{-1}$, the resulting imputed estimator is doubly robust. This is essentially the main idea of doubly robust imputation as discussed in Kim and Haziza (2014).

3. Self-efficiency

We believe that self-efficiency is defined with respect to an analyst's model and the missing data mechanism. We agree that self-efficiency is indeed a weaker requirement than self-sufficiency, but is frequently violated in common practice for multi-purpose estimation. Even in the ideal case when the imputer and the analyst's models are congenial, the requirement for the complete-data estimator to be self-efficient is restrictive. We have examined several scenarios, which are fairly common in practice; however they fail this requirement.

Example 1. Consider a simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$, where $\epsilon \sim N(0,\sigma^2)$, X is always observed, and Y is subject to missingness with MAR. Suppose the analyst is interested in estimating $\mu = E(Y)$ and $\eta = E\{I(Y < c)\}$, where c is a prespecified value. The complete-sample estimator solving $\sum_{i=1}^{n} Y_i - \mu = 0$ is self-efficient; however, the complete-sample estimator solving $\sum_{i=1}^{n} I(Y_i < c) - \eta = 0$ is not self-efficient.

Example 2. Consider the setup of Example 1 with $\beta_0 = 0$. Suppose the analyst is interested in estimating $\mu = E(Y)$ and consider the complete-sample estimator by solving $\sum_{i=1}^{n} Y_i - \mu = 0$. Yang and Kim (2016) claimed the Rubin's combining rule is not consistent in this case. There are two ways of viewing this in XM's framework: under the model $Y = \mu + \epsilon$, with $\epsilon \sim N(0,\sigma^2)$, the analyst's estimation procedure is self-efficient, but the model is not congenial with the imputer's model; under the model $Y = \beta_0 + \beta_1 X + \epsilon$, with $\epsilon \sim N(0,\sigma^2)$, the analyst' estimation procedure is not self-efficient.

Example 3. Consider a log linear regression model, $\log Y = X^T\beta + \epsilon$, where $\epsilon \sim N(0,\sigma^2)$. This model is especially useful for economic data that have skewed populations where the assumption of a normal distribution is unlikely to hold. Under this model, the analyst's complete-sample estimator of $\mu = E(Y)$ solving $\sum_{i=1}^{n} Y_i - \mu = 0$ is not self-efficient. This example is discussed in Yang and Kim (2015).

4. Confidence Validity Versus Type 2 Error

The authors suggest constructing a conservative variance estimator $2T_\infty$ for which the multiple imputation procedure has confidence validity. Our concern is how useful "confidence validity" is. Being conservative can protect Type 1 error, but how about Type 2 error? We can have a situation where the statistical power of the test based on MI is so low that it is better not to perform MI at all. To illustrate this point, we performed a simple simulation study. In the simulation, $B = 2,000$ Monte Carlo samples of size $n = 1,000$ were independently generated from

$$y_i = -1.5 + \beta_1 x_i + e_i, \tag{4.1}$$

where $\beta_1 \in \{0.05, 0.1, 0.15\}$, $x_i \sim N(2, 1)$, $e_i \sim N(0, 1.04)$, and x_i and e_i are independent. Variable x_i is always observed but the probability π_i that y_i responds follows $\text{logit}(\pi_i) = -1 + 0.5x_i$.

For each realized sample, we computed two estimators: the Complete-Case (CC) method that only uses the complete cases for the regression analysis and the MI estimator with $m = 100$. The imputer's and analyst's models are correctly specified as (4.1). The prior for the parameters is a flat prior.

From the imputed data, we computed the 95% confidence intervals for β_1. For the MI estimator, we used the conservative method $2T_\infty$. Table 1 shows that the MI method loses quite a bit of power compared to the CC method. While the point estimators are essentially the same in both methods, variance estimator in MI is positively biased and the test based on MI is less powerful.

Table 1. Results of power estimates for testing $H_0 : \beta_1 = 0$ based on $B = 2,000$ simulated datasets. CC: the complete-case estimator; MI: the multiple imputation estimator with the conservative variance estimator.

	CC	MI
$\beta_1 = 0.05$	0.2	0.04
$\beta_1 = 0.10$	0.56	0.28
$\beta_3 = 0.15$	0.90	0.66

5. Links with EM Algorithm and Fractional Imputation

The theoretical setup in Section 4 in XM's article serves as a general platform that links several important techniques, such as the EM algorithm (Dempster, Laird and Rubin (1977)), Data Augmentation (Tanner and Wong (1987)), and Fractional Imputation (Kim (2011); Yang and Kim (2015)). MI was originally motivated in a Bayesian prospective, but its frequentist properties have been

studied by a number of researchers via the Bernstein-von Mises theorem. See for example, Robins and Wang (2000); Yang and Kim (2016). Following the authors' notation, $\bar{\theta}_\infty$ is the solution to

$$E\{S^A(Z_{com};\theta^A) \mid Z_{obs};\widehat{\theta}^I_{obs}\} = 0. \tag{5.1}$$

Here, $S^A(Z_{com};\theta^A)$ is not necessarily the score function, rather, it is the estimating function that defines the parameter. That is, θ is defined through $E\{S^A(Z_{com};\theta^A)\} = 0$. If $S^A(Z_{com};\theta^A)$ is chosen to be the score function, the method is equivalent to the EM algorithm.

Fractional imputation is another effective imputation tool for general-purpose estimation with its advantage of not requiring the congeniality condition. With $m = \infty$, the fractional imputation estimator of θ^A is also the solution to (5.1), where $\widehat{\theta}^I_{obs}$ is a consistent estimator of θ^I in the imputation model. Rubin's approach of multiple imputation conducts separate analyses and then combining them, whereas fractional imputation creates a single weighted imputed dataset for analysis. To investigate the asymptotic variance of $\bar{\theta}^A_\infty$, we can view $\bar{\theta}^A_\infty = \bar{\theta}^A_\infty(\widehat{\theta}^I_{obs})$ and apply Taylor linearization:

$$\begin{aligned}\bar{\theta}^A_\infty(\widehat{\theta}^I_{obs}) &\cong \bar{\theta}^A_\infty(\theta^I_0) + E(\frac{\partial\bar{\theta}^A_\infty}{\partial\theta^I})(\widehat{\theta}^I_{obs} - \theta^I_0)\\ &\cong \bar{\theta}^A_\infty(\theta^I_0) - E(\frac{\partial\bar{\theta}^A_\infty}{\partial\theta^I})E\{\frac{\partial S^I(Z_{obs};\theta^I_0)}{\partial\theta^I}\}^{-1}S^I(Z_{obs};\theta^I_0),\end{aligned}$$

where $\widehat{\theta}^I_{obs}$ is the solution to $S^I(Z_{obs};\theta^I) = 0$. Thus, the variance of $\bar{\theta}^A_\infty(\widehat{\theta}^I_{obs})$ is approximated by the variance of $\bar{\theta}^A_\infty(\theta^I_0) - BS^I(Z_{obs};\theta^I_0)$, where $B = E(\partial\bar{\theta}^A_\infty/\partial\theta^I)$ $E\{\partial S^I(Z_{obs};\theta^I_0)/\partial\theta^I\}^{-1}$. This is the standard linearization method for imputation variance estimation, as discussed by Clayton et al. (1998), Robins and Wang (2000), Kim (2011), and Yang and Kim (2015). Resampling method will also provide valid variance estimation. Therefore, the fractionally imputed dataset coupled with replicated resampling weights provide another basis for consistent inference for multi-purpose usage. Of course, this may come at the price of a larger data storage space and more complex analysis.

