Statistica Sinica Preprint No: SS-2019-0314	
Title	Multiple Improvements of Multiple Imputation Likelihood Ratio Tests
Manuscript ID	SS-2019-0314
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0314
Complete List of Authors	Kin Wai Chan and Xiao-Li Meng
Corresponding Author	Kin Wai Chan
E-mail	kinwaichan@cuhk.edu.hk

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

Statistica Sinica

MULTIPLE IMPROVEMENTS OF MULTIPLE IMPUTATION LIKELIHOOD RATIO TESTS

Kin Wai Chan[1] and Xiao-Li Meng[2]

Department of Statistics, The Chinese University of Hong Kong[1]

Department of Statistics, Harvard University[2]

Abstract: Multiple imputation (MI) inference handles missing data by imputing the missing values m times, and then combining the results from the m complete-data analyses. However, the existing method for combining likelihood ratio tests (LRTs) has multiple defects: (i) the combined test statistic can be negative, but its null distribution is approximated by an F-distribution; (ii) it is not invariant to re-parametrization; (iii) it fails to ensure monotonic power owing to its use of an inconsistent estimator of the fraction of missing information (FMI) under the alternative hypothesis; and (iv) it requires nontrivial access to the LRT statistic as a function of parameters instead of data sets. We show, using both theoretical derivations and empirical investigations, that essentially all of these problems can be straightforwardly addressed if we are willing to perform an additional LRT by stacking the m completed data sets as one big completed data set. This enables users to implement the MI LRT without modifying the complete-data procedure. A particularly intriguing finding is that the FMI can be estimated consistently by an LRT statistic for testing whether the m completed data sets can be regarded

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

effectively as samples coming from a common model. Practical guidelines are provided based on an extensive comparison of existing MI tests. Issues related to nuisance parameters are also discussed.

Key words and phrases: Fraction of missing information, missing data, invariant test, monotonic power, robust estimation.

1. Historical Successes and Failures

1.1 The Need for Multiple Imputation Likelihood-Ratio Tests

Missing-data problems are ubiquitous in practice, to the extent that the absence of any missingness is often a strong indication that the data have been pre-processed or manipulated in some way (e.g., Blocker and Meng, 2013). Multiple imputation (MI) (Rubin, 1978, 2004) has been a preferred method, especially by those who are ill-equipped to handle missingness on their own, owing to a lack of information or skills or resources. MI relies on the data collector (e.g., a census bureau) building a reliable imputation model to fill in the missing data $m(\geqslant 2)$ times. In this way, users can apply their preferred software or procedures designed for complete data, and do so m times. MI inference is then performed by appropriately combining these m complete-data results. Note that in a typical analysis of public MI data, the analyst has no control over or understanding of how the imputation was done, including the choice of the model and m, which is

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

often small in reality (e.g., $3 \leqslant m \leqslant 10$). The analyst's job is to analyze the given m completed data sets as appropriately as possible, but only using complete-data procedures or software routines.

Although MI was designed initially for public-use data sets, over the years, it has become a method of choice in general, because it separates handling the missingness from the analysis (e.g., Tu *et al.*, 1993; Rubin, 1996, 2004; Schafer, 1999; King *et al.*, 2001; Peugh and Enders, 2004; Kenward and Carpenter, 2007; Rose and Fraser, 2008; Holan *et al.*, 2010; Kim and Yang, 2017). Software routines for performing MI are now available in R (Su *et al.*, 2011), Stata (Royston and White, 2011), SAS (Berglund and Heeringa, 2014), and SPSS; see Harel and Zhou (2007) and Horton and Kleinman (2007) for summaries.

This convenient separation, however, creates an issue of uncongeniality, that is, an incompatibility between the imputation model and the subsequent analysis procedures (Meng, 1994a). This issue is examined in detail by Xie and Meng (2017), who show that uncongeniality is easiest to deal with when the imputer's model is more saturated than the user's model/procedure, and when the user is conducting an efficient analysis, such as a likelihood inference. Therefore, this study focuses on conducting MI likelihood ratio tests (LRTs), assuming the imputation model is sufficiently saturated to render the common assumptions made in the literature about conducting LRTs with MI valid.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

Like many hypothesis testing procedures in common practice, the exact null distributions of various MI test statistics, LRTs or not, are intractable. This intractability is not computational, but rather statistical, owing to the well-known issue of a nuisance parameter, that is, the lack of a pivotal quantity, as highlighted by the Behrens–Fisher problem (Wallace, 1980). Indeed, the nuisance parameter in the MI context is the so-called "fraction of missing information" (FMI), which is determined by the ratio of the between-imputation variance to the within-imputation variance (and its multi-variate counterparts). Hence, the challenge we face is almost identical to the one faced by the Behrens–Fisher problem, as shown in Meng (1994b). Currently the most successful strategy has been to reduce the number of nuisance parameters to one by assuming an equal fraction of missing information (EFMI), a strategy we follow as well because our simulation study indicates that it achieves a better compromise between type-I and type-II errors than other strategies we (and others) have tried.

An added challenge in the MI context is that the user's complete-data procedures can be very restrictive. What is available to the user could vary from the entire likelihood function to point estimators such as the MLE and Fisher information to a single p-value. Therefore, there have been a variety of procedures proposed in the literature, depending on what quantities we assume the user has access to, as we review shortly.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

Among them, a promising idea is to directly combine LRT statistics. However, the current execution of this idea (Meng and Rubin, 1992) relies too heavily on the asymptotic equivalence (in terms of the data size, not the number of imputations, m) between the LRT and Wald test *under the null*. Its asymptotic validity, unfortunately, does not protect it from quick deterioration for small data sizes, such as delivering a negative "F test statistic" or FMI. Worst of all, the test can have essentially zero power because the estimator of the FMI can be badly inconsistent under some alternative hypotheses. The combining rule of Meng and Rubin (1992) also requires access to the LRT as a function of parameter values, not just as a function of the data. The former is often unavailable from standard software packages. This defective MI LRT, however, has been adopted by textbooks (e.g., van Buuren S, 2012; Kim and Shao, 2013) and popular software, for example, the function `pool.compare` in the R package `mice` (van Buuren and Groothuis-Oudshoorn, 2011), the function `testModels` in the R package `mitml` (Grund *et al.*, 2017), and the function `milrtest` (Medeiros, 2008) in the Stata module `mim` (Carlin *et al.*, 2008).

To minimize the negative impact of this defective LRT test, this study derives MI LRTs that are free of these defects, as detailed in Section 1.5. We achieve this mainly by switching the order of two main operators in the combining rule of Meng and Rubin (1992): we maximize the average of the m log-likelihoods

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

instead of averaging their maximizers. This switch, guided by the likelihood principle, renders positivity, invariance, and monotonic power. Other judicious uses of the likelihood functions permit us to overcome the remaining defects.

1.2 Summary of the Major Findings

Our major contributions are four-fold:

- In terms of statistical principles, we propose switching the order of two operations, namely maximization and averaging, in the existing MI LRT statistic, as suggested by the likelihood principle. This operation retrieves the non-negativity and invariance to the re-parametrization of the MI statistic.

- In terms of theoretical properties, a new estimator of the fraction of missing information is proposed. It is consistent, regardless of the validity of the null hypothesis, so that the proposed test is monotonically powerful with respect to the discrepancy between the null and alternative hypotheses.

- In terms of computational properties, the proposed test only requires that users have a standard subroutine for performing a complete-data LRT. Thus, unlike the existing MI LRT, users do not need to modify the subroutine in order to evaluate the likelihood function at arbitrary parameter values.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

- In terms of practical impact, the proposed test can be implemented easily to replace the flawed MI LRT procedures in the aforementioned software packages and beyond. It immediately resolves the issue of returning a negative F-test value. In addition, the power loss due to the flaws in the MI LRT procedure can be retrieved.

The remainder of Section 1 provides background and notation. Section 2 discusses the defects of the existing MI LRT and our remedies. Section 3 investigates the computational requirements, including theoretical considerations and comparisons. In particular, Algorithm 1 of Section 3.1 computes our most recommended test. Section 4 provides empirical evidence. Section 5 concludes the paper. Appendices A and B provide additional investigations, real-life data examples, and proofs.

1.3 Notation and Complete-Data Tests

Let X_{obs} and X_{mis} be, respectively, the observed and missing parts of an intended complete data set $X = X_{\text{com}} = \{X_{\text{obs}}, X_{\text{mis}}\}$ consisting of n observations. Denote the sampling model — probability or density, depending on the data type — of X by $f(\cdot \mid \psi)$, where $\psi \in \Psi \subseteq \mathbb{R}^h$ is a vector of parameters. Suppose that we are interested in inferring $\theta = \theta(\psi) \in \Theta \subseteq \mathbb{R}^k$, which is expressed as a function of ψ. This definition of θ is very general. For example, θ can be a sub-vector of

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

$\psi = (\theta^\intercal, \eta^\intercal)^\intercal$, or a transformation (not necessarily one-to-one) of ψ; see Section 4.4 of Serfling (2001) and Section 6.4.2 of Shao (1998).

The goal is to test $H_0 : \theta = \theta_0$ when only X_{obs} is available, where θ_0 is a specified vector. For example, if H_0 puts a k-dimensional restriction $R(\psi) = \mathbf{0}$ on the model parameter ψ, then $\theta = R(\psi)$ and $\theta_0 = \mathbf{0}$. For simplicity, we focus on a two-sided alternative, but our approach adapts to general LRTs. Here, we assume X_{obs} is rich enough that the missing data mechanism is ignorable (Rubin, 1976), or it has been properly incorporated by the imputer, who may have access to additional confidential data.

Let $\widehat{\theta} = \widehat{\theta}(X)$, $\widehat{\psi} = \widehat{\psi}(X)$, and $\widehat{\psi}_0 = \widehat{\psi}_0(X)$ be the complete-data MLE of θ, complete-data MLE of ψ, and H_0-constrained complete-data MLE of ψ, respectively. Furthermore, let $U = U_\theta = U_\theta(X)$ and $U_\psi = U_\psi(X)$ be efficient estimators of $\text{Var}(\widehat{\theta})$ and $\text{Var}(\widehat{\psi})$, respectively, for example, the inverse of the observed Fisher information. Common test statistics for H_0 include the Wald statistic $D_{\text{W}} = d_{\text{W}}(\widehat{\theta}, U)/k$ and the LRT statistic $D_{\text{L}} = d_{\text{L}}(\widehat{\psi}_0, \widehat{\psi} \mid X)/k$, where

$$d_{\text{W}}(\widehat{\theta}, U) = (\widehat{\theta} - \theta_0)^\intercal U^{-1} (\widehat{\theta} - \theta_0), \qquad d_{\text{L}}(\widehat{\psi}_0, \widehat{\psi} \mid X) = 2 \log \frac{f(X \mid \widehat{\psi})}{f(X \mid \widehat{\psi}_0)}.$$

Under regularity conditions, such as those in Section 4.2.2 and Section 4.4.2 of Serfling (2001), we have the following classical results.

