MULTIPLE IMPROVEMENTS OF MULTIPLE IMPUTATION LIKELIHOOD RATIO TESTS

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

Department of Statistics, The Chinese University of Hong Kong[1]
Department of Statistics, Harvard University[2]

Supplementary Material

A Supplementary Results

A.1 A Complication Caused by Nuisance Parameter

This section supplement the discussion of Section 2.2 in the main article. Recall that the likelihood function $L^{(\ell)}(\cdot)$ is based on both observed data X_{obs} and imputed data $X_{\text{mis}}^{(\ell)}$, which varies across ℓ. Hence, each imputed likelihood $L^{(\ell)}(\cdot)$ is associated with a (imputation-specific) pseudo parameter $\psi^{(\ell)}$, may vary across $\ell = 1, \ldots, m$.

To see the source of the negativity of $\widehat{r}_{\text{L}}$, we extend $\overline{L}(\psi)$ in (2.1) to

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

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 A.1. Each of them consists of zero, one or two of the constraints $\mathscr{C}_0 : \theta^{(1)} = \cdots = \theta^{(m)} = \theta_0$ and $\mathscr{C}^0 : \psi^{(1)} = \cdots = \psi^{(m)}$, where $\theta^{(\ell)} = \theta(\psi^{(\ell)})$ is the interested part of $\psi^{(\ell)}$ for each ℓ. The constraint $\mathscr{C}_0$ is equivalent to H_0, and the constraint $\mathscr{C}^0$ means that all $\psi^{(\ell)}$s 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 A.2. Define the maximized value of $\overline{L}(\psi^{(1)}, \ldots, \psi^{(m)})$

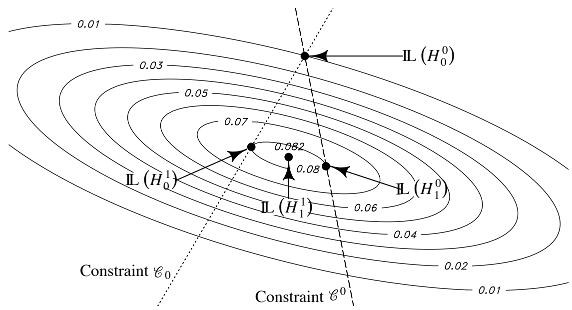

Figure A.1: A schematic illustration of the sign of (A.2). The contour lines of $\overline{L}(\psi^{(1)},\ldots,\psi^{(m)})$ are plotted. The two straight lines refer to constraints $\mathscr{C}_0$ and $\mathscr{C}^0$. Since $\mathbb{L}(H_1^1) = 0.082$, $\mathbb{L}(H_0^1) = \mathbb{L}(H_1^0) = 0.08$, 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 (A.1) is at least 4-dimensional (i.e., $\theta^{(1)},\theta^{(2)},\eta^{(1)},\eta^{(2)}$) generally, so this illustration in a 2-dimension space is just conceptual.

Table A.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., $\mathscr{r} = 0$)	$\mathscr{C}^1 : \psi^{(1)},\ldots,\psi^{(m)} \in \Psi$ (i.e., $\mathscr{r} \geqslant 0$)
$\mathscr{C}_0 : \theta^{(1)} = \cdots = \theta^{(m)} = \theta_0 \in \Theta$ (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)	$H_1^0 = \mathscr{C}_1 \cap \mathscr{C}^0$	$H_1^1 = \mathscr{C}_1 \cap \mathscr{C}^1$

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

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

Whereas the two bracketed terms in (A.2) are non-negative as they correspond to two LRT statistics, their difference can be negative.

A simple example illustrates this well. For the regression model $[Y \mid X_1, X_2] \sim \mathcal{N}(\beta_0 + \beta_1 X_1 + \beta_2 X_2, \sigma^2)$, 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 $H_0^1 : \beta_1 \in \mathbb{R}, \beta_2 = 0$; see Figure A.1 for a schematic illustration.

The decomposition (A.2) 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 $\mathscr{r} = 0$ in

$$
\begin{tikzcd}
H_0^0 \arrow[r, Rightarrow] \arrow[d, Rightarrow] \arrow[dr, Rightarrow] & H_1^1 \arrow[d, Rightarrow] \\
H_1^0 \arrow[r, Rightarrow] & H_1^1
\end{tikzcd}
$$

Figure A.2: 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.

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 (A.2) can be viewed as a measure of evidence against $\varkappa = 0$ solely from θ; Proposition 1 already hinted this possibility. However, the "test statistic" (A.2) has a fatal flaw. Because $\mathscr{C}_0$ requires all $\theta^{(\ell)}$s to coincide with a specific θ_0, $\mathscr{C}_0$ is not nested within $\mathscr{C}^0$, i.e., $\mathscr{C}^0 \not\Rightarrow \mathscr{C}_0$. Hence $\widehat{r}_{\mathrm{L}}$ is guaranteed to consistently estimate $\varkappa_m$ only under H_0. This explains Corollary 1, and leads to an improvement in Section 2.2. In it not hard to see that our new estimator $\widehat{r}_{\mathrm{L}}^{\diamondsuit}$ simply drops the second term in (A.2).

A.2 Another Motivation for $\widehat{r}_{\mathrm{L}}^{\diamondsuit}$

The definition of $\widehat{r}_{\mathrm{L}}^{\diamondsuit}$ can also be motivated by the following observation. First, observe that one simple method to construct an always non-negative estimator of $\varkappa_m$ 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 Section 3. However, it gives us some insights on how to construct a potentially better estimator. Note that, in (A.3), 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 $\boldsymbol{r}_m$, we may seek for an estimator which admits this form. Indeed, our estimator $\widehat{r}_{\mathrm{L}}^{\diamondsuit}$ also takes the same form:

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

A.3 Additional result for Section 2.3

This section presents the additional simulation result for Section 2.3. The performance of different approximations to the reference null distribution when $\alpha = 5\%$ is shown in Figure A.3.

A.4 Results for Dependent Data

This is a supplement for Section 3.1. If the data are not independent, then (3.1) is no longer true. In other words, $\overline{L}(\psi) \not\equiv \overline{L}^{\mathrm{S}}(\psi)$, where $\overline{L}(\psi) = \sum_{\ell=1}^{m} L^{(\ell)}(\psi)/m$ is defined in (2.1), and

$$\overline{L}^{\mathrm{S}}(\psi) = \frac{1}{m} \log f_{mn}(X^{(1:m)} \mid \psi). \tag{A.3}$$

In principle, $\overline{L}(\psi)$ should be used instead of the "stacked version" $\overline{L}^{\mathrm{S}}(\psi)$, however, the stacked one is much easier to compute. Because of this reason, it is of interest to see whether the stacked version can be used generally.

