UNCERTAINTY-AWARE RANK TEST WITH MISSING DATA AND APPLICATION IN CAUSAL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Covariance-rank constraints are widely used in causal discovery, including settings with latent variables, but missing observations complicate their statistical testing. Pairwise covariance estimates use different, overlapping sets of rows; yet, plugging these estimates into a classical rank test with an effective sample size correction need not properly control Type-I error. In this paper, we propose the Missing-value uncertainty-Aware Rank Test (MART), which uses partially observed rows beyond those retained by test-wise deletion. MART consistently estimates the joint uncertainty of the pairwise covariance estimates and measures their weighted distance to the rank-constrained set. A Kronecker approximation makes this distance solvable by a whitened singular value decomposition, while the full estimated uncertainty calibrates the statistic through a weighted chi-square mixture as the asymptotic null distribution. For causal discovery with many rank test queries, the covariance and uncertainty estimates can be computed once and reused to reduce the computational cost. Experiments on both synthetic and real-world data validate our theoretical claims and the effectiveness of our method (our code will be available).

1 INTRODUCTION AND RELATED WORK

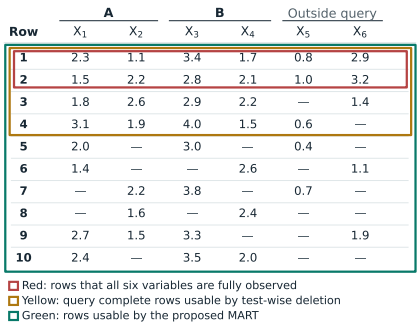
In causal discovery (Spirtes et al., 2000), covariance-rank constraints provide useful graphical information for identifying causal structure. In linear Gaussian models, Conditional Independences (CI) implied by d-separation (Pearl, 1988) correspond to zero partial covariances, which can be expressed as rank constraints on augmented covariance blocks (Sullivant et al., 2010; Di, 2009). More generally, a low-rank cross-covariance block can reveal a graphical bottleneck characterized by trek separation, including one involving latent variables (Sullivant et al., 2010; Spirtes, 2013). This information has been used to recover latent causal structure from observed variables (Huang et al., 2022; Dong et al., 2024; Dai et al., 2025). Appendix J discusses the related work in more detail.

Classical tests of cross-covariance rank are commonly formulated through canonical correlation analysis (CCA) computed from complete observations (Anderson, 1984; Robin & Smith, 2000). In practice, the variables needed for a rank query may be incompletely observed. Missing entries occur, for example, in financial panels (Bryzgalova et al., 2025), gene-expression data (Troyanskaya et al., 2001), electronic health records (Beaulieu-Jones et al., 2018), and psychological and educational research (Peugh & Enders, 2004). These settings motivate rank tests that can use information in incomplete records; here we study this problem under missing completely at random (MCAR).

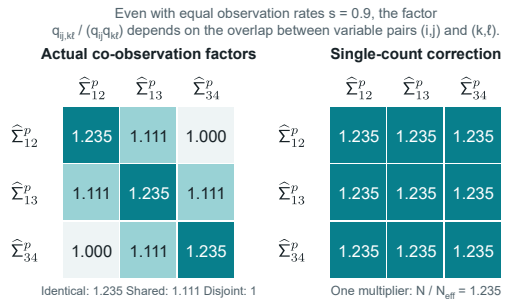
A common way to handle missing data is test-wise deletion: each rank test uses only rows with no missing values among the variables involved in that test (Strobl et al., 2018; Witte et al., 2022). When a query involves many variables, few rows may remain, even though excluded rows contain observed pairs that could contribute to covariance estimation (Figure 1(a)). Pairwise covariance estimation instead uses these additional rows by estimating each covariance from the rows in which its two variables are jointly observed (Xu et al., 2022). A tempting approach is then to insert the pairwise estimates into a classical rank statistic and replace its sample size with one effective count, analogous to the adjustments in (Marsh, 1998). Yet the estimates come from different, overlapping subsets of rows. Their joint sampling uncertainty cannot generally be represented by a single scalar, even when variables have the same missingness rate (Figure 1(b); Section 2.3).

To this end, we propose the Missing-value uncertainty-Aware Rank Test (MART). MART estimates each covariance using all rows in which the corresponding variable pair is observed, including

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Rows contributing to different test methods with query $\mathbf{A} = \{\mathsf{X}_1, \mathsf{X}_2\}$, $\mathbf{B} = \{\mathsf{X}_3, \mathsf{X}_4\}$.



(b) The co-observation factor matrix cannot be represented by a single scalar.

Figure 1: MART uses more observed data while accounting for missingness-induced uncertainty.

incomplete rows discarded by test-wise deletion. It then estimates the joint sampling uncertainty of these covariances under the missingness pattern and gives less weight to directions of rank violation with greater uncertainty. Our contributions are as follows.

- We introduce Missing-value uncertainty-Aware Rank Test (MART), a covariance-rank test that uses partially observed rows beyond those retained by test-wise deletion. Rank constraints include vanishing partial correlation as a special case and play a crucial role in causal discovery in fields like psychology and econometrics, where missing data are common. To the best of our knowledge, MART is the first asymptotically valid covariance-rank test for missing data that uses partially observed rows beyond test-wise deletion.
- We show why plugging pairwise covariance estimates into a classical rank test need not control Type-I error, even with an effective sample-size correction. The estimates draw on different, overlapping sets of rows, and the joint sampling uncertainty cannot generally be captured by one scalar. We consistently estimate this uncertainty (Thm. 3) and formulate the test statistic as a minimal weighted distance problem where the null violation directions with greater uncertainty have less weight.
- We theoretically show that the statistic's asymptotic null distribution is a weighted chi-square mixture (Thm. 5) and thus provides valid asymptotic Type-I error control. The statistic also has a likelihood-ratio interpretation under a fixed Gaussian working model (Prop. 1). Finally, in an explicit local-alternative setting, we prove that MART has strictly greater limiting power than a classical test-wise deletion rank test because it uses additional incomplete rows (Thm. 6).

2 Preliminaries

2.1 Problem Setting

Let $\mathbf{X} = (\mathsf{X}_1, \ldots, \mathsf{X}_M)^\mathsf{T} \in \mathbb{R}^M$ have mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. For two subvectors $\mathbf{A}, \mathbf{B} \subseteq \mathbf{X}$ of sizes a and b, respectively, write $\boldsymbol{\Sigma}_{\mathbf{A},\mathbf{B}} = \mathrm{Cov}(\mathbf{A}, \mathbf{B})$. The subvectors may overlap. Given a proposed rank r, our goal is to test

$$\mathcal{H}_0 : \mathrm{rank}(\boldsymbol{\Sigma}_{\mathbf{A},\mathbf{B}}) \le r \quad \text{against} \quad \mathcal{H}_1 : \mathrm{rank}(\boldsymbol{\Sigma}_{\mathbf{A},\mathbf{B}}) > r, \qquad 0 \le r < \min(a, b). \tag{1}$$

A population covariance matrix has a well-defined rank, but its sample estimate is typically full rank even when the null holds. A statistical test must distinguish a genuine rank violation of the null hypothesis from estimation noise.

We observe N i.i.d. copies $(\mathbf{X}_t, \mathbf{R}_t)$ of $(\mathbf{X}, \mathbf{R})$, where $\mathbf{R}_t = (\mathsf{R}_{t1}, \ldots, \mathsf{R}_{tM})^\mathsf{T} \in \{0, 1\}^M$ records which entries of row t are observed: $\mathsf{R}_{ti} = 1$ means X_{ti} is available. Equivalently, the observed data are $(\mathbf{R}_t, \mathbf{R}_t \odot \mathbf{X}_t)$, with the mask retained so that an unobserved entry is not confused with an observed zero. Our base assumption is as follows.

Assumption 1 (Missing completely at random (MCAR))**.** $\mathbf{R}$ *is independent of* $\mathbf{X}$*, and* $q_{ij} := \mathbb{P}(\mathsf{R}_i\mathsf{R}_j = 1) > 0$ *for every covariance pair needed by the test; in particular,* $q_{ii} = \mathbb{P}(\mathsf{R}_i = 1) > 0$.

The mask coordinates themselves may be dependent. MCAR assumption is essential as it makes the observed-pair covariance target the population $\boldsymbol{\Sigma}$. Under a general missing-at-random (MAR) mechanism, observation can depend on observed values (Rubin, 1976); unadjusted pairwise moments then need not target $\boldsymbol{\Sigma}$. A pair with $q_{ij} = 0$ cannot be estimated by the procedure studied here.

Notation. Random vectors use bold upright letters ($\mathbf{X}, \mathbf{R}$), scalar random variables use sans-serif letters ($\mathsf{X}_i, \mathsf{R}_{ti}$), and matrices use bold mathematical symbols ($\boldsymbol{\Sigma}, \boldsymbol{V}$). We use column-major vec throughout: $\mathrm{vec}(\boldsymbol{ECF}) = (\boldsymbol{F}^{\mathsf{T}} \otimes \boldsymbol{E})\,\mathrm{vec}(\boldsymbol{C})$. The tested block is $\boldsymbol{\Sigma}_{\mathbf{A},\mathbf{B}}$; its pairwise estimate, defined below, is $\widehat{\boldsymbol{\Sigma}}^{p}_{\mathbf{A},\mathbf{B}}$. Hats denote sample estimates. Table 2 collects the main symbols.

2.2 Why Covariance Rank Matters in Causal Discovery

CI provides a classical way to learn a causal graph. Its graphical counterpart is *d-separation*. Under the Markov and faithfulness assumptions, d-separation and CI agree (Pearl, 1988; Spirtes et al., 2000). For Gaussian variables, the CI $\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$ is equivalent to a zero partial cross-covariance. When the three sets are disjoint, a block-matrix identity gives

$$\mathrm{rank}(\boldsymbol{\Sigma}_{\mathbf{A}\cup\mathbf{C},\mathbf{B}\cup\mathbf{C}}) = |\mathbf{C}| + \mathrm{rank}\Big(\boldsymbol{\Sigma}_{\mathbf{A},\mathbf{B}} - \boldsymbol{\Sigma}_{\mathbf{A},\mathbf{C}}\boldsymbol{\Sigma}^{-1}_{\mathbf{C},\mathbf{C}}\boldsymbol{\Sigma}_{\mathbf{C},\mathbf{B}}\Big).$$

Thus the usual zero partial-correlation test is a special case of a covariance-rank test. To read rank constraints from a linear causal model, we assume *rank faithfulness*: special choices of model coefficients do not create extra rank deficiencies beyond those implied by the graph (Dong et al., 2024). Under this assumption, the graphical statement is as follows.