Property 1.1. *Under H_0, (i) $D_{\text{W}} \Rightarrow \chi^2_k/k$ and $D_{\text{L}} \Rightarrow \chi^2_k/k$; and (ii) $n(D_{\text{W}} -$*

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

$D_{\mathrm{L}}) \xrightarrow{\mathrm{pr}} 0$ *as* $n \to \infty$, *where "*$\Rightarrow$*" and "*$\xrightarrow{\mathrm{pr}}$*" denote convergence in distribution and in probability, respectively.*

Testing H_0 based on X_{obs} is more involved. For MI, let $X^{(\ell)} = \{X_{\mathrm{obs}}, X_{\mathrm{mis}}^{(\ell)}\}$, for $\ell = 1, \ldots, m$, be the m completed data sets, where $X_{\mathrm{mis}}^{(\ell)}$ are drawn from a proper imputation model (Rubin, 2004). We then carry out a complete-data estimation or testing procedure on $X^{(\ell)}$, for $\ell = 1, \ldots, m$, resulting in a set of m quantities. The so-called MI inference combines them to obtain a single answer. Note that the setting of MI is such that the user is unable or unwilling to carry out the test based directly on the observed data X_{obs}.

1.4 MI Wald Test and Fraction of Missing Information

Let $d_{\mathrm{W}}^{(\ell)} = d_{\mathrm{W}}(\widehat{\theta}^{(\ell)}, U^{(\ell)})$, $\widehat{\theta}^{(\ell)} = \widehat{\theta}(X^{(\ell)})$, and $U^{(\ell)} = U(X^{(\ell)})$ be the imputed counterparts of $d_{\mathrm{W}}(\widehat{\theta}, U)$, $\widehat{\theta}$, and U, respectively, for each ℓ. In addition, let

$$\overline{d}_{\mathrm{W}} = \frac{1}{m}\sum_{\ell=1}^{m} d_{\mathrm{W}}^{(\ell)}, \qquad \overline{\theta} = \frac{1}{m}\sum_{\ell=1}^{m} \widehat{\theta}^{(\ell)}, \qquad \overline{U} = \frac{1}{m}\sum_{\ell=1}^{m} U^{(\ell)}. \tag{1.1}$$

Under congeniality (Meng, 1994a), one can show that asymptotically (Rubin and Schenker, 1986) $\mathrm{Var}(\overline{\theta})$ can be consistently estimated by

$$T = \overline{U} + (1 + 1/m)B, \quad \text{where} \quad B = \frac{1}{m-1}\sum_{\ell=1}^{m}(\widehat{\theta}^{(\ell)} - \overline{\theta})(\widehat{\theta}^{(\ell)} - \overline{\theta})^{\intercal} \tag{1.2}$$

is known as the *between-imputation variance*, in contrast to $\bar{U}$ in (1.1), which measures the *within-imputation variance*. Intriguingly, $2T$ serves as a universal

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

(estimated) upper bound of $\mathrm{Var}(\overline{\theta})$ under uncongeniality (Xie and Meng, 2017). Under regularity conditions, we have that, as $m, n \to \infty$,

$$n(\overline{U} - \mathscr{U}_\theta) \xrightarrow{\mathrm{pr}} \mathbf{0}, \qquad n(T - \mathscr{T}_\theta) \xrightarrow{\mathrm{pr}} \mathbf{0}, \qquad n(B - \mathscr{B}_\theta) \xrightarrow{\mathrm{pr}} \mathbf{0},$$

for some deterministic matrices $\mathscr{U}_\theta$, $\mathscr{T}_\theta$, and $\mathscr{B}_\theta = \mathscr{T}_\theta - \mathscr{U}_\theta$, where $\mathbf{0}$ denotes a matrix of zeros, and the subscript θ highlights that these matrices are for estimating θ, because there are also corresponding $\mathscr{T}_\psi$, $\mathscr{B}_\psi$, and $\mathscr{U}_\psi$ for the entire parameter ψ. Similar to $\overline{U}$, T, and B, we define $\overline{U}_\psi$, T_ψ, and B_ψ for the parameter ψ. If $\widehat{\theta}_{\mathrm{com}}$ and $\widehat{\theta}_{\mathrm{obs}}$ are the MLEs of θ based on X_{com} and X_{obs} (under congeniality), respectively, then $\mathscr{U}_\theta \doteq \mathrm{Var}(\widehat{\theta}_{\mathrm{com}})$ and $\mathscr{T}_\theta \doteq \mathrm{Var}(\widehat{\theta}_{\mathrm{obs}})$ as $n \to \infty$, where $A_n \doteq B_n$ means that $A_n - B_n = o_p\{\min(A_n, B_n)\}$. Note that the relation $A_n \doteq B_n$ means that the difference between A_n and B_n is of a smaller order than A_n or B_n, when both $A_n \geqslant 0$ and $B_n \geqslant 0$ approach zero. This notation (or its variants) is also used in, for example, Meng and Rubin (1992), Li *et al.* (1991b), and Kim and Shao (2013).

The straightforward MI Wald test $D_{\mathrm{W}}(T) = d_{\mathrm{W}}(\overline{\theta}, T)/k$ is not practical because T is singular when $m < k$ (usually $3 \leqslant m \leqslant 10$). Even when it is not singular, it is usually not a very stable estimator of $\mathscr{T}_\theta$ because m is small. To circumvent this problem, Rubin (1978) adopted the following assumption of an EFMI.

Assumption 1 (EFMI of θ)**.** *There is $\varkappa \geqslant 0$ such that $\mathscr{T}_\theta = (1 + \varkappa)\mathscr{U}_\theta$.*

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

EFMI is a strong assumption, implying that the missing data have caused an equal loss of information for estimating every component of θ. However, as we shall see shortly, adopting this assumption *for the purpose of hypothesis testing* is essentially the same as summarizing the impact of (at least) k nuisance parameters due to FMI by a single nuisance parameter, this is, the average FMI across different components. How well this reduction strategy works has a great effect on the power of the test than on its validity, as long as we can construct an approximate null distribution that is more robust to the EFMI assumption. The issue of power turns out to be a rather tricky one, because without the reduction strategy, we also lose power when m/k is small or even modest. This is because we simply do not have enough degrees of freedom to estimate all the nuisance parameters well or at all. We illustrate this point in Section 4.2. (To clarify some confusion in literature, γ in Assumption 1 is the *odds of the missing information*, not the FMI, which is $\lambda = \gamma/(1+\gamma)$.) We also denote $\gamma_m = (1+1/m)\gamma$ as the finite-m adjusted value of γ.

Under EFMI, Rubin (2004) replaced T by $(1+\widetilde{r}'_{\mathrm{W}})\overline{U}$, where

$$\widetilde{r}'_{\mathrm{W}} = \frac{(m+1)}{k(m-1)}(\overline{d}'_{\mathrm{W}} - \widetilde{d}'_{\mathrm{W}}); \qquad \overline{d}'_{\mathrm{W}} = \frac{1}{m}\sum_{\ell=1}^{m} d_{\mathrm{W}}(\widehat{\theta}^{(\ell)}, \overline{U}), \tag{1.3}$$

$\widetilde{d}'_{\mathrm{W}} = d_{\mathrm{W}}(\overline{\theta}, \overline{U})$, and the prime "$\prime$" indicates that $\overline{U}$ is used instead of individual

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

$\{U^{(\ell)}\}_{\ell=1}^{m}$. Then, a simple MI Wald test statistic (Rubin, 2004) is

$$\widetilde{D}'_{\mathrm{W}} = \frac{\widetilde{d}'_{\mathrm{W}}}{k(1+\widetilde{r}'_{\mathrm{W}})}. \tag{1.4}$$

The intuition behind (1.3)–(1.4) is important because it forms the building blocks for virtually all the subsequent developments. The "obvious" Wald statistic $\widetilde{d}'_{\mathrm{W}}/k$ is too large (compared to the usual χ^2_k/k), because it fails to take into account the missing information. The $(1+\widetilde{r}'_{\mathrm{W}})$ factor attempts to correct this, with the amount of correction determined by the amount of between-imputation variance relative to the within-imputation variance. This relative amount can be estimated by contrasting the average of individual Wald statistics and the Wald statistic based on an average of individual estimates, as in (1.3). Using the difference between the "average of functions" and the "function of average," namely,

$$\mathrm{Ave}\{G(x)\} - G(\mathrm{Ave}\{x\}), \tag{1.5}$$

is a common practice, for example, $G(x) = x^2$ for variance; see Meng (2002).

Because the exact null distribution of $\widetilde{D}'_{\mathrm{W}}$ is intractable, Li *et al.* (1991b) proposed approximating it by $F_{k,\widetilde{\mathrm{df}}(\widetilde{r}'_{\mathrm{W}},k)}$, the F distribution with degrees of freedom k and $\widetilde{\mathrm{df}}(\widetilde{r}'_{\mathrm{W}},k)$, where, denoting $K_m = k(m-1)$,

$$\widetilde{\mathrm{df}}(\curlyvee_m, k) = \begin{cases} 4 + (K_m - 4)\{1 + (1-2/K_m)/\curlyvee_m\}^2, & \text{if } K_m > 4; \\ (m-1)(1+1/\curlyvee_m)^2(k+1)/2, & \text{otherwise.} \end{cases} \tag{1.6}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

In (1.6), n is assumed to be sufficiently large so that the asymptotic χ^2 distribution in Property 1.1 can be used. If n is small, the small sample degree of freedom in Barnard and Rubin (1999) should be used.

1.5 The Current MI Likelihood Ratio Test and Its Defect

Let $d_{\mathrm{L}}^{(\ell)} = d_{\mathrm{L}}(\widehat{\psi}_0^{(\ell)}, \widehat{\psi}^{(\ell)} \mid X^{(\ell)})$, $\widehat{\psi}_0^{(\ell)} = \widehat{\psi}_0(X^{(\ell)})$ and $\widehat{\psi}^{(\ell)} = \widehat{\psi}(X^{(\ell)})$ be the imputed counterparts of $d_{\mathrm{L}}(\widehat{\psi}_0, \widehat{\psi} \mid X)$, $\widehat{\psi}_0$ and $\widehat{\psi}$, respectively, for each ℓ. Define

$$\overline{d}_{\mathrm{L}} = \frac{1}{m}\sum_{\ell=1}^{m} d_{\mathrm{L}}^{(\ell)}, \qquad \overline{\psi}_0 = \frac{1}{m}\sum_{\ell=1}^{m} \widehat{\psi}_0^{(\ell)}, \qquad \overline{\psi} = \frac{1}{m}\sum_{\ell=1}^{m} \widehat{\psi}^{(\ell)}. \tag{1.7}$$

Similar to $\widetilde{r}'_{\mathrm{W}}$, Meng and Rubin (1992) proposed estimating ν_m by

$$\widetilde{r}_{\mathrm{L}} = \frac{m+1}{k(m-1)}(\overline{d}_{\mathrm{L}} - \widetilde{d}_{\mathrm{L}}), \quad \text{where} \quad \widetilde{d}_{\mathrm{L}} = \frac{1}{m}\sum_{\ell=1}^{m} d_{\mathrm{L}}(\overline{\psi}_0, \overline{\psi} \mid X^{(\ell)}), \tag{1.8}$$

and hence it is again in the form of (1.5). The computation of $\widetilde{r}_{\mathrm{L}}$ requires that users have access to (i) a subroutine for $(X, \psi_0, \psi) \mapsto d_{\mathrm{L}}(\psi_0, \psi \mid X)$, and (ii) the estimates $\widehat{\psi}_0^{(\ell)}$ and $\widehat{\psi}^{(\ell)}$, rather than the matrices $\overline{U}$ and B. Therefore, computing $\widetilde{r}_{\mathrm{L}}$ is easier than computing $\widetilde{r}'_{\mathrm{W}}$. The resulting MI LRT is

$$\widetilde{D}_{\mathrm{L}} = \frac{\widetilde{d}_{\mathrm{L}}}{k(1+\widetilde{r}_{\mathrm{L}})}, \tag{1.9}$$

the null distribution of which can be approximated by $F_{k,\widetilde{\mathrm{df}}(\widetilde{r}_{\mathrm{L}},k)}$. Its main theoretical justification (and motivation) is the asymptotic equivalence between the complete-data Wald test statistic and the LRT statistic *under the null*, as stated

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

in Property 1.1. This equivalence permitted the replacement of $\overline{d}'_{\mathrm{W}}$ and $\widetilde{d}'_{\mathrm{W}}$ in (1.3) by $\overline{d}_{\mathrm{L}}$ and $\widetilde{d}_{\mathrm{L}}$, respectively, in (1.8). However, this is also where the problems lie.