To begin with, we define the stacked version of all MI statistics when $\overline{L}^{\mathrm{S}}(\psi)$ is used instead of $\overline{L}(\psi)$. Let

$$\widehat{\psi}_0^{\mathrm{S}} = \operatorname*{arg\,max}_{\psi \in \Psi \,:\, \theta(\psi) = \theta_0} \overline{L}^{\mathrm{S}}(\psi), \qquad \widehat{\psi}^{\mathrm{S}} = \operatorname*{arg\,max}_{\psi \in \Psi} \overline{L}^{\mathrm{S}}(\psi); \tag{A.4}$$

$$\widehat{\delta}_{0,\mathrm{S}} = 2\overline{L}^{\mathrm{S}}(\widehat{\psi}_0^{\mathrm{S}}), \qquad \widehat{\delta}_{\mathrm{S}} = 2\overline{L}^{\mathrm{S}}(\widehat{\psi}^{\mathrm{S}}). \tag{A.5}$$

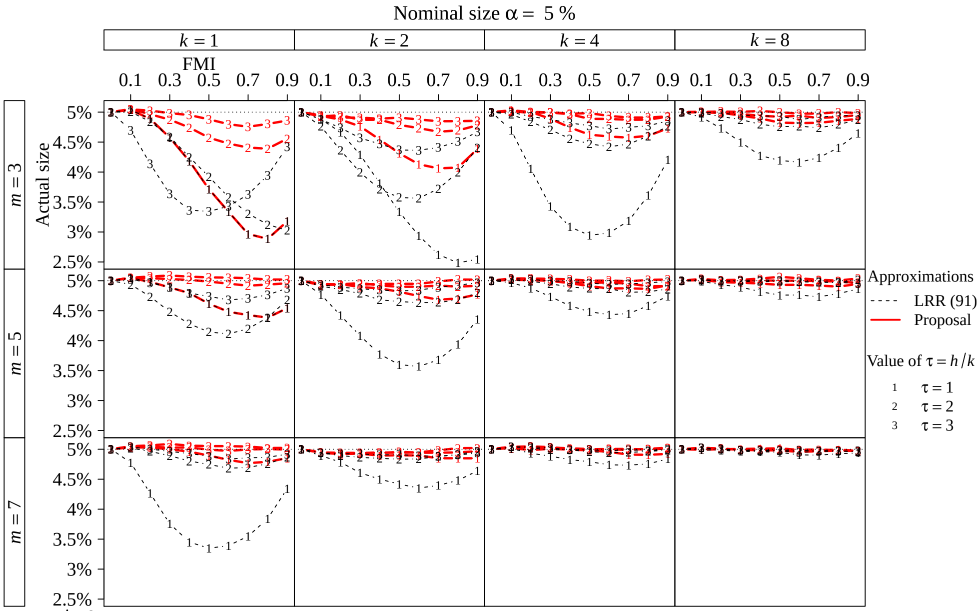

Figure A.3: The performance of two approximate null distributions when the nominal size is $\alpha = 5\%$. The vertical axis denotes $\widehat{\alpha}$ or $\widetilde{\alpha}$, and the horizontal axis denotes the value of $\mathscr{r}_m$. The number attached to each line denotes the value of $\tau = h/k$. 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.

and

$$\widehat{D}_{\mathrm{S}}(\mathscr{r}_m) = \frac{\widehat{d}_{\mathrm{S}}}{k(1+\mathscr{r}_m)}, \qquad \text{with } \widehat{d}_{\mathrm{S}} = \widehat{\delta}_{\mathrm{S}} - \widehat{\delta}_{0,\mathrm{S}} \text{ of (A.5);} \tag{A.6}$$

$$\widehat{r}_{\mathrm{S}} = \frac{m+1}{k(m-1)}(\overline{d}_{\mathrm{S}} - \widehat{d}_{\mathrm{S}}), \qquad \text{with } \overline{d}_{\mathrm{S}} = \overline{d}_{\mathrm{L}} \text{ of (1.7);} \tag{A.7}$$

$$\widehat{r}_{\mathrm{S}}^{\Diamond} = \frac{m+1}{h(m-1)}(\overline{\delta}_{\mathrm{S}} - \widehat{\delta}_{\mathrm{S}}), \qquad \text{with } \overline{\delta}_{\mathrm{S}} = \overline{\delta}_{\mathrm{L}} \text{ of (2.10);} \tag{A.8}$$

and $\widehat{r}_{\mathrm{S}}^{+} = \max(0, \widehat{r}_{\mathrm{S}})$. The stacked counterparts of $\widehat{D}_{\mathrm{L}}^{\Diamond}$ and its existing counterparts $\widehat{D}_{\mathrm{L}}$ and $\widehat{D}_{\mathrm{L}}^{+}$ (see (2.11)) then are given by

$$\widehat{D}_{\mathrm{S}}^{\Diamond} = \widehat{D}_{\mathrm{S}}(\widehat{r}_{\mathrm{S}}^{\Diamond}), \qquad \widehat{D}_{\mathrm{S}} = \widehat{D}_{\mathrm{S}}(\widehat{r}_{\mathrm{S}}), \qquad \widehat{D}_{\mathrm{S}}^{+} = \widehat{D}_{\mathrm{S}}(\widehat{r}_{\mathrm{S}}^{+}). \tag{A.9}$$

The approximation $\widehat{d}_{\mathrm{L}} \simeq \widehat{d}_{\mathrm{S}}$ is still true under the following conditions.

Assumption A.1. *(a) Define* $R(\psi) = \underline{\overline{L}}^{\mathrm{S}}(\psi) - \underline{\overline{L}}(\psi)$, *where*

$$\underline{\overline{L}}(\psi) = (mn)^{-1} \sum_{\ell=1}^{m} \log f(X^{(\ell)} \mid \psi) \quad \text{and} \quad \underline{\overline{L}}^{\mathrm{S}}(\psi) = (mn)^{-1} \log f(X^{\mathrm{S}} \mid \psi).$$

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* $\psi \mapsto \underline{\overline{\mathscr{L}}}(\psi)$, *which is free of* n *but may depend on* m, *such that, as* $n \to \infty$,

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

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

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

Conditions (b) and (c) in Assumption A.1 are standard RCs that are usually assumed for M-estimators (see Section 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 A.1. 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 A.1 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

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

Then, it is easy to see that condition (a) of Assumption A.1 is satisfied.

A.5 Other existing MI tests

First, we list some existing estimators of r_m. Let $s^2_{\mathrm{W},a}$ be the sample variances of $\{(d_{\mathrm{W}}^{(\ell)})^a\}_{\ell=1}^m$ for $a > 0$. Rubin (2004) and Li *et al.* (1991) proposed

$$\widetilde{r}_{\mathrm{W},1} = \frac{(1+1/m)s^2_{\mathrm{W},1}}{2\overline{d}_{\mathrm{W}} + \sqrt{\max\left\{0, 4\overline{d}_{\mathrm{W}}^2 - 2ks^2_{\mathrm{W},1}\right\}}}, \tag{A.10}$$

$$\widetilde{r}_{\mathrm{W},1/2} = (1+1/m)s^2_{\mathrm{W},1/2}, \tag{A.11}$$

respectively. When k is large and m is small, using (A.10) or (A.11) may lead to power loss. 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.

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

$$\widetilde{D}'_{\mathrm{W}}(r_m) = \frac{\widetilde{d}'_{\mathrm{W}}}{k(1+r_m)}, \ \widetilde{D}_{\mathrm{L}}(r_m) = \frac{\widetilde{d}_{\mathrm{L}}}{k(1+r_m)}, \ \widetilde{D}_{\mathrm{L}}^+(r_m) = \left\{\widetilde{D}_{\mathrm{L}}(r_m)\right\}^+. \tag{A.12}$$

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

$$\overline{D}'_{\mathrm{W}}(r_m) = \frac{\overline{d}'_{\mathrm{W}} - \frac{k(m-1)}{m+1}r_m}{k(1+r_m)}, \qquad \overline{D}_{\mathrm{L}}(r_m) = \frac{\overline{d}_{\mathrm{L}} - \frac{k(m-1)}{m+1}r_m}{k(1+r_m)}; \tag{A.13}$$

see, e.g., Li *et al.* (1991). The combining rule $\overline{D}'_{\mathrm{W}}(\varkappa_m)$ is useful when computing $\overline{d}'_{\mathrm{W}}$ and estimating $\varkappa_m$ are simple, but the resulting power may deteriorate. If $\widetilde{r}_{\mathrm{W},1}$ or $\widetilde{r}_{\mathrm{W},1/2}$ is used for estimating $\varkappa_m$, the null distribution of (A.12) and (A.13) can be approximated by $F_{k,\widetilde{\mathrm{df}}'(\varkappa_m,k)}$, where $\widetilde{\mathrm{df}}'(\varkappa_m,k) = (m-1)(1+\varkappa_m^{-1})^2 k^{-3/m}$; see Li *et al.* (1991).

