Statistica Sinica Preprint No: SS-2017-0069	
Title	Rejoinder: Please Visit the Wild Arboretum of Multi-Phase Inference
Manuscript ID	SS-2017-0069
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0069
Complete List of Authors	Xiao-Li Meng
Corresponding Author	Xiao-Li Meng
E-mail	meng@stat.harvard.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0069

Rejoinder: Please Visit the Wild Arboretum of Multi-Phase Inference[1]

Xianchao Xie and Xiao-Li Meng

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

[1]We thank NSF and JTF for partial financial support, and Steven Finch for proofreading.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0069

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 confidence 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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0069

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,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0069

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 [1], 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 common 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 [2]. Yang-Kim is also correct that when self-efficiency is violated, it is possible to recast the problem as an uncongeniality 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.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0069

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 [1])—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:

$$\hat{\theta}_{\text{obs}} = \frac{\sum_{i=1}^n R_i g(Y_i)}{\sum_{i=1}^n R_i}. \tag{1}$$

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,

$$\hat{\theta}_{\text{com}} = \frac{1}{n}\sum_{i=1}^n g(Y_i), \tag{2}$$

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 (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 [3], 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 (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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0069

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,

$$\text{Cov}(\hat{\theta}_{\text{com}}, \hat{\theta}_{\text{obs}} - \hat{\theta}_{\text{com}}) = 0. \tag{3}$$

But the linearity of (1) renders the linear decomposition

$$\hat{\theta}_{\text{com}} = r\hat{\theta}_{\text{obs}} + (1-r)\hat{\theta}_{\text{mis}}, \quad \text{where} \quad \hat{\theta}_{\text{mis}} = \frac{\sum_{i=1}^{n}(1-R_i)g(Y_i)}{\sum_{i=1}^{n}(1-R_i)} \tag{4}$$

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, (3) is equivalent to

$$r\text{Var}(\hat{\theta}_{\text{obs}}) = \text{Var}(\hat{\theta}_{\text{com}}) \Longleftrightarrow \text{Var}(\hat{\theta}_{\text{obs}}) \propto \frac{1}{n_{\text{obs}}}. \tag{5}$$

That is, for the sample average (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 (1) is self-efficient only in the first case of their first example. Since (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 (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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0069

is not linear in X (e.g., the log-linear example in their Example 3, where $\mathrm{E}(Y|X) = \exp\{X^\top\beta + \sigma^2/2\}$)).

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 (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 $S(\theta; Z_{\mathrm{com}})$ with respect to $p(Z_{\mathrm{mis}}|Z_{\mathrm{obs}}, \theta = \theta^{(t)})$, where $Z_{\mathrm{com}} = \{Z_{\mathrm{obs}}, Z_{\mathrm{mis}}\}$ 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 [4] for an overview):

$$Q(\theta|\theta^{(t)}) = \mathrm{E}\left[S(\theta; Z_{\mathrm{com}})|Z_{\mathrm{obs}}, \theta = \theta^{(t)}\right]. \tag{6}$$

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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0069

$p(Z_{\text{mis}}|Z_{\text{obs}}, \theta = \theta^*)$, 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

$$U(\theta, ; Z_{\text{com}}) = 0, \tag{7}$$

then, as shown in our paper, the point estimator from MI is asymptotically equivalent to solving the following observed-data estimating equation:

$$E(U(\theta; Z_{\text{com}})|Z_{\text{obs}}) = 0, \tag{8}$$

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

$$E(U(\theta; Z_{\text{com}})|Z_{\text{obs}}; \theta = \theta^*) = 0, \tag{9}$$

where θ^* is the observed-data MLE.

The fractional imputation, as described in Yang and Kim [3], seems to accomplish the same task as MI but via importance sampling. Specifically, it seeks to approximate (9) by

$$E(U(\theta; Z_{\text{com}})|Z_{\text{obs}}; \theta = \theta^*)) \approx \sum_j w_j \cdot U(\theta; Z_{\text{com}}^{(j)}|Z_{\text{obs}}; \theta = \theta^*) = 0, \tag{10}$$

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 et al. [5]; not to be confused with the SEM

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0069

algorithm of Meng and Rubin [6] 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 (6), 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 (6) 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 [7], which replaces (9) by

$$E(U(\theta;Z_{\text{com}})|Z_{\text{obs}};\theta=\theta^{(t)})=0, \tag{11}$$

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 [8], as Desmond cited; we also fully agree with Desmond that projection of the estimating equation, as in (11), 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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0069

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 [9], 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 [9] 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 preprocessor 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

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0069

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 downstream 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 [10] 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 [11].)

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0069

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

[1] K. Liu, X.-L. Meng, There is individualized treatment. Why not individualized inference?, Annual Review of Statistics and Its Applications 3 (2016) 79–111.

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

[3] S. Yang, J. K. Kim, Fractional imputation in survey sampling: A comparative review, Statistical Science 31 (2016) 415–432.

[4] D. A. van Dyk, X.-L. Meng, Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: A graphical guide book, Statistical Science 25 (2010) 429–449.

[5] G. Celeux, D. Chauveau, J. Diebolt, Stochastic versions of the EM algorithm: an experimental study in the mixture case, Journal of Statistical Computation and Simulation 55 (1996) 287–314.

[6] X.-L. Meng, D. B. Rubin, Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm, Journal of the American Statistical Association 86 (1991) 899–909.

[7] M. Elashoff, L. Ryan, An EM algorithm for estimating equations, Journal of Computational and Graphical Statistics 13 (2004) 48–65.

[8] C. Heyde, R. Morton, Quasi-likelihood and generalizing the EM algorithm, Journal of the Royal Statistical Society. Series B (Methodological) (1996) 317–327.

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

[10] X.-L. Meng, 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, 2014, pp. 537–562.

[11] X.-L. Meng, Statistical paradises and paradoxes in big data (I): The bigger the data, the surer we miss our target?, Technical Report, Department of Statistics, Harvard University (2017).