AI Review Report

Paper: Model selection after multiple imputation: a deviance correction for AIC, BIC, and likelihood-ratio tests

Paper ID: 88

Review Service: Reviewer3.com

Generated: July 28, 2026 at 05:37 UTC

Round 1 - Completed

Paper version: v1

Submitted: July 01, 2026 at 21:20 UTC

Completed: July 01, 2026 at 21:25 UTC

reviewer3

Although the title explicitly promises a deviance correction for BIC, the manuscript focuses entirely on AIC and likelihood-ratio tests. The formulation of an MI-BIC is completely absent from the text. Applying the derived deviance correction to BIC is not trivial because the BIC penalty relies on the sample size, and defining the effective sample size for multiply imputed data is a recognized challenge in the missing data literature. The authors must explicitly formulate the corrected MI-BIC, discuss the appropriate sample size penalty in this context, and ideally include it in the simulation studies to support the title's claim.

reviewer3

The manuscript relies exclusively on a four-dimensional multivariate normal distribution for all empirical validations. While the authors acknowledge this limitation, the lack of evaluation for non-normal data (such as categorical or skewed continuous variables commonly modeled via Generalized Linear Models) leaves the practical generalizability of the design-imbalance term unclear. The authors should discuss how the derived trace penalty and design-imbalance terms are expected to behave under non-normal likelihoods, and whether the current findings can be safely generalized to standard GLM frameworks used in applied research.

reviewer3

The authors correctly identify that uncongeniality causes the proposed correction to overshoot and recommend diagnosing congeniality before applying the method. However, the manuscript does not provide, nor does it cite, any practical methods or established literature for how an applied researcher should actually perform this diagnosis. Given that strict congeniality is rarely achieved in applied settings, the authors should expand the discussion to include specific, actionable diagnostic checks or cite relevant literature that guides users on evaluating imputation model congeniality.

reviewer2

Throughout the simulation results, the manuscript uses the "±" symbol without defining the measure of dispersion in the main text. Although the figure captions mention 95 percent intervals, the text must explicitly state whether these values represent standard errors, standard deviations, or the half-width of a 95% confidence interval to ensure unambiguous interpretation of the precision.

reviewer2

The authors define the extremes of the candidate set (diagonal and saturated) but omit the definitions of the two intermediate nested models used in the selection-sweep design. Without knowing the specific constraints applied to these intermediate models (e.g., compound symmetry, autoregressive, block-diagonal), readers cannot fully interpret the selection rates or independently replicate the simulation. Please explicitly define all four models included in the candidate set.

reviewer2

The manuscript does not specify which of the six possible coordinate pairs in the four-variate normal distribution are assigned the correlations of 0.6, 0.3, and 0.5. Because the missingness mechanism depends specifically on the third and fourth coordinates, the exact placement of these correlations within the covariance matrix will alter the missing information structure and the resulting relative-increase-in-variance trace. Please explicitly state the full covariance matrix or specify which variables correspond to each correlation value to ensure the simulation can be reproduced from the text.

reviewer1

The authors justify the failure of the N=1000 cell to hit the analytic target by claiming the standard deviation of the per-repetition statistic grows with sample size due to heavy tails. If the estimand is an expectation (the mean bias) and the variance is finite, the standard error of the mean must decrease with sample size. If the underlying distribution has infinite variance, then the standard error bars presented in Figure 1 are mathematically invalid and standard asymptotic confidence intervals cannot be used. The authors must either clarify why the variance of this specific statistic scales positively with N, or re-evaluate the N=1000 cell using robust estimators or a larger repetition count to achieve convergence.

reviewer1

The manuscript claims to validate the trace correction for proper multiple imputation, but the simulations rely exclusively on an approximately Bayesian bootstrap expectation-maximization approach rather than exact data augmentation. Because the theoretical derivations for proper MI depend on exact posterior draws, the lack of a fully Bayesian Markov chain Monte Carlo (MCMC) control experiment undermines the claim that the correction holds for proper MI, especially under high missingness. I recommend running the headline simulation studies using a standard data augmentation engine to confirm that the analytic results hold without the artifacts of the bootstrap approximation.

reviewer1

In the selection-sweep design, the candidate models are nested, meaning that model complexity (parameter count) is perfectly confounded with the amount of missing information. Consequently, it is impossible to determine whether the uncorrected MI-AIC misclassifies toward the saturated model specifically because of the missing information bias, or simply because the uncorrected deviance generally favors more complex models when under-penalized. To logically support the claim that the bias is driven by missing information, the authors should include a control experiment comparing non-nested candidate models that have the same number of parameters but differ significantly in their fractions of missing information.