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

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

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

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

Table A.2 is the full version of Table 1 in the main text. It summarizes the statistical and computational properties of different MI tests; see Section 3.2 for details.

Table A.2: Computational requirements and statistical properties of MI test statistics, their associated combining rules and estimators of FMI $\mathscr{r}_m$. The symbols "+" and "−" mean that the test statistic (or estimator) is equipped and not equipped with the indicated property, respectively; see the end of Section 3.2 for heading descriptions. The reference papers/book are abbreviated as follows: Rubin (2004) (R04), Li *et al.* (1991) (LMRR91) and Meng and Rubin (1992) (MR92).

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

[a] In actual computation, the $\mathscr{r}_m$ in the denominator degree of freedom of F is replaced by its corresponding estimator.

[b] Computing the test statistic $D_{\mathrm{W}}(T) = d_{\mathrm{W}}(\overline{\theta}, T)/k$ does not require estimating $\mathscr{r}_m$.

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

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

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

[f] The originally proposed combining rule is $\overline{D}'_{\mathrm{W}}(\mathscr{r}_m)$; see (A.13). Although $\overline{D}'_{\mathrm{W}}(\mathscr{r}_m)$ is more computational feasible, the power loss is more significant than $\tilde{D}'_{\mathrm{W}}(\mathscr{r}_m)$ after inserting an inefficient estimator $\tilde{r}'_{\mathrm{W},1}$ for $\mathscr{r}_m$. This footnote also applies to WT-3.

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

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

Table A.3: The values of parameters used in the simulation experiment in Section 4.1.

Experiment		**Fixed Parameters**			**Variable Parameter**				
No.	Variable Parameter	ρ	p	$\mathscr{f}$	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 $\mathscr{f}$	0.4	2	–	0.1	0.3	0.5	0.7	0.9

A.6 Supplement for Section 4.1

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 data set can be produced by the following procedure, for $\ell = 1, \ldots, m$.

1. Draw $(\Sigma^{(\ell)})^{-1}$ from a Wishart distribution with $(n_{\text{obs}} - 1)$ degrees of freedom and scale matrix S_{obs}.
2. Draw $\mu^{(\ell)}$ 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 data set, 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

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

We first study the distribution of p-values of each test under H_0. We use $n = 100$, $m = 3$, $\sigma^2 = 5$ and $\mu = \mathbf{1}_p$, with various values of ρ, p and $\mathscr{f}$ specified in Table A.3. The results under parametrizations (i), (ii) and (iii) are shown in Figures A.4, A.5 and A.6, respectively. Note that, for Wald tests under parametrization (ii), the matrix $U^{(\ell)}$ is singular in 0.25% of the replications, and those cases are removed from the analysis (which should favor the Wald tests).

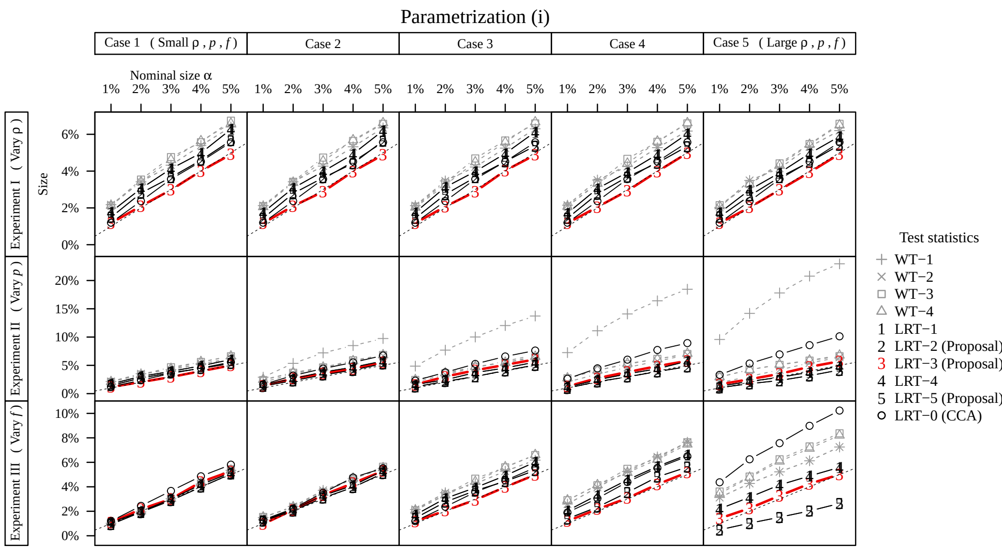

Figure A.4: The comparison between empirical size and nominal size α under parametrization (ii) for $\alpha \in (0, 5\%]$. Our most recommended proposal is LRT-3, which is highlighted red.

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 ℓ 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 ℓ are large. 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

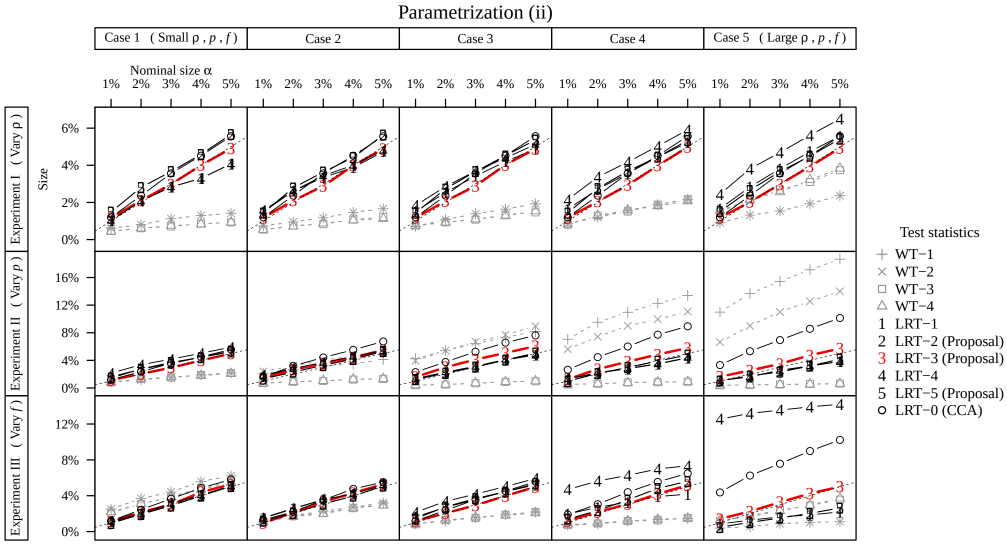

Figure A.5: The comparison between empirical size and nominal size α under parametrization (i) for $\alpha \in (0, 5\%]$. The legend in Figure A.4 also applies here.

$m = 3$.

Interestingly, as seen clearly in Figure A.5, the benchmark L-0 performs very badly for large p and $\boldsymbol{f}$. This is because 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 (see Barnard and Rubin, 1999). But this is exactly what is being used for MI tests, albeit with different choices of the denominator degrees of freedom. Note also that, in some cases, nearly half of the simulated values of $\widetilde{r}_{\mathrm{L}}$ and $\widetilde{D}_{\mathrm{L}}$ are negative; see Table A.4. In contrast, $\widehat{r}_{\mathrm{S}}$ is always non-negative in our simulation, despite the fact that it can be negative in theory.

