Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

MULTIPLE IMPROVEMENTS OF MULTIPLE IMPUTATION LIKELIHOOD RATIO TESTS

By Kin Wai Chan and Xiao-Li Meng

The Chinese University of Hong Kong and Harvard University

Multiple imputation (MI) inference handles missing data by first properly 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 has multiple defects: (i) the combined test statistic can be negative in practice but its null distribution is approximated by a standard F distribution; (ii) it is not invariant to re-parametrization; (iii) it fails to ensure monotonic power due to its use of an inconsistent estimator of the fraction of missing information (FMI) under the alternative hypothesis; and (iv) it requires non-trivial access to the likelihood ratio test statistic as a function of estimated parameters instead of datasets. This paper shows, via both theoretical derivations and empirical investigations, that essentially all of these problems can be straightforwardly addressed if we are willing to perform an additional likelihood ratio test by stacking the m completed datasets as one big completed dataset. A particularly intriguing finding is that the FMI itself can be estimated consistently by a likelihood ratio statistic for testing whether the m completed datasets produced by MI can be regarded effectively as samples coming from a common model. Practical guidelines are provided based on an extensive comparison of existing MI tests. Intrigued issues regarding nuisance parameters are also discussed.

1. Historical Successes and Failures. Missing-data problems are ubiquitous in practice, to the extent that the absence of any missingness often is 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 by many practitioners, especially those who are ill-equipped to handle missingness on their own, due to lack of information or skills or resources. MI relies on the data collector (e.g., a census bureau) to build a reliable imputation model to fill in the missing data $m(\geqslant 2)$ times, so the users of the data can apply their favorite software or procedures that are designed to handle complete data,

AMS 2000 subject classifications: Primary 62D05; Secondary 62F03, 62E20.
Keywords and phrases: Fraction of missing information; missing data; invariant test; monotonic power; robust estimation.
Meng thanks NSF and JTF for partial financial support. He is also embarrassed by the research immaturity displayed in his thesis, and is very thankful to Keith's (Kin Wai) creativity and diligence which led to the beautiful remedies presented here.

and do so m times. MI inference, e.g., hypothesis testing, is then performed by appropriately combining these m complete-data results.

Although MI was designed initially for public-use datasets, over the past 30 years or so, it has become a method of choice for handling missing data in general, because it separates the handling missingness from conducting 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 (van Buuren and Groothuis-Oudshoorn, 2011; Su *et al.*, 2011), Stata (Royston and White, 2011), SAS (Berglund and Heeringa, 2014) and SPSS; also see Harel and Zhou (2007) and Horton and Kleinman (2007) for summaries on software that utilize MI.

This convenient separation, however, creates the thorny issue of uncongeniality, i.e., the incompatibility between the imputation model and the subsequent analysis procedures (Meng, 1994a). This issue is examined in detail by Xie and Meng (2017), which shows 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 efficient analysis, such as likelihood inference. The current paper, therefore, focuses on conducting MI likelihood ratio tests (LRTs), assuming the imputation model is sufficiently saturated to render the validity of the common assumptions, which we shall review, made in the literature about conducting LRTs with MI.

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 due to the well-known issue of nuisance parameter, that is, the lack of pivotal quantity, as highlighted, historically, by the Behrens-Fisher problem (Wallace, 1980). Indeed, the nuisance parameter in the MI context is the so-called "the fraction of missing information" (FMI), which is determined by the ratio of the between-imputation variance to within-imputation variance (and its multivariate counterparts), and hence the challenge we face is almost identical to the one faced by the Behrens-Fisher problem, as shown in Meng (1994b).

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 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 shall review shortly.

Among them, a promising idea was to directly combine LRT statistics. However, the execution of this idea as presented in Meng and Rubin (1992)

relied too heavily on the usual 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 negative "F test statistic" or FMI. Worst of all, the test can have essentially zero power because the estimator of FMI can be badly inconsistent under some alternative hypotheses. In addition, the combining rule of Meng and Rubin (1992) requires the user to have access to the LRT as a function of parameter values, not just as a function of the data. The former one 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, e.g., 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), 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 paper derives MI LRTs that are free of the defects as outlined in the abstract and detailed in § 1.3 below. We achieve this mainly by switching the order of two main operators in the combining rule of Meng and Rubin (1992): Maximizing the average of the m log-likelihoods instead of averaging the maximizers of them. This switching, guided by the likelihood principle, automatically renders positivity, invariance and monotonic power. Other judicious uses of the likelihood functions permit us to overcome the remaining defects.

The remainder of Section 1 provides background and notation. Section 2 then discusses the defects of the existing MI LRT and our remedies. Section 3 investigates computational requirements for our proposals, including theoretical considerations and comparisons. In particular, Algorithm 1 of Section 3.1 computes our most recommended test. Section 4 provides empirical evidence with simulated and real data. Section 5 calls for future work. Appendices A–C supplement with proofs, and additional investigations.

1.1. *Notation and Complete-data Tests.* Let X_{obs} and X_{mis} be, respectively, the observed and missing parts of an intended complete dataset $X = X_{\text{com}} = \{X_{\text{obs}}, X_{\text{mis}}\}$, which consists of n observations. Denote the sampling model — probability or density, depending on the data type — of X by $f(\cdot \mid \psi)$, where $\psi \in \Psi \subset \mathbb{R}^h$ is a vector of parameters. The goal is to test $H_0 : \theta = \theta_0$ when only X_{obs} is available, where $\theta = \theta(\psi) \in \Theta \subset \mathbb{R}^k$ is a function of ψ, and θ_0 is a specified vector. For simplicity, we will focus on the standard two-sided alternative, but our approach adapts to general complete-data LRTs. Denote the true values of ψ and θ by $\psi^\star$ and $\theta^\star$. Here

we assume X_{obs} is rich enough that the missing data mechanism is ignorable (Rubin, 1976), or it has been properly incorporated into the imputation model by the imputer, who may have access to additional confidential data.

Let $\widehat{\theta} = \widehat{\theta}(X)$ and $U = U_\theta = U_\theta(X)$ be respectively the complete-data MLE of θ and an efficient estimator of $\mathsf{Var}(\widehat{\theta})$ (e.g., the inverse of the observed Fisher information). Also, let $\widehat{\psi}_0 = \widehat{\psi}_0(X)$ and $\widehat{\psi} = \widehat{\psi}(X)$ be respectively the H_0-constrained and unconstrained complete-data MLEs of ψ, and $U_\psi = U_\psi(X)$ be an efficient estimator of $\mathsf{Var}(\widehat{\psi})$. For testing H_0 against H_1, the common choices 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 (RCs), such as those in § 4.2.2 and § 4.4.2 of Serfling (2001) when the rows of X are independent and identically distributed, 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) $n(D_{\text{W}} - D_{\text{L}}) \xrightarrow{\text{pr}} 0$ as $n \to \infty$ where "$\Rightarrow$" and "$\xrightarrow{\text{pr}}$" denote convergence in distribution and in probability, respectively.*

Testing $\theta = \theta_0$ based on X_{obs} is more involved. For MI, let $X^\ell = \{X_{\text{obs}}, X^\ell_{\text{mis}}\}$, $\ell = 1, \ldots, m$, be the m completed datasets, where X^ℓ_{mis} are drawn conditionally independently from a proper imputation model given X_{obs}; see Rubin (2004). We then carry out a complete-data estimation or testing procedure on $X^\ell, \ell = 1, \ldots, m$, resulting in a set of m quantities. The so-called MI inference is to appropriately combine them to obtain a single answer.

1.2. *MI Wald Test and Fraction of Missing Information.* Let $d^\ell_{\text{W}} = d_{\text{W}}(\widehat{\theta}^\ell, U^\ell)$, $\widehat{\theta}^\ell = \widehat{\theta}(X^\ell)$ and $U^\ell = U(X^\ell)$ be the imputed counterparts of $d_{\text{W}}(\widehat{\theta}, U)$, $\widehat{\theta}$ and U, respectively, for each ℓ. Also, write their averages as

$$\overline{d}_{\text{W}} = \frac{1}{m} \sum_{\ell=1}^{m} d^\ell_{\text{W}}, \qquad \overline{\theta} = \frac{1}{m} \sum_{\ell=1}^{m} \widehat{\theta}^\ell, \qquad \overline{U} = \frac{1}{m} \sum_{\ell=1}^{m} U^\ell. \tag{1.1}$$

Under congeniality (Meng, 1994a), one can show that asymptotically (Rubin and Schenker, 1986) $\mathsf{Var}(\overline{\theta})$ 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 *within-imputation variance*. Intriguingly, $2T$ serves as a universal (estimated) upper bound of $\mathsf{Var}(\overline{\theta})$ under uncongeniality (Xie and Meng, 2017). Under RCs, we have that, as $m, n \to \infty$,

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

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 $\mathscr{T}_\psi, \mathscr{B}_\psi, \mathscr{U}_\psi$ for the entire parameter ψ. Similar to $\overline{U}$, T and B, we define $\overline{U}_\psi$, T_ψ and B_ψ for the component ψ. Note that if $\widehat{\theta}_{\text{com}}$ and $\widehat{\theta}_{\text{obs}}$ are the MLEs of θ based on X_{com} and X_{obs} (under congeniality), respectively, then $\mathscr{U}_\theta \doteq \mathsf{Var}(\widehat{\theta}_{\text{com}})$ and $\mathscr{T}_\theta \doteq \mathsf{Var}(\widehat{\theta}_{\text{obs}})$ as $n \to \infty$, where $A_n \doteq B_n$ means that $A_n - B_n = o_p(A_n + B_n)$.

The straightforward MI Wald test $D_{\text{W}}(T) = d_{\text{W}}(\overline{\theta}, T)/k$ 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 equal fraction of missing information (EFMI).

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

EFMI clearly is a very strong assumption, implying that the missing data have caused an equal amount of loss of information for estimating every component of θ. However, as we shall see shortly, the adoption of this assumption, *for the purpose of hypothesis testing*, is essentially the same as to summarize the impact of (at least) k nuisance parameters due to FMI by a single nuisance parameter, i.e., the average FMI across different components. How well this reduction strategy works therefore will affect more the power of the test than 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 would also lose power when m/k is small or even modest. It is because we simply do not have enough degrees of freedom to estimate all the nuisance parameters well or at all. We will illustrate this point in § 4.2. (To clarify some confusions in literature, $\varkappa$ in Assumption 1 is the *odds of the missing information*, not the FMI, which is $\varkappa = \varkappa/(1 + \varkappa)$.)

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

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

$\widetilde{d}'_{\mathrm{W}} = d_{\mathrm{W}}(\overline{\theta}, \overline{U})$, and the prime "$\prime$" indicates that $\overline{U}$ is used instead of individual $\{U^\ell\}_{\ell=1}^m$. Then, Rubin (2004) proposed a simple MI Wald test statistic:

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

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

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

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

Since the exact null distribution of $\widetilde{D}'_{\mathrm{W}}$ is intractable, Li *et al.* (1991b) proposed to approximate 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 $K_m = k(m-1)$,

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

This approximation assumes n is sufficiently large so that the standard 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.

1.3. *The Current MI Likelihood Ratio Test and Its Defect.* Let $d^\ell_{\mathrm{L}} = d_{\mathrm{L}}(\widehat{\psi}^\ell_0, \widehat{\psi}^\ell \mid X^\ell)$, $\widehat{\psi}^\ell_0 = \widehat{\psi}_0(X^\ell)$ 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. Let their averages be

$$\overline{d}_{\mathrm{L}} = \frac{1}{m}\sum_{\ell=1}^m d^\ell_{\mathrm{L}}, \qquad \overline{\psi}_0 = \frac{1}{m}\sum_{\ell=1}^m \widehat{\psi}^\ell_0, \qquad \overline{\psi} = \frac{1}{m}\sum_{\ell=1}^m \widehat{\psi}^\ell. \tag{1.7}$$

Similar to $\widetilde{r}'_{\mathrm{W}}$, Meng and Rubin (1992) proposed to estimate $\varkappa_m$ 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). Computation of $\widetilde{r}_{\mathrm{L}}$ requires users to 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

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

whose null distribution can be approximated by $F_{k,\widetilde{\mathrm{df}}(\widetilde{r}_{\mathrm{L}},k)}$.

The main theoretical justification (and motivation) was the asymptotic equivalence between the complete-data Wald test statistic and LRT statistic *under the null*, as stated in Property 1.1. This equivalence permitted the replacement of $\overline{d}'_{\mathrm{W}}$ and $\widetilde{d}'_{\mathrm{W}}$ in (1.3) respectively by $\overline{d}_{\mathrm{L}}$ and $\widetilde{d}_{\mathrm{L}}$ 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 $\widetilde{r}_{\mathrm{L}} \geqslant 0$. Since $\widetilde{D}_{\mathrm{L}}$ is referred to an F distribution and $\widetilde{r}_{\mathrm{L}}$ estimates $\text{ɾ}_m \geqslant 0$, clearly negative values of $\widetilde{D}_{\mathrm{L}}$ or $\widetilde{r}_{\mathrm{L}}$ will cause trouble.

Second, $\widetilde{D}_{\mathrm{L}}$ is not invariant to re-parameterization of ψ. For each individual LRT statistic d_{L}^ℓ and bijective map g such that $\varphi = g(\psi)$, we have

$$d_{\mathrm{L}}^\ell = d_{\mathrm{L}}(\widehat{\psi}_0^\ell, \widehat{\psi}^\ell \mid X^\ell) = d_{\mathrm{L}}(g^{-1}(\widehat{\varphi}_0^\ell), g^{-1}(\widehat{\varphi}^\ell) \mid X^\ell), \tag{1.10}$$

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

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

for most bijective maps g, where $\overline{\varphi}_0 = m^{-1}\sum_{\ell=1}^m \widehat{\varphi}_0^\ell$ and $\overline{\varphi} = m^{-1}\sum_{\ell=1}^m \widehat{\varphi}^\ell$. See § 4 how $\widetilde{D}_{\mathrm{L}}$ vary dramatically with parametrizations in finite samples.

Third, the estimator $\widetilde{r}_{\mathrm{L}}$ involves the estimators of ψ under H_0, i.e., $\widehat{\psi}_0^\ell$ and $\overline{\psi}_0$. When H_0 fails, they may not be consistent for ψ. As a result, $\widetilde{r}_{\mathrm{L}}$ is no longer consistent for ɾ_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 § 3.2) show that under certain parametrizations, the power may nearly vanish for obviously false H_0.

Fourth, in order to compute $\widetilde{d}_{\mathrm{L}}$ in (1.8), users need to have access to the LRT function $\widetilde{\mathscr{D}}_{\mathrm{L}}$, but, in most software, the function is built as $\mathscr{D}_{\mathrm{L}}$, where

$$\widetilde{\mathscr{D}}_{\mathrm{L}} : (X, \psi_0, \psi) \mapsto d_{\mathrm{L}}(\psi_0, \psi \mid X), \quad \mathscr{D}_{\mathrm{L}} : X \mapsto d_{\mathrm{L}}(\widehat{\psi}_0(X), \widehat{\psi}(X) \mid X). \tag{1.11}$$

Hence users would need to write themselves a subroutine for evaluating $\widetilde{\mathscr{D}}_{\mathrm{L}}$. This may not be feasible for users because of lack of information or skills.

In short, four problems need to be resolved: (i) lack of non-negativity, (ii) lack of invariance, (iii) lack of consistency and power, and (iv) lack of a computationally feasible algorithm. Problems (i)–(iii) are resolved in § 2 below, where § 2.1 presents an invariant combining rule, which fully resolves (ii). Next, we propose two estimators of $\varkappa_m$ (or equivalently $\varkappa$) in § 2.2 and § 2.4. We start with a quick ad hoc fix that requires slightly less assumption but only addresses (i), and then construct a test that fully resolves (i) and (iii). Finally, in § 3, we derive a very handy algorithm, which resolves (iv).

2. Improved MI Likelihood Ratio Tests.

2.1. *An Invariant Combining Rule.* To derive a MI LRT that is invariant to re-parametrization, we replace $\widetilde{d}_{\mathrm{L}}$ by an asymptotically equivalent version that behaves like a standard LRT statistic. Specifically, let

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

We emphasize that $\overline{L}(\psi)$ is not a real log-likelihood (even if we drop the divider m), because it does not properly model the completed datasets: $\mathbb{X} = \{X^1, \ldots, X^m\}$ (e.g., addressing the issue that all X^{ℓ}s 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 datasets through averaging MI LRT functions as in (2.1), rather than averaging 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 than individual MLEs, since the former has a much better chance to capture functional features of the real log-likelihood than any of their (local) maximizers can.

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

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

(a) $\psi \mapsto \underline{L}(\psi) = n^{-1} \log f(X \mid \psi)$ *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) if $\underline{I}(\psi) = -\partial^2 \underline{L}(\psi)/\partial\psi\partial\psi^\intercal$*, then for each* ψ*, there exists a positive definite matrix* $\underline{\mathscr{I}}(\psi) = \mathscr{U}_\psi^{-1}$ *such that* $\underline{I}(\psi) \xrightarrow{\text{pr}} \underline{\mathscr{I}}(\psi)$ *as* $n \to \infty$*; and*
(d) the observed-data MLE $\widehat{\psi}_{\text{obs}}$ *of* ψ *obeys*

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

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

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

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

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

independently for $\ell = 1, \ldots, m$*, as* $n \to \infty$*, provided that* $\mathscr{B}_\psi^{-1}$ *is well-defined.*

Assumption 2 holds under the usual RCs that guarantee normality and consistency of MLEs. When the imputations $X_{\text{mis}}^1, \ldots, X_{\text{mis}}^m$ are drawn independently from (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 sub-vector θ in Assumptions 2 and 3. These θ-version assumptions are sufficient to guarantee the validity of the following Theorem 2.4 and Corollary 2.3. For simplicity, Assumption 1, the θ-version of Assumptions 2 and 3, and conditions that are strong enough to guarantee Property 1.1 are collectively written as RC_θ, 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

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

for a given value of $\varkappa_m$. It follows the forms of (1.4) and of (1.9). The issue is then how to estimate $\varkappa_m$ that avoids the defects of $\widetilde{r}_{\text{L}}$ of (1.8).

2.2. *An Improved Estimator of $\varkappa_m$.* Using $\widehat{d}_{\mathrm{L}}$ in (2.3), we can modify $\widetilde{r}_{\mathrm{L}}$ in (1.8) to a potentially better estimator:

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

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

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. Then, the profile likelihood estimator of η given θ, i.e., $\widehat{\eta}_{\theta} = \arg\max_{\eta} L(\theta, \eta \mid X)$, does not depend on θ. Trivially, if there is no nuisance parameter η, the separation condition is satisfied. More generally, we have

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 $\theta^{\star}$.*

Corollary 2.3 ensures that, under H_0, $\widehat{r}_{\mathrm{L}}$ is non-negative asymptotically and converges in probability to the true $\varkappa$. But it also reveals another fundamental defect of $\widehat{r}_{\mathrm{L}}$: under H_1, the limit $\varkappa_0$ may not equal to $\varkappa$, a problem we will address in § 2.2. Fortunately, since $\widehat{d}_{\mathrm{L}} \xrightarrow{\mathrm{pr}} \infty$ under H_1, the LRT statistic $\widehat{D}_{\mathrm{L}}(\widehat{r}_{\mathrm{L}})$ is still powerful, albeit the power may be somewhat 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 of ψ for small m, e.g., in some cases, $\widetilde{r}_{\mathrm{L}}$ becomes more negative as H_1 moves away from H_1; see § 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.

2.3. *A Complication Caused by Nuisance Parameter.* To better understand the source of the negativity of $\widehat{r}_{\mathrm{L}}$, we extend $\overline{L}(\psi)$ in (2.1) to allow it take m different arguments:

$$\overline{L}(\psi^1, \ldots, \psi^m) = \frac{1}{m}\sum_{\ell=1}^{m} L^{\ell}(\psi^{\ell}). \tag{2.9}$$

TABLE 1
The definitions of hypotheses H_0^0, H_0^1, H_1^0, H_1^1.