Theorem 1 (D-separation as a rank constraint)**.** *For disjoint nonempty variable sets $\mathbf{A}, \mathbf{B}$ and a disjoint set $\mathbf{C}$ with nonsingular $\boldsymbol{\Sigma}_{\mathbf{C},\mathbf{C}}$, $\mathbf{C}$ d-separates $\mathbf{A}$ from $\mathbf{B}$ in the causal graph if and only if* $\mathrm{rank}(\boldsymbol{\Sigma}_{\mathbf{A}\cup\mathbf{C},\mathbf{B}\cup\mathbf{C}}) = |\mathbf{C}|$ *(Sullivant et al., 2010).*

Rank constraints can also reveal structure when no observed-variable CI is available. Consider four measurements of a single unobserved factor: $\mathsf{X}_i = \lambda_i \mathsf{L} + \mathsf{E}_i$ for $i = 1, \ldots, 4$, where L and the E_i are mutually independent and $\mathrm{Var}(\mathsf{L}) > 0$. For $\mathbf{A} = (\mathsf{X}_1, \mathsf{X}_2)^{\mathsf{T}}$ and $\mathbf{B} = (\mathsf{X}_3, \mathsf{X}_4)^{\mathsf{T}}$,

$$\boldsymbol{\Sigma}_{\mathbf{A},\mathbf{B}} = \mathrm{Var}(\mathsf{L}) \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} (\lambda_3 \quad \lambda_4), \qquad \mathrm{rank}(\boldsymbol{\Sigma}_{\mathbf{A},\mathbf{B}}) = 1,$$

when both loading vectors are nonzero. Generically, the four observed variables are mutually dependent, and thus no CI can be revealed from data to inform the structure, but their 2×2 cross-block has a vanishing determinant. The low rank reflects that there exists at least one latent variable connecting the two groups.

More generally, rank constraints can reveal the existence and the cardinality of a graphical bottleneck even when it involves an unobserved variable. Such constraints reveal t-separations Sullivant et al. (2010) and help causal-discovery methods infer latent variable structures (Huang et al., 2022; Dong et al., 2024). More detailed discussion can be found in Appendix B.

2.3 Existing Rank Tests and the Missingness Problem

Complete-data Rank Test. The classical Gaussian rank test uses canonical correlation analysis (CCA) to calculate the statistic. Under $\mathcal{H}_0$, let $\widehat{\boldsymbol{\Sigma}}$ be the usual complete-data sample covariance, and let $\widehat{\mathsf{c}}_j$ denote the ordered singular values of $\widehat{\boldsymbol{\Sigma}}^{-1/2}_{\mathbf{A},\mathbf{A}}\widehat{\boldsymbol{\Sigma}}_{\mathbf{A},\mathbf{B}}\widehat{\boldsymbol{\Sigma}}^{-1/2}_{\mathbf{B},\mathbf{B}}$. A Bartlett-corrected statistic is

$$\mathsf{T}_{\mathrm{CCA}} = -\left(N - \frac{a+b+3}{2}\right) \sum_{j=r+1}^{\min(a,b)} \log(1 - \widehat{\mathsf{c}}_j^2), \tag{2}$$

with an asymptotic null distribution of $\chi^2_{(a-r)(b-r)}$ (Anderson, 1951; 1984).

Pairwise covariance estimation. For each (i, j), define the number of jointly observed rows $\mathsf{N}_{ij} = \sum_{t=1}^{N} \mathsf{R}_{ti}\mathsf{R}_{tj}$ and the observed fraction $\widehat{\mathsf{q}}_{ij} = \mathsf{N}_{ij}/N$. We estimate the mean of each variable from its available entries and then use each observed pair:

$$\widehat{\mu}_i = \frac{\sum_{t=1}^{N} \mathsf{R}_{ti}\mathsf{X}_{ti}}{\sum_{t=1}^{N} \mathsf{R}_{ti}}, \qquad \widehat{\Sigma}^{p}_{ij} = \frac{1}{\mathsf{N}_{ij}} \sum_{t=1}^{N} \mathsf{R}_{ti}\mathsf{R}_{tj}(\mathsf{X}_{ti} - \widehat{\mu}_i)(\mathsf{X}_{tj} - \widehat{\mu}_j). \tag{3}$$

The estimator requires $\mathsf{N}_{ij} > 0$ for every needed pair. Under Assumption 1, its population target is Σ_{ij}. Using $\mathsf{N}_{ij} - 1$ as the denominator instead changes the estimator by a smaller-order term when $q_{ij} > 0$; we use N_{ij} for its convenient moment equations (Xu et al., 2022).

A tempting plug-in method places $\widehat{\Sigma}^p_{ij}$ into a complete-data rank test and replaces N by one effective sample size. For a set $\mathcal{P}$ of distinct covariance pairs required by a query, examples are

$$\mathsf{N}_{\min} = \min_{(i,j)\in\mathcal{P}} \mathsf{N}_{ij}, \qquad \mathsf{N}_{\text{mean}} = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \mathsf{N}_{ij}, \qquad \mathsf{N}_{\max} = \max_{(i,j)\in\mathcal{P}} \mathsf{N}_{ij}. \tag{4}$$

Related corrections have been studied for structural-equation-model fit statistics in (Marsh, 1998) and Eq. (4) defines analogous plug-in choices for our rank test problem.

Why a single count does not suffice. A single scalar generatlly cannot capture the co-observation factors, even variables have the same missing rate. Let $\mathsf{Y}_i = \mathsf{X}_i - \mu_i$ and $q_{ij,k\ell} = \mathbb{E}[\mathsf{R}_i\mathsf{R}_j\mathsf{R}_k\mathsf{R}_\ell]$. For a pair (i, j), the *estimation error* between estimate and the population counterpart is $\widehat{\Sigma}^p_{ij} - \Sigma_{ij}$. This error changes across repeated draws of the data and the missingness masks. We call the joint variability of these errors *sampling uncertainty*; for a fixed query, it is quantified by the covariance of $\mathrm{vec}(\widehat{\mathbf{\Sigma}}^p_{\mathbf{A},\mathbf{B}})$, approximately $\boldsymbol{V}/N$. Equivalently, $\boldsymbol{V}$ is the asymptotic covariance of $\sqrt{N}\ \mathrm{vec}(\widehat{\mathbf{\Sigma}}^p_{\mathbf{A},\mathbf{B}} - \mathbf{\Sigma}_{\mathbf{A},\mathbf{B}})$. The entries of $\boldsymbol{V}$ are

$$\boldsymbol{V}_{(ij),(k\ell)} = \frac{q_{ij,k\ell}}{q_{ij}q_{k\ell}}\left\{\mathbb{E}[\mathsf{Y}_i\mathsf{Y}_j\mathsf{Y}_k\mathsf{Y}_\ell] - \Sigma_{ij}\Sigma_{k\ell}\right\}. \tag{5}$$

Appendix D.1 derives this expression and a related available-case covariance formulation can also be found in Xu et al. (2022).

In Eq 5, the prefactor $\frac{q_{ij,k\ell}}{q_{ij}q_{k\ell}}$ records how often the *two* covariance pairs are observed together, in addition to their individual observation rates. As shown in Figure 1(b), even when variables have the same individual missing rate, no single sample-size substitution generally reproduces the complete matrix. Thus, plug-in CCA can consequently use an incorrect null reference and need not control Type I error at the nominal level.

Other approaches. Mean imputation generally changes both the target covariance and its sampling uncertainty. A Gaussian observed-data likelihood, fitted for example by EM (Dempster et al., 1977), can use information beyond pairwise moments when the Gaussian model is appropriate. However, inserting its covariance estimate into a complete-data CCA statistic does not directly lead to a valid test (also shown in Section 4.1.1). Our method requires MCAR and finite fourth moments, but not Gaussianity; its covariance and uncertainty estimates can be computed once and reused across queries.

3 MART: Missing-value uncertainty-Aware Rank Test

3.1 Estimating Joint Uncertainty Once

To better make use of the pair-wise covariance estimation, we first characterize the sampling uncertainty of $\mathrm{vec}(\widehat{\mathbf{\Sigma}}^p_{\mathbf{A},\mathbf{B}} - \mathbf{\Sigma}_{\mathbf{A},\mathbf{B}})$. This can be done by formulating $\widehat{\mathbf{\Sigma}}^p$ as a Z-estimator. Specifically, the observed-variable means and pairwise covariances in Eq. (3) solve the sample moment equations:

$$\frac{1}{N}\sum_{t=1}^{N}\mathsf{R}_{ti}(\mathsf{X}_{ti} - \widehat{\mu}_i) = 0, \qquad \frac{1}{N}\sum_{t=1}^{N}\mathsf{R}_{ti}\mathsf{R}_{tj}\big\{(\mathsf{X}_{ti} - \widehat{\mu}_i)(\mathsf{X}_{tj} - \widehat{\mu}_j) - \widehat{\Sigma}^p_{ij}\big\} = 0. \tag{6}$$

Linearizing its moment equation under MCAR gives its first-order contribution from row t:

$$\mathsf{h}_{ij,t} = \frac{\mathsf{R}_{ti}\mathsf{R}_{tj}}{q_{ij}}\big\{(\mathsf{X}_{ti} - \mu_i)(\mathsf{X}_{tj} - \mu_j) - \Sigma_{ij}\big\}. \tag{7}$$

Collecting these contributions for all entries of $\mathbf{\Sigma}_{\mathbf{A},\mathbf{B}}$ gives a vector $\mathbf{h}_{\mathbf{A},\mathbf{B},t}$ in column-major order, with repeated coordinates whenever the block contains the same covariance entry twice. This leads us to Theorem 2 under Assumption 2 (detailed proof in Appendix D).

Assumption 2 (Finite fourth moment)**.** $\mathbb{E}\|\mathbf{X}\|^4 < \infty$.