The power curves under nominal size 0.5% and 5% are shown in Figure 2 of the main text and Figure A.7, respectively. Note that the trivial modifications LRT-2 of LRT-1 cannot retrieve all the power it should have. Tables A.5 and A.6 show the minimum and maximum of the empirical sizes over the three parametrizations considered in each test — and only one value is needed for

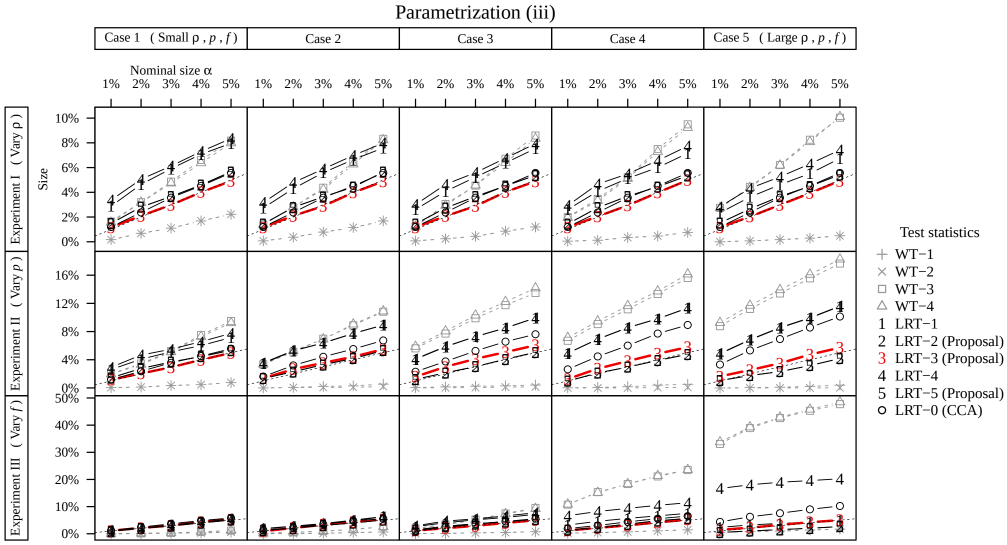

Figure A.6: The comparison between empirical size and nominal size α under parametrization (iii) for $\alpha \in (0, 5\%]$. The legend in Figure A.5 also applies here.

Table A.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

those tests that are invariant to parametrization — when the nominal size is 0.5% and 5%, respectively. 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.*,

Table A.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 parametrization-invariant tests. The nominal size is $\alpha = 0.5\%$. The results under nominal size $\alpha = 5\%$ are shown in Figure A.6.

	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

2016). Figures A.8 and A.9 plot the result for nominal size 0.5% and 5%, respectively. 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 performances of all size 0.5% tests are less satisfactory than those for size 5% tests because larger sample sizes n are required to approximate the tail behavior well.

We also compare the performance of estimators of $\mathscr{r}_m$ for different δ and parametrizations. In our experiment, we have $\mathscr{r}_m = 1 + 1/m$ because we have set $\mathscr{r} = 1$. The MSEs of estimators $\widehat{f} = \widehat{r}/(1 + \widehat{r})$ of $\mathscr{f}_m = \mathscr{r}_m/(1 + \mathscr{r}_m)$ are shown in Figure A.10, in log scale. Clearly, the only estimator that is consistent, invariant to parametrization and robust against δ is our proposal $\widehat{f}^{\diamondsuit}_{\mathrm{L}} = \widehat{r}^{\diamondsuit}_{\mathrm{L}}/(1 + \widehat{r}^{\diamondsuit}_{\mathrm{L}})$. It concentrates at the true value $\mathscr{f}_m$ quite closely even for small m and n. 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 a large MSE when $\delta \neq 0$. It explains why L-1 is not powerful.

Table A.6: 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 parametrization-invariant tests. 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

A.7 Supplements for Section 4.2

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

$$\left[X_{i1} \mid \beta_1, \tau_1^2\right] \sim \mathscr{N}(\beta_1, \tau_1^2) \qquad \text{and} \qquad \left[X_{ij} \mid X_{i,1:(j-1)}, \beta_j, \tau_j^2\right] \sim \mathscr{N}(\beta_j^{\intercal} Z_{ij}, \tau_j^2),$$

for $j = 2, \ldots, p$, 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 respectively by

$$\widehat{\beta}_j = (Z_j^{\intercal} Z_j)^{-1} Z_j^{\intercal} W_j \quad \text{and} \quad \widehat{\tau}_j^2 = \frac{(W_j - Z_j\widehat{\beta}_j)^{\intercal}(W_j - Z_j\widehat{\beta}_j)}{n_j - j},$$

where $Z_j = (Z_{1j}, \ldots, Z_{n_j j})^{\intercal}$ and $W_j = (X_{1j}, \ldots, X_{n_j j})^{\intercal}$.

We assume a Bayesian imputation 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 $\ell = 1, \ldots, m$, denote the ℓth imputed data set by $X^{(\ell)}$, whose (i, j)th element is $X_{ij}^{(\ell)}$. If $1 \leqslant j \leqslant p$ and $i \leqslant n_j$, then $X_{ij}^{(\ell)} = X_{ij}$, otherwise $X_{ij}^{(\ell)}$ is filled in by recursing the following steps for $j = 2, \ldots, p$.

1. Draw a sample $(\tau_j^{(\ell)})^2$ from $\widehat{\tau}_j^2(n_j - j)/\chi^2_{n_j - j}$.

2. Draw a sample $\beta_j^{(\ell)}$ from $\mathcal{N}_j(\widehat{\beta}_j, (\tau_j^{(\ell)})^2(Z_j^\intercal Z_j)^{-1})$.
3. Draw a sample $X_{ij}^{(\ell)}$ from $\mathcal{N}((\beta_j^{(\ell)})^\intercal Z_{ij}^{(\ell)}, (\tau_j^{(\ell)})^2)$ for $i = n_j+1, \ldots, n$, where $Z_{ij}^{(\ell)} = (1, (X_{i,1:(j-1)}^{(\ell)})^\intercal)^\intercal$.

With the ℓth imputed data set, 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 $\widehat{\Sigma}^{(\ell)} = (X^{(\ell)} - \widehat{\mu}^{(\ell)})^\intercal(X^{(\ell)} - \widehat{\mu}^{(\ell)})/n$.

The partial result is shown in Figure 3 of the main text, whereas the full version is shown in Figure A.11.

A.8 Applications to a Care-Survival Data

Meng and Rubin (1992) considered the data given in Table A.7, where i, j and k index, respectively, amount of parental care (less or more, corresponding to $i = 1, 2$), and survival status (died or survived, corresponding to $j = 1, 2$), and clinic (A or B, corresponding to $k = 1, 2$). The label k is missing for some observations. The missing mechanism was assumed to be ignorable. We consider two null hypotheses: (H_0) the clinic and parental care are conditionally independent given the survival status, and (H_0') all three variables are independent. It is remarked that testing the conditional independence model (i.e., H_0) is useful from a modeling perceptive. If H_0 cannot be rejected, then one may be tempted to adopt the more parsimonious null model (for the cell probabilities). The same model is also suggested in Little and Rubin (2002) and Meng and Rubin (1992).

Our aim is to investigate the impact on $\{\widetilde{D}_{\mathrm{S}}, \widehat{D}_{\mathrm{S}}^+, \widehat{D}_{\mathrm{S}}^\lozenge\}$ by the parametrization of the cell probabilities

$$\pi_{ijk} = \mathsf{P}(\text{parental care} = i, \text{survival status} = j, \text{clinic label} = k)$$

for $i, j, k \in \{1, 2\}$; and the impact on $\{\widetilde{r}_{\mathrm{L}}, \widehat{r}_{\mathrm{S}}^+, \widehat{r}_{\mathrm{S}}^\lozenge\}$ under different null hypotheses. Here the full model parameter vector can be expressed as $\psi = (\pi_{111}, \pi_{112}, \pi_{121}, \pi_{122}, \pi_{211}, \pi_{212}, \pi_{221})^\intercal$. Since the restrictions imposed by H_0 are $\pi_{ijk} = (\pi_{1jk} + \pi_{2jk})(\pi_{ij1} + \pi_{ij2})$ for $j = 1, 2$, one may express the parameter of interest as $\theta = (\theta_1, \theta_2)^\intercal$, where $\theta_j = \pi_{ijk} - (\pi_{1jk} + \pi_{2jk})(\pi_{ij1} + \pi_{ij2})$ for

$j = 1, 2$. Then H_0 can be equivalently stated as $\theta = \theta_0$, where $\theta_0 = (0, 0)^\intercal$. Similarly, the parameter of interest under H_0' can be defined.