	$\mathscr{C}^0 : \psi^1 = \cdots = \psi^m \in \Psi$ (i.e., $\varkappa = 0$)	$\mathscr{C}^1 : \psi^1, \ldots, \psi^m \in \Psi$ (i.e., $\varkappa \geqslant 0$)
$\mathscr{C}_0 : \theta^1 = \cdots = \theta^m = \theta_0 \in \Theta$ (i.e., H_0-constrained)	$H_0^0 = \mathscr{C}_0 \cap \mathscr{C}^0$	$H_0^1 = \mathscr{C}_0 \cap \mathscr{C}^1$
$\mathscr{C}_1 : \theta^1, \ldots, \theta^m \in \Theta$ (i.e., not H_0-constrained)	$H_1^0 = \mathscr{C}_1 \cap \mathscr{C}^0$	$H_1^1 = \mathscr{C}_1 \cap \mathscr{C}^1$

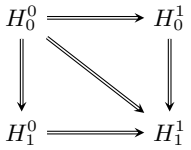

FIG 1. *The relationships between the four hypotheses H_0^0, H_0^1, H_1^0, H_1^1. Each arrow denotes an implication, e.g., $H_0^0 \Rightarrow H_0^1$ means that H_0^0 implies H_0^1.*

Using the "log-likelihood" $\overline{L}(\psi^1, \ldots, \psi^m)$, we can construct, at least conceptually, four hypotheses H_0^0, H_0^1, H_1^0, H_1^1 defined in Table 1. Each of them consists of zero, one or two of the constraints

$$\mathscr{C}_0 : \theta^1 = \cdots = \theta^m = \theta_0 \qquad \text{and} \qquad \mathscr{C}^0 : \psi^1 = \cdots = \psi^m.$$

The constraint $\mathscr{C}_0$ is equivalent to H_0, and the constraint $\mathscr{C}^0$ means that all ψ^ℓs are equal, and hence it is effectively equivalent to $\varkappa = 0$, i.e., no missing information. The relationships among H_0^0, H_0^1, H_1^0, H_1^1 can be visualized in Figure 1. Define the maximized value of $\overline{L}(\psi^1, \ldots, \psi^m)$ under hypothesis $H \in \{H_0^0, H_0^1, H_1^0, H_1^1\}$ by $\mathbb{L}(H)$. Then we can re-express $(\overline{d}_{\mathrm{L}} - \widehat{d}_{\mathrm{L}})/2$ as

$$(\overline{d}_{\mathrm{L}} - \widehat{d}_{\mathrm{L}})/2 = \left\{\mathbb{L}(H_1^1) - \mathbb{L}(H_1^0)\right\} - \left\{\mathbb{L}(H_0^1) - \mathbb{L}(H_0^0)\right\}. \tag{2.10}$$

Whereas the two bracketed terms in (2.10) are non-negative because they correspond to two LRT statistics, the difference between these two terms is not guaranteed to be non-negative. A simple example illustrates this well. For the regression model $[Y \mid X_1, X_2] \sim \mathscr{N}(\beta_0 + \beta_1 X_1 + \beta_2 X_2, \sigma^2)$, the LRT statistic for testing $H_1^0 : \beta_1 = 0, \beta_2 \in \mathbb{R}$ against $H_1^1 : \beta_1, \beta_2 \in \mathbb{R}$ is not necessarily larger (or smaller) than that for testing $H_0^0 : \beta_1 = \beta_2 = 0$ against $H_0^1 : \beta_1 \in \mathbb{R}, \beta_2 = 0$. A schematic illustration is provided in Figure 2.

The decomposition (2.10) provides another interpretation of $\widehat{r}_{\mathrm{L}}$. The test statistic $\mathbb{L}(H_1^1) - \mathbb{L}(H_1^0)$ seeks evidence for detecting the falsity of $\varkappa = 0$ in

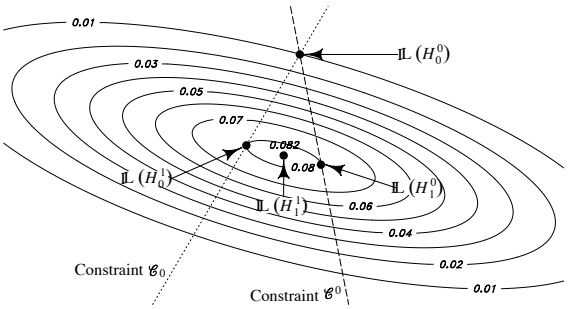

FIG 2. *A schematic illustration of the sign of (2.10). The contour lines of* $\overline{L}(\psi^1,\ldots,\psi^m)$ *are plotted. The two straight lines refer to constraints* $\mathscr{C}_0$ *and* $\mathscr{C}^0$. *Since* $\mathbb{L}(H_1^1) = 0.082$, $\mathbb{L}(H_0^1) = \mathbb{L}(H_1^0) = 0.08$, *and* $\mathbb{L}(H_0^0) = 0.01$, *we have* $\{\mathbb{L}(H_1^1) - \mathbb{L}(H_1^0)\} - \{\mathbb{L}(H_0^1) - \mathbb{L}(H_0^0)\} = 0.002 - 0.007 < 0$. *Note that the function* $\overline{L}(\psi^1,\ldots,\psi^m)$ *in (2.9) is at least* 4*-dimensional (i.e.,* $\theta^1,\theta^2,\eta^1,\eta^2$*) generally, so the above illustration in a* 2*-dimension space is just conceptual.*

both θ and η, whereas $\mathbb{L}(H_0^1) - \mathbb{L}(H_0^0)$ seeks evidence only in η. For cases where θ and η are orthogonal (at least locally), the left-hand side of (2.10) can be viewed as a measure of evidence against $\curlyvee = 0$ solely from θ; Proposition 2.2 already hinted this possibility. However, the "test statistic" (2.10) has a very serious problem apart from being possibly negative. Because $\mathscr{C}_0$ requires all θ^ℓs to coincide with a specific θ_0, $\mathscr{C}_0$ is not nested within $\mathscr{C}^0$, i.e., $\mathscr{C}^0 \nRightarrow \mathscr{C}_0$. Hence $\widehat{r}_{\mathrm{L}}$ is guaranteed to consistently estimate $\curlyvee_m$ only under H_0. This explains Corollary 2.3, and leads to an improvement below.

2.4. *A Consistent and Non-negative Estimator of* $\curlyvee_m$. Our new estimator simply drops the second term in (2.10), that is, we estimate $\curlyvee_m$ by

$$\widehat{r}_{\mathrm{L}}^{\,\diamondsuit} = \frac{m+1}{h(m-1)}(\overline{\delta}_{\mathrm{L}} - \widehat{\delta}_{\mathrm{L}}), \quad \text{where} \tag{2.11}$$

$$\overline{\delta}_{\mathrm{L}} = 2\overline{L}(\widehat{\psi}^1,\ldots,\widehat{\psi}^m), \qquad \widehat{\delta}_{\mathrm{L}} = 2\overline{L}(\widehat{\psi}^*,\ldots,\widehat{\psi}^*), \tag{2.12}$$

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 the entire parameter ψ, a stronger requirement than Assumption 1. 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 will be referred to as RC_ψ.

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

THEOREM 2.4. *Assume* RC_ψ. *Then 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)* $\widehat{r}_{\mathrm{L}}^{\diamondsuit} \xrightarrow{\mathrm{pr}} \varkappa$ *as* $m, n \to \infty$, *where* $\varkappa$ *is given in Assumption 4.*

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

$$\widehat{D}_{\mathrm{L}}^{+} = \widehat{D}_{\mathrm{L}}(\widehat{r}_{\mathrm{L}}^{+}) \qquad \text{and} \qquad \widehat{D}_{\mathrm{L}}^{\diamondsuit} = \widehat{D}_{\mathrm{L}}(\widehat{r}_{\mathrm{L}}^{\diamondsuit}). \tag{2.13}$$

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

2.5. *Reference Null Distributions.* The estimators $\widehat{r}_{\mathrm{L}}^{+}$ and $\widetilde{r}_{\mathrm{L}}$ have the same functional form asymptotically ($n \to \infty$) and rely on the same set of assumptions, hence they have the same asymptotic distribution.

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

$$\frac{\widehat{r}_{\mathrm{L}}^{+}}{\varkappa_m} \Rightarrow M_2 \qquad \textit{and} \qquad \widehat{D}_{\mathrm{L}}^{+} \Rightarrow \frac{(1+\varkappa_m)\, M_1}{1+\varkappa_m M_2} \tag{2.14}$$

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, the null distribution of $\widehat{D}_{\mathrm{L}}^{+} = \widehat{D}_{\mathrm{L}}(\widehat{r}_{\mathrm{L}}^{+})$ can be approximated by $F_{k,\widetilde{\mathrm{df}}(\widehat{r}_{\mathrm{L}}^{+},k)}$, but a better approximation will be provided shortly.

For the other proposal, although $\widehat{r}_{\mathrm{L}}^{+} - \widehat{r}_{\mathrm{L}}^{\diamondsuit} \xrightarrow{\mathrm{pr}} 0$ as $n \to \infty$ under H_0, their non-degenerated distributions (after proper scaling) are different because $\widehat{r}_{\mathrm{L}}^{\diamondsuit}$ relies on an average FMI in ψ, but $\widehat{r}_{\mathrm{L}}^{+}$ only on an average FMI in θ.

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

$$\frac{\widehat{r}_{\mathrm{L}}^{\diamondsuit}}{\varkappa_m} \Rightarrow M_3 \sim \frac{\chi^2_{h(m-1)}}{h(m-1)} \tag{2.15}$$

as $n \to \infty$, *where* M_3 *is independent of the* M_1 *defined in (2.14).*

Theorem 2.6 implies that, if n can be regarded as infinity and $\widehat{r}_{\mathrm{L}}^{\diamondsuit}$ is uniformly integrable in $\mathscr{L}^2$, then

$$\mathsf{Bias}(\widehat{r}_{\mathrm{L}}^{\diamondsuit}) = \mathsf{E}(\widehat{r}_{\mathrm{L}}^{\diamondsuit}) - \varkappa_m = 0 \qquad \text{and} \qquad \mathsf{Var}(\widehat{r}_{\mathrm{L}}^{\diamondsuit}) = \frac{2\varkappa_m^2}{h(m-1)} = O(m^{-1})$$

as $m \to \infty$. Therefore $\widehat{r}_{\mathrm{L}}^{\lozenge}$ is a $\sqrt{m}$-consistent estimator of $\boldsymbol{\mathit{r}}$ in $\mathscr{L}^2$. Moreover, for each $m > 1$ and as $n \to \infty$, we have

$$\frac{\mathsf{Bias}(\widehat{r}_{\mathrm{L}}^{+})}{\mathsf{Bias}(\widehat{r}_{\mathrm{L}}^{\lozenge})} \to 1 \qquad \text{and} \qquad \frac{\mathsf{Var}(\widehat{r}_{\mathrm{L}}^{+})}{\mathsf{Var}(\widehat{r}_{\mathrm{L}}^{\lozenge})} \to \frac{h}{k} \geqslant 1,$$

which implies that $\widehat{r}_{\mathrm{L}}^{\lozenge}$ is no less efficient than $\widehat{r}_{\mathrm{L}}^{+}$ when RC_{ψ} holds. This is of no surprise because of the extra information in $\widehat{r}_{\mathrm{L}}^{\lozenge}$ brought in by the stronger Assumption 4. Result (2.15) also gives us the exact (i.e., for any $m > 1$, but assuming $n \to \infty$) reference null distribution of $\widehat{D}_{\mathrm{L}}^{\lozenge}$, as given below.

THEOREM 2.7. *Assume* RC_{ψ} *and* $m > 1$. *Under* H_0, *we have*

$$\widehat{D}_{\mathrm{L}}^{\lozenge} \Rightarrow \frac{(1+\boldsymbol{\mathit{r}}_m)\, M_1}{1+\boldsymbol{\mathit{r}}_m M_3} \equiv D \tag{2.16}$$

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 $\boldsymbol{\mathit{r}}_m$ *on the null distribution* diminishes with m, because $\widehat{D}_{\mathrm{L}}^{\lozenge}$ and $\widehat{D}_{\mathrm{L}}^{+}$ converge in distribution to $M_1 = \chi^2_k/k$ as $m, n \to \infty$. Since $M_3 \xrightarrow{\mathrm{pr}} 1$ faster than $M_2 \xrightarrow{\mathrm{pr}} 1$, $\widehat{D}_{\mathrm{L}}^{\lozenge}$ is expected to be more robust to $\boldsymbol{\mathit{r}}_m$. Nevertheless, because m typically is small in practice (e.g., $m \leqslant 10$), we cannot ignore the impact of $\boldsymbol{\mathit{r}}_m$. This issue has been largely dealt with in the literature by seeking an $F_{k,\mathrm{df}}$ distribution as an approximate null distribution, as in Li *et al.* (1991b). However, directly adopting their $\widetilde{\mathrm{df}}$ of (1.6) leads to poorer approximation for our purposes; see below. A better approximation is to match the first two moments of the denominator of (2.16), $1+\boldsymbol{\mathit{r}}_m M_3$, with that of a scaled χ^2: $a\chi^2_b/b$. This yields $a = 1+\boldsymbol{\mathit{r}}_m$ and $b = (1+\boldsymbol{\mathit{r}}_m^{-1})^2 h(m-1)$, and the approximated $F_{k,\widehat{\mathrm{df}}(\boldsymbol{\mathit{r}}_m,h)}$, where

$$\widehat{\mathrm{df}}(\boldsymbol{\mathit{r}}_m, h) = \left\{\frac{1+\boldsymbol{\mathit{r}}_m}{\boldsymbol{\mathit{r}}_m}\right\}^2 h(m-1) = \frac{h(m-1)}{\boldsymbol{\mathit{f}}_m^2}. \tag{2.17}$$

This degrees of freedom is appealing because it simply inflates the denominator degrees of freedom M_3 by dividing it by $\boldsymbol{\mathit{f}}_m^2$, where $\boldsymbol{\mathit{f}}_m = \boldsymbol{\mathit{r}}_m/(1+\boldsymbol{\mathit{r}}_m)$ the finite imputation corrected FMI. Intuitively, the less missing information, the closer $F_{k,\widehat{\mathrm{df}}(\boldsymbol{\mathit{r}}_m,h)}$ should be to χ^2_k/k, the usual large-n asymptotic χ^2 test; as mentioned earlier, for small n, see Barnard and Rubin (1999).

To compare $F_{k,\widehat{\mathrm{df}}(\boldsymbol{\mathit{r}}_m,h)}$ with $F_{k,\widetilde{\mathrm{df}}(\boldsymbol{\mathit{r}}_m,h)}$ as approximations to the distribution of D given in (2.16), we compute via simulations

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

where $F^{-1}_{k,\mathrm{df}}(q)$ denotes the q-quantile of $F_{k,\mathrm{df}}$. We draw $N = 2^{18}$ independent copies D for each of the following possible combinations: $m \in \{3, 5, 7\}$, $k \in \{1, 2, 4, 8\}$, $\tau = h/k \in \{1, 2, 3\}$, $\text{ŕ}_m \in \{0, 0.1, \ldots, 0.9\}$, and following Benjamin *et al.* (2018)'s recommendation, we use both $\alpha \in \{0.5\%, 5\%\}$. The results for $\alpha = 0.5\%$ and for $\alpha = 5\%$ are shown respectively in Figure 3 and in the Appendix. In general, $\widehat{\alpha}$ approximates α much better than $\widetilde{\alpha}$, especially when m, k, h are small. When m, h are larger, their performances are similar because both $F_{k,\widetilde{\mathrm{df}}(\text{ŕ}_m,h)}$ and $F_{k,\widehat{\mathrm{df}}(\text{ŕ}_m,h)}$ get close to χ^2_k/k. But the performances of $\widetilde{\alpha}$ and $\widehat{\alpha}$ are not monotonic in ŕ_m. Generally speaking, the performance of $F_{k,\widehat{\mathrm{df}}(\text{ŕ}_m,h)}$ is particularly good for $0\% \lesssim \text{ŕ}_m \lesssim 30\%$. 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 uncertainties in estimating ŕ_m. Since the $F_{k,\mathrm{df}}$ 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 context, leading to acceptable approximations as empirically demonstrated in Section 4. Nevertheless, further improvements are likely and should be sought.

3. Computational Considerations and Comparisons. The statistic $\overline{d}_{\mathrm{L}}$ of (1.7) is easy to compute because only the standard complete-data procedure $\mathscr{D}_{\mathrm{L}} : X \mapsto d_{\mathrm{L}}(\widehat{\psi}_0(X), \widehat{\psi}(X) \mid X)$ is needed. However, $\widehat{d}_{\mathrm{L}}$ of (2.3) and $\widehat{r}^{\diamondsuit}_{\mathrm{L}}$ of (2.8) in general cannot be computed solely by $\mathscr{D}_{\mathrm{L}}$, e.g., $\widehat{d}_{\mathrm{L}}$ requires

$$\overline{\mathscr{D}}_{\mathrm{L}} : \mathbb{X} \mapsto \frac{1}{m}\sum_{\ell=1}^{m} d_{\mathrm{L}}(\widehat{\psi}^*_0(\mathbb{X}), \widehat{\psi}^*(\mathbb{X}) \mid X^\ell).$$

Creating a subroutine for this computation requires additional effort and information that may beyond a user's capacity. Here we show how to compute or approximate $\widehat{d}_{\mathrm{L}}$ and $\widehat{r}^{\diamondsuit}_{\mathrm{L}}$ solely by $\mathscr{D}_{\mathrm{L}}$ or a trivial modification of $\mathscr{D}_{\mathrm{L}}$.

3.1. *Computationally Feasible Combining Rule.* We begin with precise notation for our complete data X and its sampling model $f(X|\psi)$. For the vast majority of real-world datasets, X is of the form of an $n \times p$ matrix, with rows indicating subjects and columns variables/attributes. We then write $X = (X_1, \ldots, X_n)^\intercal$, and its sampling model by $f_n(X \mid \psi)$. Correspondingly, the ℓth completed-dataset by MI is $X^\ell = (X^\ell_1, \ldots, X^\ell_n)^\intercal$. Define the stacked dataset by $X^{\mathrm{S}} = [(X^1)^\intercal, \ldots, (X^m)^\intercal]^\intercal$, a matrix having mn rows, which is conceptually different from the collection of datasets $\mathbb{X} = \{X^1, \ldots, X^m\}$.

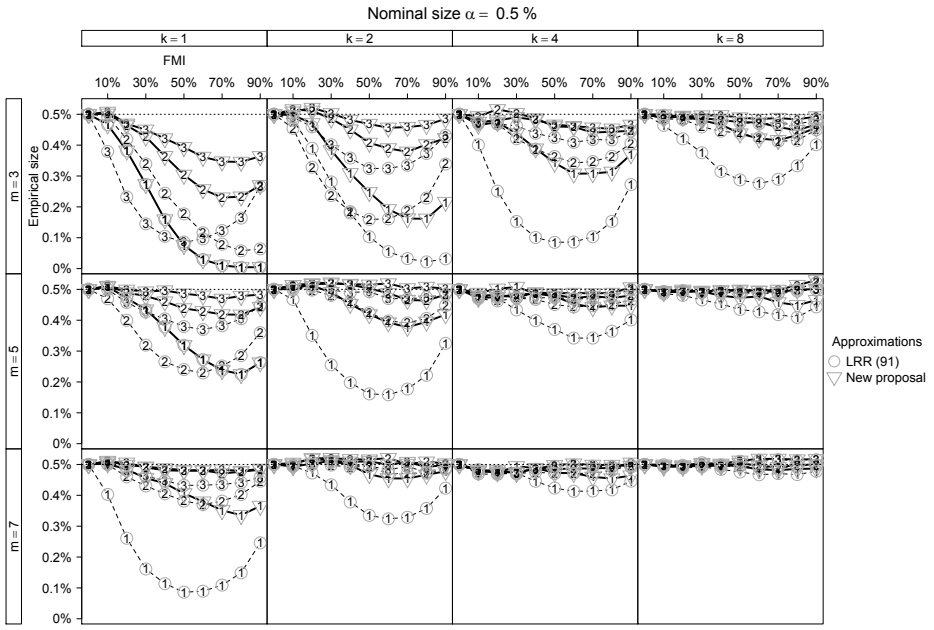

FIG 3. *The performances of two different approximated null distributions when the nominal size is* $\alpha = 0.5\%$. *The vertical axis denotes* $\widehat{\alpha}$ *or* $\widetilde{\alpha}$, *and the horizontal axis denotes the value of* $\not{f}_m$. *The number attached to each line denotes the value of* $\tau = h/k$. *The proposed approximation* $\widehat{\alpha}$ *is denoted by thick solid lines with triangles, and the existing approximation* $\widetilde{\alpha}$ *is denoted by thin dashed lines with circles.*

Treating X^{S} as a dataset with size mn, we can define

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

which, other than the scaling factor $1/m$, is just the ordinary log-likelihood function of ψ based on the dataset X^{S} (for computation purposes). Consequently, as long as the user's complete-data procedure can handle size mn instead of just n, the user can apply it to X^{S} to obtain

$$\widehat{\psi}_0^{\mathrm{S}} = \underset{\psi \in \Psi \,:\, \theta(\psi) = \theta_0}{\arg\max}\ \overline{L}^{\mathrm{S}}(\psi) \qquad \text{and} \qquad \widehat{\psi}^{\mathrm{S}} = \underset{\psi \in \Psi}{\arg\max}\, \overline{L}^{\mathrm{S}}(\psi). \tag{3.2}$$

Consequently, the quantities

$$\widehat{\delta}_{0,\mathrm{S}} = 2\overline{L}^{\mathrm{S}}(\widehat{\psi}_0^{\mathrm{S}}) \qquad \text{and} \qquad \widehat{\delta}_{\mathrm{S}} = 2\overline{L}^{\mathrm{S}}(\widehat{\psi}^{\mathrm{S}}) \tag{3.3}$$

are readily available from the user's complete-data procedure. It is then desirable if we can replace $\overline{L}(\psi)$ by $\overline{L}^{\mathrm{S}}(\psi)$ in the proposed test statistics.