6. Concluding Remarks

We conclude by thanking XM for their enlightening article, and we appreciate the opportunity to offer our viewpoints on this interesting problem. We look forward to their responses to our major points regarding robustness, self-efficiency, confidence validity, and the links with fractional imputation.

References

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973.

Clayton, D., Spiegelhalter, D., Dunn, G. and Pickles, A. (1998). Analysis of longitudinal binary data from multiphase sampling. *J. R. Stat. Soc. Ser. B.* **60**, 71–87.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B.* **39**, 1–38.

Han, P. (2015). Combining inverse probability weighting and multiple imputation to improve robustness of estimation. *Scandinavian J. Stat.* **43**, 246–260.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistic. Sci.* **22**, 523–539.

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

Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing survey data. *Statistica Sinica* **24**, 375–394.

Kim, J. K. and Rao, J. N. K. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika* **99**, 85–100.

Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113–124.

Seaman, S. R., White, I. R., Copas, A. J. and Li, L. (2012). Combining multiple imputation and inverse-probability weighting. *Biometrics* **68**, 129–137.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82**, 528–540.

Yang, S. and Kim, J. K. (2015). Fractional imputation in survey sampling: A comparative review. *Statistical Science* **31**, 415–432.

Yang, S. and Kim, J. K. (2016). A note on multiple imputation for method of moments estimation. *Biometrika* **103**, 244–251.

Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA.

E-mail: syang24@ncsu.edu

Department of Statistics, Iowa State University, Ames, IA 50011, USA.

E-mail: jkim@iastate.edu

(Received March 2016; accepted April 2016)

DISCUSSION

Roderick Little and Tingting Zhou

University of Michigan

Xie and Meng's paper is a theoretical tour de force, providing further insight

into the performance of multiple imputation combining rules when the imputer and analysis models differ. Implications for practice are not entirely clear, at least to us; one conclusion is to continue to use the MI combining rules, while seeking to minimize differences between the imputer and analyst models, or attempting to ensure that the differences are in the direction of making the MI combining rule conservative. Another conclusion is to abandon the Rubin's combining rules in favor of Xie and Meng's more conservative ones, although the penalties in increased width of confidence intervals seem stiff. The choice is an example of a basic question that applies to all statistics, namely what aspects of potential model misspecification should be formally reflected in measures of uncertainty. Xie and Meng's examples are instructive but perhaps more illustrative than realistic, and we describe here an extension of Example 1 that is very relevant to an applied setting.

An area where multiple imputation is receiving increased attention is in handling missing data in clinical trials. A National Research Council study (National Research Council (2010); Little et al. (2012)) advocates sensitivity analysis as an important component of the analysis of clinical trial data, and since that report there has been much activity to develop new methods and software (e.g. Mallinckrodt, Lin and Molenberghs (2013); Ratitch, O'Kelly and Tosiello (2013); Liublinska and Rubin (2014); Little et al. (2016)). The tricky modeling problem is to decide the appropriate range of models to consider in such an analysis: a narrow class may miss important possible scenarios, whereas a broad class that includes implausible models, such as "worst case" scenarios where dropouts are all considered treatment failures in the treated group and treatment successes in the control group, leads to excessively high ranges of uncertainty.

A convenient approach to sensitivity analysis, which is relatively easy to implement and convey to clinicians, models departures from missing at random via one or more sensitivity parameters that characterize differences between participants who do and do not drop out in each treatment group, after controlling for observed characteristics. This approach leads naturally to pattern-mixture models (Little (1993)) where distributions of trial outcomes are modeled conditional on the dropout indicator. Formally let D be a variable with value 1 for dropouts and 0 for participants who do not drop out. The joint distribution of D and trial outcomes Y is factored as:

$$f_{Y,D}(Y, D|Z, X, \phi, \delta) = f_{Y|D}(Y|D, Z, X, \phi, \delta) f_D(D|Z, X, \pi), \tag{1}$$

where Z is a treatment indicator, and X represents other fully observed co-

variates. More generally, D may have more than two values, corresponding to different drop-out times. The sensitivity analysis involves varying sensitivity parameters δ, a low (one or two-) dimensional parameter that characterize differences between $f_{Y|D}(Y|D=0,Z,X,\phi,\delta)$ and $f_{Y|D}(Y|D=1,Z,X,\phi,\delta)$; δ is generally not identified from the data, so the sensitivity analysis assesses the treatment effect over a range of plausible values of δ, or the size of δ is computed and assessed at the "tipping point" where statistical significance of the treatment effect is lost.

A practical approach to implementing this sensitivity analysis is to multiply-impute values of Y after dropout for each preset value of δ, and provide inferences for the parameters characterizing the treatment effect using Rubin's MI combining rules. This leads to potential uncongeniality, since the natural analysis model for Y is a model $f_Y(Y|Z,X,\theta)$ for the distribution of Y given Z, X in the absence of missing data; this analysis model is often incompatible with the imputation model of form (1). If the imputation model does not correspond to the model that generated data, resulting inferences clearly have the potential for bias. A more subtle question is the validity of MI inferences based on the model $f_Y(Y|Z,X,\theta)$ when imputations are generated under the correct model that generated the data. To shed light on this issue, we describe the results of a small simulation study, based on a realistic extension of the Xie and Meng's Example 1.

Repeated univariate samples of size $N=50$ for an outcome Y and drop-out indicator D are generated from a simple version of (1), with Z and X null, $\phi=(\mu_0,\sigma^2)$, and $\delta=(\delta_1,\delta_2)$:

$$\begin{aligned} &D \sim Bernoulli(\pi), \\ &Y|D=0 \sim \ N(\mu_0,\sigma^2), \\ &Y|D=1 \sim \ N(\mu_0+\delta_1,\delta_2^2\sigma^2). \end{aligned} \tag{2}$$

The resulting missing-data mechanism is missing not at random unless $\delta_1=0,\delta_2=1$; the sensitivity parameters are $\delta=(\delta_1,\delta_2)$, which model differences in the mean and variance of the distribution of Y for respondents and drop-outs. The marginal distribution of Y based on (2) is a mixture of normals, with mean $\theta=\mu_0+\pi\delta_1$ and variance $\tau^2=(1-\pi)\sigma^2+\pi\sigma^2\delta_2^2+\pi(1-\pi)\delta_1^2$. The target parameter is the overall population mean θ. A sample thus has n respondents with Y measured and $N-n$ dropouts with Y missing, where n is Binomial with index N and probability π.