Table A.7: Data from Meng and Rubin (1992). The notation "?" indicates missing label.

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

The computation of the stacked MI estimators of $\{\pi_{ijk}\}$ is presented in A.8 of the Appendix. We consider three parametrizations: (i) $\psi_{ijk} = \pi_{ijk}$; (ii) $\psi_{ijk} = \log\{\pi_{ijk}/(1 - \pi_{ijk})\}$; and (iii) $\psi_{ij1} = \pi_{ij1}$ and $\psi_{ij2} = \pi_{ij2}/\pi_{ij1}$. Denote the p-values of tests $\{\widetilde{D}_{\mathrm{L}}, \widehat{D}_{\mathrm{S}}^{+}, \widehat{D}_{\mathrm{S}}^{\diamondsuit}\}$ by $\{\widetilde{p}_{\mathrm{L}}, \widehat{p}_{\mathrm{S}}^{+}, \widehat{p}_{\mathrm{S}}^{\diamondsuit}\}$, respectively. The results are summarized in Table A.8. Clearly, only $\widehat{r}_{\mathrm{S}}, \widehat{r}_{\diamondsuit}, \widehat{D}_{\mathrm{S}}^{+}, \widehat{D}_{\mathrm{S}}^{\diamondsuit}$ are always non-negative and parametrization-invariant. Some of the values of $\widetilde{r}_{\mathrm{L}}$ and $\widetilde{D}_{\mathrm{L}}$ are negative, leading to the meaningless $\widetilde{p}_{\mathrm{L}} = 1$. For testing H_0, we have $\widehat{D}_{\mathrm{S}}^{+} \approx \widehat{D}_{\mathrm{S}}^{\diamondsuit}$. For testing H_0', $\widehat{D}_{\mathrm{S}}^{+}$ and $\widehat{D}_{\mathrm{S}}^{\diamondsuit}$ are not very close to each other, but they both lead to essentially zero p-value. These results reconfirm the conclusions in Meng and Rubin (1992). Moreover, only $\widehat{r}_{\mathrm{S}}^{\diamondsuit}$ does not change under different null hypotheses.

The MI data sets are generated from a Bayesian model in Section 4.2 of Meng and Rubin (1992). The ℓth imputed log-likelihood function is $\log f(X^{(\ell)} \mid \pi) = \sum_c n_c^{(\ell)} \log \pi_c$, where $X^{(\ell)}$ are the cell counts $n_c^{(\ell)}$ in the ℓth imputed data set. Hence the unconstrained MLE of π_c is $\widehat{\pi}_c^{(\ell)} = n_c^{(\ell)}/n_+^{(\ell)}$, where $n_+^{(\ell)} = \sum_c n_c^{(\ell)}$. Let $n_c^+ = \sum_{\ell=1}^m n_c^{(\ell)}$. Consequently, the joint log-likelihood based on the stacked data is

$$\log f(X^{\mathrm{S}} \mid \pi) = \sum_{\ell=1}^{m} \sum_{c} n_c^{(\ell)} \log \pi_c = \sum_{c} n_c^{+} \log \pi_c, \tag{A.14}$$

Thus the unconstrained MLE with respect to (A.14) 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.

Table A.8: The LRTs using $\widetilde{D}_{\mathrm{L}}$, $\widehat{D}_{\mathrm{S}}^{+}$ and $\widehat{D}_{\mathrm{S}}^{\Diamond}$ under different parametrizations in Section A.8.

	Parametrization (i): identity map					
	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): logit transformation					
	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): ratios of probabilities					
	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

B Proofs

Proof of Theorem 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 $\mathrm{Cov}(\widehat{\theta}^{(\ell)}, \widehat{\eta}^{(\ell)}) \doteq \mathbf{0}$ by taking suitable linear transformation of ψ. Also write $U_{\eta}^{(\ell)}$ as an efficient estimator of $\mathrm{Var}(\widehat{\eta})$ based on $X^{(\ell)}$; and recall that $U_\theta^{(\ell)} = U^{(\ell)}$ is an efficient estimator of $\mathrm{Var}(\widehat{\theta})$ based on $X^{(\ell)}$.

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 $\psi \simeq \widehat{\psi}^*$,

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

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

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

with $\overline{U}_\eta = m^{-1}\sum_{i=1}^m U_\eta^{(\ell)}$. Under the null, $\widehat{\psi}^* \simeq \widehat{\psi}_0^*$. So, using (B.1), we have

$$\begin{aligned} \widehat{d}_{\mathrm{L}} &\simeq \left(\widehat{\psi}_0^* - \widehat{\psi}^*\right)^\intercal \overline{I}(\widehat{\psi}^*)\left(\widehat{\psi}_0^* - \widehat{\psi}^*\right), \\ &\simeq \begin{pmatrix} \theta_0 - \widehat{\theta}^* \\ \widehat{\eta}(\theta_0) - \widehat{\eta}(\widehat{\theta}^*) \end{pmatrix}^\intercal \begin{pmatrix} \overline{U}_\theta^{-1} & \mathbf{0} \\ \mathbf{0} & \overline{U}_\eta^{-1} \end{pmatrix} \begin{pmatrix} \theta_0 - \widehat{\theta}^* \\ \widehat{\eta}(\theta_0) - \widehat{\eta}(\widehat{\theta}^*) \end{pmatrix} \\ &\simeq (\overline{\theta}^\intercal - \theta_0)\overline{U}_\theta^{-1}(\overline{\theta}^\intercal - \theta_0) = \widetilde{d}'_{\mathrm{W}}, \end{aligned} \tag{B.3}$$

where we have used (a) $\widehat{\theta}^* \simeq \overline{\theta}$; 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 $\widetilde{d}'_{\mathrm{W}} \simeq \widetilde{d}_{\mathrm{L}}$ (Meng and Rubin, 1992), we have $\widehat{d}_{\mathrm{L}} \simeq \widetilde{d}_{\mathrm{L}}$. □

Proof of Proposition 1. The given condition implies that

$$\begin{aligned} \widehat{\psi}^{(\ell)} &= ((\widehat{\theta}^{(\ell)})^\intercal, (\widehat{\eta}^{(\ell)})^\intercal)^\intercal, \qquad & \widehat{\psi}_0^{(\ell)} &= (\theta_0^\intercal, (\widehat{\eta}^{(\ell)})^\intercal)^\intercal, \\ \widehat{\psi}^* &= ((\widehat{\theta}^*)^\intercal, (\widehat{\eta}^*)^\intercal)^\intercal, & \widehat{\psi}_0^* &= (\theta_0^\intercal, (\widehat{\eta}^*)^\intercal)^\intercal. \end{aligned}$$

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 1. 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 $I^{(\ell)}(\breve{\psi}^{(\ell)}) \simeq I^{(\ell)}(\breve{\psi}^{*})$ for all ℓ. Then, using similar techniques as in (B.2) and (B.3), we have

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

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

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

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

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

as $m, n \to \infty$, where $\mathscr{U}_{\theta,0}$ is a deterministic matrix that depends on both θ_0 and the true value of θ, and satisfies $n(\breve{U} - \mathscr{U}_{\theta,0}) \xrightarrow{\mathrm{pr}} 0$. Note that $\mathrm{tr}(\mathscr{U}_{\theta,0}^{-1}\mathscr{B}_{\theta}) = k\mathscr{r}_0$, for some finite $\mathscr{r}_0$ by Assumption 2. Then $\widehat{r}_{\mathrm{L}} \xrightarrow{\mathrm{pr}} \mathscr{r}_0 = \mathrm{tr}(\mathscr{U}_{\theta,0}^{-1}\mathscr{B}_{\theta})/k$, proving (ii). (But $\mathscr{U}_{\theta,0}$ may not equal to $\mathscr{U}_{\theta}$, and hence $\widehat{r}_{\mathrm{L}}$ may not be consistent for $\mathscr{r}_m$.)