Precisely, in parallel to (2.7), (2.8) and (2.11), we define

$$\widehat{D}_{\mathrm{S}}(\varkappa_m) = \frac{\widehat{d}_{\mathrm{S}}}{k(1+\varkappa_m)}, \qquad \text{with } \widehat{d}_{\mathrm{S}} = \widehat{\delta}_{\mathrm{S}} - \widehat{\delta}_{0,\mathrm{S}} \text{ of (3.3)}; \tag{3.4}$$

$$\widehat{r}_{\mathrm{S}} = \frac{m+1}{k(m-1)}(\overline{d}_{\mathrm{S}} - \widehat{d}_{\mathrm{S}}), \qquad \text{with } \overline{d}_{\mathrm{S}} = \overline{d}_{\mathrm{L}} \text{ of (1.7)}; \tag{3.5}$$

$$\widehat{r}^{\,\diamond}_{\mathrm{S}} = \frac{m+1}{h(m-1)}(\overline{\delta}_{\mathrm{S}} - \widehat{\delta}_{\mathrm{S}}), \qquad \text{with } \overline{\delta}_{\mathrm{S}} = \overline{\delta}_{\mathrm{L}} \text{ of (2.12)}; \tag{3.6}$$

and $\widehat{r}^{\,+}_{\mathrm{S}} = \max(0, \widehat{r}_{\mathrm{S}})$. The "stacked" counterparts of $\widehat{D}^{\diamond}_{\mathrm{L}}$ and its existing counterparts $\widehat{D}_{\mathrm{L}}$ and $\widehat{D}^{+}_{\mathrm{L}}$ (see (2.13)) then are given by

$$\widehat{D}^{\diamond}_{\mathrm{S}} = \widehat{D}_{\mathrm{S}}(\widehat{r}^{\,\diamond}_{\mathrm{S}}), \qquad \widehat{D}_{\mathrm{S}} = \widehat{D}_{\mathrm{S}}(\widehat{r}_{\mathrm{S}}), \qquad \widehat{D}^{+}_{\mathrm{S}} = \widehat{D}_{\mathrm{S}}(\widehat{r}^{\,+}_{\mathrm{S}}). \tag{3.7}$$

PROPOSITION 3.1. *If $X = (X_1, \ldots, X_n)^{\intercal}$ is row-independent for arbitrary n, i.e., $f(X \mid \psi) = \prod_{i=1}^{n} f(X_i \mid \psi)$, then (2.1) and (3.1) are the same: $\overline{L}(\psi) \equiv \overline{L}^{\mathrm{S}}(\psi)$. Consequently, $\widehat{D}_{\mathrm{S}} \equiv \widehat{D}_{\mathrm{L}}$ and $\widehat{D}^{\diamond}_{\mathrm{S}} \equiv \widehat{D}^{\diamond}_{\mathrm{L}}$.*

Since for many applications, the rows correspond to individual subjects, 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 it does not hold, we can still have $\widehat{d}_{\mathrm{L}} \doteq \widehat{d}_{\mathrm{S}}$ under some RCs that guarantee $\overline{L}(\psi)$ and $\overline{L}^{\mathrm{S}}(\psi)$ are close; see Appendix A, where we also reveal a subtle but important difference between $\widehat{r}^{\,\diamond}_{\mathrm{L}}$ and $\widehat{r}^{\,\diamond}_{\mathrm{S}}$.

Similar to $\mathscr{D}_{\mathrm{L}}$ in (1.11), we define complete-data functions (with data X being the only input)

$$\mathscr{D}_{\mathrm{L},0}(X) = 2\log f(X \mid \widehat{\psi}_0(X)), \quad \mathscr{D}_{\mathrm{L},1}(X) = 2\log f(X \mid \widehat{\psi}(X)). \tag{3.8}$$

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 complete data log-likelihood for objects belonging to classes `"glm"`, `"lm"`, `"nls"` and `"Arima"`.

The Algorithms 1 and 2 listed below compute tests given $\widehat{D}^{\diamond}_{\mathrm{S}}$ and $\widehat{D}^{+}_{\mathrm{S}}$, respectively. Whenever possible, we recommend the use of the robust MI LRT given by Algorithm 1, since it has the best theoretical guarantee. The second test can be useful when $\mathscr{D}_{\mathrm{L}}$ but not $\mathscr{D}_{\mathrm{L},1}$ is available.

3.2. *Computational Comparison with Existing Tests.* First, we list some existing estimators of $\varkappa_m$ and their computation. Let $s^2_{\mathrm{W},1}$ and $s^2_{\mathrm{W},1/2}$ be the sample variances of $\{d^{\ell}_{\mathrm{W}}\}_{\ell=1}^{m}$ and $\{\sqrt{d^{\ell}_{\mathrm{W}}}\}_{\ell=1}^{m}$, respectively. By the method of

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

Input: Datasets $X^1, \ldots, X^m$; dimensions h, k; functions $\mathscr{D}_{\mathrm{L},1}$, $\mathscr{D}_{\mathrm{L}}$ in (3.8), (1.11).
begin
Compute $\overline{\delta}_{\mathrm{S}} = m^{-1}\{\mathscr{D}_{\mathrm{L},1}(X^1) + \cdots + \mathscr{D}_{\mathrm{L},1}(X^m)\}$.
(i) Stack the datasets to form $X^{\mathrm{S}} = [(X^1)^\intercal, \ldots, (X^m)^\intercal]^\intercal$.
(ii) Compute $\widehat{d}_{\mathrm{S}} = m^{-1}\mathscr{D}_{\mathrm{L}}(X^{\mathrm{S}})$ and $\widehat{\delta}_{\mathrm{S}} = m^{-1}\mathscr{D}_{\mathrm{L},1}(X^{\mathrm{S}})$.
Calculate $\widehat{r}_{\mathrm{S}}^{\Diamond}$ according to (3.6), and $\widehat{D}_{\mathrm{S}}^{\Diamond}$ according to (3.4) and (3.7).
Calculate $\widehat{\mathrm{df}}(\widehat{r}_{\mathrm{S}}^{\Diamond}, h)$ according to (2.17).
Compute the p-value as $1 - F_{k,\widehat{\mathrm{df}}(\widehat{r}_{\mathrm{S}}^{\Diamond},h)}(\widehat{D}_{\mathrm{S}}^{\Diamond})$.

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

Input: Datasets $X^1, \ldots, X^m$; dimension k; function $\mathscr{D}_{\mathrm{L}}$ in (1.11).
begin
Compute $\overline{d}_{\mathrm{L}} = m^{-1}\{\mathscr{D}_{\mathrm{L}}(X^1) + \cdots + \mathscr{D}_{\mathrm{L}}(X^m)\}$.
(i) Stack the datasets to form $X^{\mathrm{S}} = [(X^1)^\intercal, \ldots, (X^m)^\intercal]^\intercal$.
(ii) Compute $\widehat{d}_{\mathrm{S}} = m^{-1}\mathscr{D}_{\mathrm{L}}(X^{\mathrm{S}})$.
Calculate $\widehat{r}_{\mathrm{S}}^{+}$ according to (3.5), and $\widehat{D}_{\mathrm{S}}^{+}$ according to (3.4) and (3.7).
Calculate $\widehat{\mathrm{df}}(\widehat{r}_{\mathrm{S}}^{+}, k)$ according to (2.17).
Compute the p-value as $1 - F_{k,\widehat{\mathrm{df}}(\widehat{r}_{\mathrm{S}}^{+},k)(\widehat{D}_{\mathrm{S}}^{+})}$.

moments concerning $s^2_{\mathrm{W},1}$ and $s^2_{\mathrm{W},1/2}$, Rubin (2004) and Li *et al.* (1991a) respectively proposed estimating $\curlyvee_m$ by

$$\widetilde{r}_{\mathrm{W},1} = \frac{(1+1/m)s^2_{\mathrm{W},1}}{2\overline{d}_{\mathrm{W}} + \sqrt{\left\{4\overline{d}^2_{\mathrm{W}} - 2ks^2_{\mathrm{W},1}\right\}^+}}, \qquad \widetilde{r}_{\mathrm{W},1/2} = (1+1/m)s^2_{\mathrm{W},1/2}, \tag{3.9}$$

where $\{a\}^+ = \max(0, a)$. Note that when k is large and m is small, using (3.9) may lead to power loss, although the users have no choice when they are given only $\{d^\ell_{\mathrm{W}}\}^m_{\ell=1}$. A trivial modification of $\widetilde{r}_{\mathrm{L}}$ of (1.8), i.e., $\widetilde{r}^+_{\mathrm{L}} = \max(0, \widetilde{r}_{\mathrm{L}})$, is a better alternative if the user is able to compute $\widetilde{r}_{\mathrm{L}}$.

Second, we list some alternative MI combining rules. Having the above estimators of $\curlyvee_m$, we can insert them into the following combining rules:

$$\widetilde{D}'_{\mathrm{W}}(\curlyvee_m) = \frac{\widetilde{d}'_{\mathrm{W}}}{k(1+\curlyvee_m)}, \quad \widetilde{D}_{\mathrm{L}}(\curlyvee_m) = \frac{\widetilde{d}_{\mathrm{L}}}{k(1+\curlyvee_m)}, \quad \widetilde{D}^+_{\mathrm{L}}(\curlyvee_m) = \left\{\widetilde{D}_{\mathrm{L}}(\curlyvee_m)\right\}^+.$$

Using (1.3) and (1.8), we can also define the following combining rules:

$$\overline{D}'_{\mathrm{W}}(\curlyvee_m) = \frac{\overline{d}'_{\mathrm{W}} - \frac{k(m-1)}{m+1}\curlyvee_m}{k(1+\curlyvee_m)}, \qquad \overline{D}_{\mathrm{L}}(\curlyvee_m) = \frac{\overline{d}_{\mathrm{L}} - \frac{k(m-1)}{m+1}\curlyvee_m}{k(1+\curlyvee_m)}. \tag{3.10}$$

The combining rule $\overline{D}'_{\mathrm{W}}(\text{ɾ}_m)$ is useful when computing $\widetilde{d}'_{\mathrm{W}}$ is difficult but computing $\overline{d}'_{\mathrm{W}}$ and estimating ɾ_m are simple. However, the resulting power may deteriorate if the estimator of ɾ_m is inefficient or inaccurate. This type of test statistics is also mentioned in Li *et al.* (1991a). Indeed, there are infinitely many asymptotically equivalent test statistics, e.g., any convex combination of $\widetilde{D}_{\mathrm{W}}(\text{ɾ}_m)$ and $\overline{D}_{\mathrm{W}}(\text{ɾ}_m)$, i.e., $\phi\widetilde{D}_{\mathrm{W}}(\text{ɾ}_m) + (1-\phi)\overline{D}_{\mathrm{W}}(\text{ɾ}_m)$, for $\phi \in [0,1]$. When $\widetilde{r}_{\mathrm{W},1}$ or $\widetilde{r}_{\mathrm{W},1/2}$ is used for estimating ɾ_m, the null distributions of the resulting MI test statistics can be approximated by $F_{k,\widetilde{\mathrm{df}}'(\text{ɾ}_m,k)}$, where $\widetilde{\mathrm{df}}'(\text{ɾ}_m,k) = (m-1)(1+\text{ɾ}_m^{-1})^2 k^{-3/m}$; see Li *et al.* (1991a).

Next, we introduce and recall some notation to facilitate the comparison: (a) standard complete-data moments estimation ($\mathscr{M}_{\mathrm{W}}$ and $\mathscr{M}_{\mathrm{L}}$) and testing ($\mathscr{D}_{\mathrm{W}}$ and $\mathscr{D}_{\mathrm{L}}$) procedures, and (b) non-standard complete-data procedures ($\widetilde{\mathscr{D}}_{\mathrm{L}}$, $\overline{\mathscr{D}}_{\mathrm{L}}$, $\mathscr{D}_{\mathrm{L},1}$ and $\overline{\mathscr{D}}_{\mathrm{L},1}$), where

$$\mathscr{M}_{\mathrm{W}}(X) = \left\{\widehat{\theta}(X), U(X)\right\}, \quad \mathscr{M}_{\mathrm{L}}(X) = \left\{\widehat{\psi}(X), \widehat{\psi}_0(X)\right\},$$
$$\mathscr{D}_{\mathrm{W}}(X) = d_{\mathrm{W}}(\widehat{\theta}(X), U(X)), \quad \overline{\mathscr{D}}_{\mathrm{L},1}(\mathbb{X}) = \frac{2}{m}\sum_{\ell=1}^{m} \log f(X^{\ell} \mid \widehat{\psi}^*(\mathbb{X})).$$

Clearly, $\mathscr{M}_{\mathrm{W}}$ produces $\mathscr{D}_{\mathrm{W}}$, and $\{\mathscr{M}_{\mathrm{L}}, \widetilde{\mathscr{D}}_{\mathrm{L}}\}$ produces $\mathscr{D}_{\mathrm{L}}$. If users can perform optimization, $\widetilde{\mathscr{D}}_{\mathrm{L}}$ produces $\{\mathscr{M}_{\mathrm{L}}, \mathscr{D}_{\mathrm{L}}, \mathscr{D}_{\mathrm{L},1}\}$. Note that an un-normalized density can be used in $\mathscr{D}_{\mathrm{L},1}$ and $\overline{\mathscr{D}}_{\mathrm{L},1}$.

Table 2 summarizes whether a particular pair of $D(\cdot)$ and r, resulting the statistic $D(r)$, has the following statistical or computational properties.

- (Inv) $D(r)$ and r are invariant to re-parametrization of ψ;
- (Rob) r is robust against θ_0, i.e., consistent under both H_0 and H_1;
- ($\geqslant 0$) $D(r)$ and r are non-negative for all m and n;
- (Pow) the test has high power to reject H_0 under H_1;
- (Def) $D(r)$ and r are always well-defined and numerically well-conditioned;
- (Sca) the MI procedure requires users only to deal with scalars;
- (Dep) $X_1, \ldots, X_n$ can be dependent; and
- (EFMI) whether EFMI is assumed for θ or for ψ.

In summary, $\widehat{D}_{\mathrm{S}}(\widehat{r}_{\mathrm{S}}^{+})$ is the most computationally attractive test statistic. If the user is willing to make stronger assumptions, $\widehat{D}_{\mathrm{S}}(\widehat{r}_{\mathrm{S}}^{\diamondsuit})$ has better statistical properties, and is still computationally feasible. Nevertheless, $\widehat{D}_{\mathrm{L}}(\widehat{r}_{\mathrm{L}}^{\diamondsuit})$ is the most general test statistic and has the best statistical properties.

TABLE 2

Computational requirements and statistical properties of MI test statistics, their associated combining rules and estimators of FMI r_m. The symbols "+" and "−" mean that the test statistic (or estimator) is equipped and not equipped with the indicated property, respectively; see the end of § 3.2 for heading descriptions. The reference papers/book are abbreviated as follows: Rubin (2004) (R04), Li et al. (1991a) (LMRR91) and Meng and Rubin (1992) (MR92).

		Combining Rule		Estimator of r_m		Approx. null distribution[a]			Properties								
Test	No.	Formula	Routine	Formula	Routine	Original	Proposed	Reference	Inv	Rob	$\geqslant 0$	Pow	Def	Sca	Dep	EFMI
WT	W-1	$\widetilde{D}'_{\text{W}}(\text{r}_m)$	$\mathscr{M}_{\text{W}}$	$\widetilde{\text{r}}'_{\text{W}}$	$\mathscr{M}_{\text{W}}$	$F_{k,\widetilde{\text{df}}(\text{r}_m,k)}$[b]	$F_{k,\widehat{\text{df}}(\text{r}_m,k)}$	R04	−	+[c]	+	−	−	−	+	θ
	W-2	$\widetilde{D}'_{\text{W}}(\text{r}_m)$[d]	$\mathscr{M}_{\text{W}}$	$\widetilde{\text{r}}'_{\text{W},1}$	$\mathscr{D}_{\text{W}}$	$F_{k,\widetilde{\text{df}}'(\text{r}_m,k)}$	NA	R04	−	−	+	−	−	−	+	θ
	W-3	$\widetilde{D}'_{\text{W}}(\text{r}_m)$	$\mathscr{M}_{\text{W}}$	$\widetilde{\text{r}}'_{\text{W},1/2}$	$\mathscr{D}_{\text{W}}$	$F_{k,\widetilde{\text{df}}'(\text{r}_m,k)}$	NA	LMRR91	−	−	+	−	−	−	+	θ
	W-4	$D_{\text{W}}(T)$[e]	$\mathscr{M}_{\text{W}}$	$\widetilde{\text{r}}'_{\text{W}}$	$\mathscr{D}_{\text{W}}$	$F_{k,\widetilde{\text{df}}(\text{r}_m,k)}$	$F_{k,\widehat{\text{df}}(\text{r}_m,k)}$	R04	−	+	+	−	−	−	+	θ[f]
	W-5	$\overline{D}'_{\text{W}}(\text{r}_m)$	$\mathscr{D}_{\text{W}}$	$\widetilde{\text{r}}'_{\text{W},1}$	$\mathscr{D}_{\text{W}}$	$F_{k,\widetilde{\text{df}}'(\text{r}_m,k)}$	NA	R04	−	−	−	−	−	+	+	θ
	W-6	$\overline{D}'_{\text{W}}(\text{r}_m)$	$\mathscr{D}_{\text{W}}$	$\widetilde{\text{r}}'_{\text{W},1/2}$	$\mathscr{D}_{\text{W}}$	$F_{k,\widetilde{\text{df}}'(\text{r}_m,k)}$	NA	LMRR91	−	−	−	−	−	+	+	θ
LRT	L-1	$\widetilde{D}_{\text{L}}(\text{r}_m)$	$\mathscr{M}_{\text{L}}, \widetilde{\mathscr{D}}_{\text{L}}$	$\widetilde{\text{r}}_{\text{L}}$	$\mathscr{M}_{\text{L}}, \widetilde{\mathscr{D}}_{\text{L}}$	$F_{k,\widetilde{\text{df}}(\text{r}_m,k)}$	$F_{k,\widehat{\text{df}}(\text{r}_m,k)}$	MR92	−	−	−	−	+	−[g]	+	θ
	L-2	$\widetilde{D}^{+}_{\text{L}}(\text{r}_m)$	$\mathscr{M}_{\text{L}}, \widetilde{\mathscr{D}}_{\text{L}}$	$\widetilde{\text{r}}^{+}_{\text{L}}$	$\mathscr{M}_{\text{L}}, \widetilde{\mathscr{D}}_{\text{L}}$	$F_{k,\widetilde{\text{df}}(\text{r}_m,k)}$	$F_{k,\widehat{\text{df}}(\text{r}_m,k)}$	MR92[h]	−	−	+	−	+	−	+	θ
	L-3	$\widehat{D}_{\text{S}}(\text{r}_m)$	$\mathscr{D}_{\text{L}}$	$\widehat{\text{r}}_{\text{S}}$	$\mathscr{D}_{\text{L}}$	$F_{k,\widetilde{\text{df}}(\text{r}_m,k)}$	$F_{k,\widehat{\text{df}}(\text{r}_m,k)}$	Proposal	+	−	−	−	+	+	−	θ
	L-4	$\widehat{D}_{\text{S}}(\text{r}_m)$	$\mathscr{D}_{\text{L}}$	$\widehat{\text{r}}^{+}_{\text{S}}$	$\mathscr{D}_{\text{L}}$	$F_{k,\widetilde{\text{df}}(\text{r}_m,k)}$	$F_{k,\widehat{\text{df}}(\text{r}_m,k)}$	Proposal	+	−	+	−	+	+	−	θ
	L-5	$\widehat{D}_{\text{S}}(\text{r}_m)$	$\mathscr{D}_{\text{L}}$	$\widehat{\text{r}}^{\diamond}_{\text{S}}$	$\mathscr{D}_{\text{L},1}$	$F_{k,\widetilde{\text{df}}(\text{r}_m,h)}$	$F_{k,\widehat{\text{df}}(\text{r}_m,h)}$	Proposal	+	+	+	+	+	+	−	ψ
	L-6[i]	$\widehat{D}_{\text{L}}(\text{r}_m)$	$\overline{\mathscr{D}}_{\text{L}}$	$\widehat{\text{r}}^{+}_{\text{L}}$	$\overline{\mathscr{D}}_{\text{L}}$	$F_{k,\widetilde{\text{df}}(\text{r}_m,k)}$	$F_{k,\widehat{\text{df}}(\text{r}_m,k)}$	Proposal	+	−	+	−	+	+	+	θ
	L-7[j]	$\widehat{D}_{\text{L}}(\text{r}_m)$	$\overline{\mathscr{D}}_{\text{L}}$	$\widehat{\text{r}}^{\diamond}_{\text{L}}$	$\overline{\mathscr{D}}_{\text{L},1}$	$F_{k,\widetilde{\text{df}}(\text{r}_m,h)}$	$F_{k,\widehat{\text{df}}(\text{r}_m,h)}$	Proposal	+	+	+	+	+	+	+	ψ

[a] In actual computation, the r_m in the denominator degree of freedom of F is replaced by its corresponding estimator given in the previous column.

[b] The original approximate null distribution documented in Rubin (2004) was modified by Li *et al.* (1991a). This footnote also applies to W-2,4,5.

[c] The estimator $\widetilde{\text{r}}'_{\text{W}}$ does not depend on θ_0, but its MSE may be inflated under H_1 if a bad parametrization of θ is used.

[d] The originally proposed combining rule is $\overline{D}'_{\text{W}}(\text{r}_m)$; see (3.10). Although $\overline{D}'_{\text{W}}(\text{r}_m)$ is more computational feasible, the power loss is more significant than $\widetilde{D}'_{\text{W}}(\text{r}_m)$ after inserting an inefficient estimator $\widetilde{\text{r}}'_{\text{W},1}$ for r_m. This footnote also applies to W-3.

[e] Computing the test statistic $D_{\text{W}}(T) = d_{\text{W}}(\bar{\theta}, T)/k$ does not require estimating r_m.

[f] EFMI is not required for the test statistic $D_{\text{W}}(T)$, but it is required for its approximate null distribution.

[g] Averaging and processing vector estimators of ψ, but not their covariance matrixes, is needed. This footnote also applies to L-2.

[h] It is a trivial modification of the original proposal in MR92 by replacing $\widetilde{\text{r}}_{\text{L}}$ with $\widetilde{\text{r}}^{+}_{\text{L}} = \max\{0, \widetilde{\text{r}}_{\text{L}}\}$.

[i] L-6 is equivalent to L-4 when the rows of X are independent.

[j] L-7 is equivalent to L-5 when the rows of X are independent.

TABLE 3
The values of parameters used in the simulation experiment in § 4.1.