First, with finite samples, $0 \leqslant \widetilde{d}_{\mathrm{L}} \leqslant \overline{d}_{\mathrm{L}}$ is not guaranteed; consequently, nor is $\widetilde{D}_{\mathrm{L}} \geqslant 0$ or $\widetilde{r}_{\mathrm{L}} \geqslant 0$. Because $\widetilde{D}_{\mathrm{L}}$ is referred to as an F distribution and $\widetilde{r}_{\mathrm{L}}$ estimates $\nu_m \geqslant 0$, clearly, negative values of $\widetilde{D}_{\mathrm{L}}$ or $\widetilde{r}_{\mathrm{L}}$ will cause trouble. Second, the MI LRT statistic $\widetilde{D}_{\mathrm{L}}$ is not invariant to re-parameterization of ψ, although invariance is a natural property of the standard LRT; see, for example, Dagenais and Dufour (1991). This invariance principle is an appealing property because it requires that problems with the same formal structure should produce the same statistical results; see Chapter 6 of Berger (1985) and Chapter 3.2 of Lehmann and Casella (1998). Formally, we say that $\varphi = g(\psi)$ is a re-parametrization of ψ if g is a bijective map. The classical LRT statistic is invariant to re-parametrization because

$$d_{\mathrm{L}}(\widehat{\psi}_0, \widehat{\psi} \mid X) = d_{\mathrm{L}}(g^{-1}(\widehat{\varphi}_0), g^{-1}(\widehat{\varphi}) \mid X),$$

where $\widehat{\varphi}_0$ and $\widehat{\varphi}$ are the constrained and unconstrained MLEs, respectively, of φ based on X. However, the MI (pooled) LRT statistic $\widetilde{d}_{\mathrm{L}}$ no longer has this property because

$$\sum_{\ell=1}^{m} d_{\mathrm{L}}(\overline{\psi}_0, \overline{\psi} \mid X^{(\ell)}) \neq \sum_{\ell=1}^{m} d_{\mathrm{L}}(g^{-1}(\overline{\varphi}_0), g^{-1}(\overline{\varphi}) \mid X^{(\ell)}),$$

in general, where $\widehat{\varphi}_0^{(\ell)}$ and $\widehat{\varphi}^{(\ell)}$ are the constrained and unconstrained MLEs,

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

respectively, of φ based on $X^{(\ell)}$, and $\overline{\varphi}_0 = m^{-1}\sum_{\ell=1}^{m}\widehat{\varphi}_0^{(\ell)}$ and $\overline{\varphi} = m^{-1}\sum_{\ell=1}^{m}\widehat{\varphi}^{(\ell)}$. Section 4 shows how the MI LRT results vary dramatically with parametrizations in finite samples.

Third, the estimator $\widetilde{r}_{\mathrm{L}}$ involves the estimators of ψ under H_0, this is, $\widehat{\psi}_0^{(\ell)}$ and $\overline{\psi}_0$. When H_0 fails, they may be inconsistent for ψ. Thus, $\widetilde{r}_{\mathrm{L}}$ is no longer consistent for $\mathscr{r}_m$. A serious consequence is that the power of the test statistic $\widetilde{D}_{\mathrm{L}}$ is not guaranteed to monotonically increase as H_1 moves away from H_0. Indeed, our simulations (see Section 3.2) show that under certain parametrizations, the power may nearly vanish for obviously false H_0. Fourth, computing $\widetilde{d}_{\mathrm{L}}$ in (1.8) requires that users have access to $\widetilde{\mathscr{D}}_{\mathrm{L}}$, a function of both data and parameters. However, in most software, the available function is $\mathscr{D}_{\mathrm{L}}$, a function of data only; that is,

$$\widetilde{\mathscr{D}}_{\mathrm{L}} : (X, \psi_0, \psi) \mapsto d_{\mathrm{L}}(\psi_0, \psi \mid X), \quad \mathscr{D}_{\mathrm{L}} : X \mapsto d_{\mathrm{L}}(\widehat{\psi}_0(X), \widehat{\psi}(X) \mid X). \tag{1.10}$$

It is not always feasible for users to write themselves a subroutine $\widetilde{\mathscr{D}}_{\mathrm{L}}$.

In short, four problems need to be resolved: (i) the lack of non-negativity, (ii) the lack of invariance, (iii) the lack of consistency and power, and (iv) the lack of a feasible algorithm. Problems (i)–(iii) are resolved in Section 2; (iv) is resolved in Section 3.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

2. Improved MI Likelihood Ratio Tests

2.1 Invariant Combining Rule and Estimator of κ_m

To derive a parametrization-invariant MI LRT, we replace $\widetilde{d}_{\mathrm{L}}$ by an asymptotically equivalent version that behaves like a standard LRT statistic. Let

$$\overline{L}(\psi) = \frac{1}{m}\sum_{\ell=1}^{m} L^{(\ell)}(\psi), \qquad \text{where} \quad L^{(\ell)}(\psi) = \log f(X^{(\ell)} \mid \psi). \tag{2.1}$$

Here, $\overline{L}(\psi)$ is *not* a real log-likelihood, because it does not properly model the completed data sets: $\mathbb{X} = \{X^1, \ldots, X^m\}$ (e.g., all X^ℓ share the same X_{obs}). Nevertheless, $\overline{L}(\psi)$ can be treated as a log-likelihood for computational purposes. In particular, we can maximize it to obtain

$$\widehat{\psi}_0^* = \widehat{\psi}_0^*(\mathbb{X}) = \underset{\psi\in\Psi\,:\,\theta(\psi)=\theta_0}{\arg\max}\ \overline{L}(\psi), \qquad \widehat{\psi}^* = \widehat{\psi}^*(\mathbb{X}) = \underset{\psi\in\Psi}{\arg\max}\,\overline{L}(\psi). \tag{2.2}$$

The corresponding log-likelihood ratio test statistic is given by

$$\widehat{d}_{\mathrm{L}} = 2\left\{\overline{L}(\widehat{\psi}^*) - \overline{L}(\widehat{\psi}_0^*)\right\} = \frac{1}{m}\sum_{\ell=1}^{m} d_{\mathrm{L}}(\widehat{\psi}_0^*, \widehat{\psi}^* \mid X^{(\ell)}). \tag{2.3}$$

Thus, in contrast to $\widetilde{d}_{\mathrm{L}}$ of (1.8), $\widehat{d}_{\mathrm{L}}$ aggregates MI data sets by averaging the MI LRT functions, as in (2.1), rather than averaging the MI test statistics and moments, as in (1.7). Although $\sqrt{n}(\widehat{\psi}_0^* - \overline{\psi}_0) \xrightarrow{\mathrm{pr}} \mathbf{0}$ and $\sqrt{n}(\widehat{\psi}^* - \overline{\psi}) \xrightarrow{\mathrm{pr}} \mathbf{0}$ as $n \to \infty$ for each m, only $\widehat{d}_{\mathrm{L}}$, not $\widetilde{d}_{\mathrm{L}}$, is guaranteed to be non-negative and invariant to parametrization of ψ for all m, n. Indeed, the likelihood principle guides us to consider averaging individual log-likelihoods rather than individual

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

MLEs, because the former has a much better chance of capturing the functional features of the real log-likelihood than any of their (local) maximizers can.

To derive the properties of $\widehat{d}_{\mathrm{L}}$, we need the usual regularity conditions on the MLE and MI.

Assumption 2. *The sampling model $f(X \mid \psi)$ satisfies the following:*

(a) The map $\psi \mapsto \underline{L}(\psi) = n^{-1} \log f(X \mid \psi)$ is twice continuously differentiable;

(b) The complete-data MLE $\widehat{\psi}(X)$ is the unique solution of $\partial \underline{L}(\psi)/\partial\psi = \mathbf{0}$;

(c) Let $\underline{I}(\psi) = -\partial^2 \underline{L}(\psi)/\partial\psi\partial\psi^{\intercal}$; then, for each ψ, there exists a positive-definite matrix $\underline{\mathscr{I}}(\psi) = \mathscr{U}_{\psi}^{-1}$ such that $\underline{I}(\psi) \xrightarrow{\mathrm{pr}} \underline{\mathscr{I}}(\psi)$ as $n \to \infty$; and

(d) The observed-data MLE $\widehat{\psi}_{\mathrm{obs}}$ of ψ obeys

$$\left[\mathscr{T}_{\psi}^{-1/2}\left(\widehat{\psi}_{\mathrm{obs}} - \psi\right) \middle| \psi\right] \Rightarrow \mathscr{N}_h(\mathbf{0}, I_h) \tag{2.4}$$

as $n \to \infty$, where I_h is the $h \times h$ identity matrix.

Assumption 3. *The imputation model is proper (Rubin, 2004):*

$$\left[\mathscr{B}_{\psi}^{-1/2}\left(\widehat{\psi}^{(\ell)} - \widehat{\psi}_{\mathrm{obs}}\right) \middle| X_{\mathrm{obs}}\right] \Rightarrow \mathscr{N}_h(\mathbf{0}, I_h), \tag{2.5}$$

$$\left[\mathscr{T}_{\psi}^{-1}\left(U_{\psi}^{(\ell)} - \mathscr{U}_{\psi}\right) \middle| X_{\mathrm{obs}}\right] \xrightarrow{\mathrm{pr}} \mathbf{0}, \qquad \left[\mathscr{T}_{\psi}^{-1}\left(B_{\psi} - \mathscr{B}_{\psi}\right) \middle| X_{\mathrm{obs}}\right] \xrightarrow{\mathrm{pr}} \mathbf{0} \tag{2.6}$$

independently for each ℓ, as $n \to \infty$, provided that $\mathscr{B}_{\psi}^{-1}$ is well defined.

Assumption 2 holds under the usual regularity conditions that guarantee the normality and consistency of MLEs. When $X_{\mathrm{mis}}^{(1)}, \ldots, X_{\mathrm{mis}}^{(m)}$ are drawn inde-

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

pendently from a (correctly specified) posterior predictive distribution $f(X_{\text{mis}} \mid X_{\text{obs}})$, Assumption 3 is typically satisfied. Clearly, we can replace ψ by its subvector θ in Assumptions 2 and 3. These θ-version assumptions are sufficient to guarantee the validity of Theorem 2.4 and Corollary 2.3. For simplicity, Assumption 1, the θ-version of Assumptions 2 and 3, and the conditions that guarantees Property 1.1 are collectively written as RC_θ (RC denotes "regularity conditions"), which are commonly assumed for MI inference.