Theorem 2 (Joint limit of pairwise covariances)**.** *Under Assumptions 1 and 2 for query* $(\mathbf{A}, \mathbf{B})$*,*

$$\sqrt{N}\ \mathrm{vec}\big(\widehat{\mathbf{\Sigma}}^p_{\mathbf{A},\mathbf{B}} - \mathbf{\Sigma}_{\mathbf{A},\mathbf{B}}\big) = \frac{1}{\sqrt{N}}\sum_{t=1}^{N}\mathbf{h}_{\mathbf{A},\mathbf{B},t} + o_{\mathbb{P}}(1) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{V}), \tag{8}$$

where $\boldsymbol{V} = \mathbb{E}[\mathbf{h}_{\mathbf{A},\mathbf{B}}\mathbf{h}_{\mathbf{A},\mathbf{B}}^{\mathsf{T}}]$*, whose entries are given by Eq.* (5)*.*

To estimate $\boldsymbol{V}$, we replace the unknown quantities in Eq. (7) by their pairwise estimates and set

$$\widehat{\mathsf{h}}_{ij,t} = \frac{\mathsf{R}_{ti}\mathsf{R}_{tj}}{\widehat{\mathsf{q}}_{ij}}\big\{(\mathsf{X}_{ti} - \widehat{\mu}_i)(\mathsf{X}_{tj} - \widehat{\mu}_j) - \widehat{\Sigma}^p_{ij}\big\}, \qquad \widehat{\boldsymbol{V}} = \frac{1}{N}\sum_{t=1}^{N} \widehat{\mathbf{h}}_{\mathbf{A},\mathbf{B},t}\widehat{\mathbf{h}}^{\mathsf{T}}_{\mathbf{A},\mathbf{B},t}, \tag{9}$$

where $\widehat{\boldsymbol{V}}$ consistently estimates $\boldsymbol{V}$ by the following theorem.

Theorem 3 (Consistency of joint uncertainty estimation)**.** *Under Assumptions 1 and 2, for query* $(\mathbf{A}, \mathbf{B})$, $\widehat{\boldsymbol{V}} \xrightarrow{\mathbb{P}} \boldsymbol{V}$.

Eq. (8) identifies how $\widehat{\boldsymbol{\Sigma}}^p_{\mathbf{A},\mathbf{B}}$ fluctuates, while Eq. (9) estimates that uncertainty from the observed data. Note that a larger deviation from null-rank-hypothesis is more compelling when it lies in a precisely estimated direction with small uncertainty than when it lies in a noisy one, which is why the next subsection we use this information to construct our test statistics. For a fixed dataset, one can first calculate the pairwise covariance estimation over all variables in the dataset and then corresponding $\widehat{\boldsymbol{V}}_{\text{all}}$, and cache them for later use; the same cached estimates can serve different $\mathbf{A}$, $\mathbf{B}$, and r.

3.2 From Full Inverse-Covariance Weighting to Kronecker Minimum Distance

The estimate $\widehat{\boldsymbol{V}}$ above provides a natural way to measure distance from $\widehat{\boldsymbol{\Sigma}}^p_{\mathbf{A},\mathbf{B}}$ to the rank-null set. If $\widehat{\boldsymbol{V}}$ is positive definite, a test statistics T_{full} as the distance weighted by $\widehat{\boldsymbol{V}}^{-1}$ is

$$\mathsf{T}_{\text{full}} = N \min_{\text{rank}(\boldsymbol{C})\le r} \text{vec}(\widehat{\boldsymbol{\Sigma}}^p_{\mathbf{A},\mathbf{B}} - \boldsymbol{C})^{\mathsf{T}}\widehat{\boldsymbol{V}}^{-1}\,\text{vec}(\widehat{\boldsymbol{\Sigma}}^p_{\mathbf{A},\mathbf{B}} - \boldsymbol{C}). \tag{10}$$

Since $\widehat{\boldsymbol{V}}/N$ estimates the covariance of the unscaled pairwise estimates, $N\widehat{\boldsymbol{V}}^{-1}$ is their inverse-covariance weight. This formulation assigns less weight to directions with greater estimation uncertainty (Cragg & Donald, 1997).

However, two difficulties arise. First, direct optimization of the weighted distance over rank-null matrices is nonconvex and costly, while the rank test queries in causal discovery can be exponential in the number of variables in the worst case. Second, $\widehat{\boldsymbol{V}}$ can be singular; a pseudoinverse would simply assign zero cost to some residual directions without resolving the rank-constrained optimization. Thus, we propose to use Kronecker approximation of $\widehat{\boldsymbol{V}}$, with reasons as follows.

Kronecker approximation of $\widehat{\boldsymbol{V}}$. Left and right multiplication by invertible matrices preserve rank. Accordingly, we fit an $a \times a$ row factor and a $b \times b$ column factor to $\widehat{\boldsymbol{V}}$, as:

$$(\widehat{\boldsymbol{K}}_{\mathbf{A}}, \widehat{\boldsymbol{K}}_{\mathbf{B}}) \in \underset{\boldsymbol{K}_{\mathbf{A}}\in\mathbb{R}^{a\times a},\ \boldsymbol{K}_{\mathbf{B}}\in\mathbb{R}^{b\times b}}{\arg\min} \Big\|\widehat{\boldsymbol{V}} - \boldsymbol{K}_{\mathbf{B}} \otimes \boldsymbol{K}_{\mathbf{A}}\Big\|_F^2. \tag{11}$$

The Kronecker product $\boldsymbol{K}_{\mathbf{B}} \otimes \boldsymbol{K}_{\mathbf{A}}$ is a $b \times b$ grid of $a \times a$ blocks, with block (j, ℓ) equal to $(\boldsymbol{K}_{\mathbf{B}})_{j\ell}\boldsymbol{K}_{\mathbf{A}}$. Rearranging $\widehat{\boldsymbol{V}}$ into an $a^2 \times b^2$ matrix turns Eq. (11) into a best rank-one approximation, directly solvable by an SVD (Van Loan & Pitsianis, 1993). We orient and symmetrize the fitted factors, applying positive-definiteness stabilization only when needed to define numerically stable inverse square roots. We continue to denote the resulting factors by $\widehat{\boldsymbol{K}}_{\mathbf{A}}$ and $\widehat{\boldsymbol{K}}_{\mathbf{B}}$, and set

$$\widehat{\boldsymbol{K}} = \widehat{\boldsymbol{K}}_{\mathbf{B}} \otimes \widehat{\boldsymbol{K}}_{\mathbf{A}}. \tag{12}$$

Thus, using the working weight $\widehat{\boldsymbol{K}}$ instead of $\widehat{\boldsymbol{V}}$ in Eq. 10 gives the test statistics of MART, as

$$\mathsf{T}_K = N \min_{\text{rank}(\boldsymbol{C})\le r} \text{vec}(\widehat{\boldsymbol{\Sigma}}^p_{\mathbf{A},\mathbf{B}} - \boldsymbol{C})^{\mathsf{T}}\widehat{\boldsymbol{K}}^{-1}\,\text{vec}(\widehat{\boldsymbol{\Sigma}}^p_{\mathbf{A},\mathbf{B}} - \boldsymbol{C}), \tag{13}$$

whose rank-constrained minimization also has an exact SVD solution, summarized in Theorem 4.

Theorem 4 (Exact SVD solution for T_K)**.** *Define the whitened block* $\widehat{\boldsymbol{H}} = \widehat{\boldsymbol{K}}_{\mathbf{A}}^{-1/2}\widehat{\boldsymbol{\Sigma}}^p_{\mathbf{A},\mathbf{B}}\widehat{\boldsymbol{K}}_{\mathbf{B}}^{-1/2}$. *Let* $\mathsf{s}_1 \ge \cdots \ge \mathsf{s}_{\min(a,b)}$ *be the singular values of* $\widehat{\boldsymbol{H}}$. *Then*

$$\mathsf{T}_K = N \min_{\text{rank}(\boldsymbol{C})\le r} \|\widehat{\boldsymbol{K}}_{\mathbf{A}}^{-1/2}(\widehat{\boldsymbol{\Sigma}}^p_{\mathbf{A},\mathbf{B}} - \boldsymbol{C})\widehat{\boldsymbol{K}}_{\mathbf{B}}^{-1/2}\|_F^2 = N\sum_{j=r+1}^{\min(a,b)} \mathsf{s}_j^2. \tag{14}$$

3.3 Asymptotic Null Distribution of T_K

We have shown that the test statistics T_K, though itself involves an optimization problem, can be solved directly by SVD, as in Theorem 4. Yet, its asymptotic null distribution is still unknown. In particular, the fitted $\widehat{\boldsymbol{K}}$ need not equal the sampling covariance $\widehat{\boldsymbol{V}}$. We therefore use $\widehat{\boldsymbol{V}}$ to calibrate the distance in the directions that distinguish rank r from nearby higher-rank matrices.

Let $\widehat{\boldsymbol{U}}_\perp \in \mathbb{R}^{a\times(a-r)}$ and $\widehat{\boldsymbol{W}}_\perp \in \mathbb{R}^{b\times(b-r)}$ contain the trailing left and right singular vectors of $\widehat{\boldsymbol{H}}$. The corresponding normal directions is as follows

$$\widehat{\boldsymbol{P}} = (\widehat{\boldsymbol{K}}_{\mathbf{B}}^{-1/2}\widehat{\boldsymbol{W}}_\perp) \otimes (\widehat{\boldsymbol{K}}_{\mathbf{A}}^{-1/2}\widehat{\boldsymbol{U}}_\perp), \qquad \widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{P}}^\mathsf{T}\widehat{\boldsymbol{V}}\widehat{\boldsymbol{P}}. \tag{15}$$

The $d = (a-r)(b-r)$ columns of $\widehat{\boldsymbol{P}}$ retain the estimation errors normal to the rank-r set. Crucially, the middle matrix in $\widehat{\boldsymbol{\Omega}}$ is $\widehat{\boldsymbol{V}}$, not the Kronecker working weight. If we substituted $\widehat{\boldsymbol{K}}$, the result would always be $\boldsymbol{I}_d$, concealing any discrepancy between the weight and the actual estimation uncertainty.

Theorem 5 (Asymptotic Null Distribution of T_K)**.** *Under Assumptions 1 and 2, consider a fixed query with* $\operatorname{rank}(\boldsymbol{\Sigma}_{\mathbf{A},\mathbf{B}}) = r$*. Suppose the fitted factors converge to positive definite limits* $\boldsymbol{K}_{\mathbf{A}}$ *and* $\boldsymbol{K}_{\mathbf{B}}$*. Let* $\boldsymbol{P}$ *be the population limit of* $\widehat{\boldsymbol{P}}$*, up to rotations of its normal-space basis, and let* $\mathsf{w}_1, \ldots, \mathsf{w}_d \geq 0$ *be the eigenvalues of* $\boldsymbol{\Omega} = \boldsymbol{P}^\mathsf{T}\boldsymbol{V}\boldsymbol{P}$*. Then*

$$\mathsf{T}_K \xrightarrow{d} \sum_{j=1}^{d} \mathsf{w}_j \mathsf{Z}_j^2, \qquad \mathsf{Z}_j \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \tag{16}$$

where the weights w_j *can be consistently estimated by the eigenvalues* $\widehat{\mathsf{w}}_j$ *of* $\widehat{\boldsymbol{\Omega}}$*.*

For example, if $d = 2$ and the projected uncertainty has eigenvalues 1 and 4, the reference is $\mathsf{Z}_1^2 + 4\mathsf{Z}_2^2$, rather than χ_2^2. If $\boldsymbol{\Omega} = \boldsymbol{I}_d$, it reduces to χ_d^2. We evaluate the upper tail of Eq. (16) at the observed T_K, using the estimated weights $\widehat{\mathsf{w}}_j$ to calculate the p-values. Algorithm 1 summarizes the procedure of MART. Appendix I provides the implementation details and Appendix C gives the per-query bound, the option to cache only needed pairs, implementation overhead, and numerical-stability details. Table 4 and Table 5 provides the theoretical and empirical run time analysis respectively.