Missing values of Y were multiply imputed (with 100 imputations) using their posterior predictive distribution, based on the correct model (2) that generated

Table 1. Empirical Bias*1,000, Root Mean Squared Error *1,000 and Confidence Interval Noncoverage (Nominal = 50) over 1000 simulated data sets of sample size of 50, for MI Inferences Under (a) PMM = the Pattern-Mixture model (2) that generated the data, with correct choice of δ, and (b) NOR = the normal complete-data model (3). $\delta_2 = 1$ and δ_1 varied from 0 to 3.

	δ_1					
$\delta_2 = 1$	0	0.5	1.0	1.5	2.0	3.0
Bias PMM	−2	0	2	4	6	10
Bias NOR	−1	−1	−1	−1	−1	−1
RMSE PMM	251	252	252	252	252	253
RMSE NOR	252	252	252	252	252	252
Noncov PMM	48	45	40	30	23	9
Noncov NOR	43	40	32	23	21	6

the data, assuming in particular the correct choice of sensitivity parameters δ, with Jeffreys' prior distributions for the parameters ϕ. The resulting MI data sets were analyzed using Rubin's combining rules, for two choices of analysis models:

(a) the pattern-mixture model that generated the data, again with the correct choice of δ, and

(b) the standard univariate normal model for the complete data,

$$Y \sim \ N(\theta, \tau^2). \tag{3}$$

Tables 1 and 2 show empirical bias, bias, root mean squared error, and 95% confidence interval coverage for the two MI analyses, over 1,000 replicate data sets. In Table 1, we set $\delta_2 = 1$ and varied δ_1 from 0 to 3, thus varying the difference in means for respondents and dropouts. In Table 2, we set $\delta_1 = 0$ and varied δ_2 from 0.2 to 5, thus varying the differences in variances for respondents and dropouts.

In both sets of simulations in Tables 1 and 2, Bayes inference based on the pattern-mixture model that generated the data had small empirical bias and confidence coverage that was close to nominal or conservative. In Table 1, where the mean is being varied but the variance is held constant, Bayes inference for the normal model yielded small empirical bias, and confidence coverage close to nominal or conservative. However in Table 2, the MI inference for the normal model had close to nominal coverage when $\delta_2 = 1$, conservative coverage when δ_2 was much less than 1, and anti-conservative when δ_2 was much greater than 1. These results are consistent with the results in Example 1 of Xie and Meng, and suggest that analyses under the normal model are robust to sensitivity analyses that concern deviations in the means between respondents and nonrespondents, but are less robust to sensitivity analyses that concern deviations in the variances. Bas-

Table 2. Empirical Bias*1,000, Root Mean Squared Error *1,000 and Confidence Interval Noncoverage (Nominal = 50) over 1,000 simulated data sets of sample size of 50, for MI Inferences Under (a) PMM = the Pattern-Mixture model (2) that generated the data, with correct choice of δ, and (b) NOR = the normal complete-data model (3). $\delta_1 = 0$ and δ_2 varied from 0.2 to 5.

	δ_2				
$\delta_1 = 0$	0.2	0.5	1	2	5
Bias PMM	−2	−2	−2	−2	−2
Bias NOR	−2	−2	−1	−1	0
RMSE PMM	251	251	251	251	251
RMSE NOR	252	252	252	253	260
Noncov PMM	48	48	48	48	48
Noncov NOR	115	98	43	1	0

ing the inference on the pattern-mixture models that generated the imputations yields more coherent results, although it deviates from current practice.

References

Little, R. J. (1993). Pattern mixture models for multivariate incomplete data. *J. Amer.Statist. Assoc.*, **88**, 125–134.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., Rotnitsky, A., Scharfstein, D., Neaton, J. D., Shih, W., Siegel, J. P. and Stern, H. (2012). Special report: The prevention and treatment of missing data in clinical trials. *New England J. Medicine*, **367**, 1355–1360.

Little, R. J., Wang, J., Sun, X., Tian, H., Suh, E-Y., Lee, M., Sarich, T., Oppenheimer, L., Plotnikov, A., Wittes, J., Cook-Bruns, N., Burton, P., Gibson, M., and Mohanty, S. (2016). The treatment of missing data in a large cardiovascular clinical outcomes study. *Clin. Trials* **13**, 344–351.

Liublinska, V. and Rubin, D. B. (2014). Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. *Statist. Med.* **33**, 4170–4185.

Mallinckrodt, C. H., Lin, Q. and Molenberghs, M. (2013). A structured framework for assessing sensitivity to missing data assumptions in longitudinal clinical trials. *Pharm Stat.* **12**(1), 1–6.

National Research Council. (2010). The prevention and treatment of missing data in clinical trials. *National Academy Press: Washington DC.*

Ratitch, B., O'Kelly, M. and Tosiello, R. (2013). Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern-mixture models. *Pharmaceut. Statist.* **12**, 337–347.

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA

E-mail: rlittle@umich.edu
Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA
E-mail: tkzhou@umich.edu

(Received July 2016; accepted July 2016)

DISCUSSION

Jerome P. Reiter

Duke University

1. Introduction

I congratulate Dr. Xie and Dr. Meng, henceforth XM, on a fascinating and deep investigation of multi-phase inference and multiple imputation. The forest that they encourage us to enter is indeed intimidating, but one could not ask for more knowledgeable and insightful guides than XM. In my discussion, I make additional connections to multi-phase inference and offer some thoughts on XM's findings on multiple imputation. I do so primarily through the lens of a government statistics agency disseminating data to the public which, as I shall describe, is a setting full of opportunities to use multi-phase inference and multiple imputation.

2. Multi-Phase Inference

Most government statistics agencies view disseminating data to the public for secondary analyses as a core mission. However, agencies do not simply dump what was collected into a public use file. Often the reported data include values that are implausible or logically inconsistent, such as a pregnant male or married three-year old, due to respondent or processing error. Including faulty values in a public use file would complicate secondary analyses, as well as undermine public trust in the quality of the data and the agency. Therefore, agencies typically "correct" faulty values through a process known as edit-imputation, in which they (1) blank some subset of values deemed responsible for making the record faulty, where the subset is selected according to some (usually unverifiable) assumption about the error-generating process, and (2) impute corrected values based on assumptions about the distribution of error-free values; see Kim et al. (2015) for

examples of this process. Missing data usually are handled as part of the edit-imputation routines. Essentially, missing values are blanked by the respondent rather than the agency.

Agencies often put data through another phase of preparation before releasing them as public use files. Most agencies are ethically and legally obligated to protect the confidentiality of data subjects' identities and sensitive attributes. Simply stripping direct identifiers like names and addresses does not suffice to protect confidentiality. Ill-intentioned individuals might be able to link the records in the public use file to identified records in some external database by matching on variables common to both files, such as demographic variables. To reduce the risks of such unintended disclosures, agencies perturb confidential values before release; see Reiter (2012) for a review of common techniques.

In many if not most datasets, agencies use both edit-imputation and redaction before releasing public use files. Typically, the edit-imputation is done in one phase, and the disclosure limitation is done in another phase, usually by a different group in the agency. Often agencies release a single dataset constructed from methods that imply restrictive assumptions about the distributions of the data. Under such approaches, it is practically impossible for secondary analysts to account for the uncertainty resulting from the data preparation phases, and, therefore, unlikely that their inferences will be confidence valid generally.