Theorem 2.1. *Assume* RC_θ. *Under* H_0, *we have (i)* $\widehat{d}_{\text{L}} \geqslant 0$ *for all* m, n; *(ii)* $\widehat{d}_{\text{L}}$ *is invariant to parametrization of* ψ *for all* m, n; *and (iii)* $\widehat{d}_{\text{L}} \doteq \widetilde{d}_{\text{L}}$ *as* $n \to \infty$ *for each* m.

Consequently, an improved combining rule is defined as

$$\widehat{D}_{\text{L}}(\kappa_m) = \frac{\widehat{d}_{\text{L}}}{k(1+\kappa_m)}, \tag{2.7}$$

for a given value of κ_m. The forms of (1.4) and (1.9) follow. Using $\widehat{d}_{\text{L}}$ in (2.3), we can modify $\widetilde{r}_{\text{L}}$ in (1.8) to a potentially better estimator:

$$\widehat{r}_{\text{L}} = \frac{m+1}{k(m-1)}(\overline{d}_{\text{L}} - \widehat{d}_{\text{L}}). \tag{2.8}$$

Although $\widehat{d}_{\text{L}} \geqslant 0$ is guaranteed by our construction, $\widehat{r}_{\text{L}} \geqslant 0$ does not hold in general for a finite m. However, it is guaranteed in the following situation.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

Proposition 2.2. *Write $\psi = (\theta^\intercal, \eta^\intercal)^\intercal$, where η represents a nuisance parameter that is distinct from θ. If there exist functions $L_\dagger$ and $L_\ddagger$ such that, for all X, the log-likelihood function $L(\psi \mid X) = \log f(X \mid \psi)$ is of the form $L(\psi \mid X) = L_\dagger(\theta \mid X) + L_\ddagger(\eta \mid X)$, then $\widehat{r}_{\mathrm{L}} \geqslant 0$ for all m, n.*

The condition in Proposition 2.2 means that the likelihood function of ψ is separable, which ensures that the profile likelihood estimator of η given θ, this is, $\widehat{\eta}_\theta = \arg\max_\eta L(\theta, \eta \mid X)$, is free of θ. Clearly, in the absence of the nuisance parameter η, the separation condition holds trivially. More generally, we have the following.

Corollary 2.3. *Assume* RC_θ. *We have (i) under H_0, $\widehat{r}_{\mathrm{L}} \xrightarrow{\mathrm{pr}} \varkappa$ as $m, n \to \infty$; and (ii) under H_1, $\widehat{r}_{\mathrm{L}} \xrightarrow{\mathrm{pr}} \varkappa_0$ as $m, n \to \infty$, where $\varkappa_0 \geqslant 0$ is some finite value depending on θ_0 and the true value of θ.*

Corollary 2.3 ensures that, under H_0, $\widehat{r}_{\mathrm{L}}$ is non-negative asymptotically and converges in probability to the true $\varkappa$. However, it also reveals another fundamental defect of $\widehat{r}_{\mathrm{L}}$: under H_1, the limit $\varkappa_0$ may not equal $\varkappa$, a problem we address in Section 2.2. Fortunately, because $\widehat{d}_{\mathrm{L}} \xrightarrow{\mathrm{pr}} \infty$ under H_1, the LRT statistic $\widehat{D}_{\mathrm{L}}(\widehat{r}_{\mathrm{L}})$ is still powerful, albeit the power may be reduced. Similarly, $\widetilde{r}_{\mathrm{L}}$ of (1.8) has the same asymptotic properties and defects, but $\widehat{r}_{\mathrm{L}}$ behaves more nicely than $\widetilde{r}_{\mathrm{L}}$ for finite m. This hinges closely on the high sensitivity of $\widetilde{r}_{\mathrm{L}}$ to the parametrization

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

of ψ; for example, $\widetilde{r}_{\mathrm{L}}$ may become more negative as H_1 moves away from H_0; see Section 4.1.

Whereas we can fix the occasional negativeness of $\widehat{r}_{\mathrm{L}}$ by using $\widehat{r}_{\mathrm{L}}^{+} = \max(0, \widehat{r}_{\mathrm{L}})$, such an ad hoc fix misses the opportunity to improve upon $\widehat{r}_{\mathrm{L}}$, and indeed it cannot fix the inconsistency of $\widehat{r}_{\mathrm{L}}$ under H_1.

2.2 A Consistent and Non-Negative Estimator of $\curlyvee_m$

Proposition 2.2 already hinted that the source of the negativity and inconsistency of $\widehat{r}_{\mathrm{L}}$ is related to the existence of the nuisance parameter η. By definition, $\overline{d}_{\mathrm{L}}$ and $\widehat{d}_{\mathrm{L}}$ depend on the specification of θ_0. In general, the effect of θ_0 may not be cancelled out by their difference $\overline{d}_{\mathrm{L}} - \widehat{d}_{\mathrm{L}}$, unless a certain type of orthogonality assumption is made on η and θ; see Proposition 2.2 for an example. Consequently, the validity of the estimator $\widehat{r}_{\mathrm{L}}$ depends on the correctness of H_0. A more elaborate discussion can be found in Appendix A.1. In order to principally resolve the aforementioned problem, we need to eliminate the dependence on θ_0 in our estimator for the odds of missing information, $\curlyvee_m$. We achieve this goal by estimating these odds for the entire ψ, resulting in the following estimator for $\curlyvee_m$:

$$\widehat{r}_{\mathrm{L}}^{\diamondsuit} = \frac{m+1}{h(m-1)}(\overline{\delta}_{\mathrm{L}} - \widehat{\delta}_{\mathrm{L}}), \quad \text{where} \tag{2.9}$$

$$\overline{\delta}_{\mathrm{L}} = 2\overline{L}(\widehat{\psi}^{(1)}, \ldots, \widehat{\psi}^{(m)}), \qquad \widehat{\delta}_{\mathrm{L}} = 2\overline{L}(\widehat{\psi}^{*}, \ldots, \widehat{\psi}^{*}), \tag{2.10}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

where h is the dimension of ψ, and the rhombus "$\diamondsuit$" symbolizes a robust estimator. It is robust because it is consistent under either H_0 or H_1, as long as we are willing to impose the EFMI assumption on ψ, this is, Assumption 4. This expansion from θ to ψ is inevitable because the LRT must handle the entire ψ, not just θ. The collection of Assumptions 2–4 are referred to as RC_ψ.

Assumption 4 (EFMI of ψ)**.** *There is $\mathscr{r} \geqslant 0$ such that $\mathscr{T}_\psi = (1+\mathscr{r})\mathscr{U}_\psi$.*

Theorem 2.4. *Assume* RC_ψ*. For any value of ψ, we have (i) $\widehat{r}_{\mathrm{L}}^{\diamondsuit} \geqslant 0$ for all m, n; (ii) $\widehat{r}_{\mathrm{L}}^{\diamondsuit}$ is invariant to parametrization of ψ for all m, n; and (iii) $\widehat{r}_{\mathrm{L}}^{\diamondsuit} \xrightarrow{\mathrm{pr}} \mathscr{r}$ as $m, n \to \infty$, where $\mathscr{r}$ is given in Assumption 4.*

With the improved combining rule $\widehat{D}_{\mathrm{L}}(\mathscr{r}_m)$ of (2.7) and improved estimators for $\mathscr{r}_m$, we are ready to propose two MI LRT statistics:

$$\widehat{D}_{\mathrm{L}}^{+} = \widehat{D}_{\mathrm{L}}(\widehat{r}_{\mathrm{L}}^{+}) \qquad \text{and} \qquad \widehat{D}_{\mathrm{L}}^{\diamondsuit} = \widehat{D}_{\mathrm{L}}(\widehat{r}_{\mathrm{L}}^{\diamondsuit}). \tag{2.11}$$

For comparison, we also study the test statistic $\widehat{D}_{\mathrm{L}} = \widehat{D}_{\mathrm{L}}(\widehat{r}_{\mathrm{L}})$.

2.3 Reference Null Distributions

The estimators $\widehat{r}_{\mathrm{L}}^{+}$ and $\widetilde{r}_{\mathrm{L}}$ have the same functional form asymptotically ($n \to \infty$). Hence, they have the same asymptotic distribution.

Lemma 2.5. *Suppose* RC_θ *and $m > 1$. Under H_0, we have, jointly,*

$$\frac{\widehat{r}_{\mathrm{L}}^{+}}{\mathscr{r}_m} \Rightarrow M_2 \qquad \textit{and} \qquad \widehat{D}_{\mathrm{L}}^{+} \Rightarrow \frac{(1+\mathscr{r}_m)\, M_1}{1+\mathscr{r}_m M_2} \tag{2.12}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

as $n \to \infty$*, where* $M_1 \sim \chi^2_k/k$ *and* $M_2 \sim \chi^2_{k(m-1)}/\{k(m-1)\}$ *are independent.*

Consequently, $\widehat{D}^+_{\mathrm{L}} = \widehat{D}_{\mathrm{L}}(\widehat{r}^+_{\mathrm{L}})$ approximately follows $F_{k,\widetilde{\mathrm{df}}(\widehat{r}^+_{\mathrm{L}},k)}$ under H_0, but a better approximation is provided shortly. For the other proposal, although $\widehat{r}^+_{\mathrm{L}} - \widehat{r}^{\diamondsuit}_{\mathrm{L}} \xrightarrow{\mathrm{pr}} 0$ as $n \to \infty$ under H_0, their non-degenerated limiting distributions are different because $\widehat{r}^{\diamondsuit}_{\mathrm{L}}$ and $\widehat{r}^+_{\mathrm{L}}$ rely on an average FMI in ψ and θ, respectively.

Theorem 2.6. *Suppose* RC_ψ *and* $m > 1$*. Then, for any value of* ψ*,*

$$\frac{\widehat{r}^{\diamondsuit}_{\mathrm{L}}}{\varkappa_m} \Rightarrow M_3 \sim \frac{\chi^2_{h(m-1)}}{h(m-1)} \tag{2.13}$$

as $n \to \infty$*, where* M_3 *is independent of the* M_1 *defined in (2.12).*

Theorem 2.6 implies that, if n can be regarded as infinity and $\widehat{r}^{\diamondsuit}_{\mathrm{L}}$ is uniformly integrable in $\mathscr{L}^2$, then $\mathsf{Bias}(\widehat{r}^{\diamondsuit}_{\mathrm{L}})=\mathrm{E}(\widehat{r}^{\diamondsuit}_{\mathrm{L}})-\varkappa_m=0$ and $\mathrm{Var}(\widehat{r}^{\diamondsuit}_{\mathrm{L}})=2\varkappa^2_m/\{h(m-1)\} = O(m^{-1})$ as $m \to \infty$. Hen $\widehat{r}^{\diamondsuit}_{\mathrm{L}}$ is a $\sqrt{m}$-consistent estimator of $\varkappa$ in $\mathscr{L}^2$. Moreover, for each $m > 1$ and as $n \to \infty$, we have $\mathsf{Bias}(\widehat{r}^+_{\mathrm{L}})/\mathsf{Bias}(\widehat{r}^{\diamondsuit}_{\mathrm{L}}) \to 1$ and $\mathrm{Var}(\widehat{r}^+_{\mathrm{L}})/\mathrm{Var}(\widehat{r}^{\diamondsuit}_{\mathrm{L}}) \to h/k \geqslant 1$, which imply that $\widehat{r}^{\diamondsuit}_{\mathrm{L}}$ is no less efficient than $\widehat{r}^+_{\mathrm{L}}$ when RC_ψ holds. This is not surprising because of the extra information brought in by the stronger Assumption 4. Result (2.13) also gives us the exact (i.e., for any $m > 1$, but assuming $n \to \infty$) reference null distribution of $\widehat{D}^{\diamondsuit}_{\mathrm{L}}$, as given below.