Theorem 5 concerns a fixed query at a rank-exactly-r null. At a lower-rank null, the rank-r set is singular and the limit requires a separate analysis. The classical CCA rank tests has a similar qualification: when testing rank at most r, its regular rank-r limit need not be the chi-square distribution at a lower-rank null (Nielsen, 1999; Chen & Fang, 2019). In sequential rank estimation, one often tests $r = 0, 1, \ldots$ and stops at the first nonrejection. If the true rank is k, reaching any later test $r > k$ requires first rejecting the correctly calibrated test at k; hence the lower-rank issue is less consequential for that workflow. We also empirically find that even at a lower-rank null, the type-I error can still be properly controlled (as in Section 4.1.1). The finite-fourth-moment Assumption 2 allows non-Gaussian observations as input, including uniform, Laplace, and Student-t_5 variables, but it does not cover Student-t_ν variables with $\nu \leq 4$ (model misspecifications experiments in Section 4.1.4).

3.4 Likelihood Ratio Interpretation and Power Gain Over Test-Wise Deletion

The distance in Eq. (14) has a likelihood-ratio interpretation if we regard $\mathbf{Y} = \operatorname{vec}(\widehat{\boldsymbol{\Sigma}}^p_{\mathbf{A},\mathbf{B}})$ as a the observation of Gaussian *summary* with a fixed covariance.

Proposition 1 (Likelihood Ratio Interpretation.)**.** *For any fixed positive-definite factors* $\boldsymbol{K}_{\mathbf{A}}, \boldsymbol{K}_{\mathbf{B}}$*, let* $\boldsymbol{K} = \boldsymbol{K}_{\mathbf{B}} \otimes \boldsymbol{K}_{\mathbf{A}}$ *and consider* $\mathbf{Y} \sim \mathcal{N}(\operatorname{vec}(\boldsymbol{C}), \boldsymbol{K}/N)$*, with* $\boldsymbol{C} \in \mathbb{R}^{a\times b}$ *unrestricted under the alternative and* $\operatorname{rank}(\boldsymbol{C}) \leq r$ *under the null. Write* $\mathcal{L}_K(\boldsymbol{C}; \mathbf{Y})$ *for its Gaussian likelihood and define*

$$\Lambda = \frac{\sup_{\operatorname{rank}(\boldsymbol{C})\leq r} \mathcal{L}_K(\boldsymbol{C}; \mathbf{Y})}{\sup_{\boldsymbol{C}\in\mathbb{R}^{a\times b}} \mathcal{L}_K(\boldsymbol{C}; \mathbf{Y})}.$$

Then the likelihood ratio statistic $-2\log\Lambda$ *is exactly Eq. 13.*

Further, the following Theorem 6 gives an explicit sufficient setting where MART has strict power gain over test-wise deletion by making use of additional incomplete rows.

Theorem 6 (Strict Local Power Gain)**.** *Let* $r \geq 1$*, and let disjoint subvectors* $\mathbf{A} = (\mathbf{A}_s^\mathsf{T}, \mathsf{X}_*)^\mathsf{T}$ *and* $\mathbf{B} = (\mathbf{B}_s^\mathsf{T}, \mathsf{Y}_*)^\mathsf{T}$ *have sizes* $a = b = r + 1$*. Suppose their joint distribution is centered Gaussian with*

positive-definite covariance. The signal subvectors have cross-covariance $\boldsymbol{D} \in \mathbb{R}^{r\times r}$ of rank r. The pair of unit-variance variables $(\mathsf{X}_*, \mathsf{Y}_*)$ is independent of the signal subvectors, and its covariance is $\tau/\sqrt{N}$, where $\tau \in \mathbb{R}$ is a fixed local signal amplitude. Thus $\tau = 0$ is the rank-r null; for fixed $\tau \neq 0$, this covariance approaches zero at rate $N^{-1/2}$, and the cross-block is

$$\boldsymbol{\Sigma}^{(N)}_{\mathbf{A},\mathbf{B}} = \begin{pmatrix} \boldsymbol{D} & \mathbf{0} \\ \mathbf{0}^{\mathsf{T}} & \tau/\sqrt{N} \end{pmatrix}. \tag{17}$$

Assume a fixed MCAR mechanism with positive observation probabilities for all required pairs. Write q_* for the probability of jointly observing $\mathsf{X}_*, \mathsf{Y}_*$ and q_{all} for the probability of observing all of $\mathbf{A} \cup \mathbf{B}$, and suppose $0 < q_{\mathrm{all}} < q_*$. Assume the effective working factors converge to positive-definite limits under these local alternatives, and mixture-evaluation error vanishes. For $0 < \alpha < 1$, let c_α be the $(1-\alpha)$ quantile of χ^2_1. The limiting powers of the proposed calibrated test and the classical Gaussian CCA test using only test-wise complete rows satisfy

$$\beta_{\mathsf{T}_K}(\tau) = \mathbb{P}\{\chi^2_1(q_*\tau^2) > c_\alpha\}, \qquad \beta_{\mathsf{T}_{\mathrm{CCA}}}(\tau) = \mathbb{P}\{\chi^2_1(q_{\mathrm{all}}\tau^2) > c_\alpha\}. \tag{18}$$

Thus, $\beta_{\mathsf{T}_K} > \beta_{\mathsf{T}_{\mathrm{CCA}}}$ for every fixed $\tau \neq 0$, and both limiting rejection probabilities equal α at $\tau = 0$.

4 Experiments

We evaluate the proposed method MART at two levels: statistical type-I control and calibrated power for individual rank test queries, and structure recovery when the test is used within causal discovery methods. Unless specified otherwise, tests use significance level $\alpha = 0.05$, and competing methods receive the same data and missingness masks within each replication.

Compared methods. We compare MART with (i) test-wise deletion based on classical CCA rank test, (ii) plug-in pairwise covariance with the minimum effective count, (iii) plug-in pairwise covariance with the average effective count, (iv) column-mean imputation followed by classical CCA rank test, (v) Gaussian EM conditional-mean imputation followed by classical CCA rank test, and (vi) Classical CCA rank test with complete data before masking, which provides an reference.

4.1 Statistical performance of rank tests

Data-generating model. Let $A = A_0 \cup C$ and $B = B_0 \cup C$, where A_0, B_0, and C are disjoint, $|A_0| = |B_0| = 5$, and $|C|$ ranges from 10 to 20. For each replication we draw a new positive-definite population covariance and set the true rank for $\Sigma_{A,B}$ as $r = |C| + k$ where $k \in \{0, 1, 2\}$. We generate the data by a linear causal model with different kinds of independent noise terms: Gaussian, Laplace, and Student t-5. Sample sizes are $n \in \{500, 1000, 2000, 5000\}$ and nominal missing rates are $m \in \{0.10, 0.15\}$. For each dataset and variable j, a missingness probability π_j is sampled independently from $\mathrm{Unif}(0.9m, 1.1m)$; entries are then masked independently conditional on these probabilities.

4.1.1 Type I error under heterogeneous missingness

We test $H_0 : \mathrm{rank}(\Sigma_{A,B}) \leq r$. To assess Type I error throughout the composite null, rather than only at its boundary, we draw the population rank uniformly from $\{r, r-1, r-2\}$ in each replication. We do this for every combination of innovation distribution, sample size, missingness rate, and tested rank in the shared experimental design, with 500 independent replications per setting. Figures 2 and 6 report the empirical Type I error rates at the nominal level $\alpha = 0.05$. MART, test-wise deletion, and the complete-data test maintain error rates near the nominal level across the settings considered. In contrast, pairwise minimum, pairwise average, mean imputation, and EM imputation do not consistently control Type I error. The error rates of the two imputation-based methods even increase with sample size.

4.1.2 Ordinary and size-calibrated power

To evaluate power, we test $H_0 : \mathrm{rank}(\Sigma_{A,B}) \leq r - 1$ using populations of rank r. We report calibrated power at $\alpha = 0.05$ in Figures 3 and 7. For calibrated-power, we estimated a method-specific critical value from 300 independent null datasets within each simulation setting and evaluated power on separate alternative datasets. Calibration matched the covariance-generating and missingness distributions within each setting. As expected, the complete-data rank test has the highest power because it uses fully observed samples. MART and the EM-imputation-based test have similar

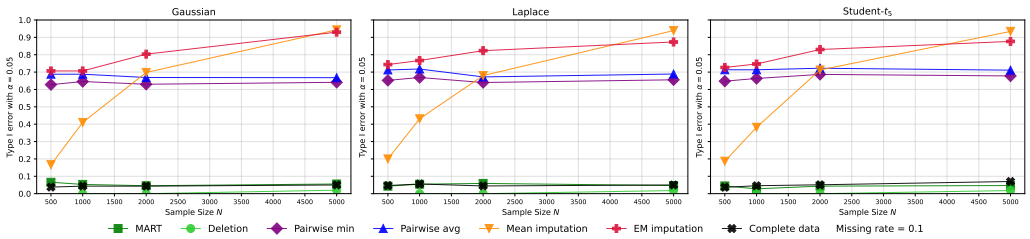


Figure 2: Type-I error of each method under different sample sizes with missing rate 0.1.