Multiple imputation (MI), however, is ideally suited for this two-phase task. In the first phase, the agency creates $m > 1$ completed datasets with all missing/faulty values filled in by MI routines. In the second phase, the agency creates $r > 1$ synthetic datasets for each completed dataset, where each synthetic dataset is generated by replacing confidential values with draws from predictive distributions estimated with the corresponding completed dataset. The result is mr released datasets, including labels indicating the nest that each synthetic dataset belongs to. Reiter (2004) shows that this two-stage imputation procedure requires a combining rule that includes three variance terms, including within-nest and between-nest variance components. In this way, the analyst (under perfect congeniality) can appropriately account for the uncertainty due to the missing/faulty data and due to the replacement of collected values with simulated ones.

It is not difficult to imagine, at least conceptually, extending this nested imputation scheme to three or more stages, with layers and nest indicators for each phase of a multi-phase data preparation process. This could enable valid multi-phase inference for multi-phase data dissemination, at least under the agency's

data preparation process and some heretofore unexplored conditions on congeniality. Of course, multi-stage data preparation and corresponding MI combining rules do not solve the problems caused by uncongeniality—indeed, they make apparent the many opportunities for mismatches in the analysis and preparation phase. The analyst's model might be uncongenial with the edit-imputation model, the disclosure limitation model, or both. This suggests an important area for research: how do we adjust multi-stage MI variance estimators to ensure confidence valid inferences (under the agency's data preparation models)? With multi-stage MI, one can imagine adjustments targeted to individual stages or, more practically, applied to a single stage in a way that ensures sufficient variance inflation. The survey sampling literature offers motivation for the latter approach. Most survey analysts estimate variances in complex, multi-stage probability samples by considering only the variance in the first stage of the sampling, ignoring variability from later stages and acting as if the data at the first stage were sampled with replacement.

Multi-stage imputation also makes apparent the multiple opportunities for the agency to make poor modeling decisions in the data preparation process. This issue is a particularly pressing concern in settings where heavy data redaction is necessary to ensure sufficient disclosure protection. There is high potential for sizable differences in the inferences the analyst makes using the redacted data and the inferences he or she could have made if given the agency's data (after missing/faulty values have been dealt with), and possibly even bigger differences from the inferences based on God's data. For many redaction strategies as applied in practice, it is very difficult for analysts to know the magnitudes of these differences for their specific analysis of interest. One solution is to let analysts have a peek under the hood in one or more of the phases. Specifically, agencies can provide analysts access to a verification server (Reiter, Oganian and Karr (2009)) that has the agency's (not God's) data and the redacted data. Analysts request that the server run a specific analysis on both the redacted and confidential data, and the server reports back measures that reflect the similarity of the two sets of inferences, e.g., how far apart are the point estimates or how much do the confidence intervals overlap. Given such feedback, analysts can decide whether or not the results from the redacted data are of sufficient quality to publish in the broad sense.

The verification server allows analysts to touch data from an earlier phase in the data preparation, enabling them to assess how the actions of a later phase impact their inferences. This strategy could be applied for other types of phases

in multi-phase data preparation. To use one of XM's examples, suppose that an agency releases a constructed variable comprising a sum of q responses, but the analyst wishes to define the variable using $p < q$ responses. The analyst could request that the server re-run the analysis using the newly defined variable. Such "earlier-phase sensitivity analyses" also could be used to assess the impact of different ways of handling missing values, as I describe at the end of the next section.

3. Multiple Imputation

XM's theoretical insights on MI solidify the rationale for long-revered advice given to imputers (make the imputation models as general as possible) and analysts (use sensible complete-data estimators). The examples used to demonstrate these conclusions involve parametric models, for which estimators with the desirable property of self-efficiency are known to exist. Often, however, analyses of public use files are design-based, for example, Horvitz and Thompson (1952) estimators of means and totals. It is not clear how well design-based estimators fit into the theoretical framework. It is well known that there is no minimum variance unbiased estimator in finite population surveys (Godambe (1955)). Given this, presumably some design-based estimators could fail to satisfy self-efficiency (even assuming a sensible re-weighting of the observed cases) in some finite populations. This suggests an intriguing question: is there any hope of general results on the consistency of the MI variance in design-based estimation? Certainly simulation evidence suggests that MI can yield consistent variance estimators and confidence valid inferences, provided that the survey design is accounted for in the imputation modeling and inferences (e.g., Reiter, Raghunathan and Kinney (2006)), but this seems a quite important trail to follow in the multi-phase inference forest.

XM's suggestions of doubling the MI variance and adding the standard deviations of the variance components are brilliant. They offer insurance against under-estimation of variance (assuming the imputation model accurately describes the data). Suppose, however, that the complete data comprise $n = 1{,}000$ randomly sampled individuals from a large population with unknown mean θ, and the missingness mechanism blanks two randomly selected values. In this case, the true repeated sampling variance of $\bar{\theta}_\infty$, the MI point estimate of the unknown θ, generally is very close to the complete data variance; that is, the true between imputation variance $E(B_\infty)$ generally is much smaller than the true within-imputation variance $E(\bar{U}_\infty)$, where the expectations are over repeated

draws from God's data. In this case, doubling the estimated MI variance (and to a lesser extent adding the standard deviations of the MI variance terms) is a heavy price to pay, as the realized $\bar{U}_\infty$ by itself is likely to be a reasonably accurate estimate of the MI variance. There may be ways to refine the rule of thumb by tuning adjustments to the magnitude of B_∞. I do not have a suggestion for how to do so, but this seems a promising path to explore. Alternatively, and more abstractly, perhaps one could give up on always bounding the true MI variance in favor of a rule that works (results in a conservative estimate of variance) a theoretically known, high percentage of times. Effectively, one could make confidence statements on whether or not confidence validity holds.

This speculation raises a philosophical question. Should confidence validity always be the primary desideratum, and if not when should we eschew it? In settings like the one above, the coverage rate of the usual MI confidence interval (without doubling the estimated variance) may be close enough to 95% that it is worth sacrificing a slight failure of confidence validity for a much shorter interval length. After all, the goal of the inference is to learn a plausible region for θ; a slightly too short interval based an unbiased estimate of θ might be deemed more useful for decision-making than a very wide, confidence valid interval based on the same unbiased estimate of θ. This suggests evaluation of MI confidence intervals (not just $\bar{\theta}_\infty$) by means of decision-theoretic frameworks rather than confidence validity alone.

Finally, in my experience, very low coverage rates in MI confidence intervals arise more often from the imputation procedure generating bias in $\bar{\theta}_\infty$ than from bias in the MI variance estimator. I have seen this especially in default applications of MI methods, for example, using main effects only in parametric conditional models in MI by chained equations, which can force convenient and possibly inaccurate distributions on the imputed values. As with the analysis of heavily redacted data, it is generally quite difficult for analysts to determine how the imputation model assumptions impact their particular inferences of interest from the released data alone.