Theorem 2.7. *Assume* RC_ψ *and* $m > 1$*. Under* H_0*, we have*

$$\widehat{D}^{\diamondsuit}_{\mathrm{L}} \Rightarrow \frac{(1+\varkappa_m)\, M_1}{1+\varkappa_m M_3} \equiv D \tag{2.14}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

as $n \to \infty$*, where* $M_1 \sim \chi^2_k/k$ *and* $M_3 \sim \chi^2_{h(m-1)}/\{h(m-1)\}$ *are independent.*

The impact of the nuisance parameter ν_m on D diminishes with m because $\widehat{D}_{\mathrm{L}}^{\diamondsuit}$ and $\widehat{D}_{\mathrm{L}}^{+}$ converge in distribution to $M_1 = \chi^2_k/k$ as $m, n \to \infty$. Because $M_3 \xrightarrow{\mathrm{pr}} 1$ faster than $M_2 \xrightarrow{\mathrm{pr}} 1$, $\widehat{D}_{\mathrm{L}}^{\diamondsuit}$ is expected to be more robust to ν_m. Nevertheless, m typically is small in practice (e.g., $m \leqslant 10$), so we cannot ignore the impact of ν_m. This issue has been largely dealt with in the literature by seeking an $F_{k,\mathrm{df}}$ distribution to approximate D, as in Li *et al.* (1991b). However, directly adopting their $\widetilde{\mathrm{df}}$ of (1.6) leads to a poorer approximation for our purposes; see below. A better approximation is to match the first two moments of the denominator of (2.14), $1 + \nu_m M_3$, with that of a scaled χ^2: $a\chi^2_b/b$. This yields $a = 1 + \nu_m$ and $b = (1 + \nu_m^{-1})^2 h(m-1)$, and the approximated $F_{k,\widehat{\mathrm{df}}(\nu_m,h)}$, where

$$\widehat{\mathrm{df}}(\nu_m, h) = \left\{\frac{1+\nu_m}{\nu_m}\right\}^2 h(m-1) = \frac{h(m-1)}{\ell_m^2}, \tag{2.15}$$

which is appealing because it simply inflates the denominator degrees of freedom $h(m-1)$ by dividing it by the square of the finite-m corrected FMI $\ell_m = \nu_m/(1+\nu_m)$. The less missing information, the closer $F_{k,\widehat{\mathrm{df}}(\nu_m,h)}$ is to χ^2_k/k, the usual large-n χ^2 test; as mentioned earlier, for small n, see Barnard and Rubin (1999).

To compare the performance of $F_{k,\widehat{\mathrm{df}}(\nu_m,h)}$ in (2.15) with the existing best approximation $F_{k,\widetilde{\mathrm{df}}(\nu_m,h)}$, as approximations to the limiting distribution of D

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

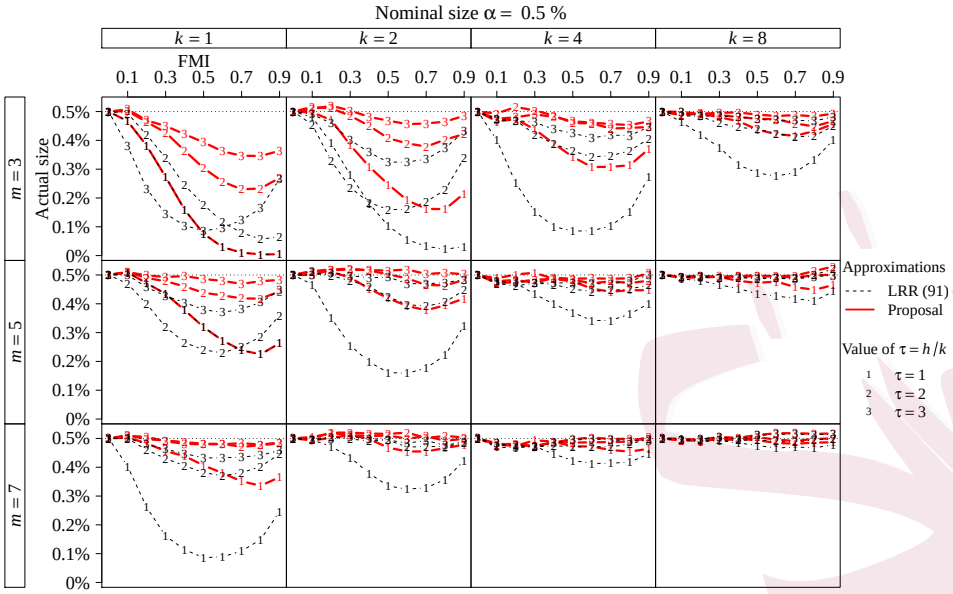

Figure 1: The performance of two approximated null distributions when the nominal size is $\alpha = 0.5\%$. The vertical axis denotes $\widehat{\alpha}$ or $\widetilde{\alpha}$, and the horizontal axis denotes the value of ℓ_m. The number attached to each line denotes the value of $\tau = h/k$.

given in (2.14), we compute via simulations

$$\widetilde{\alpha} = \mathrm{P}\left\{D > F^{-1}_{k,\widetilde{\mathrm{df}}(\ell_m,h)}(1-\alpha)\right\} \quad \text{and} \quad \widehat{\alpha} = \mathrm{P}\left\{D > F^{-1}_{k,\widehat{\mathrm{df}}(\ell_m,h)}(1-\alpha)\right\},$$

where $F^{-1}_{k,\mathrm{df}}(q)$ denotes the q-quantile of $F_{k,\mathrm{df}}$. Note that the experiments assess solely the performance of the finite-m approximation instead of the performance of the large-n χ^2-approximation of the asymptotic LRT statistics. We draw $N = 2^{18}$ independent copies D for each of the following possible combinations: $m \in \{3, 5, 7\}$, $k \in \{1, 2, 4, 8\}$, $\tau = h/k \in \{1, 2, 3\}$, $\ell_m \in \{0, 0.1, \ldots, 0.9\}$, and following

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

the recommendation of Benjamin *et al.* (2018), we use both $\alpha \in \{0.5\%, 5\%\}$. The results for $\alpha = 0.5\%$ and $\alpha = 5\%$ are shown in Figure 1 and Figure A.3 of the Appendix, respectively. In general, $\widehat{\alpha}$ approximates α much better than $\widetilde{\alpha}$ does, especially when m, k, h are small. When m, h are larger, they perform similarly because both $F_{k,\widetilde{\mathrm{df}}(\varkappa_m,h)}$ and $F_{k,\widehat{\mathrm{df}}(\varkappa_m,h)}$ get closer to χ^2_k/k. However, the performance of $\widetilde{\alpha}$ and $\widehat{\alpha}$ is not monotonic in $\not\ell_m$. The performance of $F_{k,\widehat{\mathrm{df}}(\varkappa_m,h)}$ is particularly good for $0\% \lesssim \not\ell_m \lesssim 30\%$. Consequently, we recommend using $F_{k,\widehat{\mathrm{df}}(\widehat{r}^{\diamondsuit}_{\mathrm{L}},h)}$ as an approximate null distribution for $\widehat{D}^{\diamondsuit}_{\mathrm{L}}$, and $F_{k,\widehat{\mathrm{df}}(\widehat{r}^{+}_{\mathrm{L}},k)}$ for $\widehat{D}^{+}_{\mathrm{L}}$, as employed in the rest of this paper. However, these approximations obviously suffer from the usual "plug-in problem" by ignoring the uncertainty in estimating $\varkappa_m$. Because $F_{k,\mathrm{df}}$ is not too sensitive to the value of df once it is reasonably large ($\mathrm{df} \geqslant 20$), the "plug-in problem" is less an issue here than in many other contexts, leading to acceptable approximations, as empirically demonstrated in Section 4. Nevertheless, further improvements should be sought, especially for dealing with the violation of the EFMI assumption, which would likely make the performance of our tests deteriorate with large k or h, in contrast to the results shown in Figure 1; see Chan (2020) for a possible remedy.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

3. Computational Considerations and Comparisons

3.1 Computationally Feasible Combining Rule

For many real-world data sets, X is an $n \times p$ matrix, with rows indicating subjects and columns indicating attributes. We write $X = (X_1, \ldots, X_n)^\intercal$, and its sampling model by $f_n(X \mid \psi)$. Correspondingly, the ℓth imputed data set is $X^{(\ell)} = (X_1^{(\ell)}, \ldots, X_n^{(\ell)})^\intercal$. Define the stacked data set by $X^{(1:m)} = [(X^{(1)})^\intercal, \ldots, (X^{(m)})^\intercal]^\intercal$, a $mn \times p$ matrix, which is conceptually different from the collection of data sets $\{X^{(1)}, \ldots, X^{(m)}\}$. Assuming that the rows of X are independent, we can compute (2.1) as

$$\overline{L}(\psi) = \frac{1}{m} \log f_{mn}(X^{(1:m)} \mid \psi). \tag{3.1}$$

Consequently, as long as the user's complete-data procedures can handle size mn instead of n, the user can apply them to $X^{(1:m)}$ to obtain $\widehat{D}_{\mathrm{L}}^{+}$ and $\widehat{D}_{\mathrm{L}}^{\Diamond}$ in (2.11).

In many applications, the rows correspond to individual subjects. Thus, the row-independence assumption typically holds for *arbitrary* n. Hence, we can extend from n to mn, assuming the user's complete-data procedure is not size-limited. Even if this is not true, (3.1) can still hold approximately under some regularity conditions; see Appendix A, where we also reveal a subtle, but important difference between the computation formulae (2.1) and (3.1).

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

Similar to $\mathscr{D}_{\mathrm{L}}$ in (1.10), we define complete-data functions

$$\mathscr{D}_{\mathrm{L},0}(X) = 2\log f(X \mid \widehat{\psi}_0(X)), \quad \mathscr{D}_{\mathrm{L},1}(X) = 2\log f(X \mid \widehat{\psi}(X)), \tag{3.2}$$

the only input of which is the data set X. Clearly, $\mathscr{D}_{\mathrm{L}}(X) = \mathscr{D}_{\mathrm{L},1}(X) - \mathscr{D}_{\mathrm{L},0}(X)$. The subroutine for evaluating the complete-data LRT function $X \mapsto \mathscr{D}_{\mathrm{L}}(X)$ is usually available, as is the subroutine for $X \mapsto \mathscr{D}_{\mathrm{L},1}(X)$, for example, the function `logLik` in R extracts the maximum of the complete data log-likelihood for objects belonging to classes `"glm"`, `"lm"`, `"nls"`, and `"Arima"`.