Experiment		Parameters							
		Fixed			Variable				
No.	Varying	ρ	p	ν	Case 1	Case 2	Case 3	Case 4	Case 5
I	Correlation ρ	–	2	0.5	-0.8	-0.4	0	0.4	0.8
II	Dimension p	0.4	–	0.5	2	3	4	5	6
III	FMI ν	0.4	2	–	0.1	0.3	0.5	0.7	0.9

4. Empirical Investigation and Findings.

4.1. *Simulation Studies.* Suppose that $X_1, \ldots, X_n \sim \mathcal{N}_p(\mu, \Sigma)$ independently, where $\Sigma = \sigma^2\{(1-\rho)I_p + \rho \mathbf{1}_p \mathbf{1}_p^\intercal\}$, and $\mathbf{1}_p$ is the p-vector of ones. The values of p, σ^2, ρ and μ are specified below. Further assume that only $n_{\text{obs}} = \lfloor (1-\nu)n \rfloor$ data points are observed, where $\nu \in (0,1)$ is the FMI. Let $X_{\text{obs}} = \{X_i : i = 1, \ldots, n_{\text{obs}}\}$ and $X_{\text{mis}} = \{X_i : i = n_{\text{obs}}+1, \ldots, n\}$. Suppose that we want to test whether the means of all components are equal, i.e., $H_0 : \mu = \mu_0 \mathbf{1}_p$, where $\mu_0 \in \mathbb{R}$ is an unknown constant.

Obviously, one may directly use the observed dataset to construct the LRT statistic D_{L} without MI. The test D_{L} (denoted by L-0) is regarded as a benchmark for comparison. Throughout this subsection, W-1,2,3,4 and L-1,2,3,4,5 listed in Table 2 are compared. In the imputation step, a Bayesian model is employed for imputation. Assume a multivariate Jeffreys prior on (μ, Σ), i.e., $f(\mu, \Sigma) \propto |\Sigma|^{-(p+1)/2}$. Let $\overline{X}_{\text{obs}}$ and S_{obs} be the sample mean and sample covariance matrix based on X_{obs}. Then, the ℓth imputed missing dataset can be produced by the following procedure, for $\ell = 1, \ldots, m$.

1. Draw a posterior sample Σ^ℓ from the inverse-Wishart distribution with $(n_{\text{obs}} - 1)$ degrees of freedom and scale matrix S_{obs}^{-1}.
2. Draw one posterior sample μ^ℓ from $\mathcal{N}_p(\overline{X}_{\text{obs}}, \Sigma^\ell / n_{\text{obs}})$.
3. Draw $(n - n_{\text{obs}})$ imputed missing values $\{X_i^\ell : i = n_{\text{obs}}+1, \ldots, n\}$ from $\mathcal{N}_p(\mu^\ell, \Sigma^\ell)$ independently. Also, denote $X_i^\ell = X_i$ for $i = 1, \ldots, n_{\text{obs}}$.

With the ℓth completed dataset, the unconstrained MLEs for μ and Σ are

$$\widehat{\mu}^\ell = \frac{1}{n}\sum_{i=1}^{n} X_i^\ell, \qquad \widehat{\Sigma}^\ell = \frac{1}{n}\sum_{i=1}^{n} \left(X_i^\ell - \widehat{\mu}^\ell\right)\left(X_i^\ell - \widehat{\mu}^\ell\right)^\intercal.$$

Whereas we generate data using a covariance matrix with common variance and correlation, our model does not assume any structure for Σ. The only restriction we can impose is the common-mean assumption under the null,

for which the constrained MLEs are

$$\widehat{\mu}_0^\ell = \left\{ \frac{\mathbf{1}_p^\intercal (\widehat{\Sigma}^\ell)^{-1} \widehat{\mu}^\ell}{\mathbf{1}_p^\intercal (\widehat{\Sigma}^\ell)^{-1} \mathbf{1}_p} \right\} \mathbf{1}_p, \qquad \widehat{\Sigma}_0^\ell = \widehat{\Sigma}^\ell + \left(\widehat{\mu}^\ell - \widehat{\mu}_0^\ell\right) \left(\widehat{\mu}^\ell - \widehat{\mu}_0^\ell\right)^\intercal.$$

In the experiment, we study the impact of parametrization on different test statistics. For the Wald tests, three parametrizations of θ are examined:

(i) $\theta = (\mu_2 - \mu_1, \ldots, \mu_p - \mu_{p-1})^\intercal$ — differences of means,
(ii) $\theta = (\mu_2/\mu_1, \ldots, \mu_p/\mu_{p-1})^\intercal - \mathbf{1}_{p-1}$ — relative differences of means, and
(iii) $\theta = \left(\mu_2^3 - \mu_1^3, \ldots, \mu_p^3 - \mu_{p-1}^3\right)^\intercal$ — differences of cubic means.

For any case above, H_0 can be re-expressed as $\theta_0 = \mathbf{0}_{p-1}$, an $(p-1)$-vector of zeros. For LRTs, the following parametrizations of ψ are used:

(i) $\psi = \{\mu; \Sigma\}$ — means and covariances,
(ii) $\psi = \{\sqrt{\sigma_{ii}}/\mu_i, 1 \leqslant i \leqslant p; \Sigma\}$ — noise-to-signal and covariances, and
(iii) $\psi = \left\{\mu^\intercal \Sigma^{-1/2}; \Sigma^{-1}\right\}$ — standardized means and precisions,

where $\Sigma = (\sigma_{ij})$ and $\Sigma^{1/2}$ is the symmetric square root of Σ. The dimension of ψ is $h = (p^2 + 3p)/2$.

In the first part of the experiment, we study the distribution of p-values derived from each test under H_0. In particular, we use $n = 100$, $m = 3$, $\sigma^2 = 5$ and $\mu = \mathbf{1}_p$, with various values of ρ, p and ν specified in Table 3. All simulations are repeated 2^{12} times. The comparison under parametrization (ii) is shown in Figure 4; whereas those under parametrizations (i) and (iii) are deferred to Appendix C. Note that, for Wald tests under parametrization (ii), the matrix U^ℓ is singular in less than 0.25% of the replications, and those cases are removed from the analysis (which should favor the Wald tests).

The empirical sizes (i.e., type-I errors) of the MI Wald tests generally deviate from the nominal size α under parametrization (ii). In contrast, the sizes of all LRTs are closer to α. However, the original L-1 and its trivial modification L-2 do not have accurate sizes when $|\rho|$ or ν is large. They can be over-sized or under-sized depending on which parametrization is used. Moreover, the trivial modification L-2 does not help to correct the size, and it may even worsen the test. For our test statistics L-3 and L-4, they are invariant to parametrizations and have quite accurate sizes, although they are under-sized in challenging cases where both p and ν are large. Moreover, they are identical throughout our simulation experiments, i.e., we never observed $\widehat{r}_{\mathrm{L}} < 0$. For our recommended statistic L-5, it gives the most satisfactory overall results. It generally has very accurate size, except that it is slightly over-sized for large p, a problem that should diminish when we use m beyond the smallest recommended $m = 3$.

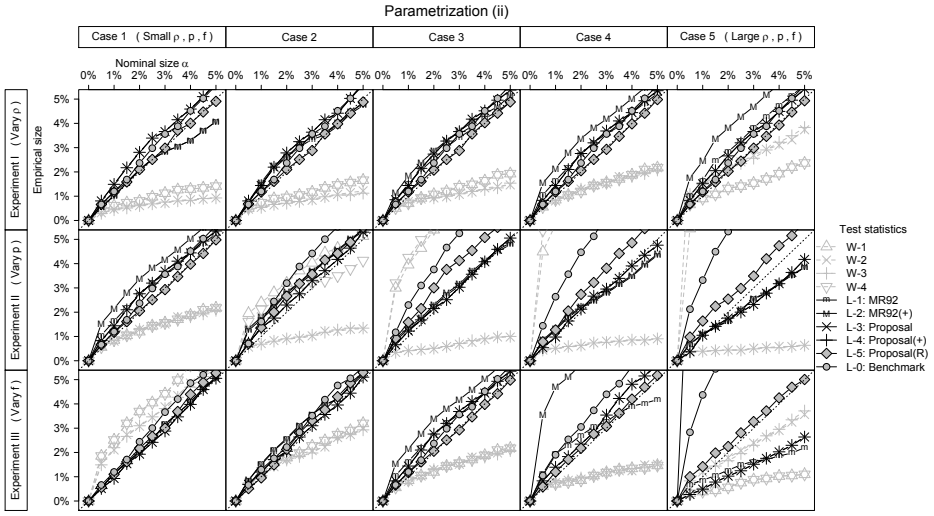

FIG 4. *The comparison between empirical size and nominal size α under parametrization (ii) for $\alpha \in (0, 5\%]$. The Wald tests (W-1,2,3,4) and LRTs (L-0,1,2,3,4,5) are represented by grey dashed and black solid lines, respectively. The LRT statistic $\widetilde{D}_{\mathrm{L}}$ (L-1: MR92) and its modification $\widetilde{D}_{\mathrm{L}}^{+}$ (L-2: MR92(+)) are the tests that greatly improved upon by our proposals $\widehat{D}_{\mathrm{S}}$ (L-3: Proposal), $\widehat{D}_{\mathrm{S}}^{+}$ (L-4: Proposal(+)) and $\widehat{D}_{\mathrm{S}}^{\Diamond}$ (L-5: Proposal(R)).*

Interestingly, as seen clearly in Figure 4, the benchmark L-0 performs very badly for large p and ν. The sample size per parameter, n/h, is small; for $p \geqslant 4$, $n/h \leqslant 100/14 < 8$. The asymptotic null distribution χ^2_k/k then can fail badly under arbitrary or even all parametrizations; (ii) apparently falls into this category. An F approximation would be more appropriate. But this is exactly what is being used for MI tests, albeit with different choices of the denominator degrees of freedom. Table 4 documents how often $\widetilde{r}_{\mathrm{L}}$, $\widetilde{D}_{\mathrm{L}}$ and $\widehat{r}_{\mathrm{S}}$ are negative. In some cases, nearly half of the simulated values of $\widetilde{r}_{\mathrm{L}}$ and $\widetilde{D}_{\mathrm{L}}$ are negative. In contrast, $\widehat{r}_{\mathrm{S}}$ is always non-negative in our simulation, despite the fact that it can be negative in theory.

To study the power of each test, we set $\nu = 0.5$, $p = 2$, $\rho = 0.8$, $\sigma^2 = 5$ and $\mu = (-2 + \delta, -2 + 2\delta)^{\intercal}$ for different values of $m \in \{3, 10, 30\}$, $n \in \{100, 400, 1600\}$ and $\delta = \mu_2 - \mu_1 \in [0, 4]$. The empirical power functions for size 0.5% tests under parametrizations (i), (ii) and (iii) are plotted in Figure 5. The results for size 5% tests are deferred to Table 12 of the Appendix. Generally, none of the Wald tests exhibits monotonically increasing power as δ increases, and their performance is affected significantly by parametrization. In particular, the powers can be as low as zero when $1 \lesssim \delta \lesssim 2$ under

TABLE 4

The empirical proportions of negative $\widetilde{r}_{\mathrm{L}}$ and $\widetilde{D}_{\mathrm{L}}$. The results under parametrizations (ii) and (iii) are shown. For parametrization (i), $\widetilde{r}_{\mathrm{L}} \geqslant 0$ and $\widetilde{D}_{\mathrm{L}} \geqslant 0$ in the experiments.

		Case									
		1	2	3	4	5	1	2	3	4	5
Experiment	**Parametrization**	% of $\widetilde{r}_{\mathrm{L}} < 0$					% of $\widetilde{D}_{\mathrm{L}} < 0$				
I	(ii)	1	2	3	4	5	26	16	13	12	12
	(iii)	6	6	7	7	7	1	1	1	1	2
II	(ii)	4	1	0	0	0	12	5	3	4	3
	(iii)	7	3	1	1	1	1	0	0	0	0
III	(ii)	13	6	4	4	3	55	25	12	5	2
	(iii)	18	9	7	5	4	20	5	1	1	0

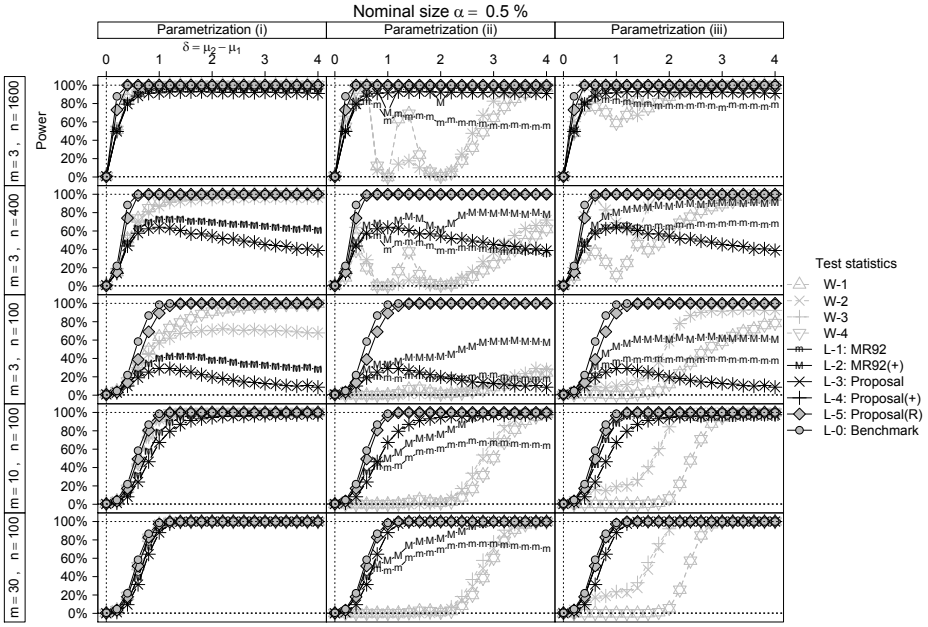

FIG 5. *The power curves under different parametrizations. The nominal size is $\alpha = 0.5\%$. In each plot, the vertical axis denotes the power, whereas the horizontal axis denotes the value of $\delta = \mu_2 - \mu_1$. The legend in Figure 4 also applies here.*

parametrizations (ii) and (iii). Under parametrization (ii), L-1 is not powerful even for large δ. Moreover, its trivial modifications L-2 cannot retrieve all the power it should have. On the other hand, our first proposed test

TABLE 5

The range of empirical size $[\min \widehat{\alpha}, \max \widehat{\alpha}]$ *in percentage, where* max *and* min *are taken over the three parametrizations. Only one value is recorded for those tests that are invariant to parametrization. The nominal size is* $\alpha = 0.5\%$.

	Range of empirical size: $[\min \widehat{\alpha}, \max \widehat{\alpha}]/\%$				
(n, m)	$(1600, 3)$	$(400, 3)$	$(100, 3)$	$(100, 10)$	$(100, 30)$
W-1	[0.90, 1.05]	[0.76, 1.05]	[0.20, 1.22]	[0.07, 0.56]	[0.02, 0.49]
W-2	[0.90, 1.05]	[0.98, 1.22]	[0.93, 1.25]	[0.32, 0.73]	[0.20, 0.85]
W-3	[0.98, 1.05]	[0.98, 1.25]	[0.90, 1.29]	[0.34, 0.71]	[0.22, 0.73]
W-4	[0.90, 1.05]	[0.76, 1.05]	[0.20, 1.22]	[0.07, 0.56]	[0.02, 0.49]
L-1	[0.90, 1.03]	[1.10, 1.64]	[1.15, 1.49]	[0.37, 1.05]	[0.10, 0.46]
L-2	[0.90, 1.05]	[1.10, 1.76]	[1.15, 2.37]	[0.37, 0.98]	[0.10, 0.49]
L-3	0.90	1.10	0.83	0.24	0.07
L-4	0.90	1.10	0.83	0.24	0.07
L-5	0.46	0.44	0.68	0.46	0.42
L-0	0.39	0.66	0.66	0.66	0.66

statistics L-3 and L-4 perform better than L-1 and L-2 at least for large m, however, they also lose a significant amount of power when m is small.

Compared with all these, our recommended test statistic L-5 performs extremely well for all m and n, with power very close to the benchmark L-0 even for small m. To ensure the comparisons of power are fair, we also investigate the empirical (actual) size, $\widehat{\alpha}$, in comparison to the nominal type-I error α. Table 5 shows the minimum and maximum of the empirical sizes over the three parametrizations considered in each test — and only one value is needed for those tests that are invariant to parametrization — when the nominal size $\alpha = 0.5\%$. We see the deviations from the nominal α can be noticeable, especially when $m = 3$. To take that into account, we report the empirical size adjusted power, that is, $O = \text{power}/\widehat{\alpha}$, which also has the interpretation as (an approximated) posterior odds of H_1 to H_0 (Bayarri *et al.*, 2016). Figure 6 plots the result for size 0.5% tests. Compared with the benchmark L-0, the odds O of the proposed robust MI test (L-5) is closer to the nominal value $1/\alpha$ as $\delta \to \infty$. Nevertheless, the finite sample performances of all size 0.5% tests are less satisfactory than those for size 5% tests (given in Appendix) because a very large sample size n is required in order to approximate the tail behavior of test statistics satisfactorily.

We also compare the performance of estimators of $\curlyvee_m$ for different δ and parametrizations. In our experiment, we have $\curlyvee_m = 1 + 1/m$ because we have set $\curlyvee = 1$. The MSEs of estimators $\widehat{f} = \widehat{r}/(1 + \widehat{r})$ of $\curlywedge_m = \curlyvee_m/(1 + \curlyvee_m)$ are

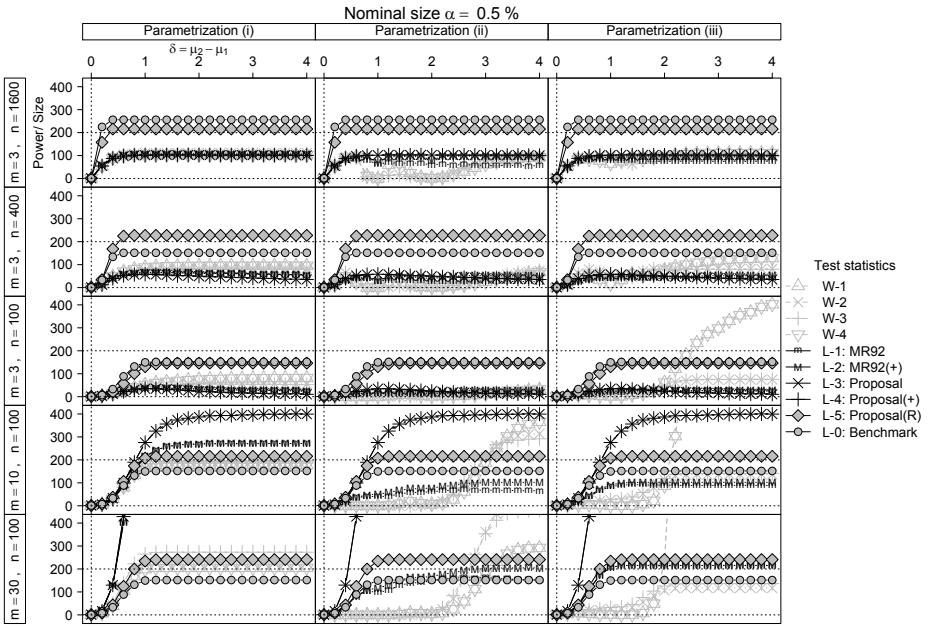

FIG 6. *The ratios of empirical power to empirical size under different parametrizations. The nominal size is* $\alpha = 0.5\%$. *In each plot, the vertical axis denotes the ratio, whereas the horizontal axis denotes* $\delta = \mu_2 - \mu_1$. *The legend in Figure 4 also applies here.*

shown in Figure 7, in log scale. Clearly, the only estimator that is consistent, invariant to parametrization and robust against δ is our proposal $\widehat{f}_{\mathrm{L}}^{\Diamond} = \widehat{r}_{\mathrm{L}}^{\Diamond}/(1+\widehat{r}_{\mathrm{L}}^{\Diamond})$. It concentrates at the true value $\not\ell_m$ quite closely even for small m and n. Since $\widehat{f}_{\mathrm{L}}^{\Diamond}$ is the only reliable estimator of $\not\ell_m$, it verifies why L-5 has the greatest power. On the other hand, the estimator $\widetilde{f}_{\mathrm{L}} = \widetilde{r}_{\mathrm{L}}/(1+\widetilde{r}_{\mathrm{L}})$ has very large MSE when $\delta \neq 0$. It also explains why L-1 is not powerful.