To help analysts make such assessments, agencies could adapt verification server approaches. For example, the agency can construct a gold standard dataset out of the complete cases, punch holes in it according to a mechanism that closely mimics the distribution of missingness patterns in the collected data D, run the imputation procedure (estimated from D) to create a large number k of completed datasets, and refit the specific analysis of interest on the completed datasets. The server can repeat this process many times, each time computing whether or not

$\bar{\theta}_k \pm 1.96\sqrt{(1+1/k)B_k}$ contains the point estimate from the complete data, or computing other measures based on an analyst-specified loss function. This is not an exact measurement of the impact of the imputation phase on inferences from D, but it at least offers the analyst some insight on this potential impact.

More interestingly, the analyst might be able to use output from the server to make a "phase correction" to the inferences. For example, and writing generically, rather than use $\bar{\theta}_\infty$ and the doubled (or summed standard deviations) MI variance estimator, the analyst could make (Bayesian) inferences for θ using

$$(\bar{\theta}_\infty + \delta) - \theta \sim N(0, 2(\bar{B}_\infty + \bar{U}_\infty)), \tag{3.1}$$

$$\delta \sim f(\cdot), \tag{3.2}$$

where the distribution $f(\cdot)$ is based on the results of the repeated sampling study done by the server. For example, when the output from the server suggests the imputations could plausibly generate a bias for θ in the range (α_1, α_2), the analyst can put reasonably high probability over that range when setting $f(\cdot)$. In this way, agencies can help analysts do a better, albeit not perfect, job of propagating uncertainty in multi-phase inferences.

References

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society* **17**, 269–278.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663 – 685.

Kim, H. J., Cox, L. H., Karr, A. F., Reiter, J. P. and Wang, Q. (2015). Simultaneous editing and imputation for continuous data. *Journal of the American Statistical Association* **110**, 987–999.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology* **30**, 235–242.

Reiter, J. P. (2012). Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public Opinion Quarterly* **76**, 163–181.

Reiter, J. P., Oganian, A. and Karr, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics and Data Analysis* **53**, 1475–1482.

Reiter, J. P., Raghunathan, T. E. and Kinney, S. K. (2006). The importance of modeling the survey design in multiple imputation for missing data. *Survey Methodology* **32**, 143–150.

Department of Statistical Science, Duke University, Durham, NC 27708-0251, USA
E-mail: jerry@stat.duke.edu

(Received June 2016; accepted June 2016)

REJOINDER
PLEASE VISIT THE WILD ARBORETUM OF MULTI-PHASE INFERENCE

Xianchao Xie and Xiao-Li Meng

Harvard University

This was not an easy article to write or to publish. As with most statistical theory, the difficulty was not in proving theorems, but in formulating the relevant ones that can convey statistical insights and provide practical guidelines. A further challenge for multi-phase inference lies in finding the most intuitive and simplest ways to illustrate and explain the intricate relationships among different phases and their consequences, especially those that are counter-intuitive. It therefore took us a while to pave an entry path into the multi-phase forest, and it took even longer for us to convince enough visitors that it is not a dangerous jungle but rather a wild arboretum with many flowers and fruits, some of which are rather low-hanging.

We are therefore very grateful to the editors of *Statistica Sinica* for organizing a general tour of this relatively new landscape of statistical foundation, and to our eight brave VIPs (Very Insightful Participants) of the tour. Judging from their comments, we see that we have had a mixed success (or failure) in our attempt to provide an informative and enticing tour guide. Some shared our desire to greatly explore this landscape because the current single-phase theory does not address the increasingly common multi-phase reality. We particularly thank Banks-Peña, Draper and Reiter for their endorsements with additional examples going beyond the multiple imputation setting. Others indicated that we need to do a better job to spell out the practical relevance of our findings (e.g., Little-Zhou) and to demystify the complex world with multiple Gods and parties (Desmond and Raghunathan). Below, by addressing some major points raised by the VIPs, we hope to improve and enhance our tour guide, although we are mindful that multi-phase reality will always be more complex than any single brochure can possibly capture.

1. How Valid is Our Concept of Validity?

Several discussants (e.g., Desmond, Draper, Yang-Kim) raised the question of the usefulness of the concept of *confidence validity*, which permits a confi-

dence procedure to cover more than its nominal coverage. We particularly thank Desmond for a very helpful investigation of the origin of, and possible motivations for, Neyman's definition of this concept; Draper's personal touch, being one of Neyman's academic grandsons, is also appreciated. We also agree with Desmond that historically the allowance for over-coverage was mostly motivated by its mathematical convenience to deal with discreteness. Nevertheless, it reflected an implicit preference of Neyman and many of our founding generations to rather err on over-covering than under-covering. For practical purposes, it is trivial to come up with many examples of harmful consequences of being either overly confident or inadequately confident. However, statistical inference is *not* a symmetric game. It is a game of exclusion and contradiction, not inclusion or confirmation.

Regardless of whether our information comes from a (reliable) prior or data or both, we use inferential tools to *sharpen our inference*. That is, we reduce the region of plausible states of our inferential target by excluding those pre-inferential states that are now deemed to be implausible because they have reached a critical level of conflict with available information, as determined by a criterion specified by our inferential procedure. From this perspective, over-covering is simply a necessary step to ensure that the actual exclusion criterion used is in itself not in conflict with what is called for by our procedure. That is, we exclude a target state only when we are *sure* that it has satisfied the exclusion criteria we adopted; otherwise we have to give it the benefit of doubt. Over-covering is therefore not as much an issue of being conservative, but rather a means to ensure rigorousness and hence replicability.

Indeed, the consideration of replicability of research is a compelling reason to prefer overestimating the uncertainties in our inference, which typically implies over-coverage, than underestimating them, when the exact assessment (and hence exact coverage) cannot be achieved. Exact assessment, such as under perfect normality, is never achieved in practice – just considering all kinds of errors and approximations we make, from data defects to modeling frailties to computational corner-cutting. Much of the current crisis of non-replicable research in sciences, especially in the medical, life and social sciences, is due to our asymmetric incentive system, which effectively encourage researches to rush into "discoveries" based on quantitative evidence that does not stand up to scrutiny. Ignoring or under-assessing uncertainties, due to a whole host of mishandling, e.g., selection bias, multiple comparisons, over-fitting, etc., is a common cause.

Handling model mis-specifications, for which uncongeniality can be viewed

as a special (though unavoidable) case, via variance doubling is not a universal recipe. But it is the simplest and most applicable way of combating the common tendency of underestimating actual uncertainty, leading to many falsely significant results. Surely there will be cases where variance doubling can result in missed opportunities, due to loss of power, for example, as in Yang-Kim's simulation. This could have rather serious consequences, such as a delayed release of life-saving medication, and therefore we must be particularly cautious of applying it in those cases where false negative has graver consequences than false positive. Nevertheless, if a more sophisticated *and* justifiable approach is unavailable or non-implementable, then variance doubling is likely the lesser of the evils, the other being to ignore the issue (say) un-congeniality all together. This is because variance doubling has the added benefit of at least partially "covering" the omissions of other kinds, such as failing to take into account model uncertainty.