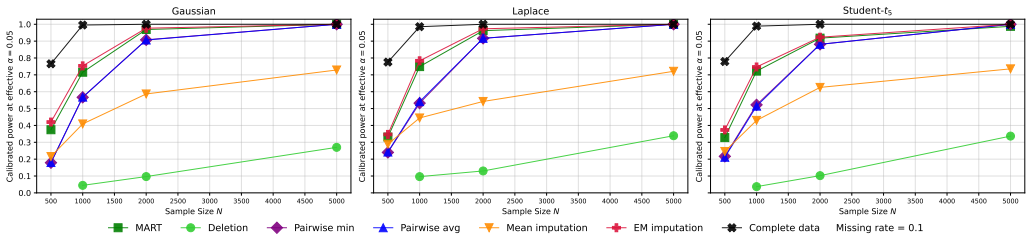


Figure 3: Calibrated power of each method under different sample sizes with missing rate 0.1.

calibrated power at missing rates of both 0.1 and 0.15. However, the EM-imputation-based test fails to control Type I error. Among the methods that do control Type I error in the presence of missing values, MART has the highest power, substantially outperforming test-wise deletion.

4.1.3 Power from additional incomplete observations

To assess the information gained from incomplete rows, we compare MART with test-wise deletion while holding 200 complete observations fixed within each replication panel. We then append nested prefixes of $b \in 0, 250, 500, 1000, 2000$ additional rows drawn from the same population. Each appended row is required to be incomplete; its coordinate-specific missingness probabilities are drawn once per panel from $\mathrm{Unif}(0.09, 0.11)$. Figure 4 shows that MART's power increases as more incomplete rows are added, demonstrating that it uses information from these rows to improve detection. At $b = 0$, MART has slightly lower power than the classical rank test. This is unsurprising: with fully observed data, the classical rank test is a likelihood-ratio test and benefits from the strong power properties of likelihood-ratio methods. MART trades a small amount of power in this complete-data setting for the ability to use incomplete observations, yielding greater power when missing values are present.

4.1.4 Robustness beyond finite fourth moments

We examine MART's behavior when the finite-fourth-moment assumption is violated. Following the preceding experimental design, we generate data with Student's t noise with three degrees of freedom and an average missing rate of 0.15. Figure 5 reports empirical Type I error at $\alpha = 0.05$ and calibrated power. Despite the infinite population fourth moment, MART maintains Type I error near the nominal level in these finite-sample experiments with all different sample sizes. Its power also exceeds that of test-wise deletion, the pairwise minimum and average corrections, and mean imputation.

4.2 Application to causal discovery

In this section we employ the final causal discovery performance to evaluate each rank test methods by using both synthetic data and real-world data.

4.2.1 PC with rank-based conditional independence tests

We first evaluate MART in causal discovery by running the PC algorithm (Spirtes et al., 2000) on synthetic data, replacing its partial-correlation test with each rank-test method, and comparing the resulting skeleton F1 scores. We use the implementation in causal-learn (Zheng et al., 2023) without a cap on conditioning-set size. The main design contains 30 observed variables. Given a topological ordering, each possible forward edge is included independently with probability 1/5. Nonzero edge coefficients have independent random signs and magnitudes drawn from $\mathrm{Unif}(0.2, 0.8)$, and the data are generated from the resulting linear causal model with independent Gaussian noise terms. We

Table 1: F1 (%) of skeleton among all variables by sample size and missingness rate of PC algorithm with different rank tests. The best-performing one, excluding complete data, is shown in bold.

N	Missingness	**MART**	Deletion	Pairwise min	Pairwise avg	Mean imputation	EM imputation	Complete data
500	10%	**92.40**	90.86	90.28	90.16	88.09	90.14	93.58
500	15%	**91.77**	90.91	88.99	88.79	88.48	88.94	93.58
1000	10%	**94.53**	92.87	92.35	92.50	89.27	91.97	95.01
1000	15%	**94.87**	92.04	92.32	92.16	87.79	90.23	95.01
2000	10%	**94.93**	92.73	93.31	93.09	88.78	91.53	95.42
2000	15%	**95.22**	92.19	92.75	92.51	86.78	90.76	95.42
5000	10%	**94.50**	92.21	92.34	92.18	85.93	90.92	95.85
5000	15%	**94.94**	92.50	92.58	92.61	83.06	89.68	95.85

consider $N \in \{500, 1000, 2000, 5000\}$, missing rates $m \in \{0.10, 0.15\}$, and 100 replications per setting. Graphs and complete samples are paired across missingness conditions, and all methods are evaluated on the same masked dataset within each replication. The significance level is 0.05. Table 1 shows that PC equipped with MART achieves the highest skeleton F1 score among methods using the same incomplete data in every setting. Its performance is also close to that of PC using complete data. For these partial-correlation queries, the classical Fisher-Z (Fisher, 1921) and CCA rank tests produce nearly identical p-values, differing by less than 10^{-6}, and consequently return the same graphs.

4.2.2 RLCD FOR LATENT VARIABLES STRUCTURE WITH RANK TESTS

We next evaluate MART in causal discovery using RLCD (Dong et al., 2024). We use the benchmark graphs from Dong et al. (2024), preserving their observed and latent structures. For each graph, nonzero edge coefficients have independent random signs and magnitudes drawn from $\mathrm{Unif}(0.2, 0.8)$, with average missing rate $m = 0.1$, All methods use the same complete samples and missingness masks. We evaluate the F1 score of the undirected skeleton over both observed and latent variables, keeping observed-variable labels fixed and allowing predicted latent nodes to be permuted for evaluation. Table 3 shows that RLCD equipped with MART achieves the highest skeleton F1 score among methods using the same incomplete data in every setting. At sample sizes of 1,000 and 2,000, test-wise deletion cannot be applied because RLCD frequently requires rank tests involving all variables, leaving too few complete rows. This highlights the benefit of MART for rank-constraint-based causal discovery methods (Huang et al., 2022; Dong et al., 2024): it can use a lot more partially observed rows that test-wise deletion simply discards.

4.3 REAL-WORLD DATA EXPERIMENT

We use the Multi-tasking Behavior dataset (Himi et al., 2019), selecting the same nine variables as Himi et al. (2019). To simulate missingness, we independently mask each variable with probability 0.1. We then compare the graphs recovered from the incomplete data using different rank tests with the graph recovered from the complete data. We treat the latter as the reference graph for this comparison. Figure 8 shows the latent-variable causal graph recovered by RLCD from the complete data. As shown in Figure 9(a), RLCD equipped with MART recovers nearly the same graph from the incomplete data, missing only one edge. EM imputation yields the next closest result, with three edges differing from the reference graph (Figure 9(c)). The remaining methods recover only one or two latent variables, compared with four in the complete-data graph. The poor performance of test-wise deletion is again unsurprising: the original dataset contains only 202 observations, and deletion leaves even fewer rows available for rank testing. MART nevertheless recovers a graph close to the complete-data reference, illustrating its ability to use limited, partially observed data effectively.

5 CONCLUSION

We introduced MART, a covariance-rank test that uses incomplete rows through pairwise covariance estimation while accounting for the estimates' joint sampling uncertainty. A Kronecker working weight makes the rank-constrained distance solvable by a whitened SVD; the full uncertainty estimate calibrates it with a weighted chi-square mixture. Under MCAR and finite fourth moments, MART provides valid asymptotic Type-I error control. We establish a strict limiting power gain over test-wise deletion in an explicit setting and show how estimates can be reused across causal-discovery queries. Extending these guarantees to broader missingness mechanisms remains future work.

AI Assistance Disclosure

Generative AI (OpenAI Codex) assisted with presentations, drafting proofs, refining hypotheses, and plotting figures. We did not use generative AI for any other task requiring disclosure under the ICLR 2027 AI policy.

References

T. W. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 22(3):327–351, 1951. URL https://doi.org/10.1214/aoms/1177729580.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis. 2nd ed*. John Wiley & Sons, 1984.

Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measure of conditional independence. *Australian and New Zealand Journal of Statistics*, 46: 657–664, 12 2004.

Brett K. Beaulieu-Jones, Daniel R. Lavage, John W. Snyder, Jason H. Moore, Sarah A. Pendergrass, and Christopher R. Bauer. Characterizing and managing missing structured data in electronic health records: Data analysis. *JMIR Medical Informatics*, 6(1):e11, 2018. doi: 10.2196/medinform.8960.

Svetlana Bryzgalova, Sven Lerner, Martin Lettau, and Markus Pelger. Missing financial data. *The Review of Financial Studies*, 38(3):803–882, 2025. doi: 10.1093/rfs/hhae036.

Qihui Chen and Zheng Fang. Improved inference on the rank of a matrix. *Quantitative Economics*, 10(4):1787–1824, 2019. URL https://doi.org/10.3982/QE1139.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pp. 294–321, 2012.

John G. Cragg and Stephen G. Donald. Inferring the rank of a matrix. *Journal of Econometrics*, 76 (1–2):223–250, 1997. URL https://doi.org/10.1016/0304-4076(95)01790-9.

Ruifei Cui, Perry Groot, and Tom Heskes. Learning causal structure from mixed data with missing values using Gaussian copula models. *Statistics and Computing*, 29(2):311–333, 2019. URL https://doi.org/10.1007/s11222-018-9810-x.

Haoyue Dai, Yiwen Qiu, Ignavier Ng, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Latent variable causal discovery under selection bias. *arXiv preprint arXiv:2512.11219*, 2025.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–38, 1977. URL https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Yanming Di. t-separation and d-separation for directed acyclic graphs. *preprint*, 2009.

Xinshuai Dong, Biwei Huang, Ignavier Ng, Xiangchen Song, Yujia Zheng, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. A versatile causal discovery framework to allow causally-related hidden variables. In *ICLR*, 2024.

Xinshuai Dong, Ignavier Ng, Boyang Sun, Haoyue Dai, Guang-Yuan Hao, Shunxing Fan, Peter Spirtes, Yumou Qiu, and Kun Zhang. Permutation-based rank test in the presence of discretization and application in causal discovery with mixed data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14137–14151. PMLR, 2025. URL https://proceedings.mlr.press/v267/dong25i.html.

R. A. Fisher. The distribution of the partial correlation coefficient. *Metron*, 3:329–332, 1924.

Ronald A Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Alexander Gain and Ilya Shpitser. Structure learning under missing data. In *Proceedings of the Ninth International Conference on Probabilistic Graphical Models*, volume 72 of *Proceedings of Machine Learning Research*, pp. 121–132, 2018. URL https://proceedings.mlr.press/v72/gain18a.html.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, 2007.

Samsad Afrin Himi, Markus Buehner, Matthias Schwaighofer, Anna Klapetek, and Sven Hilbert. Multitasking behavior and its related constructs: Executive functions, working memory capacity, relational integration, and divided attention. *Cognition*, 189:275–298, 08 2019.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

Biwei Huang, Charles Jia Han Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. *arXiv preprint arXiv:2210.01798*, 2022.

Herbert W. Marsh. Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, 5(1):22–36, 1998. URL https://doi.org/10.1080/10705519809540087.