proof-verifier

The abstract claims generally that under congenial proper multiple imputation, the averaged log-likelihood overstates its complete-data counterpart by one half the trace of the relative-increase-in-variance matrix. However, Theorem 5.1 explicitly restricts this result to "a model that estimates a scale or covariance." This condition is strictly necessary because, as shown in Equation 14 and discussed in Section 7, the bias for a known-scale model under proper MI is actually $-1/2 \operatorname{tr}(\text{RIV})$, which has the opposite sign. The abstract's formulation implies the $+1/2 \operatorname{tr}(\text{RIV})$ bias is a universal property of proper MI, whereas the proof establishes it only for the estimated-scale regime.

proof-verifier

While the abstract states that adding one trace term per candidate removes the deviance bias in expectation, Theorem 5.1 establishes that the deviance bias actually consists of two components: the trace term and a design-imbalance term $(A) + (C)$ that is $O(1)$ under missing-at-random data. The proposed correction in Equation 21 only offsets the trace component. As shown in Proposition 5.4, the uncorrected design-imbalance term remains and becomes a genuine $O(1)$ differential bias when comparing candidates with separated pseudo-true values. Readers will wonder if the full bias is eliminated, so the claim should be scoped to clarify that the correction removes only the trace component.

contradictions-reviewer

The text defines α as $-\mathcal{E}'(\theta_0)$ and describes it as 'the gradient of the conditional missing-data entropy'. If $\mathcal{E}$ represents the entropy, its gradient is $\mathcal{E}'(\theta_0)$, making α the negative gradient, not the gradient itself.

contradictions-reviewer

The main text states that the Kolmogorov-Smirnov distance falls to 0.019 under equating. However, Figure 8 displays the value 0.018 above the equating bar.

contradictions-reviewer

The symbol k is initially defined as the number of parameters (e.g., $k = p + \frac{1}{2}p(p+1)$). However, in Section 5's information-criterion selection, k is redefined as a model index in $\text{AIC}_{MI}(k)$, and q_k is used instead for the parameter count.

contradictions-reviewer

The manuscript uses both m and M to denote the number of imputations. Section 3 introduces $m \to \infty$, but Section 5 uses $M \to \infty$ in Equation 8, and Section 6 refers to $M \in \{20, 200\}$ while other passages refer to 'finite-m' or 'infinite-m' limits.

Round 2 - Approved

Paper version: v2
Submitted: July 10, 2026 at 20:43 UTC
Completed: July 10, 2026 at 20:43 UTC

Comment 1

Score: 4/4 - Fully addressed

Original comment:
Although the title explicitly promises a deviance correction for BIC, the manuscript focuses entirely on AIC and likelihood-ratio tests. The formulation of an MI-BIC is completely absent from the text. Applying the derived deviance correction to BIC is not trivial because the BIC penalty relies on the sample size, and defining the effective sample size for multiply imputed data is a recognized challenge in the missing data literature. The authors must explicitly formulate the corrected MI-BIC, discuss the appropriate sample size penalty in this context, and ideally include it in the simulation studies to support the title's claim.

Author response:
We formulated the corrected MI-BIC and added it to the same simulation grid the paper already uses for MI-AIC (60 cells, ?= 2,000 repetitions) [...] We added this as a new Discussion scope-limit paragraph (G8) rather than a full new BIC results section, since it is an honest qualification of an existing claim, not a new headline.

Reviewer assessment:
The author added a new Discussion section (G8) that explicitly formulates the MI-BIC using the complete-data sample size penalty. They also included new simulation results for BIC at various sample sizes (N=50, 100, 200), demonstrating that the correction is effective even when the BIC penalty is not yet dominant.

Comment 2

Score: 4/4 - Fully addressed

Original comment:
The manuscript relies exclusively on a four-dimensional multivariate normal distribution for all empirical validations. While the authors acknowledge this limitation, the lack of evaluation for non-normal data (such as categorical or skewed continuous variables commonly modeled via Generalized Linear Models) leaves the practical generalizability of the

design-imbalance term unclear. The authors should discuss how the derived trace penalty and design-imbalance terms are expected to behave under non-normal likelihoods, and whether the current findings can be safely generalized to standard GLM frameworks used in applied research.

Author response:

We ran checks outside the multivariate normal family rather than leaving this as a prediction-only paragraph [...] Manuscript change landed: Discussion paragraph G9 rewritten to the corrected, chronology-first account; the repository?s decision log records the correction and supersedes the earlier entry.

Reviewer assessment:

The author added a new Discussion section (G9) reporting simulation results for Bernoulli, Poisson, and Student-t models. They provided a derivation for canonical-link GLMs showing that the trace penalty remains valid, while also identifying specific conditions (entropy-gradient behavior) that determine the bias behavior in non-normal families.