Raghunathan raised a deeper question about validity for multi-phase inferences, especially in the context of multiple imputation for public data files, where there are potentially many analysts. Even assuming every analyst is perfectly trained to do absolutely the best job based on the information s/he has, we still have many model classes to contemplate and each one can lead to its own version of validity, as Raghunathan's "x-analyst" example illustrates, where x can take on many values. Which validity were/are we talking about then?

As argued in Liu and Meng (2016), to define validity meaningfully we first need to determine the relevant replication setting, over which we can then determine whether some properties are replicable. In a multi-phase setting especially with multiple analysts, there are multiple ways of defining meaningful replications, including the marginal, conditional, and joint ones articulated by Raghunathan. Furthermore, shall we treat (some of) the pre-analysis phases, such as an imputation phase, fixed, or should it be a part of our replications? As we argued in the paper, whereas it is natural to consider all kinds of replications, currently we are able only to obtain useful theory under the "grand replications", that is, with respect to God's model that generates the variations for all phases. Theories under more restrictive replications, especially permitting mis-specifications, are challenging. But we hope the more challenging a problem might be, the more enticing it is for adventurous minds.

2. How Efficient is Our Formulation of Self-Efficiency?

Yang-Kim is correct that self-efficiency can easily be violated by very com-

mon procedures, such as ordinary least squares (applied to heteroscedastic models), as we demonstrated in the on-line supplementary appendix, borrowing an example from Meng and Xie (2014). Yang-Kim is also correct that when self-efficiency is violated, it is possible to recast the problem as an un-congeniality issue, because the latter is formulated via model embedding. A *self-inefficient* procedure with respect to one model can be self-efficient with respect to another; the *model* here includes both the process that generates the original compete data and the missing-data mechanism.

The examples in Yang-Kim also provide a good demonstration of the need to be explicit about the procedure being evaluated and with respect to what models—or more generally *replications* (see Liu and Meng (2016))—the evaluation is made. If we understand the notation in Yang-Kim correctly, we surmise that the commonality of their three examples is as follows. We have i.i.d. triplets $\{(Y_i, X_i, R_i)\}_{i=1}^n$, where Y_i is the outcome subject to missingness, X_i is the covariate, which is always observed, and R_i is the missing-data indicator, taking value one when Y_i is (fully) observed and zero otherwise. Our estimand θ is the marginal mean of $g(Y)$ for some pre-specified g, and our estimator is the simple average over the observed sample:

$$\hat{\theta}_{\text{obs}} = \frac{\sum_{i=1}^n R_i g(Y_i)}{\sum_{i=1}^n R_i}. \tag{2.1}$$

We emphasize that the concept of self-efficiency is defined for the *observed-data procedure*, not the *complete-data procedure*, as stated in Yang-Kim,

$$\hat{\theta}_{\text{com}} = \frac{1}{n}\sum_{i=1}^n g(Y_i), \tag{2.2}$$

because $\hat{\theta}_{\text{obs}}$ trivially specifies $\hat{\theta}_{\text{com}}$ as a special case when all $R_i = 1$, but clearly not vice versa.

The usefulness of $\hat{\theta}_{\text{obs}}$ as defined in (2.1) is well-known to depend on the missing data mechanism (MDM). Yang-Kim invoked the safe assumption of MAR, but upon checking the cited article by Yang and Kim (2016), it seems Yang-Kim's assumption is a more restrictive (but common) one, that is, Y_i and R_i are conditionally independent given the covariate X_i, for all $i = 1, \ldots, n$. Under such an assumption, it is easy to show that $\hat{\theta}_{\text{obs}}$ is unbiased for θ, and we can rely on the asymptotic result given by Theorem 4 of our paper to determine the self-efficiency of $\hat{\theta}_{\text{obs}}$. However, for the linear form (2.1), we can derive exact results for any sample sizes, which can render statistical insights without any distraction of approximation.

Specifically, by the definition of self-efficiency as given in Section 6 of our paper, $\hat{\theta}_{\text{obs}}$ is self-efficient with respect to a given MSE norm, which is the same as Var when $\hat{\theta}_{\text{obs}}$ is unbiased, if and only if $\hat{\theta}_{\text{com}}$ is orthogonal to $\hat{\theta}_{\text{obs}} - \hat{\theta}_{\text{com}}$, that is,

$$\text{Cov}(\hat{\theta}_{\text{com}}, \hat{\theta}_{\text{obs}} - \hat{\theta}_{\text{com}}) = 0. \tag{2.3}$$

But the linearity of (2.1) renders the linear decomposition

$$\hat{\theta}_{\text{com}} = r\hat{\theta}_{\text{obs}} + (1-r)\hat{\theta}_{\text{mis}}, \quad \text{where} \quad \hat{\theta}_{\text{mis}} = \frac{\sum_{i=1}^{n}(1-R_i)g(Y_i)}{\sum_{i=1}^{n}(1-R_i)} \tag{2.4}$$

and $r = \sum_{i=1}^{n} R_i/n$ is the proportion of the observed data size. Suppose our MSE calculation is conditioning on the *missing-data pattern*, that is, the values of $\{R_i\}_{i=1}^{n}$. Then under the conditional independence assumption of Y_i and R_i given X_i, $\text{Cov}(\hat{\theta}_{\text{obs}}, \hat{\theta}_{\text{mis}}) = 0$. Consequently, (2.3) is equivalent to

$$r\text{Var}(\hat{\theta}_{\text{obs}}) = \text{Var}(\hat{\theta}_{\text{com}}) \Longleftrightarrow \text{Var}(\hat{\theta}_{\text{obs}}) \propto \frac{1}{n_{\text{obs}}}. \tag{2.5}$$

That is, for the sample average (2.1) as an estimation procedure, it is (exactly) self-efficient, with respect to the MDM as previously specified, if and only if the variance of the procedure follows (exactly) the well-known inverse-sample-size rule (for all samples sizes or a sample size sufficiently large). But this is trivially true when Y_i's are i.i.d.

We were therefore puzzled initially when we read Yang-Kim's statement that (2.1) is self-efficient only in the first case of their first example. Since (2.5) is a sufficient and necessary condition (assuming $\text{Cov}(\hat{\theta}_{\text{obs}}, \hat{\theta}_{\text{mis}}) = 0$), we know that in order for this statement to hold, we must consider a different variance operation for which (2.5) will hold only for the first case of Yang-Kim's first example. Given Yang-Kim's regression-like setting, the obvious alternative choice would be the conditional variance $\text{Var}(\hat{\theta}_{\text{obs}}|\vec{X})$, where $\vec{X} = (X_1, \ldots, X_n)$. Indeed, for this choice of replications (i.e., with $\vec{X}$ fixed), $\text{Var}(\hat{\theta}_{\text{obs}}|\vec{X})$ is free of $\vec{X}$ for the first case in Yang-Kim's Example 1, where $g(Y) = Y$ and only its conditional mean depends on X, not its conditional variance. For other cases in Yang-Kim, the conditional variance of $g(Y)$ given X is not free of X either because $g(Y)$ is not linear in Y (e.g., $g(Y) = I(Y < C)$ as in their Example 1) or $g(Y)$ is linear in Y, but $\text{E}(Y|X)$ itself is not linear in X (e.g., the log-linear example in their Example 3, where $\text{E}(Y|X) = \exp\{X^{\top}\beta + \sigma^2/2\}$).