Algorithms 1 and 2 compute $\widehat{D}_{\mathrm{L}}^{\diamondsuit}$ and $\widehat{D}_{\mathrm{L}}^{+}$, respectively. We recommend using the robust MI LRT in Algorithm 1, because it has the best theoretical guarantee. The second test can be useful when $\mathscr{D}_{\mathrm{L}}$ is available but $\mathscr{D}_{\mathrm{L},1}$ is not.

3.2 Computational Comparison with Existing Tests

Different MI tests require different computing subroutines, for example, $\mathscr{D}_{\mathrm{L}}$, $\widetilde{\mathscr{D}}_{\mathrm{L}}$, $\mathscr{D}_{\mathrm{L},1}$,

$$\mathscr{M}_{\mathrm{W}}(X) = \left\{\widehat{\theta}(X), U(X)\right\} \qquad \text{and} \qquad \mathscr{M}_{\mathrm{L}}(X) = \left\{\widehat{\psi}(X), \widehat{\psi}_0(X)\right\},$$

where the unnormalized density can be used in $\mathscr{D}_{\mathrm{L},1}$. We summarize the computing requirement in Table 1. We also compare the following statistical and computational properties of various MI test statistics and various estimators of r_m:

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

Algorithm 1: (Robust) MI LRT statistic $\widehat{D}_{\mathrm{L}}^{\Diamond}$

Input: Data sets $X^{(1)}, \ldots, X^{(m)}$; h, k; functions $\mathscr{D}_{\mathrm{L},1}$, $\mathscr{D}_{\mathrm{L}}$ in (3.2), (1.10).

begin

Stack the data sets to form $X^{(1:m)} = [(X^{(1)})^{\intercal}, \ldots, (X^{(m)})^{\intercal}]^{\intercal}$.

Find $\overline{\delta}_{\mathrm{L}} = \sum_{\ell=1}^{m} \mathscr{D}_{\mathrm{L},1}(X^{(\ell)})/m$, $\widehat{\delta}_{\mathrm{L}} = \mathscr{D}_{\mathrm{L},1}(X^{(1:m)})/m$, $\widehat{d}_{\mathrm{L}} = \mathscr{D}_{\mathrm{L}}(X^{(1:m)})/m$.

Calculate $\widehat{r}_{\mathrm{L}}^{\Diamond}$ according to (2.9), and $\widehat{D}_{\mathrm{L}}^{\Diamond}$ according to (2.7) and (2.11).

Calculate $\widehat{\mathrm{df}}(\widehat{r}_{\mathrm{L}}^{\Diamond}, h)$ according to (2.15).

Compute the p-value as $1 - F_{k,\widehat{\mathrm{df}}(\widehat{r}_{\mathrm{L}}^{\Diamond}, h)}(\widehat{D}_{\mathrm{L}}^{\Diamond})$.

Algorithm 2: MI LRT statistic $\widehat{D}_{\mathrm{L}}^{+}$

Input: Data sets $X^{(1)}, \ldots, X^{(m)}$; k; function $\mathscr{D}_{\mathrm{L}}$ in (1.10).

begin

Stack the data sets to form $X^{(1:m)} = [(X^{(1)})^{\intercal}, \ldots, (X^{(m)})^{\intercal}]^{\intercal}$.

Find $\overline{d}_{\mathrm{L}} = \sum_{\ell=1}^{m} \mathscr{D}_{\mathrm{L}}(X^{(\ell)})/m$ and $\widehat{d}_{\mathrm{L}} = m^{-1}\mathscr{D}_{\mathrm{L}}(X^{(1:m)})$.

Calculate $\widehat{r}_{\mathrm{L}}^{+}$ according to (2.8), and $\widehat{D}_{\mathrm{L}}^{+}$ according to (2.7) and (2.11).

Calculate $\widehat{\mathrm{df}}(\widehat{r}_{\mathrm{L}}^{+}, k)$ according to (2.15).

Compute the p-value as $1 - F_{k,\widehat{\mathrm{df}}(\widehat{r}_{\mathrm{L}}^{+}, k)}(\widehat{D}_{\mathrm{L}}^{+})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

- (Inv) The MI test is invariant to re-parametrization of ψ.
- (Con) The estimator of ɾ_m is consistent, regardless of whether or not H_0 is true.
- ($\geqslant 0$) The test statistic and estimator of ɾ_m are always non-negative.
- (Pow) The MI test has high power to reject H_0 under H_1.
- (Def) The MI test statistic is well defined and numerically well conditioned.
- (Sca) The MI procedure requires that users deal with scalars only.
- (EFMI) Whether EFMI is assumed for θ or for ψ.

In summary, our proposed LRT-2 is the most computationally attractive. If the user is willing to make stronger assumptions, out proposed LRT-3 has better statistical properties, and is still computationally feasible. In practice, we recommend using LRT-3. We also present other existing MI tests and compare our proposals with them in Appendix A.5.

3.3 Summary of Notation

For ease of referencing, we summarize all major notation used in the paper. Recall that $\psi \in \mathbb{R}^h$ is the model parameter, and θ is the parameter of interest. We would like to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

- Complete-data Estimators and Test Statistics:
  - $\widehat{\theta}(X)$ and $U(X)$: MLE of θ and its variance estimator.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

Table 1: Computational requirements and statistical properties of MI tests. The symbol "+" (resp. "−") means that a test has (resp. does not have) the indicated property; see Section 3.2 for detailed descriptions. WT-1 (Rubin, 2004; Li *et al.*, 1991a) and LRT-1 (Meng and Rubin, 1992) are existing tests. LRT-2 and LRT-3 are the proposed tests, which can be computed by Algorithms 2 and 1, respectively. LRT-3 is recommended.

				Properties						
Test	Statistic	Distribution	Routine	Inv	Con	$\geqslant 0$	Pow	Def	Sca	EFMI
WT-1	$D_{\mathrm{W}}(T)$	$\approx F_{k,\widetilde{\mathrm{df}}(\widetilde{r}'_{\mathrm{W}},k)}$	$\mathscr{M}_{\mathrm{W}}$	−	+	+	−	−	−	θ
LRT-1	$\widetilde{D}_{\mathrm{L}}(\widetilde{r}_{\mathrm{L}})$	$\approx F_{k,\widetilde{\mathrm{df}}(\widetilde{r}_{\mathrm{L}},k)}$	$\mathscr{M}_{\mathrm{L}}, \widetilde{\mathscr{D}}_{\mathrm{L}}$	−	−	−	−	+	−	θ
LRT-2	$\widehat{D}_{\mathrm{L}}(\widehat{r}^{+}_{\mathrm{L}})$	$\approx F_{k,\widehat{\mathrm{df}}(\widehat{r}^{+}_{\mathrm{L}},k)}$	$\mathscr{D}_{\mathrm{L}}$	+	−	+	−	+	+	θ
LRT-3	$\widehat{D}_{\mathrm{L}}(\widehat{r}^{\diamondsuit}_{\mathrm{L}})$	$\approx F_{k,\widehat{\mathrm{df}}(\widehat{r}^{\diamondsuit}_{\mathrm{L}},k)}$	$\mathscr{D}_{\mathrm{L}}, \mathscr{D}_{\mathrm{L},1}$	+	+	+	+	+	+	ψ

  - $\widehat{\psi}(X)$ and $\widehat{\psi}_0(X)$: the unrestricted and H_0-restricted MLEs of ψ.
  - $d_{\mathrm{W}}(\widehat{\theta}, U) = (\widehat{\theta} - \theta_0)^{\intercal} U^{-1} (\widehat{\theta} - \theta_0)$: the Wald test statistic.
  - $d_{\mathrm{L}}(\widehat{\psi}_0, \widehat{\psi} \mid X) = 2 \log\{f(X \mid \widehat{\psi}) / f(X \mid \widehat{\psi}_0)\}$: the LRT statistic.
- Complete-data Functions (or Software Routines):
  - $\mathscr{M}_{\mathrm{W}}(X) = \{\widehat{\theta}(X), U(X)\}$ and $\mathscr{M}_{\mathrm{L}}(X) = \{\widehat{\psi}(X), \widehat{\psi}_0(X)\}$.
  - $\widetilde{\mathscr{D}}_{\mathrm{L}}(X, \psi_0, \psi) = d_{\mathrm{L}}(\psi_0, \psi \mid X)$: a nonstandard LRT function/routine.
  - $\mathscr{D}_{\mathrm{L}}(X) = d_{\mathrm{L}}(\widehat{\psi}_0(X), \widehat{\psi}(X) \mid X)$: the standard LRT function/routine.
  - $\mathscr{D}_{\mathrm{L},1}(X) = 2 \log f(X \mid \widehat{\psi}(X))$: the (scaled) maximum log-likelihood.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

- MI Statistics:
  - $\widehat{\theta}^{(\ell)}$, $U^{(\ell)}$, $\widehat{\psi}_0^{(\ell)}$, $\widehat{\psi}^{(\ell)}$, $d_{\mathrm{W}}^{(\ell)}$, $d_{\mathrm{L}}^{(\ell)}$: the imputed values of $\widehat{\theta}$, U, $\widehat{\psi}_0$, $\widehat{\psi}$, $d_{\mathrm{W}}(\widehat{\theta}, U)$, $d_{\mathrm{L}}(\widehat{\psi}_0, \widehat{\psi} \mid X)$ using the imputed data set $X^{(\ell)}$ for each ℓ.
  - $\overline{\theta}$, $\overline{U}$, $\overline{\psi}_0$, $\overline{\psi}$, $\overline{d}_{\mathrm{W}}$, $\overline{d}_{\mathrm{L}}$: the averages (over ℓ) of $\widehat{\theta}^{(\ell)}$, $U^{(\ell)}$, $\widehat{\psi}_0^{(\ell)}$, $\widehat{\psi}^{(\ell)}$, $d_{\mathrm{W}}^{(\ell)}$, $d_{\mathrm{L}}^{(\ell)}$.
  - $T = \overline{U} + (1 + 1/m)B$, where $B = \sum_{\ell=1}^{m} (\widehat{\theta}^{(\ell)} - \overline{\theta})(\widehat{\theta}^{(\ell)} - \overline{\theta})^{\intercal}/(m-1)$.
  - $\overline{d}'_{\mathrm{W}} = \sum_{\ell=1}^{m} d_{\mathrm{W}}(\widehat{\theta}^{(\ell)}, \overline{U})/m$ and $\widetilde{d}'_{\mathrm{W}} = d_{\mathrm{W}}(\overline{\theta}, \overline{U})$.
  - $\widetilde{d}_{\mathrm{L}} = \sum_{\ell=1}^{m} \widetilde{\mathscr{D}}_{\mathrm{L}}(X^{(\ell)}, \overline{\psi}_0, \overline{\psi})/m$: an existing pooled LRT statistic.
  - $\widehat{d}_{\mathrm{L}} = \mathscr{D}_{\mathrm{L}}(X^{(1:m)})/m$: the proposed pooled LRT statistic.
  - $\overline{\delta}_{\mathrm{L}} = \sum_{\ell=1}^{m} \mathscr{D}_{\mathrm{L},1}(X^{(\ell)})/m$ and $\widehat{\delta}_{\mathrm{L}} = \mathscr{D}_{\mathrm{L},1}(X^{(1:m)})/m$: two proposed ways for pooling maximized log-likelihood.
- Estimators of r_m:
  - $\widetilde{r}'_{\mathrm{W}} = (m+1)(\overline{d}'_{\mathrm{W}} - \widetilde{d}'_{\mathrm{W}})/\{k(m-1)\}$ (Rubin, 2004).
  - $\widetilde{r}_{\mathrm{L}} = (m+1)(\overline{d}_{\mathrm{L}} - \widetilde{d}_{\mathrm{L}})/\{k(m-1)\}$ (Meng and Rubin, 1992).
  - $\widehat{r}_{\mathrm{L}}^{+} = \max[0, (m+1)(\overline{d}_{\mathrm{L}} - \widehat{d}_{\mathrm{L}})/\{k(m-1)\}]$: our first proposal.
  - $\widehat{r}_{\mathrm{L}}^{\diamondsuit} = (m+1)(\overline{\delta}_{\mathrm{L}} - \widehat{\delta}_{\mathrm{L}})/\{h(m-1)\}$: our second proposal.
- MI Test Statistics for Testing H_0 against H_1:
  - (WT-1) $D_{\mathrm{W}}(T) = d_{\mathrm{W}}(\overline{\theta}, T)/k$: the classical MI Wald test.
  - (LRT-1) $\widetilde{D}_{\mathrm{L}}(\widetilde{r}_{\mathrm{L}}) = \widetilde{d}_{\mathrm{L}}/\{k(1 + \widetilde{r}_{\mathrm{L}})\}$: the existing MI LRT.
  - (LRT-2) $\widehat{D}_{\mathrm{L}}(\widehat{r}_{\mathrm{L}}^{+}) = \widehat{d}_{\mathrm{L}}/\{k(1 + \widehat{r}_{\mathrm{L}}^{+})\}$: our first proposal.
  - (LRT-3) $\widehat{D}_{\mathrm{L}}(\widehat{r}_{\mathrm{L}}^{\diamondsuit}) = \widehat{d}_{\mathrm{L}}/\{k(1 + \widehat{r}_{\mathrm{L}}^{\diamondsuit})\}$: our second proposal.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