Bent Nielsen. The likelihood-ratio test for rank in bivariate canonical correlation analysis. *Biometrika*, 86(2):279–288, 1999. doi: 10.1093/biomet/86.2.279.

J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan kaufmann, 1988.

James L. Peugh and Craig K. Enders. Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4):525–556, 2004. doi: 10.3102/00346543074004525.

Joseph Ramsey. A scalable conditional independence test for nonlinear, Non-Gaussian data. *arXiv preprint arXiv:1401.5031*, 2014.

Thomas Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, 1996.

Jean-Marc Robin and Richard J. Smith. Tests of rank. *Econometric Theory*, 16(2):151–175, 2000. URL https://doi.org/10.1017/S0266466600162012.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. URL https://doi.org/10.1093/biomet/63.3.581.

Rajen Dinesh Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 2018.

Ricardo Silva, Richard Scheine, Clark Glymour, and Peter Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(Feb):191–246, 2006.

Peter Spirtes. Calculation of entailed rank constraints in partially non-linear and cyclic models. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 606–615. AUAI Press, 2013.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9:62–72, 1991.

Peter Spirtes, Chris Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Conference on Uncertainty in Artificial Intelligence*, 1995.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Eric V. Strobl, Shyam Visweswaran, and Peter L. Spirtes. Fast causal inference with non-random missingness by test-wise deletion. *International Journal of Data Science and Analytics*, 6(1): 47–62, 2018. doi: 10.1007/s41060-017-0094-6.

Seth Sullivant, Kelli Talaska, and Jan Draisma. Trek separation for gaussian graphical models. *arXiv:0812.1938.*, 2010.

Boyang Sun, Yu Yao, Xinshuai Dong, Zongfang Liu, Tongliang Liu, Yumou Qiu, and Kun Zhang. A sample efficient conditional independence test in the presence of discretization. In *ICML*, 2025a.

Boyang Sun, Yu Yao, Guang-Yuan Hao, Yumou Qiu, and Kun Zhang. A conditional independence test in the presence of discretization. In *ICLR*, 2025b.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Patrick Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001. doi: 10.1093/bioinformatics/17.6.520.

Ruibo Tu, Cheng Zhang, Paul Ackermann, Karthika Mohan, Hedvig Kjellström, and Kun Zhang. Causal discovery in the presence of missing data. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1762–1770, 2019. URL https://proceedings.mlr.press/v89/tu19a.html.

Charles F. Van Loan and Nikos Pitsianis. Approximation with Kronecker products. In Marc S. Moonen, Gene H. Golub, and Bart L. R. De Moor (eds.), *Linear Algebra for Large Scale and Real-Time Applications*, volume 232 of *NATO ASI Series E: Applied Sciences*, pp. 293–314. Kluwer Academic Publishers, 1993. URL https://doi.org/10.1007/978-94-015-8196-7_17.

Vy Vo, He Zhao, Trung Le, Edwin V. Bonilla, and Dinh Phung. Optimal transport for structure learning under missing data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 49605–49626, 2024. URL https://proceedings.mlr.press/v235/vo24b.html.

Janine Witte, Ronja Foraita, and Vanessa Didelez. Multiple imputation and test-wise deletion for causal discovery with incomplete cohort data. *Statistics in Medicine*, 41(23):4716–4743, 2022. doi: 10.1002/sim.9535.

Tianchen Xu, Kun Chen, and Gen Li. The more data, the better? demystifying deletion-based methods in linear regression with missing data. *Statistics and Its Interface*, 15(4):515–526, 2022. URL https://doi.org/10.4310/21-SII717.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Conference on Uncertainty in Artificial Intelligence*, 2012.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *arXiv preprint arXiv:2307.16405*, 2023.

Algorithm 1 Uncertainty-Aware Missing Value Rank Test by Kronecker Minimum Distance

Require: Data with missingness masks; query $(\mathbf{A}, \mathbf{B}, r)$; level α.

1: Once per dataset: compute $\widehat{\boldsymbol{\Sigma}}^p$, pair counts, and $\widehat{\boldsymbol{V}}_{\mathrm{all}}$ from unique pair influence functions.
2: Extract $\widehat{\boldsymbol{\Sigma}}^p_{\mathbf{A},\mathbf{B}}$ and $\widehat{\boldsymbol{V}} = \boldsymbol{J}_{\mathbf{A},\mathbf{B}} \widehat{\boldsymbol{V}}_{\mathrm{all}} \boldsymbol{J}^{\mathsf{T}}_{\mathbf{A},\mathbf{B}}$.
3: Fit and stabilize factors $\widehat{\boldsymbol{K}}_{\mathbf{A}}, \widehat{\boldsymbol{K}}_{\mathbf{B}}$ using rearrangement SVD.
4: Compute the SVD of $\widehat{\boldsymbol{H}} = \widehat{\boldsymbol{K}}_{\mathbf{A}}^{-1/2} \widehat{\boldsymbol{\Sigma}}^p_{\mathbf{A},\mathbf{B}} \widehat{\boldsymbol{K}}_{\mathbf{B}}^{-1/2}$.
5: Compute $\mathsf{T}_K = N \sum_{j>r} \mathsf{s}_j^2$ and $\widehat{\boldsymbol{P}}$ from Eq. (15).
6: Compute eigenvalues of $\widehat{\boldsymbol{P}}^{\mathsf{T}} \widehat{\boldsymbol{V}} \widehat{\boldsymbol{P}}$ and the upper-tail mixture probability $\widehat{\mathsf{p}}$.
7: **return** Reject if $\widehat{\mathsf{p}} < \alpha$; cache the result for the same query.

Table 2: Main notation. The population covariance $\boldsymbol{\Sigma}$, covariance $\boldsymbol{V}$ of the scaled estimation error, and working covariance $\boldsymbol{K}$ play different roles.

Symbol	Meaning
$N, M;\ t, i, j$	Number of rows and variables; row and variable indices.
$\mathbf{X}, \mathbf{X}_t;\ \boldsymbol{\mu}, \boldsymbol{\Sigma}$	Generic M-vector and its row-t copy; population mean and covariance.
$\mathbf{R}, \mathbf{R}_t;\ \mathsf{R}_i, \mathsf{R}_{ti}$	Generic and row-t mask vectors in $\{0,1\}^M$; their coordinate-i indicators.
$q_{ij}, \widehat{\mathsf{q}}_{ij}, \mathsf{N}_{ij}$	Pair observation probability, observed fraction, and observed-pair count; $q_{ij,k\ell}$ is the joint observation probability of two pairs.
$\mathbf{A}, \mathbf{B};\ a, b, r$	Queried random subvectors of $\mathbf{X}$, their dimensions, and the tested rank bound.
$\widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\Sigma}}^p$	Classical empirical covariance and pairwise covariance estimator.
$k, k_{\mathcal{Q}}, m, d$	All unique pairs $M(M+1)/2$, pairs needed by planned queries, block size ab, and normal dimension $(a-r)(b-r)$.
$\mathsf{h}_{ij,t}, \mathbf{h}_{\mathbf{A},\mathbf{B},t}$	Pairwise influence contribution and its column-major stack for the queried block.
$\boldsymbol{V}, \widehat{\boldsymbol{V}}$	Asymptotic covariance of the scaled block error and its empirical estimate.
$\widehat{\boldsymbol{V}}_{\mathrm{all}}, \boldsymbol{J}_{\mathbf{A},\mathbf{B}}$	Cached unique-pair covariance estimate and query selection/duplication matrix.
$\widehat{\boldsymbol{K}}_{\mathbf{A}}, \widehat{\boldsymbol{K}}_{\mathbf{B}}, \widehat{\boldsymbol{K}}$	Stabilized factors and fitted product $\widehat{\boldsymbol{K}}_{\mathbf{B}} \otimes \widehat{\boldsymbol{K}}_{\mathbf{A}}$; $\boldsymbol{K}_{\mathbf{A}}, \boldsymbol{K}_{\mathbf{B}}$ are factor limits.
$\widehat{\boldsymbol{H}}, \mathsf{s}_j$	Empirical block whitened by the inverse factor roots and its singular values.
$\boldsymbol{C}, \mathsf{T}_K, \mathsf{T}_{\mathrm{CCA}}$	Candidate rank-constrained block, proposed statistic, and classical CCA statistic.
$\widehat{\boldsymbol{U}}_\perp, \widehat{\boldsymbol{W}}_\perp$	Trailing left/right singular-vector bases of the whitened block.
$\widehat{\boldsymbol{P}}, \boldsymbol{P};\ \widehat{\boldsymbol{\Omega}}, \boldsymbol{\Omega}$	Estimated and limiting weighted normal maps and projected covariances $\widehat{\boldsymbol{P}}^{\mathsf{T}} \widehat{\boldsymbol{V}} \widehat{\boldsymbol{P}}$ and $\boldsymbol{P}^{\mathsf{T}} \boldsymbol{V} \boldsymbol{P}$.
$\widehat{\mathsf{w}}_j, \mathsf{w}_j$	Estimated and limiting eigenvalue weights of the chi-square mixture.
$\boldsymbol{K}_0, \boldsymbol{N}_0, \boldsymbol{\Omega}_{\mathrm{raw}}$	Limiting Kronecker working covariance, unwhitened normal basis, and projected sampling covariance.
$q_*, q_{\mathrm{all}}, \tau$	Residual-pair and complete-query observation probabilities and fixed local signal amplitude in Theorem 6.
$\beta_{\mathsf{T}_K}(\tau), \beta_{\mathsf{T}_{\mathrm{CCA}}}(\tau)$	Limiting powers of the proposed and complete-case CCA tests under that local alternative.
B_{MC}	Number of draws used to evaluate the chi-square mixture tail.

A Additional Experimental Results

Tables 6 and 7 show the Type-I error and the calibrated power of each method under missing rate 0.15.

Table 1 show the causal discovery result by using PC algorithm with different rank test methods (as partial correlation test is a special case of rank test).

Table 5 show the empirical computational cost per query for each method.

B Trek and T-separations

Definition 1 (Treks (Sullivant et al., 2010))**.** In $\mathcal{G}$, a trek from X to Y is an ordered pair of directed paths (P_1, P_2) where P_1 has a sink X, P_2 has a sink Y, and both P_1 and P_2 have the same source Z.