However, even if we adopt this conditional evaluation when the estimand is defined unconditionally, we still cannot conclude that the procedure in Yang-Kim's Example 2 is not self-efficient because this example is a special case of

the first case of their Example 1, by setting the regression intersection to be zero. We therefore wonder if Yang-Kim used some other variance operation for determining the procedure (2.1) is self-efficient in the first case of Example 1, but not for a special case of it as in Example 2.

Our puzzle notwithstanding, Yang-Kim's general message is the one that we share, that is, one should not take self-efficiency for granted. Fortunately, there are other ways to ensure the consistency of Rubin's variance combining rules, as Chen reported. Moreover, as we demonstrated in Section 8 of our paper, it is possible for uncongeniality to effectively cancel self-inefficiency to produce a consistent variance estimator by Rubin's combining rule, highlighting the intricate nature of multi-phase inference.

3. EM, MI, and FI – Are They Cousins?

Yang-Kim also raised the issue of the links between MI to EM and to Fractional Imputation (FI). As we stated in Section 4.2 of our paper, "performing MI with an infinite number of imputations (and with the plug-in predictive imputation) is the same as carrying out the final EM iteration." This is because the E-step of the EM algorithm evaluates the conditional expectation of the complete-data score function $S(\theta; Z_{\text{com}})$ with respect to $p(Z_{\text{mis}}|Z_{\text{obs}}, \theta = \theta^{(t)})$, where $Z_{\text{com}} = \{Z_{\text{obs}}, Z_{\text{mis}}\}$ with Z_{obs} and Z_{mis} denoting respectively the observed data and missing data. That is, at the $(t+1)th$ iteration of EM, we utilize the so-called Q-function in the EM literature (see van Dyk and Meng (2010) for an overview):

$$Q(\theta|\theta^{(t)}) = \mathrm{E}\left[S(\theta; Z_{\text{com}})|Z_{\text{obs}}, \theta = \theta^{(t)}\right]. \tag{3.1}$$

Therefore, at the last iteration of EM, we compute $Q(\theta|\theta^*)$, where $\theta^* = \lim_{t\to\infty} \theta^{(t)}$. This is equivalent to using an infinite number of draws from $p(Z_{\text{mis}}|Z_{\text{obs}}, \theta = \theta^*)$, that is, an infinite number of imputations from the "plug-in" predictive posterior to perform multiple imputation inference.

Multiple imputation, although is closely related to EM as a method to deal with missing data problems, is designed to handle more general situations where subsequent analysis with complete data can use any valid estimating method in addition to MLE. For example, if the subsequent complete-data analysis uses an estimating equation

$$U(\theta; Z_{\text{com}}) = 0, \tag{3.2}$$

then, as shown in our paper, the point estimator from MI is asymptotically

equivalent to solving the following observed-data estimating equation:

$$E(U(\theta; Z_{\text{com}})|Z_{\text{obs}}) = 0, \tag{3.3}$$

where the conditional distribution $p(Z_{\text{mis}}|Z_{\text{obs}})$ is the predictive distribution of Z_{mis} from a Bayesian model, which is also asymptotically equivalent to its frequentist's counterpart

$$E(U(\theta; Z_{\text{com}})|Z_{\text{obs}}; \theta = \theta^*) = 0, \tag{3.4}$$

where θ^* is the observed-data MLE.

The fractional imputation, as described in Yang and Kim (2016), seems to accomplish the same task as MI but via importance sampling. Specifically, it seeks to approximate (3.4) by

$$E(U(\theta; Z_{\text{com}})|Z_{\text{obs}}; \theta = \theta^*)) \approx \sum_j w_j \cdot U(\theta; Z_{\text{com}}^{(j)}|Z_{\text{obs}}; \theta = \theta^*) = 0, \tag{3.5}$$

where $w_j \propto p(Z_{\text{com}}^{(j)}|Z_{\text{obs}}; \theta = \theta^*)/h(Z_{\text{com}}^{(j)}|Z_{\text{obs}})$ is the (standardized) weight of the importance sampling with $h(Z_{\text{com}}^{(j)}|Z_{\text{obs}})$ as its (pre-chosen) proposal distribution. The accuracy of this weighting approach, as is well-understood, depends on the choice of the proposal.

If our understanding of FI is correct, then there is a link between FI to another cousin in the big family of missing-data approaches, that is, Stochastic EM (SEM; see Celeux, Chauveau and Diebolt (1996); not to be confused with the SEM algorithm of Meng and Rubin (1991) for computing variance estimators). SEM uses Monte Carlo draws from $p(Z_{\text{com}}^{(j)}|Z_{\text{obs}}; \theta = \theta^{(t)})$ to form a Monte Carlo estimator of (3.1), and then it iterates just as the standard EM. Because the resulting iterative sequence now depends on noise introduced by the Monte Carlo draws, it is stochastic. Clearly we can introduce importance sampling in approximating (3.1) as well, where the proposal density can vary with iteration—preferred for statistical efficiency, or fixed at some $h(Z_{\text{com}}^{(j)}|Z_{\text{obs}})$—preferred for computational efficiency, or a hybrid of them to achieve a sensible compromise. In that sense, SEM to FI is like EM to MI, as FI and MI can be viewed as the final iteration of SEM and EM, respectively.

Another closely related cousin is the ES algorithm investigated by Elashoff and Ryan (2004), which replaces (3.4) by

$$E(U(\theta; Z_{\text{com}})|Z_{\text{obs}}; \theta = \theta^{(t)}) = 0, \tag{3.6}$$

and then *solves* (hence the "S" in "ES") it to obtain $\hat{\theta}^{(t+1)}$. This generalizes the EM algorithm for maximizing likelihood estimation to solving a more general estimating equation with incomplete data. A special case of ES is the iterative

Projection-Solution algorithm for quasi-likelihood in Heyde and Morton (1996), as Desmond cited; we also fully agree with Desmond that projection of the estimating equation, as in (3.6), is more powerful and fruitful than projection of estimators, at least for finite-sample properties. MI then can be viewed as the final iteration of ES, but with the Expectation step carried out via Monte Carlo.

4. A Clean Theory of the Messy World of Pre-Processing?

A common theme of the multi-phase examples provided by the VIPs is that they are all *messy*. Some are necessarily so, such as protecting confidentiality, as outlined by Reiter, because it would forever be a struggle between protecting privacy and preserving information. We simply cannot have both: complete protection and full information. Others are avoidable, such as those unsettling zeros produced by the team that did not share the same VP as Draper. But the messiest of all are those cases where the analysts have little idea about what was done to their data, which is rather the rule than the exception, as in many cases of pre-processing. Could then there be any "clean" theory to deal with such messiness?