4. Empirical Investigation and Findings

4.1 Monte Carlo Experiments With EFMI

Let $X_1, \ldots, X_n \sim \mathcal{N}_p(\mu, \Sigma)$ independently, where $\mu = (\mu_1, \ldots, \mu_p)^\intercal$. Assume that only $n_{\text{obs}} = \lfloor (1-\ell\!\!\!/)n \rfloor$ data points are observed. Let $X_{\text{obs}} = \{X_i : i = 1, \ldots, n_{\text{obs}}\}$ and $X_{\text{mis}} = \{X_i : i = n_{\text{obs}} + 1, \ldots, n\}$. We want to test $H_0 : \mu_1 = \cdots = \mu_p$.

Obviously, one may directly use the observed data set to construct the LRT statistic D_{L} without MI. Thus, it is regarded as a benchmark (denoted by LRT-0). The tests WT-1 and LRT-1,2,3 listed in Table 1 are investigated. We perform MI using a Bayesian model with a multivariate Jeffreys prior on (μ, Σ), this is, $f(\mu, \Sigma) \propto |\Sigma|^{-(p+1)/2}$. The imputation procedure is detailed in Appendix A.6. We study the impact of the parametrization on different test statistics.

- Parametrizations of θ for the Wald tests: (i) $\theta = (\mu_2 - \mu_1, \ldots, \mu_p - \mu_{p-1})^\intercal$; (ii) $\theta = (\mu_2/\mu_1 - 1, \ldots, \mu_p/\mu_{p-1} - 1)^\intercal$; and (iii) $\theta = \left(\mu_2^3 - \mu_1^3, \ldots, \mu_p^3 - \mu_{p-1}^3\right)^\intercal$. For any case above, H_0 can be expressed as $\theta = (0, \ldots, 0)^\intercal$.
- Parametrizations of ψ for LRTs: (i) $\psi = \{\mu; \Sigma\}$; (ii) $\psi = \{\sqrt{\sigma_{ii}}/\mu_i, 1 \leqslant i \leqslant p; \Sigma\}$; and (iii) $\psi = \left\{\mu^\intercal \Sigma^{-1/2}; \Sigma^{-1}\right\}$, where $\Sigma = (\sigma_{ij})$ and $\Sigma^{1/2}$ is the square root of Σ via the spectral method. The dimension of ψ is $h = (p^2 + 3p)/2$.

We set $\Sigma = \sigma^2\{(1-\rho)I_p + \rho \mathbf{1}_p \mathbf{1}_p^\intercal\}$, $\ell\!\!\!/ = 0.5$, $p = 2$, $\rho = 0.8$, $\sigma^2 = 5$, and $\mu = (-2+\delta, -2+2\delta)^\intercal$ for different values of $m \in \{3, 10, 30\}$, $n \in \{100, 400, 1600\}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

and $\delta = \mu_2 - \mu_1 \in [0, 4]$. All simulations are repeated 2^{12} times. The empirical power functions for $\alpha = 0.5\%$ tests are plotted in Figure 2. The results for $\alpha = 5\%$ tests are deferred to Table A.7 of the Appendix.

In general, WT-1 exhibits monotonically increasing power as δ increases, and its performance is affected significantly by parametrization. Indeed, the power can be as low as zero when $1 \lesssim \delta \lesssim 2$ under parametrizations (ii) and (iii). Under parametrization (ii), LRT-1 is not powerful, even for large δ. On the other hand, our first proposed test statistic LRT-2 performs better than LRT-1, at least for large m; however, they also lose a significant amount of power when m is small. Our recommended proposal LRT-3 performs best in all cases. The superiority of LRT-3 is particularly striking when m is small, this is, $m = 3$.

We also investigate (a) the distribution of the p-value, (b) the empirical size $\widehat{\alpha}$ in comparison to the nominal type-I error α, (c) the empirical size-adjusted power (Bayarri *et al.*, 2016), (d) the robustness of our proposed estimators of ρ_m, and (e) the performance of other existing MI tests. The results are shown in Appendix A.6, all of which indicate that our proposed tests perform best.

4.2 Monte Carlo Experiments Without EFMI

To check how robust various tests are to the assumption of EFMI, we simulate $X_i = (X_{i1}, \ldots, X_{ip})^\intercal \sim \mathcal{N}_p(\mu, \Sigma)$ independently for $i = 1, \ldots, n$. Let

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

R_{ij} be defined by $R_{ij} = 1$ if X_{ij} is observed, and $R_{ij} = 0$ otherwise. Suppose that the first variable $X_{\cdot 1}$ is always observed, and the rest form a monotone missing pattern, as defined by a logistic model on the missing propensity: $\mathrm{P}\left(R_{ij} = 0 \mid R_{i,j-1} = a\right) = \left[1 + \exp(\alpha_0 + \alpha_1 X_{i,j-1})\right]^{-1}$ (for $j = 2, \ldots, p$) when $a = 1$. This probability is zero when $a = 0$ (i.e., nothing is missing). If $\alpha_1 = 0$, the data are missing completely at random (MCAR); otherwise they are missing at random (MAR); see Rubin (1976). The imputation procedure is given in Appendix A.7.

We test $H_0 : \mu = \mathbf{0}_p$ against $H_1 : \mu \neq \mathbf{0}_p$. We set $\mu = \delta \mathbf{1}_p$, where $\delta \in [0, 0.6]$; $\Sigma_{ij} = 0.5^{|i-j|}$, for $i, j = 1, \ldots, p$; $n = 500$; $m \in \{3, 5\}$; $p = 5$; and $(\alpha_0, \alpha_1) \in \{(2, -1), (1, 0)\}$. Our model treats Σ as unknown, and hence $k = p$ and $h = (3p + p^2)/2$. Under H_0 and MAR, the FMI, this is, the eigenvalues of $\mathscr{B}_\theta \mathscr{T}_\theta^{-1}$, are $(0, 19\%, 34\%, 45\%, 55\%)$. Thus, the assumption of EFMI does not hold.

In this experiment, we also compare the performance of WT-1 and LRT-1,2,3. For reference, the complete-case (asymptotic) LRT using $\{X_i : R_{i1} = \cdots = R_{ip} = 1\}$, denoted by LRT-0, is also computed. The results are shown in Figure 3. The size of LRT-3 is accurate when the nominal size is small. If the data are MCAR, LRT-0 is valid, but with slightly less power. (LRT-0 is typically invalid without MCAR.) The test LRT-3 has the best power-to-size ratio among all other tests.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

The power-to-size ratio of LRT-2 and LRT-3 become closer to the nominal value $1/0.5\% = 200$ as m increases. These results indicate that our proposed tests perform well and best, despite the serious violation of the EFMI assumption.

5. Conclusion, Limitation and Future Work

In addition to conducting a general comparative study of MI tests, we have proposed two particularly promising MI LRTs based on $\widehat{D}_{\mathrm{L}}^{\Diamond} = \widehat{D}_{\mathrm{L}}(\widehat{r}_{\mathrm{L}}^{\Diamond})$ and $\widehat{D}_{\mathrm{L}}^{+} = \widehat{D}_{\mathrm{L}}(\widehat{r}_{\mathrm{L}}^{+})$. Both test statistics are non-negative, invariant to parametrizations, and powerful to reject a false H_0 (at least for large enough m). The test $\widehat{D}_{\mathrm{L}}^{\Diamond}$ is the most principled, and has desirable monotonically increasing power as H_1 departs from H_0. However, it is derived under the stronger assumption of EFMI for ψ, not just for θ. Furthermore, row independence of X_{com} is needed for ease of computation. (With a slightly more computationally demanding requirement, $\widehat{D}_{\mathrm{L}}(\widehat{r}_{\mathrm{L}}^{\Diamond})$ can be used without the independence assumption.) The main advantage of $\widehat{D}_{\mathrm{L}}^{+}$ is that it is easier to compute, because it requires only standard complete-data computer subroutines for LRTs. One drawback is that the ad hoc fix $\widehat{r}_{\mathrm{L}}^{+} = \max(0, \widehat{r}_{\mathrm{L}})$ is inconsistent, in general. However, the inconsistency does not significantly affect the asymptotic power, at least in our experiments. Although $\widehat{D}_{\mathrm{L}}^{+}$ and $\widehat{D}_{\mathrm{L}}^{\Diamond}$ offer significant improvements over existing options, more research is needed, for the reasons listed below:

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

- When the missing-data mechanism is not ignorable, but the imputers fail to fully take that into account, the issue of uncongeniality becomes critical (Meng, 1994a). Xie and Meng (2017) provide theoretical tools to address this issue in the context of estimation, and research is needed to extend their findings to the setting of hypothesis testing.
- Violating the EFMI assumption may not invalidate a test, but it will affect its power. Thus, it is desirable to explore MI tests without assuming EFMI.
- The robust $\widehat{D}_{\mathrm{L}}^{\Diamond}$ relies on a stronger assumption of EFMI on ψ. We can modify it so only EFMI on θ is required, but the modification may be very difficult to compute, and may require that users have access to nontrivial complete-data procedures. Hence, a computationally feasible robust test that only assumes EFMI on θ needs to be developed.
- Because the FMI is a fundamental nuisance parameter and there is no (known) pivotal quantity, all MI tests are just approximations. If FMI is large or m is small, they may perform poorly. Thus, seeking powerful MI tests that are least affected by FMI is of both theoretical and practical interest.

