Statistica Sinica Preprint No: SS-2016-0187	
Title	Discussion on “Dissecting Multiple Imputation from a Multi-phase Inference Perspective: What Happens When God's, Imputer's and Analyst's Models Are Uncongenial?”
Manuscript ID	SS-2016-0187
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0187
Complete List of Authors	Yi-Hau Chen
Corresponding Author	Yi-Hau Chen
E-mail	yhchen@stat.sinica.edu.tw

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0187

Discussion on "Dissecting Multiple Imputation from a Multi-phase Inference Perspective: What Happens When God's, Imputer's and Analyst's Models Are Uncongenial?"

Yi-Hau Chen

Institute of Statistical Science, Academia Sinica

yhchen@stat.sinica.edu.tw

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 (2016), 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 $\hat{\theta}^A_{com}$ and $\bar{\theta}_\infty - \hat{\theta}^A_{com}$, where $\hat{\theta}^A_{com}$ 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

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0187

the variance of $\bar{\theta}_\infty$ is thus

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

the asymptotic orthogonality between $\hat{\theta}^A_{com}$ and $\bar{\theta}_\infty - \hat{\theta}^A_{com}$ (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), and hence consistency of T_∞, to hold is that

$$\hat{\theta}^A_{com} \text{ is self-consistent } (\hat{\theta}^A_{com} \succ \hat{\theta}^A_{obs}) \text{ and } \hat{\theta}^A_{obs} \succ \hat{\theta}^I_{obs} \tag{2}$$

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

My first comment is that, in practice, the condition (1) and hence consistency of T_∞ can be satisfied under more general settings than those dictated by (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 $\hat{\theta}^A_{obs}$ 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 (2016)), the results in Seaman et al. (2012) and in the GLM framework of

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0187

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

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0187

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.

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. (2016). Dissecting multiple imputation from a multi-phase inference perspective: what happens when God's, imputer's and analyst's models are uncongenial? *Statistica Sinica*, in print.