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

Proof of Theorem 2. (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

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

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

Proof of Lemma 1. 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 $[\widehat{\theta}_{\mathrm{obs}} \mid \theta] \overset{\mathscr{D}}{\approx} \mathscr{N}_k(\theta, \mathscr{T}_\theta)$, 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 $n \to \infty$,

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

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

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

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

where $Z_1, \ldots, Z_m, W \overset{\mathrm{iid}}{\sim} \mathscr{N}_k(0, I_k)$. 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 $\overline{\theta} \overset{\mathscr{D}}{\approx}$

$\widehat{\theta}_{\text{obs}} + \mathscr{B}_\theta^{1/2}\overline{Z}_\bullet$, where $\overline{Z}_\bullet = m^{-1}\sum_{\ell=1}^m Z_\ell$. Since $\mathscr{B}_\theta = \mathscr{r}\mathscr{U}_\theta$, we have

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

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

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

After some simple algebra, we obtain

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

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 $\widehat{D}_{\text{L}}^+$ to

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

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 (2.12). $\square$

Proof of Theorem 3. Similar to (B.9) and (B.10), we can have a more general representation:

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

where $Z_1, \ldots, Z_h, W \overset{\text{iid}}{\sim} \mathscr{N}_h(0, I_h)$. 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

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

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

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

which is equivalent to (2.13). Note that it is true under both H_0 and H_1. $\square$

Proof of Theorem 4. From the representations of $\widehat{d}_{\mathrm{L}}^{\diamondsuit}$ and $\widehat{r}_{\mathrm{L}}^{\diamondsuit}$ in Lemma 1 and Theorem 3, we know that they are asymptotically ($n \to \infty$) independent. The proof then follows the derivation for Lemma 1. $\square$

Proof of Theorem A.1. (i) Using the representation (A.3), we can easily see that $\widehat{r}_{\mathrm{L}}^{\triangle} \geqslant 0$. (ii) It suffices to show

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

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

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

Under the parametrization of ψ in the proof of Theorem 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 $U^{(\ell)}$, we have $\left(U^{(\ell)}\right)^{-1} \simeq \overline{U}^{-1}$ and

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

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 A.1. (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

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

which, together with condition (b), imply that

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

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,

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

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 $\nabla\underline{\overline{L}}(\widehat{\psi}^{\mathrm{S}}) = O_p(1/n)$. Also, by the definition of MLE, we have $\nabla\underline{\overline{L}}(\widehat{\psi}^{*}) = \mathbf{0}$.

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

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

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

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

Similar to (B.14), we have

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

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

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

Now consider two cases.

(i) Under H_0, we have $\widehat{d}_{\mathrm{L}} = O_p(1)$ and $\widehat{\psi}_0^{\mathrm{S}} \simeq \widehat{\psi}^{\mathrm{S}}$. 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 $\widehat{d}_{\mathrm{L}} \simeq \widehat{d}_{\mathrm{S}}$ and $\overline{d}_{\mathrm{L}} = \overline{d}_{\mathrm{S}}$, we know $\widehat{r}_{\mathrm{L}} \simeq \widehat{r}_{\mathrm{S}}$. □

Note that, even under the assumption of this theorem, $\widehat{r}_{\mathrm{S}}$ and $\widehat{r}_{\mathrm{S}}^{\diamondsuit}$ are not equivalent. From (A.7) and (A.8), $\widehat{r}_{\mathrm{S}}$ 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 $\widehat{r}_{\mathrm{S}}^{\diamondsuit}$.

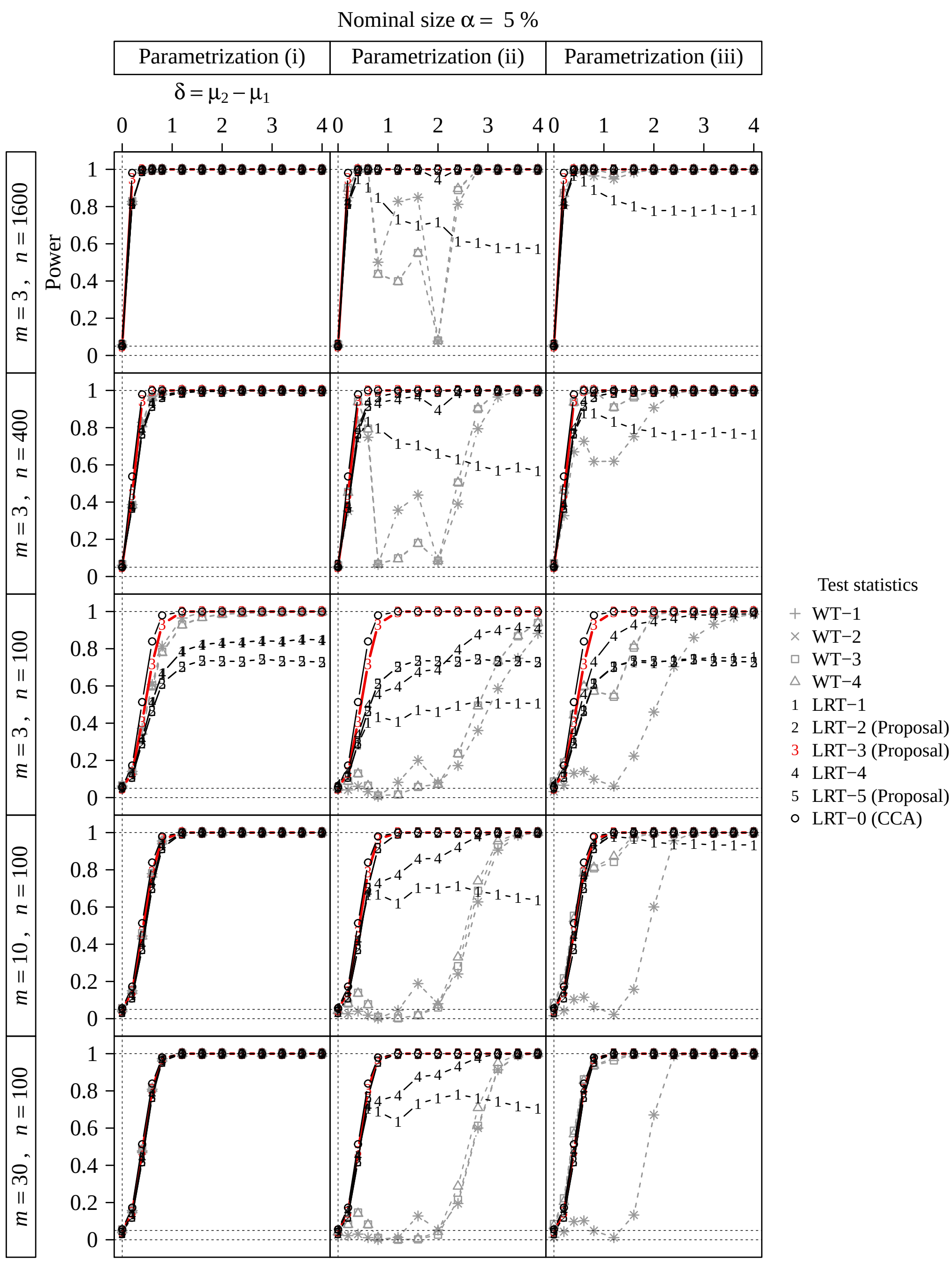

Figure A.7: The power curves under different parametrizations. The nominal size is $\alpha = 5\%$. In each plot, the vertical axis denotes the power, whereas the horizontal axis denotes the value of $\delta = \mu_2 - \mu_1$. The legend in Figure A.5 also applies here.