Supplementary Material

Appendix A contains additional theoretical results and details of numerical examples. Appendix B contains proofs of the main results. The R code is provided

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

online.

Acknowledgments

Meng thanks the NSF and JTF for their partial financial support. He is also grateful for Keith's (Kin Wai's) creativity and diligence, which led to the remedies presented here, and which are also a part of Keith's thesis. Chan thanks the University Grant Committee of HKSAR for its partial financial support.

References

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

Bayarri, M. J., Benjamin, D. J., Berger, J. O. and Sellke, T. M. (2016) Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, **72**, 90–103.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C. *et al.* (2018) Redefine statistical significance. *Nature Human Behaviour*, **2**, 6–10.

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag New York.

Berglund, P. and Heeringa, S. G. (2014) *Multiple imputation of missing data using SAS*. SAS Institute, Cary.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

Carlin, J. B., Galati, J. C. and Royston, P. (2008) A new framework for managing and analyzing multiply imputed data in stata. *The Stata Journal*, **8**, 49–67.

Chan, K. W. (2020) General and feasible tests with multiply-imputed datasets. *Submitted*.

Dagenais, M. G. and Dufour, J.-M. (1991) Invariance, nonlinear models, and asymptotic tests. *Econometrica*, **59**, 1601–1615.

Grund, S., Robitzsch, A. and Luedtke, O. (2017) *Tools for Multiple Imputation in Multilevel Modeling*.

Harel, O. and Zhou, X.-H. (2007) Multiple imputation - review of theory, implementation and software. *Statistics in medicine*, **26**, 3057–3077.

Holan, S. H., Toth, D., Ferreira, M. A. R. and Karr, A. F. (2010) Bayesian multiscale multiple imputation with implications for data confidentiality. *Journal of the American Statistical Association*, **105**, 564–577.

Horton, N. and Kleinman, K. P. (2007) Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, **61**, 79–90.

Kenward, M. G. and Carpenter, J. R. (2007) Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, **16**, 199–218.

Kim, J. K. and Shao, J. (2013) *Statistical Methods for Handling Incomplete Data*. Chapman and Hall/CRC, Boca Raton.

Kim, J. K. and Yang, S. (2017) A note on multiple imputation under complex sampling. *Biometrika*, **104**, 221–228.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

King, G., Honaker, J., Joseph, A. and Scheve, K. (2001) Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, **95**, 49–69.

Lehmann, E. L. and Casella, G. (1998) *Theory of Point Estimation*. Springer-Verlag New York.

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

Li, K. H., Raghunathan, T. E. and Rubin, D. B. (1991b) Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, **86**, 1065–1073.

Medeiros, R. (2008) Likelihood ratio tests for multiply imputed datasets: Introducing milrtest.

Meng, X.-L. (1994a) Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, **9**, 538–573.

Meng, X.-L. (1994b) Posterior predictive p-values. *The Annals of Statistics*, **22**, 1142–1160.

Meng, X.-L. (2002) Discussion of "Bayesian measures of model complexity and fit" by Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. *Journal of the Royal Statistical Society B*, **64**, 633.

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

Peugh, J. L. and Enders, C. K. (2004) Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational*

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

Research, **74**, 525–556.

Rose, R. A. and Fraser, M. W. (2008) A simplified framework for using multiple imputation in social work research. *Social Work Research*, **32**, 171–178.

Royston, P. and White, I. R. (2011) Multiple imputation by chained equations (mice): Implementation in stata. *Journal of Statistical Software*, **45**, 1–20.

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

Rubin, D. B. (1978) Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20–34.

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

Rubin, D. B. (2004) *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Rubin, D. B. and Schenker, N. (1986) Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, **81**, 366–374.

Schafer, J. L. (1999) Multiple imputation: A primer. *Statistical Methods in Medical Research*, **8**, 3–15.

Serfling, R. J. (2001) *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Shao, J. (1998) *Mathematical Statistics*. Springer-Verlag New York.

Su, Y.-S., Gelman, A., Hill, J. and Yajima, M. (2011) Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*,

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

45, 1–31.

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.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011) Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, **45**, 1–67.

van Buuren S (2012) *Flexible Imputation of Missing Data*. Chapman and Hall/CRC.

Wallace, D. L. (1980) The Behrens-Fisher and Fieller-Creasy problems. In *R. A. Fisher: An Appreciation* (eds. S. E. Fienberg and D. V. Hinkley), 119–147. Springer New York.

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? (with discussion). *Statistica Sinica*, **27**, 1485–1594.

Department of Statistics, The Chinese University of Hong Kong.

E-mail: kinwaichan@cuhk.edu.hk

Department of Statistics, Harvard University.

E-mail: meng@stat.harvard.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

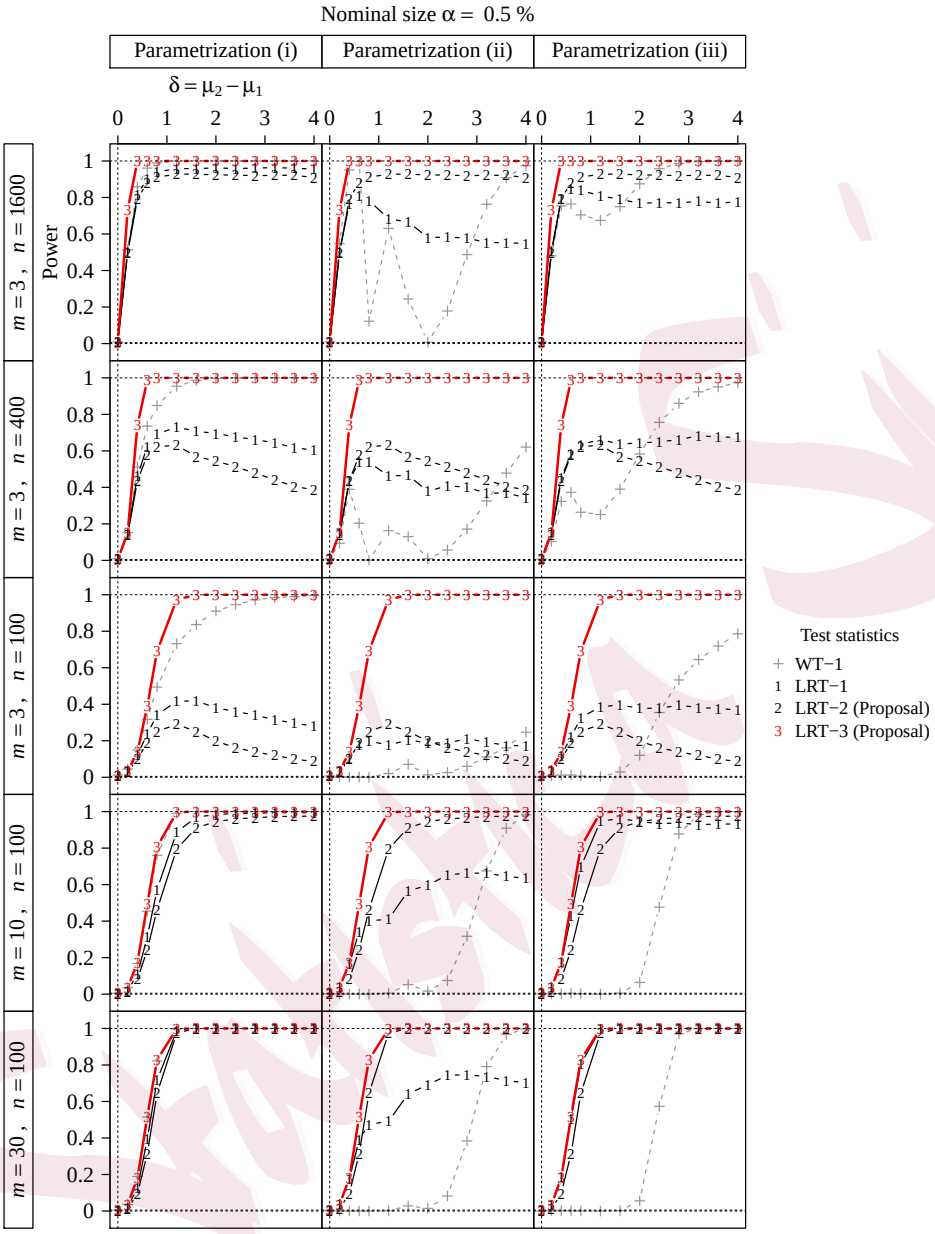

Figure 2: The power curves under nominal size $\alpha = 0.5\%$. In each plot, the vertical axis denotes the power, and the horizontal axis denotes the value of $\delta = \mu_2 - \mu_1$.

Statistica Sinica: Preprint
doi:10.5705/ss.202019.0314

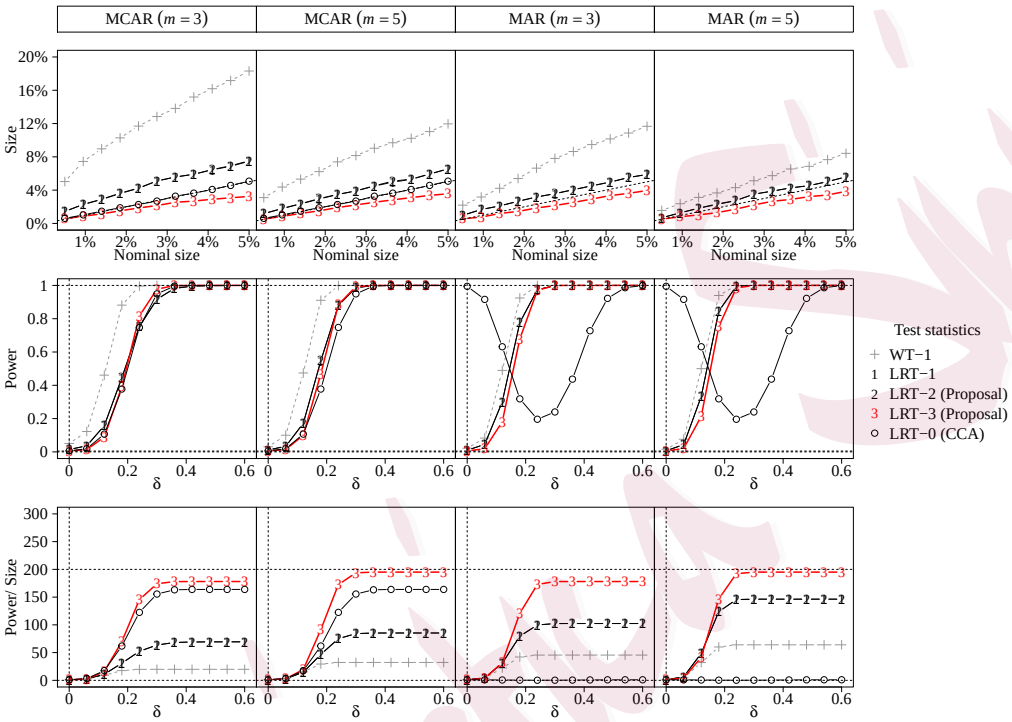

Figure 3: The empirical size, empirical power, and their ratio. The first row of plots show the empirical sizes. The size of the complete-case test (C2) under MAR is off the chart (always equal to one) because it is invalid. The second and third rows of plots show the powers and the power-to-size ratios, respectively. The nominal size is 0.5%.