Comment 3

Score: 4/4 - Fully addressed

Original comment:

The authors correctly identify that uncongeniality causes the proposed correction to overshoot and recommend diagnosing congeniality before applying the method. However, the manuscript does not provide, nor does it cite, any practical methods or established literature for how an applied researcher should actually perform this diagnosis. Given that strict congeniality is rarely achieved in applied settings, the authors should expand the discussion to include specific, actionable diagnostic checks or cite relevant literature that guides users on evaluating imputation model congeniality.

Author response:

The congeniality-limit paragraph (G7) now closes with three sentences of practical pointers, each backed by a reference acquired and read in full for this response, with the acquisition and claim-verification trail recorded in the repository

Reviewer assessment:

The revised Discussion section G7 now includes specific pointers to diagnostic methods, such as posterior predictive checks and propensity-conditional comparisons. The author cited three relevant papers (He & Zaslavsky, Bondarenko & Raghunathan, and Nguyen et al.) to guide applied researchers in evaluating model congeniality.

Comment 4

Score: 4/4 - Fully addressed

Original comment:

Throughout the simulation results, the manuscript uses the "±" symbol without defining the measure of dispersion in the main text. Although the figure captions mention 95 percent intervals, the text must explicitly state whether these values represent standard errors, standard deviations, or the half-width of a 95% confidence interval to ensure unambiguous interpretation of the precision.

Author response:

Manuscript change landed: the theorem-validation results paragraph now states at first use that its ± is one Monte Carlo standard error and explicitly distinguishes it from the figure?s 95 percent whiskers.

Reviewer assessment:

The author updated the text in the 'Theorem validation' results section to explicitly define the '±' symbol as representing one Monte Carlo standard error. They also clarified the distinction between this measure and the 95% intervals shown in the figures.

Comment 5

Score: 4/4 - Fully addressed

Original comment:

The authors define the extremes of the candidate set (diagonal and saturated) but omit the definitions of the two intermediate nested models used in the selection-sweep design. Without knowing the specific constraints applied to

these intermediate models (e.g., compound symmetry, autoregressive, block-diagonal), readers cannot fully interpret the selection rates or independently replicate the simulation. Please explicitly define all four models included in the candidate set.

Author response:

Manuscript change landed: the selection-sweep design paragraph now names all four candidates from the actual sweep code: the diagonal model (eight free parameters), a model that frees the wrong off-diagonal pair ?12, ?34 (ten free parameters), a model that frees the correct pair ?13, ?24 matching the truth (ten free parameters, the true model), and the saturated model (fourteen free parameters).

Reviewer assessment:

The revised 'Selection-sweep design' paragraph now provides explicit definitions and parameter counts for all four models in the candidate set. This change ensures the simulation design is transparent and reproducible.

Comment 6

Score: 4/4 - Fully addressed

Original comment:

The manuscript does not specify which of the six possible coordinate pairs in the four-variate normal distribution are assigned the correlations of 0.6, 0.3, and 0.5. Because the missingness mechanism depends specifically on the third and fourth coordinates, the exact placement of these correlations within the covariance matrix will alter the missing information structure and the resulting relative-increase-in-variance trace. Please explicitly state the full covariance matrix or specify which variables correspond to each correlation value to ensure the simulation can be reproduced from the text.

Author response:

Manuscript change landed: the simulation-design paragraph now states the placement from the covariance matrix the simulation code actually uses: ?12 = 0.6 between the two coordinates that carry missingness, ?13 = ?24 = 0.3 between each missing coordinate and the fully observed coordinate that drives its own missingness, ?34 = 0.5 between the two fully observed driving coordinates, and ?14 = ?23 = 0.

Reviewer assessment:

The author updated the 'Theorem-validation design' paragraph to specify the exact placement of correlations within the covariance matrix. This provides the necessary detail for independent replication of the missing information structure.

Comment 7

Score: 4/4 - Fully addressed

Original comment:

The authors justify the failure of the N=1000 cell to hit the analytic target by claiming the standard deviation of the per-repetition statistic grows with sample size due to heavy tails. If the estimand is an expectation (the mean bias) and the variance is finite, the standard error of the mean must decrease with sample size. If the underlying distribution has infinite variance, then the standard error bars presented in Figure 1 are mathematically invalid and standard asymptotic confidence intervals cannot be used. The authors must either clarify why the variance of this specific statistic scales positively with N, or re-evaluate the N=1000 cell using robust estimators or a larger repetition count to achieve convergence.

Author response:

Manuscript change landed: the results paragraph reporting the pooled theorem-validation estimate now states this explicitly ? the repetition count (1,000) is named as identical across every sample size, and the per-replicate spread (not the Monte Carlo standard error) is what grows with ?, and that spread is finite.