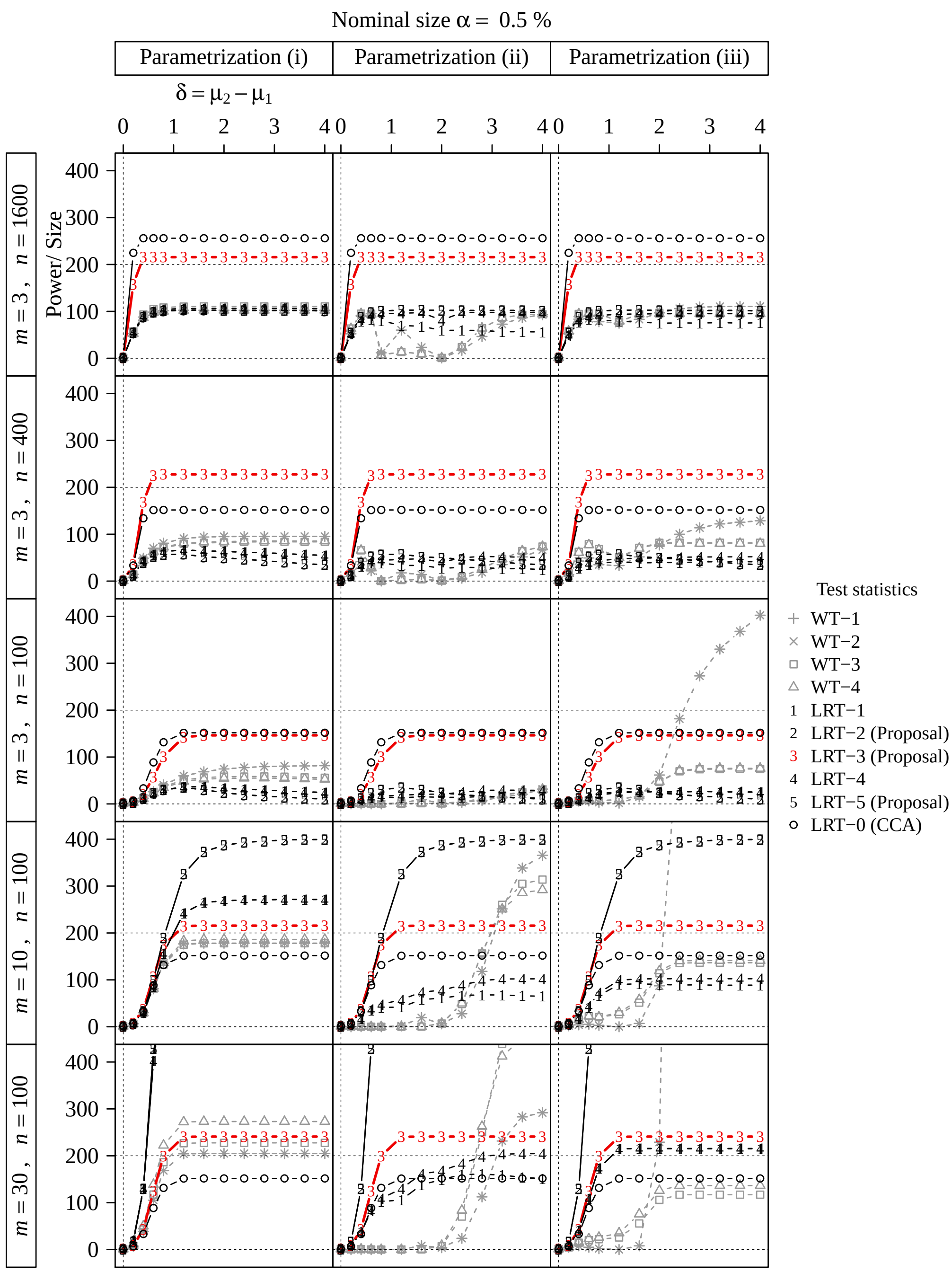

Figure A.8: The ratios of empirical power to empirical size under different parametrizations. The nominal size is $\alpha = 0.5\%$. In each plot, the vertical axis denotes the ratio, and the horizontal axis denotes $\delta = \mu_2 - \mu_1$. The legend in Figure A.5 also applies here. The results under nominal size 5% are shown in Figure A.9.

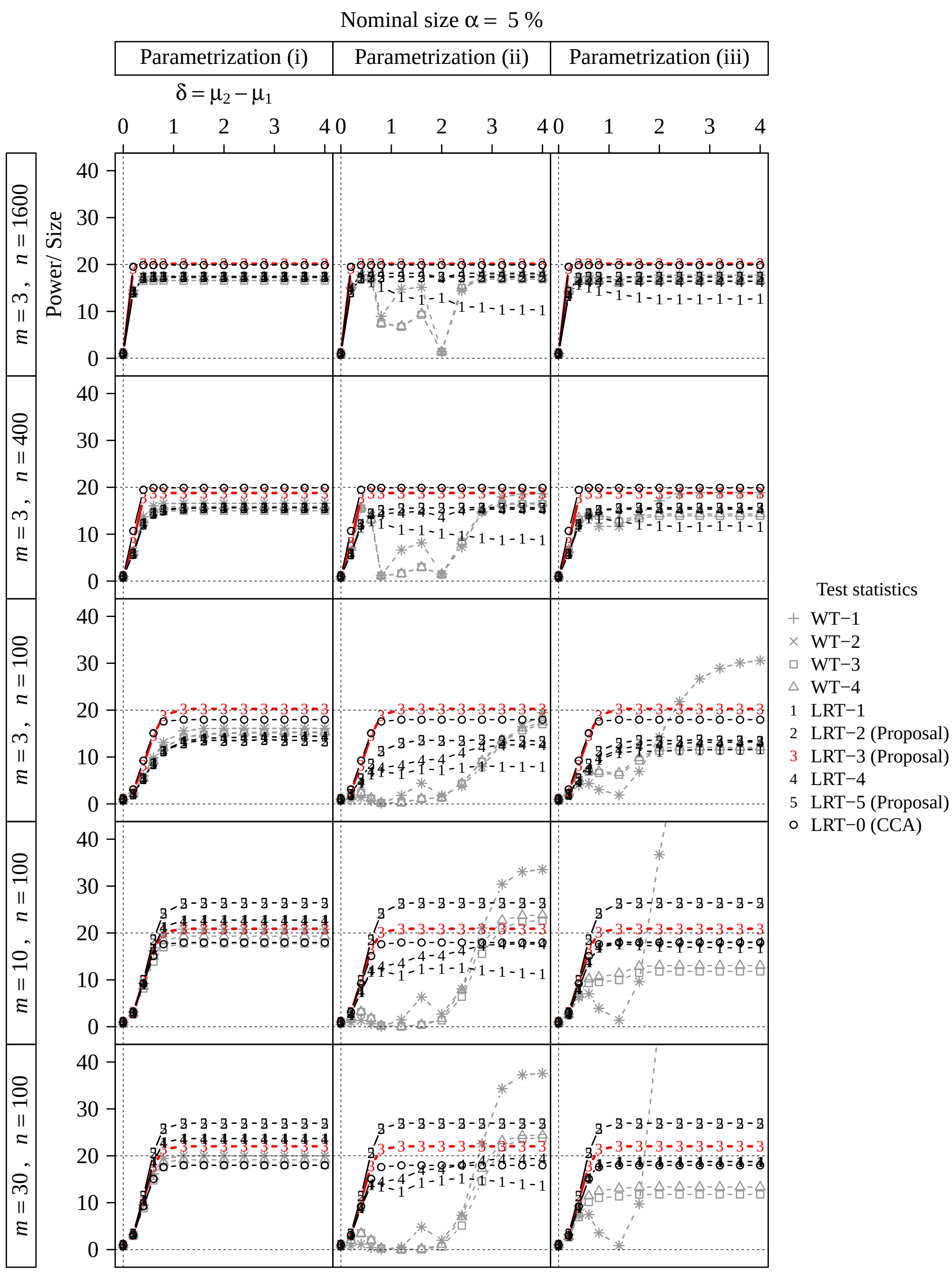

Figure A.9: The ratios of empirical power to empirical size under different parametrizations. The nominal size is $\alpha = 5\%$. In each plot, the vertical axis denotes the ratio, whereas the horizontal axis denotes $\delta = \mu_2 - \mu_1$. The legend in Figure A.5 also applies here.