Draper outlined the idea of a Bayesian composition model, borrowing the notion of function composition, $f_2(f_1(D))$, where D denotes data, asking how f_i's should be constructed to preserve as much information as possible. A similar question was asked in Blocker and Meng (2013), in the context of distributed pre-processing, that is, what the analyst received is in the form of $\{g_j(D_j)\}_{j=1}^J$ from a system with J pre-processors (e.g., one for each experiment). The question then is what are the computationally economical and yet information-preserving choices of $g_j, j = 1, \ldots, J$? We can see clearly the competing nature of our goals: computationally, the most economical choice would be (say) to set all $g_j \equiv 0$, which is ridiculous as it preserves no information. On the other hand, choosing $g_j(D_j) = D_j, j = 1, \ldots, J$ will preserve whatever information is contained in the data, but it achieves no computational saving or any other kind of desirable pre-processing (such as privacy protection). Furthermore, preserving information is not a meaningful requirement without specifying the meaning of information or for what purposes (e.g., estimation? testing? prediction?).

But even in the classic context of sufficiency with respect to a well-specified parametric family, it is not easy at all to obtain a "clean" theory for the most economical lossless data compression. Blocker and Meng (2013) obtained sufficient conditions, as well as necessary conditions, but not sufficient *and* necessary

conditions for such $g'_j s, j = 1, \ldots, I$. A simple example suffices to illustrate the difficulty. Suppose $I = 1$ and the data $D_1 = \{Y_j\}_{j=1}^J$ are i.i.d. Poisson observations with mean θ. The pre-processor however chooses the convenient (and very wrong) model $N(\mu, 1)$, and hence he preserves its sufficient statistic $\bar{Y}_n = \sum_{i=1}^n Y_i/n$. However, since $\bar{Y}_n$ is also the sufficient statistic for θ under the Poisson model, there is no information lost even if the pre-processor used an entirely wrong model, which does not even share the same support with the correct model. This indicates the difficulties with establishing if-and-only-if conditions for pre-processing, since we can obtain the same results with very different models.

The problem becomes even harder when sufficient statistics are difficult to come by, as Banks-Peña questioned, and when information in the data is hard to quantify; and most challengingly, when the pre-processor is not well-informed of, or just unable to model, the purposes of analyses by down-stream users. But we hope these challenges will help to entice those with strong adventurous spirits to join us in our search for a "clean theory" about pre-processing. By clean theories we mean those that can either shed lights on the treacherous paths, or those that can lead to practical and effective (though not necessarily optimal) procedures, such as our variance doubling rule.

5. Is Bias-Variance Trade-Off also Critical for Multi-Phase Inference?

Yes, very much so. Yang-Kim's question on robustness of modeling, by analysts and by imputer, lies at the heart of statistical inference, and to answer it sensibly one must have full grasp of one of a very few fundamental principles of statistics, namely, the ubiquity of robustness-efficiency trade-off, a.k.a, bias-variance trade-off. "Some questions" raised by Banks-Peña, especially the last one, emphasized the very trade-off. Chen's emphasis on paying attention to (analysis) model selection touches on the same issue, because the most critical balancing act of any model selection procedure is to ensure capturing replicable signals but not to overfit the idiosyncratic individualities. Not incidently, this need for balancing presents a grand challenge for building a framework toward accumulating statistical evidence underlying *individualized inference/prediction*, but that is the subject for another hard-to-write paper. An initial attempt was made in Meng (2014) for establishing a multi-resolution framework supporting individualized inference, as one of the framework trio. (The other two cover multi-phase inference, for which our current paper is a sequel, and multi-source

inference, as in Meng (2017).)

Both Banks-Peña and Desmond raised the possibility of an all-encompassing Bayesian modelling strategy for multi-phase inference. Indeed any (serious) Bayesian can, and probably is compelled to, model the entire multi-phase as a whole, which has the added benefit of being coherent. But then there is a bias-variance trade-off. Given the *uncongenial* nature of the multi-phase paradigm, literally and technically, such modeling would necessarily need critical assumptions that are known to be false or minimally cannot be confirmed by reality, because otherwise there would not be any uncongeniality in the first place. And even seemingly "good" pre-processing models can (and often) lead to provably undesirable results, as Banks-Peña's Carlo-Bob example further illustrated. This is what makes the multi-phase inference paradigm interesting, intriguing, and inspiring. The many examples from government statistics agencies, especially under the mandate of disclosure protection, as succinctly summarized by Reiter, and from industrial and business sectors, as vividly illustrated by Desmond, highlighted the urgent need of developing this paradigm.

Indeed, as Desmond correctly recognized, ultimately the multi-phase paradigm needs to handle an unholy trinity: missingness, misspecification, and uncongnialty. In comparison to this grand goal, what we presented in the current paper is only one of many needed building blocks. We are therefore humbled by the kind encouragements from the VIPs, especially the extremely flattering endorsement from Banks-Peña, Chen, Desmond and Reiter. We also particularly thank Draper for his *RSS* style vote for thanks, regardless of whether he would propose or second, especially because initially we did plan to seek such a vote. Ultimately, our long journey of dealing with uncongenality led us to the welcoming arms of *Statistica Sinica*, to which we are deeply grateful.

References

Blocker, A. W. and Meng, X.-L. (2013). The potential and perils of preprocessing: Building new foundations. *Bernoulli* **19**, 1176–1211.

Celeux, G., Chauveau, D. and Diebolt, J. (1996). Stochastic versions of the EM algorithm: an experimental study in the mixture case. *Journal of Statistical Computation and Simulation* **55**, 287–314.

Elashoff, M. and Ryan, L. (2004). An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics* **13**, 48–65.

Heyde, C. and Morton, R. (1996). Quasi-likelihood and generalizing the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 317–327.

Liu, K. and Meng, X.-L. (2016). There is individualized treatment. Why not individualized inference? *Annual Review of Statistics and Its Applications* **3**, 79–111.

Meng, X.-L. (2014). A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it). in: *Past, Present, and Future of Statistical Science (Eds: Lin et. al.)*, CRC Press, pp. 537–562.

Meng, X.-L. (2017). Statistical paradises and paradoxes in big data (I): The bigger the data, the surer we fool ourselves? *Technical Report, Department of Statistics, Harvard University.*

Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association* **86**, 899–909.

Meng, X.-L. and Xie, X. (2014). I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb? *Econometric Reviews* **33**, 218–250.

van Dyk, D. A. and Meng, X.-L. (2010). Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: A graphical guide book. *Statistical Science* **25**, 429–449.

Yang, S. and Kim, J. K. (2016). Fractional imputation in survey sampling: A comparative review. *Statistical Science* **31**, 415–432.

Department of Statistics, Harvard University, Cambridge, MA 02138-2901, USA.

E-mail: xie1981@gmail.com

Department of Statistics, Harvard University, Cambridge, MA 02138-2901, USA.

E-mail: meng@stat.harvard.edu

(Received February 2017; accepted February 2017)