4.2. *Monte Carlo Experiments Without EFMI.* To check how robust various tests are to the assumption of EFMI, we simulate $X_i = (X_{i1}, \ldots, X_{ip})^{\intercal} \overset{\text{iid}}{\sim} \mathcal{N}_p(\mu, \Sigma)$ for $i = 1, \ldots, n$. Let R_{ij} be defined by $R_{ij} = 1$ if X_{ij} is observed, otherwise $R_{ij} = 0$. 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:

$$\mathsf{P}\left(R_{ij} = 0 \mid R_{i,j-1} = a\right) = \begin{cases} [1 + \exp(\alpha_0 + \alpha_1 X_{i,j-1})]^{-1} & \text{if } a = 1; \\ 1 & \text{if } a = 0, \end{cases}$$

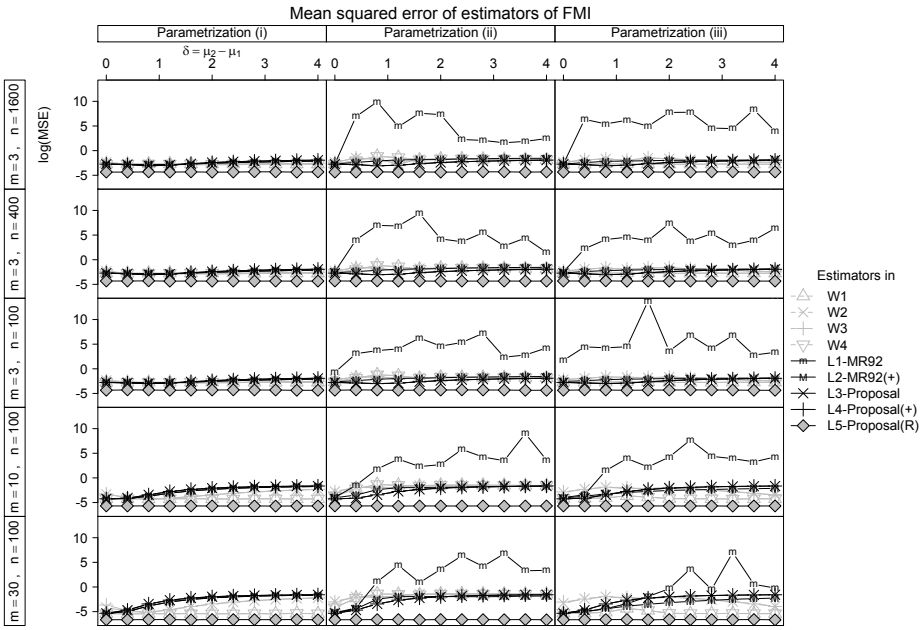

FIG 7. *The MSEs of estimators of ℓ_m used in the test statistics. In each plot, the vertical axis denotes the log of MSE, whereas the horizontal axis denotes the value of $\delta = \mu_2 - \mu_1$. The legend in Figure 4 also applies here.*

for $j = 2, \ldots, p$, where $\alpha_0, \alpha_1 \in \mathbb{R}$. If $\alpha_1 = 0$, then the data are missing completely at random (MCAR); otherwise they are missing at random (MAR), as defined in Rubin (1976). Let $n_j = \sum_{i=1}^{n} R_{ij}$ be the number of observed jth component. Without loss of generality, assume X_{obs} is arranged in such a way that $R_{ij} \geqslant R_{i'j}$ for all $i < i'$ and j.

To impute the missing data, it is useful to represent X_i by

$$\begin{aligned} \left[X_{i1} \mid \beta_1, \tau_1^2\right] &\sim \mathcal{N}(\beta_1, \tau_1^2), \\ \left[X_{ij} \mid X_{i,1:(j-1)}, \beta_j, \tau_j^2\right] &\sim \mathcal{N}(\beta_j^\intercal Z_{ij}, \tau_j^2), \qquad j = 2, \ldots, p, \end{aligned}$$

where $\tau_1^2, \ldots, \tau_p^2 \in \mathbb{R}^+$, $\beta_j \in \mathbb{R}^j$, $X_{i,1:(j-1)} = (X_{i1}, \ldots, X_{i,j-1})^\intercal$ and $Z_{ij} = (1, X_{i,1:(j-1)}^\intercal)^\intercal$ for $j \geqslant 2$. Denote the (complete-case) least squares estimators of β_j and τ_j^2 by $\widehat{\beta}_j = (Z_j^\intercal Z_j)^{-1} Z_j^\intercal W_j$ and $\widehat{\tau}_j^2 = \frac{1}{n_j - j}(W_j - Z_j\widehat{\beta}_j)^\intercal(W_j - Z_j\widehat{\beta}_j)$, where $Z_j = (Z_{1j}, \ldots, Z_{n_j j})^\intercal$ and $W_j = (X_{1j}, \ldots, X_{n_j j})^\intercal$.

To perform MI, we assume a Bayesian model with the non-informative prior $f(\beta_1, \ldots, \beta_p, \tau_1^2, \ldots, \tau_p^2) \propto 1/(\tau_1^2 \cdots \tau_p^2)$. For each $\ell = 1, \ldots, m$, the ℓth imputed dataset X^ℓ, whose (i,j)th element is X_{ij}^ℓ, is produced as follows.

1. Let $X_{ij}^{\ell} = X_{ij}$ for all $1 \leqslant j \leqslant p$ and $i \leqslant n_j$.
2. For each $j = 2, \ldots, p$, repeat Step 3 to Step 5.
3. Draw a sample $(\tau_j^{\ell})^2$ from $\widehat{\tau}_j^2(n_j - j)/\chi_{n_j-j}^2$.
4. Draw a sample β_j^{ℓ} from $\mathcal{N}_j(\widehat{\beta}_j, (\tau_j^{\ell})^2(Z_j^{\intercal}Z_j)^{-1})$.
5. Draw a sample X_{ij}^{ℓ} from $\mathcal{N}((\beta_j^{\ell})^{\intercal}Z_{ij}^{\ell}, (\tau_j^{\ell})^2)$ for $i = n_j+1, \ldots, n$, where $Z_{ij}^{\ell} = (1, (X_{i,1:(j-1)}^{\ell})^{\intercal})^{\intercal}$.

We test $H_0 : \mu = \mathbf{0}_p$ against $H_1 : \mu \neq \mathbf{0}_p$. In the experiments, we set $\mu = \delta\mathbf{1}_p$, where $\delta \in [0, 0.6]$; the (i, j)th element of Σ to be $\Sigma_{ij} = 0.5^{|i-j|}$ for $i, j = 1, \ldots, p$; $n = 500$; $m \in \{3, 5\}$; $p = 5$; $(\alpha_0, \alpha_1) \in \{(2, -1), (1, 0)\}$. Our model treats Σ unknown, and hence $k = p$ and $h = (3p+p^2)/2$. With the ℓth imputed dataset, the H_0-constrained MLEs of μ and Σ are $\widehat{\mu}_0^{\ell} = \mathbf{0}_p$ and $\widehat{\Sigma}_0^{\ell} = (X^{\ell})^{\intercal}(X^{\ell})/n$; whereas the unconstrained counterparts are $\widehat{\mu}^{\ell} = \mathbf{1}_n^{\intercal}X^{\ell}/n$ and $\widehat{\Sigma}^{\ell} = (X^{\ell} - \widehat{\mu}^{\ell})^{\intercal}(X^{\ell} - \widehat{\mu}^{\ell})/n$. Under H_0 and MAR, the fractions of missing *observations* of the five variables are $(0, 16\%, 28\%, 38\%, 47\%)$, whereas the average fractions of missing *information*, i.e, the eigenvalues of $\mathscr{B}_{\theta}\mathscr{T}_{\theta}^{-1}$, are $(0, 19\%, 34\%, 45\%, 55\%)$. So, the assumption of EFMI does not hold.

We compare (L4) $\widehat{D}_{\mathrm{L}}^{+} \approx F_{k,\widehat{\mathrm{df}}(\widehat{\tau}_{\mathrm{L}}^{+},k)}$, (L5) $\widehat{D}_{\mathrm{L}}^{\Diamond} \approx F_{k,\widehat{\mathrm{df}}(\widehat{\tau}_{\mathrm{L}}^{\Diamond},h)}$, (C1) complete-data (asymptotic) LRT using $\{X_i : i = 1, \ldots, n\}$, and (C2) complete-case (asymptotic) LRT using $\{X_i : i = 1, \ldots, n_p\}$. The results are shown in Figure 8. The size of $\widehat{D}_{\mathrm{L}}^{\Diamond}$ is quite accurate when the nominal size is small. If the data are MCAR, complete-case test C2 is valid, however, with slightly less power. (Test C2 is typically invalid without MCAR.) In terms of power-to-size ratio, the performance of $\widehat{D}_{\mathrm{L}}^{\Diamond}$ is the best among the three implementable tests L4, L5 and C2. Its performance is comparable to the (unavailable) complete-data test C1. Note also that the power-to-size ratio of $\widehat{D}_{\mathrm{L}}^{+}$ and $\widehat{D}_{\mathrm{L}}^{\Diamond}$ become closer to the nominal value 1/0.5% when m increases. All these indicate that the performance of our proposed tests are acceptable despite of the serious violation of the EFMI assumption.

4.3. *Applications to a Care-Survival Data.* Meng and Rubin (1992) applied their test to the data given in Table 6, where i, j and k index, respectively, clinic (A or B), amount of parental care (more or less) and survival status (died or survived). However, the clinic label k is missing for some of the observations (and the missing-data mechanism was assumed to be ignorable). Two hypotheses were tested in Meng and Rubin (1992). The first is whether the clinic and parental care are conditionally independent given the survival status, and the second is whether all three variables are mutually independent. The MI datasets are generated from a Bayesian model in § 4.2 of Meng and Rubin (1992). Our aim here is to investigate the impact on the

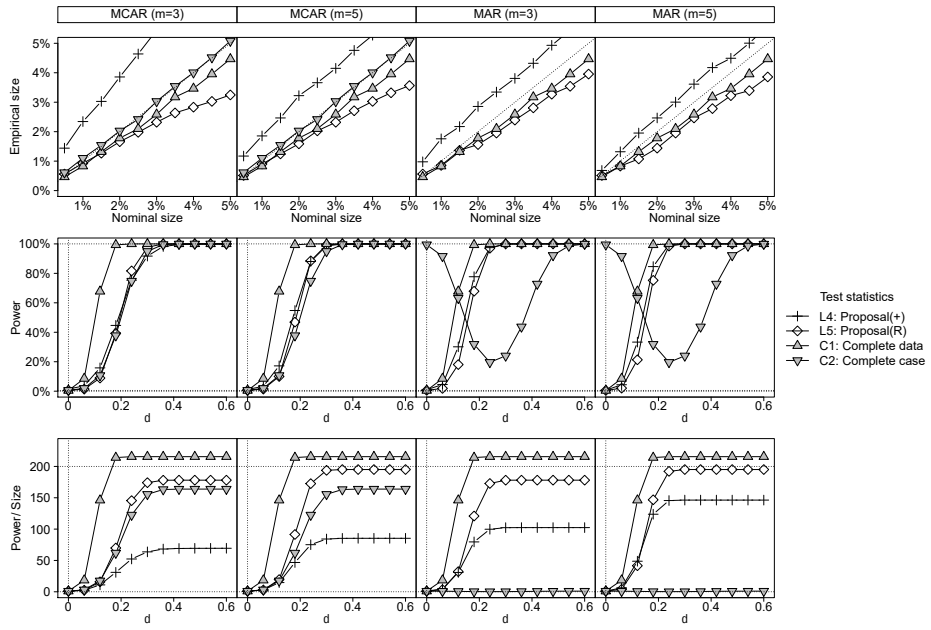

FIG 8. *The empirical size, empirical power, and their ratio. The first row of plots show the empirical sizes. The size of the complete-case test (C2) under MAR is off the chat (always equals to one) in the experiment because it is invalid. The second and third rows of plots show the powers and the power-to-size ratios, respectively, where the nominal size is* 0.5%.

test statistics $\widetilde{D}_{\mathrm{S}}$, $\widehat{D}_{\mathrm{S}}^{+}$ and $\widehat{D}_{\mathrm{S}}^{\Diamond}$ by different parametrizations of $\{\pi_{ijk}\}$; and the impact on the estimators $\widetilde{r}_{\mathrm{L}}$, $\widehat{r}_{\mathrm{S}}^{+}$ and $\widehat{r}_{\mathrm{S}}^{\Diamond}$ under different null hypotheses.

TABLE 6
Data from Meng and Rubin (1992). The notation "?" indicates missing label.

		Survival Status (j)	
Clinic (k)	**Parental care** (i)	Died	Survived
A	Less	3	176
	More	4	293
B	Less	17	197
	More	2	23
?	Less	10	150
	More	5	90

Specifically, the ℓth imputed log-likelihood function is $\log f(X^{\ell} \mid \pi) = \sum_c n_c^{\ell} \log \pi_c$, where X^{ℓ} are the cell counts n_c^{ℓ} in the ℓth imputed dataset.

Hence the unconstrained MLE of π_c is $\widehat{\pi}_c^{\ell} = n_c^{\ell}/n_+^{\ell}$, where $n_+^{\ell} = \sum_c n_c^{\ell}$. Consequently, the joint log-likelihood based on the stacked data is

$$\log f(X^{\mathrm{S}} \mid \pi) = \sum_{\ell=1}^{m} \sum_{c} n_c^{\ell} \log \pi_c = \sum_{c} n_c^{+} \log \pi_c, \tag{4.1}$$

where $n_c^{+} = \sum_{\ell=1}^{m} n_c^{\ell}$. Thus the unconstrained MLE with respect to (4.1) is $\widehat{\pi}_c^{\mathrm{S}} = n_c^{+}/n_+^{+}$, where $n_+^{+} = \sum_c n_c^{+}$. Similarly, we can find the constrained MLEs under a given null. We consider the following parametrizations:

(i) $\psi_{ijk} = \pi_{ijk}$ — the identity map,
(ii) $\psi_{ijk} = \log\{\pi_{ijk}/(1-\pi_{ijk})\}$ — the logit transformation, and
(iii) $\psi_{ij1} = \pi_{ij1}$ and $\psi_{ij2} = \pi_{ij2}/\pi_{ij1}$ — ratios of probabilities.

The p-values $\widetilde{p}_{\mathrm{L}}$, $\widehat{p}_{\mathrm{S}}^{+}$ and $\widehat{p}_{\mathrm{S}}^{\Diamond}$ of the tests $\widetilde{D}_{\mathrm{L}}$, $\widehat{D}_{\mathrm{S}}^{+}$ and $\widehat{D}_{\mathrm{S}}^{\Diamond}$ and the associated estimates of $\varkappa_m$, i.e., $\widetilde{r}_{\mathrm{L}}$, $\widehat{r}_{\mathrm{S}}^{+}$ and $\widehat{r}_{\mathrm{S}}^{\Diamond}$, are shown in Table 7. A more detailed comparison is deferred to Table 9 of the Appendix.

TABLE 7
The LRTs using $\widetilde{D}_{\mathrm{L}}$, $\widehat{D}_{\mathrm{S}}^{+}$ and $\widehat{D}_{\mathrm{S}}^{\Diamond}$ under different parametrizations in § 4.3.

	Parametrization (i)					
	H_0: Conditional independence			H_0: Full independence		
m	$\widetilde{r}_{\mathrm{L}}, \widehat{r}_{\mathrm{S}}^{+}, \widehat{r}_{\mathrm{S}}^{\Diamond}$	$\widetilde{D}_{\mathrm{L}}, \widehat{D}_{\mathrm{S}}^{+}, \widehat{D}_{\mathrm{S}}^{\Diamond}$	$\widetilde{p}_{\mathrm{L}}, \widehat{p}_{\mathrm{S}}^{+}, \widehat{p}_{\mathrm{S}}^{\Diamond}$	$\widetilde{r}_{\mathrm{L}}, \widehat{r}_{\mathrm{S}}^{+}, \widehat{r}_{\mathrm{S}}^{\Diamond}$	$\widetilde{D}_{\mathrm{L}}, \widehat{D}_{\mathrm{S}}^{+}, \widehat{D}_{\mathrm{S}}^{\Diamond}$	$\widetilde{p}_{\mathrm{L}}, \widehat{p}_{\mathrm{S}}^{+}, \widehat{p}_{\mathrm{S}}^{\Diamond}$
3	0.54, 0.54, 0.38	0.08, 0.08, 0.09	0.93, 0.93, 0.92	0.31, 0.31, 0.38	54.2, 54.2, 51.4	0, 0, 0
10	0.50, 0.50, 0.70	0.14, 0.14, 0.12	0.87, 0.87, 0.88	0.56, 0.56, 0.70	45.4, 45.4, 41.7	0, 0, 0
50	0.31, 0.31, 0.45	0.11, 0.11, 0.10	0.90, 0.90, 0.91	0.33, 0.33, 0.45	51.5, 51.5, 47.3	0, 0, 0
	Parametrization (ii)					
	H_0: Conditional independence			H_0: Full independence		
m	$\widetilde{r}_{\mathrm{L}}, \widehat{r}_{\mathrm{S}}^{+}, \widehat{r}_{\mathrm{S}}^{\Diamond}$	$\widetilde{D}_{\mathrm{L}}, \widehat{D}_{\mathrm{S}}^{+}, \widehat{D}_{\mathrm{S}}^{\Diamond}$	$\widetilde{p}_{\mathrm{L}}, \widehat{p}_{\mathrm{S}}^{+}, \widehat{p}_{\mathrm{S}}^{\Diamond}$	$\widetilde{r}_{\mathrm{L}}, \widehat{r}_{\mathrm{S}}^{+}, \widehat{r}_{\mathrm{S}}^{\Diamond}$	$\widetilde{D}_{\mathrm{L}}, \widehat{D}_{\mathrm{S}}^{+}, \widehat{D}_{\mathrm{S}}^{\Diamond}$	$\widetilde{p}_{\mathrm{L}}, \widehat{p}_{\mathrm{S}}^{+}, \widehat{p}_{\mathrm{S}}^{\Diamond}$
3	1.08, 0.54, 0.38	−0.07, 0.08, 0.09	1.00, 0.93, 0.92	0.61, 0.31, 0.38	43.9, 54.2, 51.4	0, 0, 0
10	0.99, 0.50, 0.70	−0.10, 0.14, 0.12	1.00, 0.87, 0.88	1.09, 0.56, 0.70	33.7, 45.4, 41.7	0, 0, 0
50	0.63, 0.31, 0.45	−0.10, 0.11, 0.10	1.00, 0.90, 0.91	0.65, 0.33, 0.45	41.3, 51.5, 47.3	0, 0, 0
	Parametrization (iii)					
	H_0: Conditional independence			H_0: Full independence		
m	$\widetilde{r}_{\mathrm{L}}, \widehat{r}_{\mathrm{S}}^{+}, \widehat{r}_{\mathrm{S}}^{\Diamond}$	$\widetilde{D}_{\mathrm{L}}, \widehat{D}_{\mathrm{S}}^{+}, \widehat{D}_{\mathrm{S}}^{\Diamond}$	$\widetilde{p}_{\mathrm{L}}, \widehat{p}_{\mathrm{S}}^{+}, \widehat{p}_{\mathrm{S}}^{\Diamond}$	$\widetilde{r}_{\mathrm{L}}, \widehat{r}_{\mathrm{S}}^{+}, \widehat{r}_{\mathrm{S}}^{\Diamond}$	$\widetilde{D}_{\mathrm{L}}, \widehat{D}_{\mathrm{S}}^{+}, \widehat{D}_{\mathrm{S}}^{\Diamond}$	$\widetilde{p}_{\mathrm{L}}, \widehat{p}_{\mathrm{S}}^{+}, \widehat{p}_{\mathrm{S}}^{\Diamond}$
3	−2.35, 0.54, 0.38	−1.16, 0.08, 0.09	1.00, 0.93, 0.92	−1.22, 0.31, 0.38	−321, 54.2, 51.4	1, 0, 0
10	−2.04, 0.50, 0.70	−2.20, 0.14, 0.12	1.00, 0.87, 0.88	−1.85, 0.56, 0.70	−86, 45.4, 41.7	1, 0, 0
50	−1.22, 0.31, 0.45	−7.39, 0.11, 0.10	1.00, 0.90, 0.91	−1.06, 0.33, 0.45	−1136, 51.5, 47.3	1, 0, 0

The simulation outputs demonstrate that $\widehat{D}_{\mathrm{S}}^{+}$ and $\widehat{D}_{\mathrm{S}}^{\Diamond}$ are invariant to parametrizations, whereas $\widetilde{D}_{\mathrm{L}}$ is not. Moreover, the impact on $\widetilde{D}_{\mathrm{L}}$ is large

under the parametrization (iii). In particular, the value of $\widetilde{r}_{\mathrm{L}}$ is inflated; and some of the values of $\widetilde{r}_{\mathrm{L}}$ and $\widetilde{D}_{\mathrm{L}}$ are negative, leading to the meaningless $\widetilde{p}_{\mathrm{L}} = 1$, especially under parametrization (iii). In contrast, $\widehat{r}_{\mathrm{S}} \geqslant 0$ for all cases in this example (and hence $\widehat{r}_{\mathrm{S}}^{+} = \widehat{r}_{\mathrm{S}}$). In addition, $\widehat{D}_{\mathrm{S}}^{+} \approx \widehat{D}_{\mathrm{S}}^{\Diamond}$ for testing the conditional independence, a hypothesis that is not rejected by either test. In contrast, for testing the full independence, $\widehat{D}_{\mathrm{S}}^{+}$ and $\widehat{D}_{\mathrm{S}}^{\Diamond}$ are not very close to each other, but they both lead to essentially zero p-value, and hence both reject the null hypothesis. These results reconfirm the conclusions in Meng and Rubin (1992). Last but not least, the estimator $\widehat{r}_{\mathrm{S}}^{\Diamond}$ does not change under different null hypotheses, however it is not true for $\widetilde{r}_{\mathrm{L}}$ and $\widehat{r}_{\mathrm{S}}^{+}$.

5. Conclusions, Limitations and Future Work. In addition to conducting a general comparative study of MI tests, we proposed two particularly promising MI LRT based on $\widehat{D}_{\mathrm{S}}^{\Diamond} = \widehat{D}_{\mathrm{S}}(\widehat{r}_{\mathrm{S}}^{\Diamond})$ and $\widehat{D}_{\mathrm{S}}^{+} = \widehat{D}_{\mathrm{S}}(\widehat{r}_{\mathrm{S}}^{+})$. Both test statistics are non-negative, invariant to re-parametrizations, and powerful to reject a false null hypothesis (at least for large enough m). Test $\widehat{D}_{\mathrm{S}}^{\Diamond}$ is most principled, and the resulting test has the 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; and row independence of X_{com} is needed for the ease of computation. (The computationally more demanding test based on $\widehat{D}_{\mathrm{L}}(\widehat{r}_{\mathrm{L}}^{\Diamond})$ relaxes the independence assumption.) The main advantage of $\widehat{D}_{\mathrm{S}}^{+}$ is that it is easier to compute, as it requires only standard complete-data computer subroutines for likelihood ratio tests. One drawback is that the ad hoc fix $\widehat{r}_{\mathrm{S}}^{+} = \max(0, \widehat{r}_{\mathrm{S}})$ is inconsistent in general. However, the inconsistency does not appear to significantly affect the asymptotic power, at least in our experiments. Whereas $\widehat{D}_{\mathrm{S}}^{+}$ and $\widehat{D}_{\mathrm{S}}^{\Diamond}$ significantly improve over existing counterparts, more studies are needed, for reasons listed below.