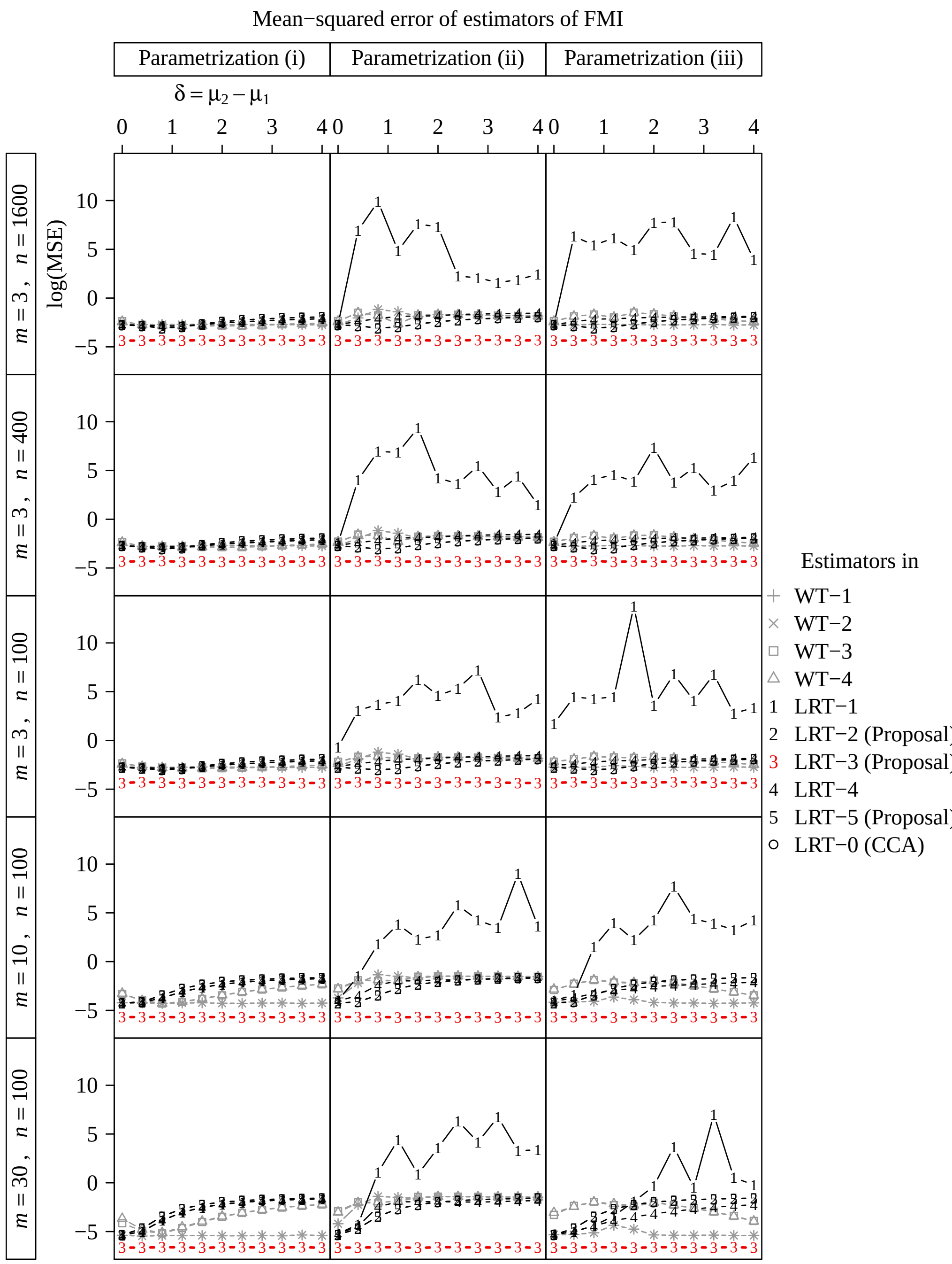

Figure A.10: The MSEs of estimators of ℓ_m used in the test statistics. The vertical axis denotes the log of MSE, whereas the horizontal axis denotes the value of $\delta = \mu_2 - \mu_1$. The legend in Figure A.5 also applies here.

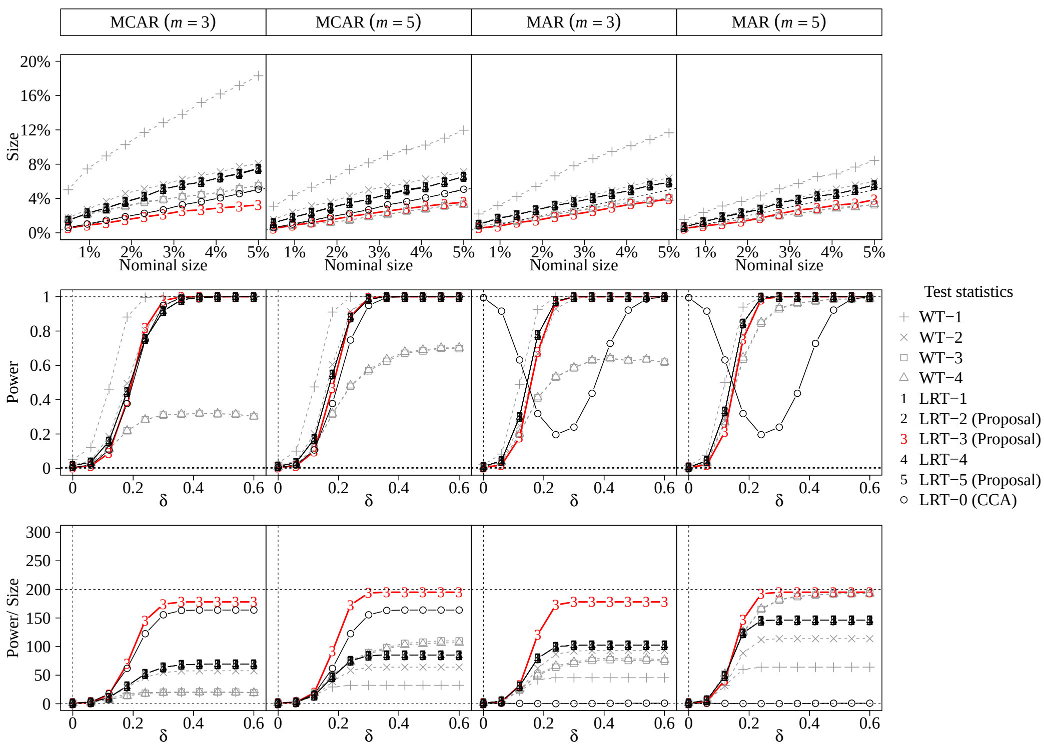

Figure A.11: The empirical size, empirical power, and their ratio. The first row of plots show the empirical sizes. The size of the complete-case test (C2) under MAR is off the chart (always equals to one) because it is invalid. The second and third rows of plots show the powers and the power-to-size ratios, respectively, where the nominal size is 0.5%.

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.

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

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

Little, R. J. A. and Rubin, D. B. (2002) *Statistical analysis with missing data.* Wiley, New York.

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

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

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

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