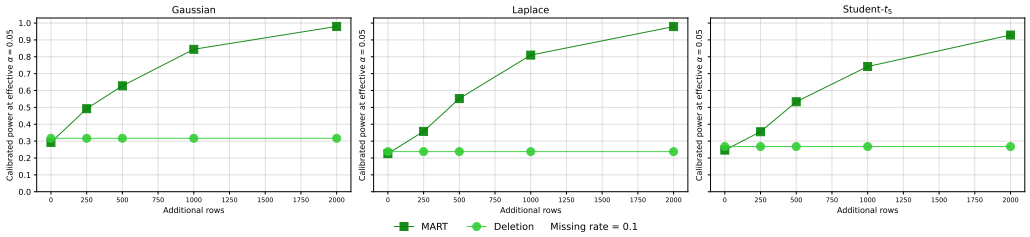


Figure 4: Calibrated power gain of MART by the use of additional incomplete rows.

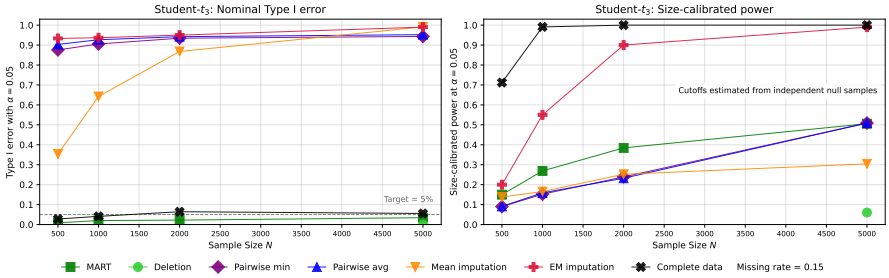


Figure 5: Performance of MART under model misspecification (violation of finite fourth moment).

A trek between two variables consists of two directed paths with a common source, one ending at each variable. A pair $(\mathbf{C_A}, \mathbf{C_B})$ *t-separates* $\mathbf{A}$ from $\mathbf{B}$ if every trek between the sets meets $\mathbf{C_A}$ on its $\mathbf{A}$ side or $\mathbf{C_B}$ on its $\mathbf{B}$ side. In the four-measurement example, every trek between the two groups passes through L; the one-variable bottleneck explains their cross-block rank of one.

Theorem 7 (Rank and t-separation (Sullivant et al., 2010))**.** *In a linear acyclic causal model with independent errors and rank faithfulness,*

$$\operatorname{rank}(\boldsymbol{\Sigma}_{\mathbf{A},\mathbf{B}}) = \min\big\{|\mathbf{C_A}| + |\mathbf{C_B}| : (\mathbf{C_A}, \mathbf{C_B}) \textit{ t-separates } \mathbf{A} \textit{ from } \mathbf{B}\big\}. \tag{19}$$

C Computational Cost and Numerical Stability

Let $k = M(M+1)/2$ be the number of distinct covariance entries among M variables, $m = ab$ the number of entries in a queried block, and $d = (a-r)(b-r)$ its normal dimension. Let B_{MC} be the number of draws used to evaluate the reference mixture. We distinguish the cost of constructing a dataset-wide cache from the cost of a new rank query.

Shared pairwise and joint covariance cache. Computing all pairwise means, counts, and covariances costs $O(NM^2)$. For each row, forming its k distinct-pair influence contributions costs $O(k)$, and adding their outer product to the unique-pair covariance estimate costs $O(k^2)$. Thus building the full cache costs $O(NM^2 + Nk^2)$ time and $O(k^2)$ space for the cached covariance estimate and moments, apart from storage of the input data. This potentially large $O(M^4)$ space cost is the price of caching the *full* joint covariance estimate. If the planned queries use only $k_{\mathcal{Q}} < k$ distinct covariance pairs, caching those pairs instead reduces the covariance estimation to $O(Nk_{\mathcal{Q}}^2)$ time and $O(k_{\mathcal{Q}}^2)$ space, in addition to computing their means, counts, and covariances. The shared work is performed once for a fixed dataset and missingness mask, regardless of how many subsequent rank queries use them.

One new query after caching. Extracting the m covariance entries and their $m \times m$ joint covariance estimate costs $O(m^2)$. With a full SVD of the $a^2 \times b^2$ rearrangement, Kronecker fitting costs $O(a^2b^2 \min\{a^2, b^2\})$. Factor roots and the whitened block SVD cost $O(a^3 + b^3 + ab(a+b))$. Forming and projecting onto the d normal directions costs $O(m^2 d + md^2)$; diagonalizing the projected covariance and simulating the mixture cost $O(d^3 + B_{\mathrm{MC}} d)$. Consequently, the dense arithmetic cost per new query is

$$O\big(m^2 + a^2b^2 \min\{a^2, b^2\} + a^3 + b^3 + ab(a+b) + m^2 d + md^2 + d^3 + B_{\mathrm{MC}} d\big)\,.$$

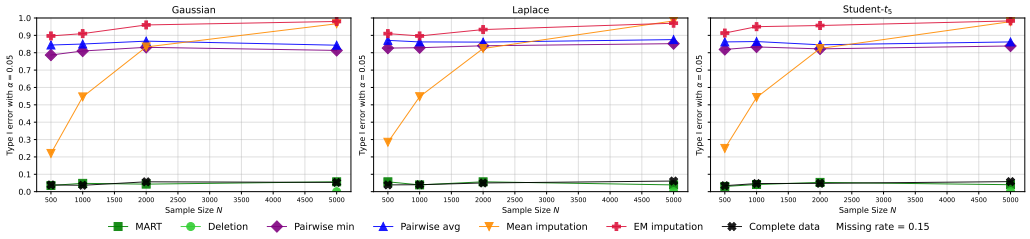


Figure 6: Type-I error of each method under different sample sizes with missing rate 0.15.

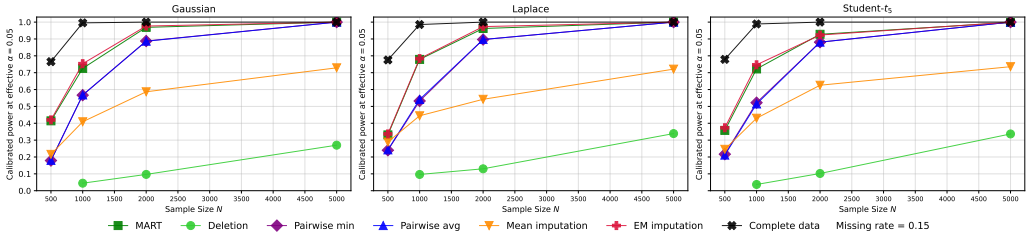


Figure 7: Calibrated power of each method under different sample sizes with missing rate 0.15.

The Kronecker factors depend on $\mathbf{A}, \mathbf{B}$, and the normal projection also depends on r. For Q queries, the total cost is the one-time $O(NM^2 + Nk^2)$ setup plus the sum of their query costs. The mathematical query operations no longer scale with N after caching; wrappers that copy selected data columns can still add $O(N|\mathbf{A} \cup \mathbf{B}|)$ time per query. For fixed $\mathbf{A}, \mathbf{B}$, factor fitting and the whitened SVD can be reused across ranks, and exact repeat queries can reuse their p-values. The method avoids repeated nonlinear rank fitting and inversion of a full $m \times m$ weight. Table 4 summarizes the terms; Section I documents the implementation conventions. Numerically, the factor roots require positive-definite stabilization when fitted factors are singular or nearly singular. Calibration still uses the unregularized covariance estimate projected onto the normal space; its eigenvalues may be clipped only to remove small negative roundoff errors.

Table 4: Dense time costs for shared setup and one new rank query. Here $k = M(M+1)/2$, $m = ab$, and $d = (a-r)(b-r)$. A full SVD is listed to match the current rearrangement implementation, not an assumed iterative leading-vector solver.

Operation	Arithmetic cost
Pairwise means, counts, and covariances, once	$O(NM^2)$
Full unique-pair covariance estimate, once	$O(Nk^2)$
Extract query covariance and its joint covariance estimate	$O(m^2)$
Rearrangement SVD, $a^2 \times b^2$	$O(a^2b^2 \min(a^2, b^2))$
Factor roots and whitened SVD	$O(a^3 + b^3 + ab(a+b))$
Covariance projection to normal space	$O(m^2d + md^2)$
Mixture eigenvalues and simulation	$O(d^3 + B_{\mathrm{MC}}d)$

Table 3: F1 scores (%) of skeleton among all variables by sample size and missingness rate of the RLCD algorithm with different rank test methods. The best-performing method, excluding complete data, is shown in bold. - means the test-wise deletion does not have enough samples to perform the test.

N	Missingness	**MART**	Deletion	Pairwise min	Pairwise avg	Mean imputation	EM imputation	Complete data
1000	10%	**43.53**	-	35.60	36.50	30.27	41.23	48.64
2000	10%	**54.93**	-	39.31	40.09	35.78	48.53	61.42
5000	10%	**59.50**	48.21	44.34	45.18	34.93	53.92	72.85
10000	10%	**65.40**	53.86	47.28	47.16	33.09	61.14	79.58

D Pairwise Influence Functions and Covariance Estimation

D.1 Derivation of the Pairwise Error Covariance

We derive Eq. (5) under Assumption 1 and establish the joint CLT and covariance estimator consistency in Theorems 2 and 3.

Proofs of Theorems 2 and 3. **Step 1: the random denominator and influence function.** Equivalently to linearizing Eq. (6), we expand the pairwise ratio directly. First center at the true means and write $\mathsf{Y}_{ti} = \mathsf{X}_{ti} - \mu_i$. The true-mean-centered pairwise estimator is $\widetilde{\Sigma}^{p}_{ij} = \mathsf{N}_{ij}^{-1} \sum_t \mathsf{R}_{ti}\mathsf{R}_{tj}\mathsf{Y}_{ti}\mathsf{Y}_{tj}$. The first equality below is exact whenever $\mathsf{N}_{ij} > 0$:

$$\begin{aligned}\sqrt{N}\left(\frac{\sum_t \mathsf{R}_{ti}\mathsf{R}_{tj}\mathsf{Y}_{ti}\mathsf{Y}_{tj}}{\mathsf{N}_{ij}} - \Sigma_{ij}\right) &= \frac{1}{\widehat{\mathsf{q}}_{ij}\sqrt{N}} \sum_t \mathsf{R}_{ti}\mathsf{R}_{tj}(\mathsf{Y}_{ti}\mathsf{Y}_{tj} - \Sigma_{ij}) \\ &= \frac{1}{\sqrt{N}} \sum_t \mathsf{h}_{ij,t} + o_{\mathbb{P}}(1).\end{aligned}$$

The second equality uses $\widehat{\mathsf{q}}_{ij} \xrightarrow{\mathbb{P}} q_{ij} > 0$ and the $O_{\mathbb{P}}(1)$ centered sum scaled by $N^{-1/2}$, yielding

$$\mathsf{h}_{ij,t} = \frac{\mathsf{R}_{ti}\mathsf{R}_{tj}}{q_{ij}}(\mathsf{Y}_{ti}\mathsf{Y}_{tj} - \Sigma_{ij}).$$

Thus the denominator's randomness is accounted for by centering the product inside the observed-pair contribution; it is not treated as a fixed count. Under MCAR, $\mathbb{E}[\mathsf{h}_{ij,t}] = 0$ because $\mathbb{E}[\mathsf{Y}_i\mathsf{Y}_j] = \Sigma_{ij}$ and $\mathbf{R} \perp \mathbf{X}$.