- 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) provides theoretical tools for addressing such an issue in the context of estimation, and research is needed to extend their findings to the setting of hypothesis testing.
- Although the violation of the EFMI assumption may not (seriously) invalidate a test, it will affect its power. It is therefore desirable to explore MI tests without this assumption.
- The robust $\widehat{D}_{\mathrm{S}}^{\Diamond}$ 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 users to have access to non-trivial complete-data procedures. Hence a computational feasible robust test that only assumes EFMI on θ needs to be developed.

- Because the FMI is a fundamental nuisance parameter here and there is no (known) pivotal quantity, all MI tests are approximate in nature. In particular, they all have the potential of doing poorly when FMI is large and/or m is small. It is therefore of both theoretical and practical interest to seek powerful MI tests that are least affected by FMI.

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.

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

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

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.

Cox, D. R. and Reid, N. (1987) Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society B*, **49**, 1–39.

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 F., K. A. (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.

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

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.

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 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*. John Wiley & Sons.

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-Interscience.

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*, **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.

van der Vaart, A. W. (2000) *Asymptotic Statistics*. Cambridge University Press.

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.

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

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.

APPENDIX A: SUPPLEMENTARY RESULTS

A.1. Another Motivation for $\widehat{r}_{\mathrm{L}}^{\diamond}$. The definition of $\widehat{r}_{\mathrm{L}}^{\diamond}$ can also be motivated by the following observation. First, observe that one simple method to construct an always non-negative estimator of $\mathscr{r}_m$ is to perturb $\widehat{\psi}_0^*$ and $\widehat{\psi}_0^\ell$ by a suitable amount, say Δ, so that the perturbed version of $\widehat{r}_{\mathrm{L}}$ is always non-negative, and is still asymptotically equivalent to the original $\widehat{r}_{\mathrm{L}}$. We show, in Theorem A.1 below, that the right amount of Δ is $\Delta = \widehat{\psi}^* - \widehat{\psi}_0^*$. Using the perturbed version of $\widehat{r}_{\mathrm{L}}$, we obtain

$$\widehat{r}_{\mathrm{L}}^{\triangle} = \frac{m+1}{k(m-1)}\widehat{\delta}_{\mathrm{L}}^{\triangle},$$

where

$$\widehat{\delta}_{\mathrm{L}}^{\triangle} = \frac{2}{m}\sum_{\ell=1}^{m}\log\left\{\frac{f(X^\ell \mid \widehat{\psi}^\ell)}{f(X^\ell \mid \widehat{\psi}^*)}\frac{f(X^\ell \mid \widehat{\psi}_0^* + \Delta)}{f(X^\ell \mid \widehat{\psi}_0^\ell + \Delta)}\right\} = \frac{1}{m}\sum_{\ell=1}^{m} d_{\mathrm{L}}(\widehat{\psi}_0^\ell + \Delta, \widehat{\psi}^\ell \mid X^\ell).$$

Then we have the following result.

THEOREM A.1. *Suppose* RC_θ. *Under* H_0, *we have (i)* $\widehat{r}_{\mathrm{L}}^{\triangle} \geqslant 0$ *for all* m, n; *and (ii)* $\widehat{r}_{\mathrm{L}}^{\triangle} \doteq \widehat{r}_{\mathrm{L}}$ *as* $n \to \infty$ *for each* m.

Although $\widehat{r}_{\mathrm{L}}^{\triangle} \geqslant 0$, it is only invariant to affine transformations, and not robust against θ_0, and less computational feasible than $\widehat{r}_{\mathrm{L}}$; see § 3. However, it gives us some insights on how to construct a potentially better estimator. Note that, in (A.1), the constrained MLE is not used in $d_{\mathrm{L}}(\cdot, \cdot \mid X^\ell)$, but it is still always non-negative. We call this a "pseudo" LRT statistics. Then, $\widehat{\delta}_{\mathrm{L}}^{\triangle}$ is just a multiple of an average of many "pseudo" LRT statistics. In order to find a good estimator of $\mathscr{r}_m$, we may seek for an estimator which admits this form. Indeed, our proposed estimator $\widehat{r}_{\mathrm{L}}^{\diamond}$ also takes the same form:

$$\widehat{r}_{\mathrm{L}}^{\diamond} = \frac{m+1}{h(m-1)}\frac{1}{m}\sum_{\ell=1}^{m} d_{\mathrm{L}}(\widehat{\psi}^*, \widehat{\psi}^\ell \mid X^\ell).$$

A.2. Results for Dependent Data. This is a supplement for § 3.1. If the data are not independent, $\widehat{d}_{\mathrm{L}} \doteq \widehat{d}_{\mathrm{S}}$ is still true under the following conditions.

ASSUMPTION 5. *(a) Define* $R(\psi) = \overline{L}^{\mathrm{S}}(\psi) - \overline{L}(\psi)$, *where* $\overline{L}(\psi) = (mn)^{-1}\sum_{\ell=1}^{m}\log f(X^\ell \mid \psi)$ *and* $\overline{L}^{\mathrm{S}}(\psi) = (mn)^{-1}\log f(X^{\mathrm{S}} \mid \psi)$. *For each* m, *as* $n \to \infty$,

$$\sup_{\psi\in\Psi}|R(\psi)| = O_p(1/n), \qquad \sup_{\psi\in\Psi}\left|\frac{\partial}{\partial\psi}R(\psi)\right| = O_p(1/n).$$

(b) For each m, there exists a continuous function $\psi \mapsto \overline{\mathscr{L}}(\psi)$, which is free of n but may depend on m, such that, as $n \to \infty$,

$$\sup_{\psi\in\Psi} \left|\overline{L}(\psi) - \overline{\mathscr{L}}(\psi)\right| = o_p(1).$$

(c) Let $\psi_0^* = \arg\max_{\psi\in\Psi\,:\,\psi(\theta)=\theta_0} \overline{\mathscr{L}}(\psi)$ and $\psi^* = \arg\max_{\psi\in\Psi} \overline{\mathscr{L}}(\psi)$. For any fixed m, and for all $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\sup_{\substack{\psi\in\Psi\,:\,|\psi_0^*-\psi|>\varepsilon \\ \theta(\psi)=\theta_0}} \left\{\overline{\mathscr{L}}(\psi_0^*) - \overline{\mathscr{L}}(\psi)\right\} \geqslant \delta, \qquad \sup_{\psi\in\Psi\,:\,|\psi^*-\psi|>\varepsilon} \left\{\overline{\mathscr{L}}(\psi^*) - \overline{\mathscr{L}}(\psi)\right\} \geqslant \delta.$$

Conditions (b) and (c) in Assumption 5 are standard RCs that are usually assumed for M-estimators (see § 5 of van der Vaart (2000)); whereas condition (a) is satisfied by many models (see Example A.1 below).

THEOREM A.2. *Suppose* RC_θ *and Assumption 5. Under both H_0 and H_1, we have (i) $\widehat{d}_{\mathrm{S}}, \widehat{r}_{\mathrm{S}} \geqslant 0$ for all m, n; (ii) $\widehat{d}_{\mathrm{S}}, \widehat{r}_{\mathrm{S}}$ are invariant to the parametrization of ψ for all m, n; and (iii) $\widehat{d}_{\mathrm{L}} \doteq \widehat{d}_{\mathrm{S}}$ and $\widehat{r}_{\mathrm{L}} \doteq \widehat{r}_{\mathrm{S}}$ as $n \to \infty$ for each m.*

Theorem A.2 implies that the handy test statistics $\widehat{D}_{\mathrm{S}}$ and $\widehat{D}_{\mathrm{S}}^+$ approximate $\widehat{D}_{\mathrm{L}}$ and $\widehat{D}_{\mathrm{L}}^+$ for dependent data, provided that Assumption 5 holds.

EXAMPLE A.1. Consider a stationary autoregressive model of order one. Suppose the complete data $X = (X_1, \ldots, X_n)^\intercal$ is generated as following: $X_1 \sim \mathscr{N}(0, v^2)$ and $[X_i | X_{i-1}] \sim \mathscr{N}(\phi X_{i-1}, \sigma^2)$ for $i \geqslant 2$, where $v^2 = \sigma^2(1+\phi)/(1-\phi)$. Then $\psi = (\phi, \sigma^2)^\intercal$, and

$$\begin{aligned}
\overline{L}(\psi) &= -\frac{1}{2}\log(2\pi) - \frac{1}{2n}\log v^2 - \frac{1}{mn}\sum_{\ell=1}^{m}\frac{X_1^\ell}{2v^2} - \frac{n-1}{2n}\log\sigma^2 \\
&\quad - \frac{1}{mn}\sum_{\ell=1}^{m}\sum_{i=2}^{n}\frac{(X_i^\ell - \phi X_{i-1}^\ell)^2}{2\sigma^2}, \\
\overline{L}^{\mathrm{S}}(\psi) &= -\frac{1}{2}\log(2\pi) - \frac{1}{2mn}\log v^2 - \frac{(X_1^1)^2}{2mnv^2} - \frac{mn-1}{2mn}\log\sigma^2 \\
&\quad - \frac{1}{mn}\sum_{\ell=1}^{m}\sum_{i=2}^{n}\frac{(X_i^\ell - \phi X_{i-1}^\ell)^2}{2\sigma^2} - \frac{1}{mn}\sum_{\ell=2}^{m}\frac{(X_1^\ell - \phi X_n^{\ell-1})^2}{2\sigma^2}.
\end{aligned}$$

Then, it is easy to see that condition (a) of Assumption 5 is satisfied.

APPENDIX B: PROOFS

PROOF OF THEOREM 2.1. (i, ii) From (2.3), we know $\widehat{d}_{\mathrm{L}} \geqslant 0$ is invariant to parametrization ψ. (iii) Since $\widehat{d}_{\mathrm{L}}$ is invariant to transformation of ψ, we assume, without loss of generality, that ψ admits a parameterization such that $\mathsf{Cov}(\widehat{\theta}^{\ell}, \widehat{\eta}^{\ell}) \doteq \mathbf{0}$ by taking suitable linear transformation of ψ. Also write U_{η}^{ℓ} as an efficient estimator of $\mathsf{Var}(\widehat{\eta})$ based on X^{ℓ}; and recall that $U_{\theta}^{\ell} = U^{\ell}$ is an efficient estimator of $\mathsf{Var}(\widehat{\theta})$ based on X^{ℓ}.

Using Taylor's expansion on $\psi \mapsto \overline{L}(\psi) = m^{-1} \sum_{\ell=1}^{m} \log f(X^{\ell} \mid \psi)$ around $\widehat{\psi}^{*} = ((\widehat{\theta}^{*})^{\intercal}, (\widehat{\eta}^{*})^{\intercal})^{\intercal}$, we know that for $\psi \doteq \widehat{\psi}^{*}$,

$$\overline{L}(\psi) \doteq \overline{L}(\widehat{\psi}^{*}) - \frac{1}{2} \left(\psi - \widehat{\psi}^{*}\right)^{\intercal} \overline{I}(\widehat{\psi}^{*}) \left(\psi - \widehat{\psi}^{*}\right), \tag{B.1}$$

where $\overline{I}(\psi) = -\partial^{2} \overline{L}(\psi) / \partial\psi\partial\psi^{\intercal}$, which satisfies

$$\overline{I}(\widehat{\psi}^{*}) \doteq \begin{pmatrix} \overline{U}_{\theta}^{-1} & \mathbf{0} \\ \mathbf{0} & \overline{U}_{\eta}^{-1} \end{pmatrix} \tag{B.2}$$

with $\overline{U}_{\eta} = m^{-1} \sum_{i=1}^{m} U_{\eta}^{\ell}$. Under the null, $\widehat{\psi}^{*} \doteq \widehat{\psi}_{0}^{*}$. So, using (B.1), we have

$$\begin{aligned} \widehat{d}_{\mathrm{L}} &\doteq \left(\widehat{\psi}_{0}^{*} - \widehat{\psi}^{*}\right)^{\intercal} \overline{I}(\widehat{\psi}^{*}) \left(\widehat{\psi}_{0}^{*} - \widehat{\psi}^{*}\right), \\ &\doteq \begin{pmatrix} \theta_{0} - \widehat{\theta}^{*} \\ \widehat{\eta}(\theta_{0}) - \widehat{\eta}(\widehat{\theta}^{*}) \end{pmatrix}^{\intercal} \begin{pmatrix} \overline{U}_{\theta}^{-1} & \mathbf{0} \\ \mathbf{0} & \overline{U}_{\eta}^{-1} \end{pmatrix} \begin{pmatrix} \theta_{0} - \widehat{\theta}^{*} \\ \widehat{\eta}(\theta_{0}) - \widehat{\eta}(\widehat{\theta}^{*}) \end{pmatrix} \\ &\doteq (\overline{\theta}^{\intercal} - \theta_{0}) \overline{U}_{\theta}^{-1} (\overline{\theta}^{\intercal} - \theta_{0}) = \widetilde{d}_{\mathrm{W}}', \end{aligned} \tag{B.3}$$

where we have used (a) $\widehat{\theta}^{*} \doteq \overline{\theta}$; see, e.g., Lemma 1 of Wang and Robins (1998), and (b) $\widehat{\eta}(\theta_{0}) - \widehat{\eta}(\widehat{\theta}^{*}) = O_{p}(1/n)$ if $\theta_{0} - \widehat{\theta}^{*} = O_{p}(1/\sqrt{n})$; see Cox and Reid (1987). Since $\widetilde{d}_{\mathrm{W}}' \doteq \widetilde{d}_{\mathrm{L}}$ (Meng and Rubin, 1992), we have $\widehat{d}_{\mathrm{L}} \doteq \widetilde{d}_{\mathrm{L}}$. □

PROOF OF PROPOSITION 2.2. The given condition implies that

$$\begin{aligned} \widehat{\psi}^{\ell} &= ((\widehat{\theta}^{\ell})^{\intercal}, (\widehat{\eta}^{\ell})^{\intercal})^{\intercal}, \qquad \widehat{\psi}_{0}^{\ell} = (\theta_{0}^{\intercal}, (\widehat{\eta}^{\ell})^{\intercal})^{\intercal}, \\ \widehat{\psi}^{*} &= ((\widehat{\theta}^{*})^{\intercal}, (\widehat{\eta}^{*})^{\intercal})^{\intercal}, \qquad \widehat{\psi}_{0}^{*} = (\theta_{0}^{\intercal}, (\widehat{\eta}^{*})^{\intercal})^{\intercal}. \end{aligned}$$

Clearly, we also have the decomposition: $L^{\ell}(\psi) = L_{\dagger}^{\ell}(\theta) + L_{\ddagger}^{\ell}(\eta)$ for all ℓ, where $L_{\dagger}^{\ell}(\theta) = L_{\dagger}(\theta \mid X^{\ell})$ and $L_{\ddagger}^{\ell}(\eta) = L_{\ddagger}(\eta \mid X^{\ell})$. Then,

$$\begin{aligned} \overline{d}_{\mathrm{L}} - \widehat{d}_{\mathrm{L}} &= \frac{2}{m} \sum_{\ell=1}^{m} \left\{ L^{\ell}(\widehat{\psi}^{\ell}) - L^{\ell}(\widehat{\psi}_{0}^{\ell}) - L^{\ell}(\widehat{\psi}^{*}) + L^{\ell}(\widehat{\psi}_{0}^{*}) \right\} \\ &= \frac{2}{m} \sum_{\ell=1}^{m} \left\{ L_{\dagger}^{\ell}(\widehat{\theta}^{\ell}) - L_{\dagger}^{\ell}(\widehat{\theta}^{*}) \right\} \geqslant 0 \end{aligned}$$

since $L_\dagger^\ell(\widehat{\theta}^\ell) \geqslant L_\dagger^\ell(\widehat{\theta}^*)$ for all ℓ. □

PROOF OF COROLLARY 2.3. Applying Taylor's expansion on $\psi \mapsto L^\ell(\psi)$, we can find $\breve{\psi}^\ell$ lying on the line segment joining $\widehat{\psi}^\ell$ and $\widehat{\psi}_0^\ell$ such that

$$L^\ell(\widehat{\psi}_0^\ell) = L^\ell(\widehat{\psi}^\ell) - \frac{1}{2}\left(\widehat{\psi}_0^\ell - \widehat{\psi}^\ell\right)^\intercal I^\ell(\breve{\psi}^\ell)\left(\widehat{\psi}_0^\ell - \widehat{\psi}^\ell\right),$$

where $I^\ell(\psi) = -\partial^2 L^\ell(\psi)/\partial\psi\partial\psi^\intercal$. By the lower order variability of $I^\ell(\breve{\psi}^\ell)$, we can find $\breve{\psi}^*$ such that $I^\ell(\breve{\psi}^\ell) \doteq I^\ell(\breve{\psi}^*)$ for all ℓ. Then, using similar techniques as in (B.2) and (B.3), we have

$$\begin{aligned} L^\ell(\widehat{\psi}^\ell) - L^\ell(\widehat{\psi}_0^\ell) &\doteq \frac{1}{2}\left(\widehat{\psi}_0^\ell - \widehat{\psi}^\ell\right)^\intercal I^\ell(\breve{\psi}^*)\left(\widehat{\psi}_0^\ell - \widehat{\psi}^\ell\right) \\ &\doteq \frac{1}{2}\left(\theta_0 - \widehat{\theta}^\ell\right)^\intercal \breve{U}^{-1}\left(\theta_0 - \widehat{\theta}^\ell\right) \end{aligned} \tag{B.4}$$

for some matrix $\breve{U}$. Similarly, we have

$$L^\ell(\widehat{\psi}^*) - L^\ell(\widehat{\psi}_0^*) \doteq \frac{1}{2}\left(\theta_0 - \widehat{\theta}^*\right)^\intercal \breve{U}^{-1}\left(\theta_0 - \widehat{\theta}^*\right). \tag{B.5}$$

Write $A^{\otimes 2} = AA^\intercal$ for any appropriate matrix A. Using (B.4), (B.5) and the cyclic property of trace, we have

$$\begin{aligned} \overline{d}_{\mathrm{L}} - \widehat{d}_{\mathrm{L}} &\doteq \frac{1}{m}\sum_{\ell=1}^m \left\{\left(\theta_0 - \widehat{\theta}^\ell\right)^\intercal \breve{U}^{-1}\left(\theta_0 - \widehat{\theta}^\ell\right) - \left(\theta_0 - \widehat{\theta}^*\right)^\intercal \breve{U}^{-1}\left(\theta_0 - \widehat{\theta}^*\right)\right\} \\ &= \mathrm{tr}\left[\breve{U}^{-1}\left\{\frac{1}{m}\sum_{\ell=1}^m \left(\theta_0 - \widehat{\theta}^\ell\right)^{\otimes 2} - \left(\theta_0 - \widehat{\theta}^*\right)^{\otimes 2}\right\}\right] \\ &\doteq \mathrm{tr}\left[\breve{U}^{-1}\frac{1}{m}\sum_{\ell=1}^m \left\{(\widehat{\theta}^\ell)^{\otimes 2} - \overline{\theta}^{\otimes 2}\right\}\right] \doteq \mathrm{tr}\left(\breve{U}^{-1}B\right) \doteq \mathrm{tr}\left(\mathcal{U}_{\theta,0}^{-1}\mathcal{B}_\theta\right) \end{aligned}$$

as $m, n \to \infty$, where $\mathcal{U}_{\theta,0}$ is a deterministic matrix that depends on both θ_0 and $\theta^\star$, and satisfies $n(\breve{U} - \mathcal{U}_{\theta,0}) \xrightarrow{\mathrm{pr}} 0$. Note that $\mathrm{tr}(\mathcal{U}_{\theta,0}^{-1}\mathcal{B}_\theta) = k\varkappa_0$, for some finite $\varkappa_0$ by Assumption 2. Then $\widehat{r}_{\mathrm{L}} \xrightarrow{\mathrm{pr}} \varkappa_0 = \mathrm{tr}(\mathcal{U}_{\theta,0}^{-1}\mathcal{B}_\theta)/k$, proving (ii). (But $\mathcal{U}_{\theta,0}$ may not equal to $\mathcal{U}_\theta$, and hence $\widehat{r}_{\mathrm{L}}$ may not be consistent for $\varkappa_m$.)

