Statistica Sinica Preprint No: SS-2016-0344	
Title	Discussion of "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-2016-0344
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0344
Complete List of Authors	Trivellore Raghunathan
Corresponding Author	Trivellore Raghunathan
E-mail	teraghu@umich.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0344

God, Devil and Guru in the Land of Multiple Imputation

Trivellore Raghunathan *

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 (1991), 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,

*Director and Research Professor, Survey Research Center, Professor of Biostatistics, University of Michigan, Ann Arbor, Michigan; e-mail:teraghu@umich.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0344

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)

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0344

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

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0344

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

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0344

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.

Additional References

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

2. 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, 3007-3020.