Reviewer assessment:

The author clarified in the text that the per-replicate spread of the statistic grows with N while remaining finite, which explains why the standard error bars widen at larger sample sizes given a fixed repetition count. This explanation resolves the perceived mathematical contradiction regarding the validity of the confidence intervals.

Comment 8

Score: 4/4 - Fully addressed

Original comment:

The manuscript claims to validate the trace correction for proper multiple imputation, but the simulations rely exclusively on an approximately Bayesian bootstrap expectation-maximization approach rather than exact data augmentation. Because the theoretical derivations for proper MI depend on exact posterior draws, the lack of a fully Bayesian Markov chain Monte Carlo (MCMC) control experiment undermines the claim that the correction holds for proper MI, especially under high missingness. I recommend running the headline simulation studies using a standard data augmentation engine to confirm that the analytic results hold without the artifacts of the bootstrap approximation.

Author response:

We ran the requested control. Two independent exact data-augmentation engines were built [...] The control surfaces a genuine, and now fully characterized, engine difference. [...] Changes already landed in the sourced-derivation companion document [...] and a new scope-and-correction callout following the theorem?s setup, stating the theorem?s scope (second-moment-matched, regular-draw imputers) and the disclosed correction for engines outside that scope

Reviewer assessment:

The author conducted the requested MCMC control experiments and identified a persistent O(1) bias difference between bootstrap and genuine posterior-draw imputation. They updated the manuscript (Discussion G3 and the 'Proper imputation' section) to define the scope of the main theorem and provided the necessary corrections for MCMC-based imputation in the companion documentation.

Comment 9

Score: 4/4 - Fully addressed

Original comment:

In the selection-sweep design, the candidate models are nested, meaning that model complexity (parameter count) is perfectly confounded with the amount of missing information. Consequently, it is impossible to determine whether the uncorrected MI-AIC misclassifies toward the saturated model specifically because of the missing information bias, or simply because the uncorrected deviance generally favors more complex models when under-penalized. To logically support the claim that the bias is driven by missing information, the authors should include a control experiment comparing non-nested candidate models that have the same number of parameters but differ significantly in their fractions of missing information.

Author response:

The control the reviewer asks for is a separate study already in the paper: the non-nested measurement of §5. [...] No new work needed: the requested control already exists in the manuscript; this response points to it rather than proposing new simulations.

Reviewer assessment:

The author clarified that the existing non-nested study (comparing models with identical parameter counts across MAR and MCAR conditions) already addresses this confound. By isolating the design-imbalance term through an MCAR control, the paper demonstrates that missing information specifically drives the bias, independent of model complexity.

Comment 10

Score: 4/4 - Fully addressed

Original comment:

The abstract claims generally that under congenial proper multiple imputation, the averaged log-likelihood overstates its complete-data counterpart by one half the trace of the relative-increase-in-variance matrix. However, Theorem 5.1 explicitly restricts this result to "a model that estimates a scale or covariance." This condition is strictly necessary because, as shown in Equation 14 and discussed in Section 7, the bias for a known-scale model under proper MI is actually $-1/2 \operatorname{tr}(\text{RIV})$, which has the opposite sign. The abstract's formulation implies the $+1/2 \operatorname{tr}(\text{RIV})$ bias is a universal property of proper MI, whereas the proof establishes it only for the estimated-scale regime.

Author response:

Manuscript change landed: the abstract now reads ?under congenial proper multiple imputation with the complete-data maximum likelihood estimate as target and an estimated scale or covariance,? naming the scope condition where the headline bias previously stated it without qualification.

Reviewer assessment:

The author updated the abstract to include the 'estimated scale or covariance' qualifier, correctly aligning the summary with the scope of Theorem 5.1. This resolves the potential confusion regarding the sign of the bias in known-scale models.

Comment 11

Score: 4/4 - Fully addressed

Original comment:

While the abstract states that adding one trace term per candidate removes the deviance bias in expectation, Theorem 5.1 establishes that the deviance bias actually consists of two components: the trace term and a design-imbalance term $(A) + (C)$ that is $O(1)$ under missing-at-random data. The proposed correction in Equation 21 only offsets the trace component. As shown in Proposition 5.4, the uncorrected design-imbalance term remains and becomes a genuine $O(1)$ differential bias when comparing candidates with separated pseudo-true values. Readers will wonder if the full bias is eliminated, so the claim should be scoped to clarify that the correction removes only the trace component.

Author response:

Manuscript change landed: the abstract?s ?adding one trace term per candidate removes the deviance bias in expectation? now reads ?removes that trace component of the deviance bias in expectation,? consistent with the abstract?s own preceding sentence naming the separate design-imbalance term the correction does not touch.

Reviewer assessment:

The abstract was revised to clarify that the correction specifically removes the 'trace component' of the bias. This accurately reflects that the design-imbalance term remains uncorrected, as established in the derivations.