If H_0 is true, then $\overline{\theta} \xrightarrow{\mathrm{pr}} \theta_0 = \theta^\star$ and $\breve{U} \doteq \overline{U} \doteq \mathcal{U}_\theta = \mathcal{U}_{\theta,0}$. Then, $\widehat{r}_{\mathrm{L}} \xrightarrow{\mathrm{pr}} \varkappa$ as $m, n \to \infty$. So, (i) follows. □

PROOF OF THEOREM 2.4. (i, ii) It is trivial by the definition of $\widehat{r}_{\mathrm{L}}^{\diamondsuit}$. (iii) Applying Taylor's expansion to $\psi \mapsto L^\ell(\psi)$ again, we know there is $\breve{\psi}^\ell$ lying

on the line segment joining $\widehat{\psi}^{\ell}$ and $\widehat{\psi}^{*}$ such that

$$L^{\ell}(\widehat{\psi}^{*}) = L^{\ell}(\widehat{\psi}^{\ell}) - \frac{1}{2}\left(\widehat{\psi}^{*} - \widehat{\psi}^{\ell}\right)^{\intercal} I^{\ell}(\breve{\psi}^{\ell})\left(\widehat{\psi}^{*} - \widehat{\psi}^{\ell}\right). \tag{B.6}$$

By the lower order variability of $I^{\ell}(\breve{\psi}^{\ell})$, we know that $I^{\ell}(\breve{\psi}^{\ell}) \doteq \overline{I}(\widehat{\psi}^{*})$ for all ℓ, where $\overline{I}(\psi) = m^{-1}\sum_{\ell=1}^{m} I^{\ell}(\psi)$. We also know that $\widehat{\psi}^{*} \doteq \overline{\psi}$. Thus

$$\begin{aligned}
\overline{\delta}_{\mathrm{L}} - \widehat{\delta}_{\mathrm{L}} &\doteq \frac{1}{m}\sum_{\ell=1}^{m}\left(\widehat{\psi}^{*} - \widehat{\psi}^{\ell}\right)^{\intercal}\overline{I}(\widehat{\psi}^{*})\left(\widehat{\psi}^{*} - \widehat{\psi}^{\ell}\right) \\
&= \operatorname{tr}\left\{\overline{I}(\widehat{\psi}^{*})\frac{1}{m}\sum_{\ell=1}^{m}\left(\widehat{\psi}^{*} - \widehat{\psi}^{\ell}\right)^{\otimes 2}\right\} \\
&\doteq \operatorname{tr}\left\{\overline{I}(\widehat{\psi}^{*})\frac{1}{m}\sum_{\ell=1}^{m}\left(\widehat{\psi}^{\ell} - \overline{\psi}\right)^{\otimes 2}\right\} \doteq \operatorname{tr}\left(\mathscr{U}_{\psi}^{-1}\mathscr{B}_{\psi}\right) \qquad \text{(B.7)}
\end{aligned}$$

as $m, n \to \infty$. By the assumption of EFMI of ψ, we have $\widehat{r}_{\mathrm{L}}^{\diamondsuit} \xrightarrow{\mathrm{pr}} \text{ɾ}$. □

Proof of Lemma 2.5. First, recall that, as $n \to \infty$, the observed data MLE $\widehat{\theta}_{\mathrm{obs}}$ of θ satisfies (2.4), which can be written as $[\widehat{\theta}_{\mathrm{obs}} \mid \theta^{\star}] \overset{\mathscr{D}}{\approx} \mathcal{N}_k(\theta^{\star}, \mathscr{T}_{\theta})$, where $A_{1,n} \overset{\mathscr{D}}{\approx} A_{2,n}$ means that $A_{1,n}$ and $A_{2,n}$ have the same asymptotic distribution, i.e., there exist deterministic sequences μ_n and Σ_n such that $(A_{1,n} - \mu_n)\Sigma_n^{-1/2} \Rightarrow A$ and $(A_{2,n} - \mu_n)\Sigma_n^{-1/2} \Rightarrow A$ for some non-degenerate random variable A. From Assumption 3, a proper imputation model is used. So, we have (2.5), which is equivalent to say that, as $n \to \infty$,

$$\left[\widehat{\theta}^{\ell} \mid X_{\mathrm{obs}}\right] \overset{\mathscr{D}}{\approx} \mathcal{N}_k(\widehat{\theta}_{\mathrm{obs}}, \mathscr{B}_{\theta}), \tag{B.8}$$

independently for for $\ell = 1, \ldots, m$. Therefore we can represent

$$\widehat{\theta}_{\mathrm{obs}} \overset{\mathscr{D}}{\approx} \theta^{\star} + \mathscr{T}_{\theta}^{1/2} W, \tag{B.9}$$

$$\widehat{\theta}^{\ell} \overset{\mathscr{D}}{\approx} \widehat{\theta}_{\mathrm{obs}} + \mathscr{B}_{\theta}^{1/2} Z_{\ell}, \qquad \ell = 1, \ldots, m \tag{B.10}$$

where $Z_1, \ldots, Z_m, W \overset{\mathrm{iid}}{\sim} \mathcal{N}_k(0, I_k)$. Also write $Z_{\ell} = (Z_{1\ell}, \ldots, Z_{k\ell})^{\intercal}$, for $\ell = 1, 2, \ldots, m$, and $W = (W_1, \ldots, W_k)^{\intercal}$. Averaging (B.10) over ℓ, we have $\overline{\theta} \overset{\mathscr{D}}{\approx} \widehat{\theta}_{\mathrm{obs}} + \mathscr{B}_{\theta}^{1/2}\overline{Z}_{\bullet}$, where $\overline{Z}_{\bullet} = m^{-1}\sum_{\ell=1}^{m} Z_{\ell}$. Since $\mathscr{B}_{\theta} = \text{ɾ}\mathscr{U}_{\theta}$, we have

$$\begin{aligned}
\mathscr{U}_{\theta}^{-1/2}(\widehat{\theta}^{\ell} - \theta^{\star}) &\overset{\mathscr{D}}{\approx} (1 + \text{ɾ})^{1/2} W + \text{ɾ}^{1/2} Z_{\ell}, \\
\mathscr{U}_{\theta}^{-1/2}(\overline{\theta} - \theta^{\star}) &\overset{\mathscr{D}}{\approx} (1 + \text{ɾ})^{1/2} W + \text{ɾ}^{1/2}\overline{Z}_{\bullet}.
\end{aligned}$$

Note that (2.6) implies $\mathscr{U}_\theta \doteq \overline{U}$. Under H_0, we have $\theta^\star = \theta_0$ and

$$\begin{aligned}
\overline{d}_{\mathrm{L}} &\doteq \overline{d}'_{\mathrm{W}} \overset{\mathscr{D}}{\approx} \sum_{i=1}^{k} \left\{(1+\varkappa)^{1/2} W_i + \varkappa^{1/2} Z_{i\ell}\right\}^2, \\
\widehat{d}_{\mathrm{L}} &\doteq \widetilde{d}_{\mathrm{L}} \doteq \widetilde{d}'_{\mathrm{W}} \overset{\mathscr{D}}{\approx} \sum_{i=1}^{k} \left\{(1+\varkappa)^{1/2} W_i + \varkappa^{1/2} \overline{Z}_i\right\}^2.
\end{aligned}$$

After some simple algebra, we obtain

$$\widehat{r}_{\mathrm{L}}^{+} \overset{\mathscr{D}}{\approx} \frac{(m+1)\varkappa}{mk} \sum_{i=1}^{k} s_{Z_i}^2 \qquad \text{and} \qquad \widehat{D}_{\mathrm{L}}^{+} \overset{\mathscr{D}}{\approx} \frac{m \sum_{i=1}^{k} \left\{(1+\varkappa)^{1/2} W_i + \varkappa^{1/2} \overline{Z}_{i\bullet}\right\}^2}{mk + (m+1)\varkappa \sum_{i=1}^{k} s_{Z_i}^2},$$

where $s_{Z_i}^2 = (m-1)^{-1} \sum_{\ell=1}^{m} (Z_{i\ell} - \overline{Z}_{i\bullet})^2$ is the sample variance of $\{Z_{i\ell}\}_{\ell=1}^{m}$. Since W_i, $\overline{Z}_{i\bullet}$ and $s_{Z_i}^2$ are mutually independent for each fixed i, we can simplify the representation of $\widehat{D}_{\mathrm{L}}^{+}$ to

$$\widehat{r}_{\mathrm{L}}^{+} \overset{\mathscr{D}}{\approx} \frac{(m+1)\varkappa}{m(m-1)k} \sum_{i=1}^{k} H_i^2 \qquad \text{and} \qquad \widehat{D}_{\mathrm{L}}^{+} \overset{\mathscr{D}}{\approx} \frac{(m-1)\{m + (m+1)\varkappa\} \sum_{i=1}^{k} G_i^2}{m(m-1)k + (m+1)\varkappa \sum_{i=1}^{k} H_i^2},$$

where $G_i^2 \overset{\text{iid}}{\sim} \chi_1^2$ and $H_i^2 \overset{\text{iid}}{\sim} \chi_{m-1}^2$, for $i = 1, \ldots, k$, are all mutually independent. Clearly, they can be further simplified to (2.14). □

Proof of Theorem 2.6. Similar to (B.9) and (B.10), we can have a more general representation:

$$\widehat{\psi}_{\mathrm{obs}} \overset{\mathscr{D}}{\approx} \psi^\star + \mathscr{T}_\psi^{1/2} W; \quad \widehat{\psi}^\ell \overset{\mathscr{D}}{\approx} \widehat{\psi}_{\mathrm{obs}} + \mathscr{B}_\psi^{1/2} Z_\ell, \qquad \ell = 1, \ldots, m,$$

where $Z_1, \ldots, Z_h, W \overset{\text{iid}}{\sim} \mathcal{N}_h(0, I_h)$. Also write $Z_\ell = (Z_{1\ell}, \ldots, Z_{h\ell})^\intercal$, for $\ell = 1, 2, \ldots, m$, and $W = (W_1, \ldots, W_h)^\intercal$. Using (B.7), we have

$$\begin{aligned}
\overline{\delta}_{\mathrm{L}} - \widehat{\delta}_{\mathrm{L}} &\doteq \operatorname{tr}\left\{\overline{I}(\widehat{\psi}^*) \frac{1}{m} \sum_{\ell=1}^{m} \left(\widehat{\psi}^\ell - \overline{\psi}\right)\left(\widehat{\psi}^\ell - \overline{\psi}\right)^\intercal\right\} \\
&\overset{\mathscr{D}}{\approx} \operatorname{tr}\left[\mathscr{U}_\psi^{-1} \frac{1}{m} \sum_{\ell=1}^{m} \left\{(\mathscr{T}_\psi - \mathscr{U}_\psi)^{1/2} \left(Z_\ell - \overline{Z}_\bullet\right)\right\}^{\otimes 2}\right] \\
&= \frac{1}{m} \sum_{\ell=1}^{m} \operatorname{tr}\left\{\varkappa I_h \left(Z_\ell - \overline{Z}_\bullet\right)^{\otimes 2}\right\} = \frac{\varkappa}{m} \sum_{\ell=1}^{m} \sum_{i=1}^{h} (Z_{i\ell} - \overline{Z}_{i\bullet})^2.
\end{aligned}$$

Equivalently, we can say $\overline{\delta}_{\mathrm{L}} - \widehat{\delta}_{\mathrm{L}} \Rightarrow \kappa \chi^2_{h(m-1)}/m$ as $n \to \infty$. Hence

$$\widehat{r}^{\,\diamondsuit}_{\mathrm{L}} \Rightarrow \kappa \cdot \frac{m+1}{hm(m-1)} \cdot \chi^2_{h(m-1)},$$

which is equivalent to (2.15). Note that it is true under both H_0 and H_1. □

Proof of Theorem 2.7. From the representations of $\widehat{d}^{\,\diamondsuit}_{\mathrm{L}}$ and $\widehat{r}^{\,\diamondsuit}_{\mathrm{L}}$ in Lemma 2.5 and Theorem 2.6, we know that they are asymptotically ($n \to \infty$) independent. The proof then follows the derivation for Lemma 2.5. □

Proof of Proposition 3.1. It is trivial. □

Proof of Theorem A.1. (i) Using the representation (A.1), we can easily see that $\widehat{r}^{\,\triangle}_{\mathrm{L}} \geqslant 0$. (ii) It suffices to show

$$m^{-1} \sum_{\ell=1}^{m} d_{\mathrm{L}}(\widehat{\psi}^{\ell}_0 + \Delta_m, \widehat{\psi}^{\ell} \mid X^{\ell}) \doteq \overline{d}_{\mathrm{L}} - \widetilde{d}_{\mathrm{L}},$$

where $\Delta_m = \widehat{\psi}^* - \widehat{\psi}^*_0$. Under H_0, $\Delta_m \doteq 0$ and $\widehat{\psi}^{\ell}_0 \doteq \widehat{\psi}^{\ell}$, so $\widehat{\psi}^{\ell}_0 + \Delta_m \doteq \widehat{\psi}^{\ell}$. Using Taylor's expansion on $\psi \mapsto L^{\ell}(\psi)$ around its maximizer $\widehat{\psi}^{\ell}$, we have for $\psi \doteq \widehat{\psi}^{\ell}$ that

$$L^{\ell}(\psi) \doteq L^{\ell}(\widehat{\psi}^{\ell}) - \frac{1}{2}\left(\psi - \widehat{\psi}^{\ell}\right)^{\intercal} I^{\ell}(\widehat{\psi}^{\ell}) \left(\psi - \widehat{\psi}^{\ell}\right).$$

Under the parametrization of ψ in the proof of Theorem 2.1, we know that the upper $k \times k$ sub-matrix of $I^{\ell}(\widehat{\psi}^{\ell})$ is $\left(U^{\ell}\right)^{-1}$. Using the lower order variability of U^{ℓ}, we have $\left(U^{\ell}\right)^{-1} \doteq \overline{U}^{-1}$ and

$$\begin{aligned}
&\frac{1}{m} \sum_{\ell=1}^{m} d_{\mathrm{L}}(\widehat{\psi}^{\ell}_0 + \Delta_m, \widehat{\psi}^{\ell} \mid X^{\ell}) \\
&\quad \doteq \frac{1}{m} \sum_{\ell=1}^{m} \left(\widehat{\psi}^{\ell}_0 + \Delta_m - \widehat{\psi}^{\ell}\right)^{\intercal} I^{\ell}(\widehat{\psi}^{\ell}) \left(\widehat{\psi}^{\ell}_0 + \Delta_m - \widehat{\psi}^{\ell}\right) \\
&\quad \doteq \frac{1}{m} \sum_{\ell=1}^{m} (\widehat{\theta}^{\ell} - \overline{\theta})^{\intercal} \overline{U}^{-1} (\widehat{\theta}^{\ell} - \overline{\theta}) = \overline{d}'_{\mathrm{W}} - \widetilde{d}'_{\mathrm{W}} \doteq \overline{d}_{\mathrm{L}} - \widehat{d}_{\mathrm{L}}.
\end{aligned}$$

Therefore, the desired result follows. □

Proof of Theorem A.2. Throughout this proof, conditions (a), (b) and (c) refer to the list given in Assumption 5. (i, ii) It trivially follows

from the definitions of $\widehat{d}_{\mathrm{S}}$ and $\widehat{r}_{\mathrm{S}}$. (iii) First, by the definition of maximizer and condition (a), we have

$$\begin{aligned}\underline{L}(\widehat{\psi}^*) - \underline{L}(\widehat{\psi}^{\mathrm{S}}) &= \underline{L}(\widehat{\psi}^*) - \overline{\underline{L}}^{\mathrm{S}}(\widehat{\psi}^{\mathrm{S}}) + \overline{\underline{L}}^{\mathrm{S}}(\widehat{\psi}^{\mathrm{S}}) - \underline{L}(\widehat{\psi}^{\mathrm{S}}) \\ &\leqslant \underline{L}(\widehat{\psi}^*) - \overline{\underline{L}}^{\mathrm{S}}(\widehat{\psi}^*) + \overline{\underline{L}}^{\mathrm{S}}(\widehat{\psi}^{\mathrm{S}}) - \underline{L}(\widehat{\psi}^{\mathrm{S}}) \\ &\leqslant 2 \sup_{\psi \in \Psi} \left| \underline{L}(\psi) - \overline{\underline{L}}^{\mathrm{S}}(\psi) \right| = O_p(1/n),\end{aligned}$$

which, together with condition (b), imply that

$$\begin{aligned}&\overline{\underline{\mathscr{L}}}(\psi^*) - \overline{\underline{\mathscr{L}}}(\widehat{\psi}^{\mathrm{S}}) \\ &\quad = \left\{ \overline{\underline{\mathscr{L}}}(\psi^*) - \underline{L}(\psi^*) \right\} + \left\{ \underline{L}(\psi^*) - \underline{L}(\widehat{\psi}^{\mathrm{S}}) \right\} + \left\{ \underline{L}(\widehat{\psi}^{\mathrm{S}}) - \overline{\underline{\mathscr{L}}}(\widehat{\psi}^{\mathrm{S}}) \right\} \\ &\quad \leqslant 2 \sup_{\psi \in \Psi} \left| \underline{L}(\psi) - \overline{\underline{\mathscr{L}}}(\psi) \right| + \left\{ \underline{L}(\widehat{\psi}^*) - \underline{L}(\widehat{\psi}^{\mathrm{S}}) \right\} = o_p(1). \qquad \text{(B.11)}\end{aligned}$$

Using (B.11) and (c), we have $\widehat{\psi}^{\mathrm{S}} \xrightarrow{\mathrm{pr}} \psi^*$. By (b) and (c), we also have $\widehat{\psi}^* \xrightarrow{\mathrm{pr}} \psi^*$. So, $\left| \widehat{\psi}^{\mathrm{S}} - \widehat{\psi}^* \right| \xrightarrow{\mathrm{pr}} \mathbf{0}$ as $n \to \infty$. By the definition of maximizer,

$$\mathbf{0} = \nabla \overline{\underline{L}}^{\mathrm{S}}(\widehat{\psi}^{\mathrm{S}}) = \nabla \underline{L}(\widehat{\psi}^{\mathrm{S}}) + \nabla R(\widehat{\psi}^{\mathrm{S}}), \qquad \text{(B.12)}$$

where $\nabla g(\psi) = \partial g(\psi)/\partial \psi$ is the gradient of $\psi \mapsto g(\psi)$. By condition (a), we know $\nabla R(\widehat{\psi}^{\mathrm{S}}) = O_p(1/n)$. Thus, together with (B.12), we have $\nabla \underline{L}(\widehat{\psi}^{\mathrm{S}}) = O_p(1/n)$. Also, by the definition of MLE, we have $\nabla \underline{L}(\widehat{\psi}^*) = \mathbf{0}$.

By Taylor's expansion, there exists $\breve{\psi}$ such that

$$\underline{L}(\widehat{\psi}^*) - \underline{L}(\widehat{\psi}^{\mathrm{S}}) = \left\{ \nabla \underline{L}(\breve{\psi}) \right\}^{\intercal} \left(\widehat{\psi}^* - \widehat{\psi}^{\mathrm{S}} \right) = o_p(1/n), \qquad \text{(B.13)}$$

where we have used the continuity of $\psi \mapsto \nabla \underline{L}(\psi)$ to yield $\nabla \underline{L}(\breve{\psi}) = O_p(1/n)$. Rewriting (B.13), we have

$$\underline{L}(\widehat{\psi}^*) - \overline{\underline{L}}^{\mathrm{S}}(\widehat{\psi}^{\mathrm{S}}) = R(\widehat{\psi}^{\mathrm{S}}) + o_p(1/n). \qquad \text{(B.14)}$$

Similar to (B.14), we have

$$\underline{L}(\widehat{\psi}_0^*) - \overline{\underline{L}}^{\mathrm{S}}(\widehat{\psi}_0^{\mathrm{S}}) = R(\widehat{\psi}_0^{\mathrm{S}}) + o_p(1/n). \qquad \text{(B.15)}$$

Then, using (B.14) and (B.15), we have

$$\begin{aligned}\left| \widehat{d}_{\mathrm{L}} - \widehat{d}_{\mathrm{S}} \right| &= 2n \left| \left\{ \underline{L}(\widehat{\psi}^*) - \overline{\underline{L}}^{\mathrm{S}}(\widehat{\psi}^{\mathrm{S}}) \right\} - \left\{ \underline{L}(\widehat{\psi}_0^*) - \overline{\underline{L}}^{\mathrm{S}}(\widehat{\psi}_0^{\mathrm{S}}) \right\} \right| \\ &= 2n \left| R(\widehat{\psi}^{\mathrm{S}}) - R(\widehat{\psi}_0^{\mathrm{S}}) + o_p(1/n) \right|.\end{aligned}$$

Now consider two cases.

(i) Under H_0, we have $\widehat{d}_{\mathrm{L}} = O_p(1)$ and $\widehat{\psi}_0^{\mathrm{S}} \doteq \widehat{\psi}^{\mathrm{S}}$. Thus condition (a) implies $R(\widehat{\psi}^{\mathrm{S}}) - R(\widehat{\psi}_0^{\mathrm{S}}) = o_p(1/n)$. Then, we have $\left|\widehat{d}_{\mathrm{L}} - \widehat{d}_{\mathrm{S}}\right| = o_p(\widehat{d}_{\mathrm{L}})$.

(ii) Under H_1, we have $\widehat{d}_{\mathrm{L}} \xrightarrow{\mathrm{pr}} \infty$. Condition (a) and (B.11) imply that $\underline{L}(\widehat{\psi}^*) - \underline{L}^{\mathrm{S}}(\widehat{\psi}^{\mathrm{S}}) = O_p(1/n)$. Similarly, we also have $\underline{L}(\widehat{\psi}_0^*) - \underline{L}^{\mathrm{S}}(\widehat{\psi}_0^{\mathrm{S}}) = O_p(1/n)$. Hence $\left|\widehat{d}_{\mathrm{L}} - \widehat{d}_{\mathrm{S}}\right| = O_p(1)$. Thus we have $\left|\widehat{d}_{\mathrm{L}} - \widehat{d}_{\mathrm{S}}\right| = o_p(\widehat{d}_{\mathrm{L}})$.

Therefore, under either H_0 or H_1, we also have $\left|\widehat{d}_{\mathrm{L}} - \widehat{d}_{\mathrm{S}}\right| = o_p(\widehat{d}_{\mathrm{L}})$. Since $\widehat{d}_{\mathrm{L}} \doteq \widehat{d}_{\mathrm{S}}$ and $\overline{d}_{\mathrm{L}} = \overline{d}_{\mathrm{S}}$, we know $\widehat{r}_{\mathrm{L}} \doteq \widehat{r}_{\mathrm{S}}$. □

Note that, even under the assumption of this theorem, $\widehat{r}_{\mathrm{S}}$ and $\widehat{r}_{\mathrm{S}}^{\diamondsuit}$ are not equivalent. From (3.5) and (3.6), $\widehat{r}_{\mathrm{S}}$ and $\widehat{r}_{\mathrm{S}}^{\diamondsuit}$ are a "difference of difference" estimator and a "difference" estimator, respectively. So, the "bias" of using $\overline{L}^{\mathrm{S}}(\psi)$ cannot be canceled out in $\widehat{r}_{\mathrm{S}}^{\diamondsuit}$.

APPENDIX C: ADDITIONAL FIGURES AND TABLES

This section presents additional figures and tables in § 2.5 and § 4

- Figure 9: the performance of different approximations to the reference null distribution when $\alpha = 5\%$; see § 2.5.
- Figures 10 and 11: the empirical distributions of the p-values under H_0 and parametrizations (i) and (iii), respectively; see § 4.1.
- Figures 12 and 13: the empirical power functions and the empirical ratio of power-to-size for size 5% tests, respectively; see § 4.1.
- Table 8: the ranges of empirical sizes over different parametrizations for size 5% tests; see § 4.1.
- Table 9: detailed results of the care-survival example in § 4.3.

Department of Statistics
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
E-mail: kinwaichan@sta.cuhk.edu.hk

Department of Statistics
Harvard University
Cambridge, Massachusetts 02138, U.S.A.
E-mail: meng@stat.harvard.edu

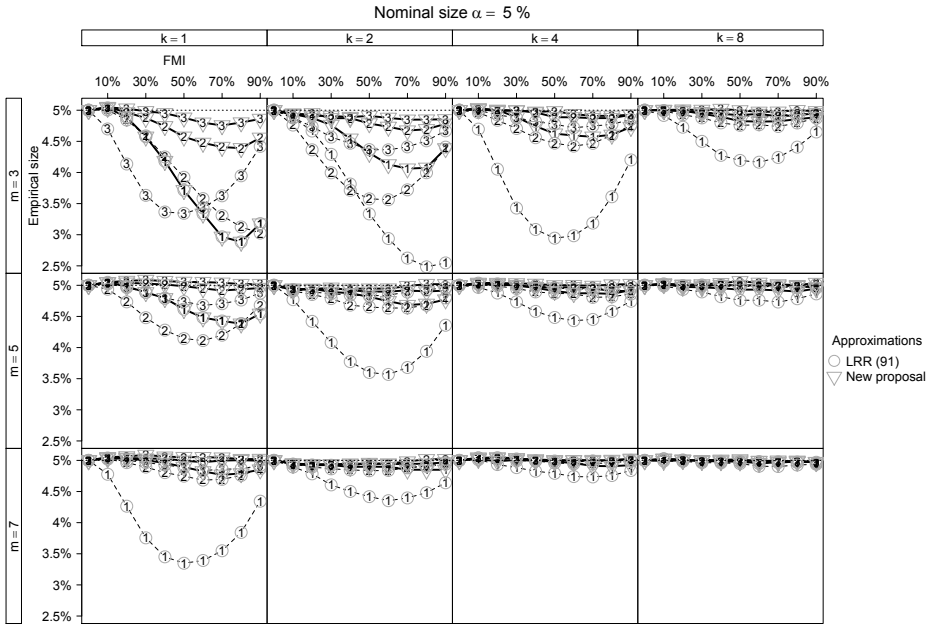

FIG 9. *The performances of two different approximate null distributions when the nominal size is* $\alpha = 5\%$. *The vertical axis denotes* $\widehat{\alpha}$ *or* $\widetilde{\alpha}$, *and the horizontal axis denotes the value of* f_m. *The number attached to each line denotes the value of* $\tau = h/k$. *The proposed approximation* $\widehat{\alpha}$ *is denoted by thick solid lines with triangles, and the existing approximation* $\widetilde{\alpha}$ *is denoted by thin dashed lines with circles.*

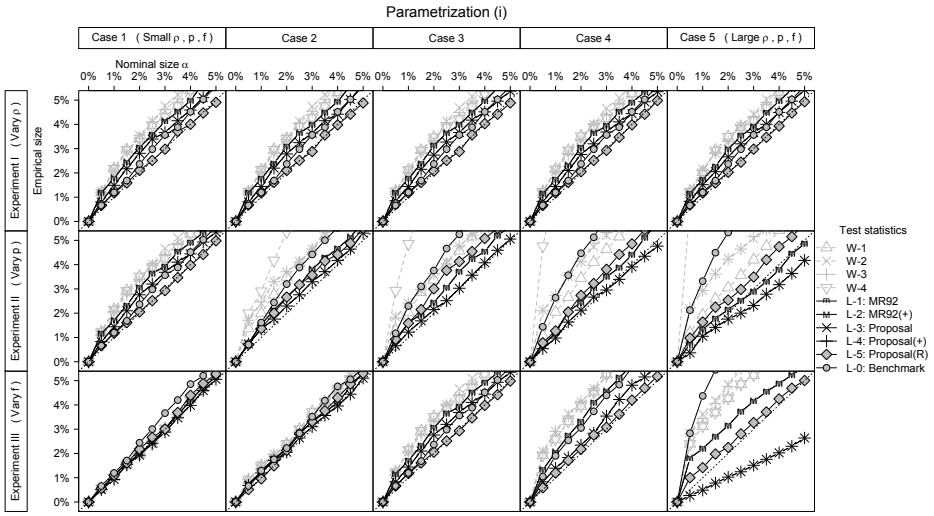

FIG 10. *The comparison between empirical size and nominal size α under parametrization (i) for $\alpha \in (0, 5\%]$. The Wald tests (W-1,2,3,4) and LRTs (L-0,1,2,3,4,5) are represented by grey dashed and black solid lines, respectively. The LRT statistic $\widetilde{D}_{\mathrm{L}}$ (L-1: MR92) (Meng and Rubin, 1992) and its modification $\widetilde{D}_{\mathrm{L}}^{+}$ (L-2: MR92(+)) are the existing counterparts of our proposals $\widehat{D}_{\mathrm{S}}$ (L-3: Proposal), $\widehat{D}_{\mathrm{S}}^{+}$ (L-4: Proposal(+)) and $\widehat{D}_{\mathrm{S}}^{\diamond}$ (L-5: Proposal(R)).*

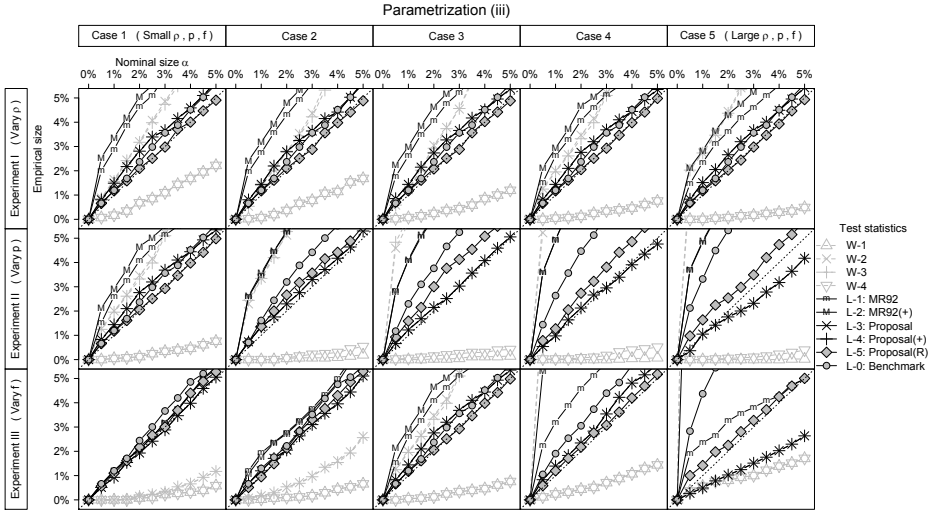

FIG 11. *The comparison between empirical size and nominal size α under parametrization (iii) for $\alpha \in (0, 5\%]$. The legend in Figure 10 also applies here.*

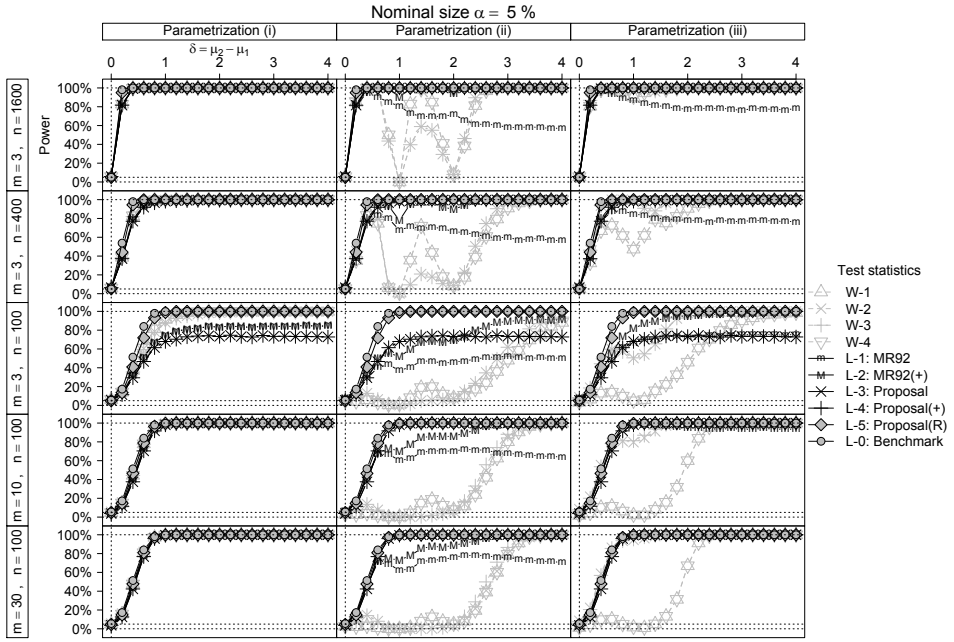

FIG 12. *The power curves under different parametrizations. The nominal size is* $\alpha = 5\%$. *In each plot, the vertical axis denotes the power, whereas the horizontal axis denotes the value of* $\delta = \mu_2 - \mu_1$. *The legend in Figure 10 also applies here.*

TABLE 8

The range of empirical size $[\min \widehat{\alpha}, \max \widehat{\alpha}]$ *in percentage, where* max *and* min *are taken over the three parametrizations. Only one value is recorded for those tests that are invariant to parametrization. The nominal size is* $\alpha = 5\%$.

	Range of empirical size: $[\min \widehat{\alpha}, \max \widehat{\alpha}]/\%$				
(n, m)	$(1600, 3)$	$(400, 3)$	$(100, 3)$	$(100, 10)$	$(100, 30)$
W-1	[5.62, 5.71]	[5.30, 6.03]	[3.22, 6.20]	[1.64, 4.81]	[1.37, 5.00]
W-2	[5.93, 6.05]	[6.08, 7.18]	[5.52, 8.69]	[4.42, 8.47]	[4.20, 8.50]
W-3	[5.81, 6.03]	[6.01, 6.98]	[5.37, 8.28]	[4.20, 7.67]	[4.10, 7.50]
W-4	[5.62, 5.71]	[5.30, 6.03]	[3.22, 6.20]	[1.64, 4.81]	[1.37, 5.00]
L-1	[5.57, 6.15]	[6.37, 6.57]	[5.88, 6.47]	[4.39, 5.66]	[4.22, 5.32]
L-2	[5.52, 6.10]	[6.37, 6.52]	[5.88, 7.47]	[4.39, 5.66]	[4.22, 5.32]
L-3	5.76	6.37	5.42	3.78	3.71
L-4	5.76	6.37	5.42	3.78	3.71
L-5	4.96	5.32	4.93	4.79	4.54
L-0	5.03	5.03	5.57	5.57	5.57

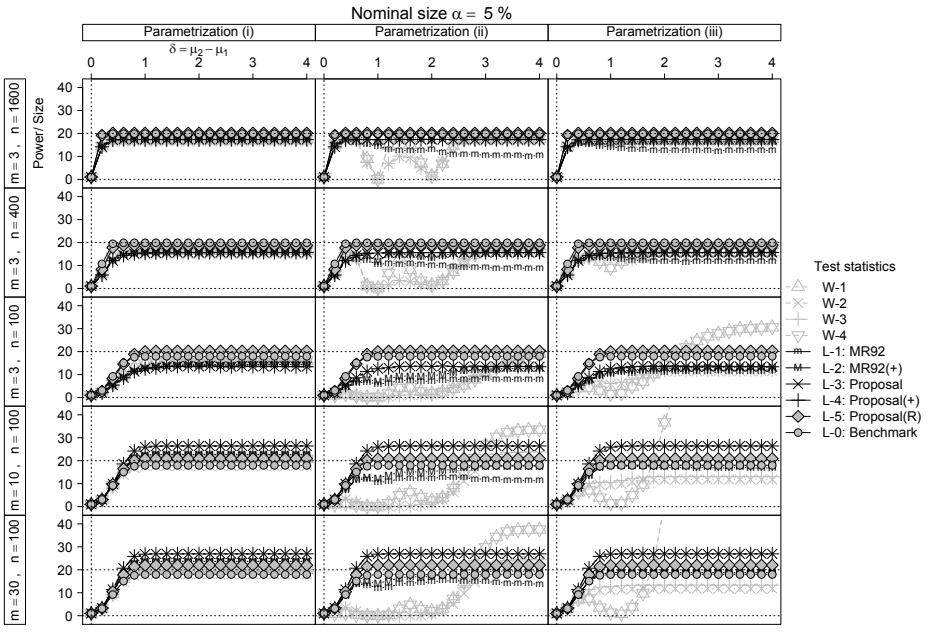

FIG 13. *The ratios of empirical power to empirical size under different parametrizations. The nominal size is $\alpha = 5\%$. In each plot, the vertical axis denotes the ratio, whereas the horizontal axis denotes $\delta = \mu_2 - \mu_1$. The legend in Figure 10 also applies here.*

TABLE 9
The LRTs using $\widetilde{D}_{\mathrm{L}}$, $\widehat{D}_{\mathrm{S}}^{+}$ and $\widehat{D}_{\mathrm{S}}^{\Diamond}$ under different parametrizations in § 4.3.

	Parametrization (i)					
	H_0: Conditional independence			H_0: Full independence		
m	$\widetilde{r}_{\mathrm{L}}, \widehat{r}_{\mathrm{S}}^{+}, \widehat{r}_{\mathrm{S}}^{\Diamond}$	$\widetilde{D}_{\mathrm{L}}, \widehat{D}_{\mathrm{S}}^{+}, \widehat{D}_{\mathrm{S}}^{\Diamond}$	$\widetilde{p}_{\mathrm{L}}, \widehat{p}_{\mathrm{S}}^{+}, \widehat{p}_{\mathrm{S}}^{\Diamond}$	$\widetilde{r}_{\mathrm{L}}, \widehat{r}_{\mathrm{S}}^{+}, \widehat{r}_{\mathrm{S}}^{\Diamond}$	$\widetilde{D}_{\mathrm{L}}, \widehat{D}_{\mathrm{S}}^{+}, \widehat{D}_{\mathrm{S}}^{\Diamond}$	$\widetilde{p}_{\mathrm{L}}, \widehat{p}_{\mathrm{S}}^{+}, \widehat{p}_{\mathrm{S}}^{\Diamond}$
2	0.63, 0.64, 0.83	0.14, 0.14, 0.12	0.87, 0.87, 0.89	0.53, 0.53, 0.83	44.4, 44.4, 37.1	0, 0, 0
3	0.54, 0.54, 0.38	0.08, 0.08, 0.09	0.93, 0.93, 0.92	0.31, 0.31, 0.38	54.2, 54.2, 51.4	0, 0, 0
5	0.49, 0.48, 0.89	0.12, 0.12, 0.10	0.89, 0.89, 0.91	0.72, 0.72, 0.89	40.8, 40.8, 37.1	0, 0, 0
7	0.23, 0.23, 0.47	0.06, 0.06, 0.05	0.94, 0.94, 0.95	0.31, 0.31, 0.47	53.2, 53.2, 47.6	0, 0, 0
10	0.50, 0.50, 0.70	0.14, 0.14, 0.12	0.87, 0.87, 0.88	0.56, 0.56, 0.70	45.4, 45.4, 41.7	0, 0, 0
25	0.35, 0.35, 0.47	0.06, 0.06, 0.06	0.94, 0.94, 0.95	0.35, 0.35, 0.47	51.4, 51.4, 47.0	0, 0, 0
50	0.31, 0.31, 0.45	0.11, 0.11, 0.10	0.90, 0.90, 0.91	0.33, 0.33, 0.45	51.5, 51.5, 47.3	0, 0, 0
	Parametrization (ii)					
	H_0: Conditional independence			H_0: Full independence		
m	$\widetilde{r}_{\mathrm{L}}, \widehat{r}_{\mathrm{S}}^{+}, \widehat{r}_{\mathrm{S}}^{\Diamond}$	$\widetilde{D}_{\mathrm{L}}, \widehat{D}_{\mathrm{S}}^{+}, \widehat{D}_{\mathrm{S}}^{\Diamond}$	$\widetilde{p}_{\mathrm{L}}, \widehat{p}_{\mathrm{S}}^{+}, \widehat{p}_{\mathrm{S}}^{\Diamond}$	$\widetilde{r}_{\mathrm{L}}, \widehat{r}_{\mathrm{S}}^{+}, \widehat{r}_{\mathrm{S}}^{\Diamond}$	$\widetilde{D}_{\mathrm{L}}, \widehat{D}_{\mathrm{S}}^{+}, \widehat{D}_{\mathrm{S}}^{\Diamond}$	$\widetilde{p}_{\mathrm{L}}, \widehat{p}_{\mathrm{S}}^{+}, \widehat{p}_{\mathrm{S}}^{\Diamond}$
2	1.23, 0.64, 0.83	0.01, 0.14, 0.12	0.99, 0.87, 0.89	0.98, 0.53, 0.83	34.2, 44.4, 37.1	0, 0, 0
3	1.08, 0.54, 0.38	−0.07, 0.08, 0.09	1.00, 0.93, 0.92	0.61, 0.31, 0.38	43.9, 54.2, 51.4	0, 0, 0
5	1.02, 0.48, 0.89	−0.09, 0.12, 0.10	1.00, 0.89, 0.91	1.40, 0.72, 0.89	29.0, 40.8, 37.1	0, 0, 0
7	0.45, 0.23, 0.47	−0.07, 0.06, 0.05	1.00, 0.94, 0.95	0.58, 0.31, 0.47	43.9, 53.2, 47.6	0, 0, 0
10	0.99, 0.50, 0.70	−0.10, 0.14, 0.12	1.00, 0.87, 0.88	1.09, 0.56, 0.70	33.7, 45.4, 41.7	0, 0, 0
25	0.71, 0.35, 0.47	−0.14, 0.06, 0.06	1.00, 0.94, 0.95	0.68, 0.35, 0.47	41.0, 51.4, 47.0	0, 0, 0
50	0.63, 0.31, 0.45	−0.10, 0.11, 0.10	1.00, 0.90, 0.91	0.65, 0.33, 0.45	41.3, 51.5, 47.3	0, 0, 0
	Parametrization (iii)					
	H_0: Conditional independence			H_0: Full independence		
m	$\widetilde{r}_{\mathrm{L}}, \widehat{r}_{\mathrm{S}}^{+}, \widehat{r}_{\mathrm{S}}^{\Diamond}$	$\widetilde{D}_{\mathrm{L}}, \widehat{D}_{\mathrm{S}}^{+}, \widehat{D}_{\mathrm{S}}^{\Diamond}$	$\widetilde{p}_{\mathrm{L}}, \widehat{p}_{\mathrm{S}}^{+}, \widehat{p}_{\mathrm{S}}^{\Diamond}$	$\widetilde{r}_{\mathrm{L}}, \widehat{r}_{\mathrm{S}}^{+}, \widehat{r}_{\mathrm{S}}^{\Diamond}$	$\widetilde{D}_{\mathrm{L}}, \widehat{D}_{\mathrm{S}}^{+}, \widehat{D}_{\mathrm{S}}^{\Diamond}$	$\widetilde{p}_{\mathrm{L}}, \widehat{p}_{\mathrm{S}}^{+}, \widehat{p}_{\mathrm{S}}^{\Diamond}$
2	1.06, 0.64, 0.83	0.04, 0.14, 0.12	0.96, 0.87, 0.88	−0.38, 0.53, 0.83	109, 44.4, 37.1	0, 0, 0
3	−2.35, 0.54, 0.38	−1.16, 0.08, 0.09	1.00, 0.93, 0.92	−1.22, 0.31, 0.38	−321, 54.2, 51.4	1, 0, 0
5	−2.64, 0.48, 0.89	−1.38, 0.12, 0.10	1.00, 0.89, 0.91	−2.24, 0.72, 0.89	−58, 40.8, 37.1	1, 0, 0
7	−0.01, 0.23, 0.47	0.25, 0.06, 0.05	0.78, 0.94, 0.95	−0.34, 0.31, 0.47	107, 53.2, 47.6	0, 0, 0
10	−2.04, 0.50, 0.70	−2.20, 0.14, 0.12	1.00, 0.87, 0.88	−1.85, 0.56, 0.70	−86, 45.4, 41.7	1, 0, 0
25	−1.39, 0.35, 0.47	−4.30, 0.06, 0.06	1.00, 0.94, 0.95	−1.12, 0.35, 0.47	−603, 51.4, 47.0	1, 0, 0
50	−1.22, 0.31, 0.45	−7.39, 0.11, 0.10	1.00, 0.90, 0.91	−1.06, 0.33, 0.45	−1136, 51.5, 47.3	1, 0, 0