Step 2: estimated means do not change the first-order term. Under MCAR the centered summand has mean zero, $\widehat{\mathsf{q}}_{ij} \xrightarrow{\mathbb{P}} q_{ij} > 0$, and $\widehat{\mu}_i - \mu_i = O_{\mathbb{P}}(N^{-1/2})$. Replacing true means by observed means changes the covariance estimate by terms such as

$$(\widehat{\mu}_i - \mu_i)\frac{\sum_t \mathsf{R}_{ti}\mathsf{R}_{tj}\mathsf{Y}_{tj}}{\mathsf{N}_{ij}} \quad \text{and} \quad (\widehat{\mu}_i - \mu_i)(\widehat{\mu}_j - \mu_j),$$

both $O_{\mathbb{P}}(N^{-1})$. Joint asymptotic normality follows from the multivariate CLT because fourth moments of $\mathbf{X}$ ensure second moments of every h_{ij}.

Step 3: covariance and the MCAR factorization. The limiting covariance of the $\sqrt{N}$-scaled errors is the covariance of one row's influence vector. Since its coordinates have mean zero,

$$\begin{aligned}\boldsymbol{V}_{(ij),(k\ell)} &= \mathrm{Cov}(\mathsf{h}_{ij}, \mathsf{h}_{k\ell}) = \mathbb{E}[\mathsf{h}_{ij}\mathsf{h}_{k\ell}] \\ &= \frac{\mathbb{E}[\mathsf{R}_i\mathsf{R}_j\mathsf{R}_k\mathsf{R}_\ell(\mathsf{Y}_i\mathsf{Y}_j - \Sigma_{ij})(\mathsf{Y}_k\mathsf{Y}_\ell - \Sigma_{k\ell})]}{q_{ij}q_{k\ell}} \\ &= \frac{q_{ij,k\ell}}{q_{ij}q_{k\ell}}\mathbb{E}[(\mathsf{Y}_i\mathsf{Y}_j - \Sigma_{ij})(\mathsf{Y}_k\mathsf{Y}_\ell - \Sigma_{k\ell})].\end{aligned}$$

The last equality uses independence of the entire mask vector and the data vector, not independence between mask coordinates. Expanding the remaining expectation gives

$$\begin{aligned}&\mathbb{E}[(\mathsf{Y}_i\mathsf{Y}_j - \Sigma_{ij})(\mathsf{Y}_k\mathsf{Y}_\ell - \Sigma_{k\ell})] \\ &\quad = \mathbb{E}[\mathsf{Y}_i\mathsf{Y}_j\mathsf{Y}_k\mathsf{Y}_\ell] - \Sigma_{ij}\mathbb{E}[\mathsf{Y}_k\mathsf{Y}_\ell] - \Sigma_{k\ell}\mathbb{E}[\mathsf{Y}_i\mathsf{Y}_j] + \Sigma_{ij}\Sigma_{k\ell} \\ &\quad = \mathbb{E}[\mathsf{Y}_i\mathsf{Y}_j\mathsf{Y}_k\mathsf{Y}_\ell] - \Sigma_{ij}\Sigma_{k\ell}.\end{aligned}$$

Substitution proves Eq. (5). The formula also covers shared indices because observation indicators satisfy $\mathsf{R}_i^2 = \mathsf{R}_i$. It does not assume Gaussian data. The first-order sampling covariance of the unscaled pairwise estimator is $\boldsymbol{V}/N$, not $\boldsymbol{V}$; this is an asymptotic description, not an exact finite-sample covariance identity.

Step 4: consistency of the empirical covariance estimate. The law of large numbers yields consistency of $N^{-1}\sum_t \mathbf{h}_t\mathbf{h}_t^{\mathsf{T}}$. Plug-in means, covariances and observation probabilities change it by $o_{\mathbb{P}}(1)$: the relevant coefficients are consistent and the required empirical moments have degree at most four. Selection and duplication by a fixed matrix preserve both the CLT and consistency. This argument does not require an invertible covariance of the influence functions. □

Table 5: Average runtime in seconds per query for disjoint blocks with $|A| = |B| = 10$, Gaussian noise, and 15% missingness. Preprocessing is cached unless otherwise indicated.

Method	$N = 500$	$N = 1000$	$N = 2000$	$N = 5000$
MART (without caching)	0.22876	0.23075	0.23465	0.24698
MART (cached)	0.02509	0.02531	0.02530	0.02568
Pairwise minimum	0.00273	0.00496	0.00422	0.00353
Pairwise average	0.00250	0.00465	0.00442	0.00350
Mean imputation	0.00257	0.00421	0.00428	0.00445
Complete-data reference	0.00265	0.00547	0.00457	0.00372
Deletion	0.00245	0.00439	0.00470	0.00338

Each entry is the arithmetic mean over successful datasets, with 30 datasets across ranks $0, 1, 2$ and three evaluations per dataset. Execution is serial with one BLAS thread.

Equivalently, with true means temporarily fixed, the covariance can be expressed through the estimating equations $\mathsf{g}_{ij,t}(\mathbf{c}) = \mathsf{R}_{ti}\mathsf{R}_{tj}(\mathsf{Y}_{ti}\mathsf{Y}_{tj} - c_{ij})$. Their derivative matrix is $-\boldsymbol{D}_q$ with $\boldsymbol{D}_q = \operatorname{diag}(q_{ij})$. The usual form is $\boldsymbol{D}_q^{-1}\mathbb{E}[\mathbf{g}\mathbf{g}^\mathsf{T}]\boldsymbol{D}_q^{-1}$, which is precisely the covariance of Eq. (7). Estimating means adds no first-order term here because the expected derivative with respect to a mean is zero under MCAR.

E Kronecker Fitting and Its Limits

Proof of Theorem 4. The vectorization identity and symmetry of the inverse roots imply

$$\operatorname{vec}(\boldsymbol{E})^\mathsf{T}(\boldsymbol{K}_\mathbf{B}^{-1} \otimes \boldsymbol{K}_\mathbf{A}^{-1}) \operatorname{vec}(\boldsymbol{E}) = \|\boldsymbol{K}_\mathbf{A}^{-1/2}\boldsymbol{E}\boldsymbol{K}_\mathbf{B}^{-1/2}\|_F^2.$$

Multiplication by invertible matrices preserves rank, so $\boldsymbol{C} \mapsto \boldsymbol{K}_\mathbf{A}^{-1/2}\boldsymbol{C}\boldsymbol{K}_\mathbf{B}^{-1/2}$ maps the rank-$\le r$ set bijectively to itself. The best rank-r approximation in Frobenius norm is the truncated SVD; its residual norm squared is the sum of squared discarded singular values. This proves the minimum-distance identity for the statistic. □

F Local Geometry and Mixture Calibration

At a rank-r matrix with SVD bases $\boldsymbol{U}_0, \boldsymbol{W}_0$, the tangent space of the rank-r manifold is

$$\mathcal{T}_0 = \{\boldsymbol{E} : \boldsymbol{U}_{0,\perp}^\mathsf{T}\boldsymbol{E}\boldsymbol{W}_{0,\perp} = 0\}. \tag{20}$$

Its Frobenius-orthogonal normal space has basis $\boldsymbol{N}_0 = \boldsymbol{W}_{0,\perp} \otimes \boldsymbol{U}_{0,\perp}$. To see the first-order constraint directly, rotate a nearby matrix into blocks whose top-left $r \times r$ block $\boldsymbol{D} + \boldsymbol{E}_{11}$ is invertible. Rank r requires $\boldsymbol{E}_{22} = \boldsymbol{E}_{21}(\boldsymbol{D} + \boldsymbol{E}_{11})^{-1}\boldsymbol{E}_{12}$, so $\boldsymbol{E}_{22} = 0$ to first order. The difference between the curved rank set and its tangent is second order.

Proof of Theorem 5. Use the fitted factors $\widehat{\boldsymbol{K}}_{\mathbf{A},N}$ and $\widehat{\boldsymbol{K}}_{\mathbf{B},N}$ to define

$$\boldsymbol{H}_{0,N} = \widehat{\boldsymbol{K}}_{\mathbf{A},N}^{-1/2}\boldsymbol{\Sigma}_{\mathbf{A},\mathbf{B}}\widehat{\boldsymbol{K}}_{\mathbf{B},N}^{-1/2}, \qquad \boldsymbol{E}_N = \widehat{\boldsymbol{K}}_{\mathbf{A},N}^{-1/2}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{A},\mathbf{B}}^{p} - \boldsymbol{\Sigma}_{\mathbf{A},\mathbf{B}})\widehat{\boldsymbol{K}}_{\mathbf{B},N}^{-1/2}.$$

Here $\widehat{\boldsymbol{H}} = \boldsymbol{H}_{0,N} + \boldsymbol{E}_N$. The first matrix has rank exactly r for every N; its retained singular values remain separated from zero with probability tending to one. Moreover, $\boldsymbol{E}_N = O_\mathbb{P}(N^{-1/2})$ by Theorem 2 and stable factors. In a neighborhood of these regular matrices, squared distance to the rank set has the expansion

$$\operatorname{dist}_F^2(\boldsymbol{H}_{0,N} + \boldsymbol{E}_N, \{\boldsymbol{C} : \operatorname{rank}(\boldsymbol{C}) \le r\}) = \|\boldsymbol{U}_{0,N,\perp}^\mathsf{T}\boldsymbol{E}_N\boldsymbol{W}_{0,N,\perp}\|_F^2 + o_\mathbb{P}(\|\boldsymbol{E}_N\|_F^2). \tag{21}$$

This follows from the tangent representation above and the smooth local SVD projection across the gap between retained and zero singular values. For $r = 0$, the formula is exact because the null set is $\{0\}$. Consequently,

$$\mathsf{T}_K = \left\| \boldsymbol{P}_{0,N}^\mathsf{T}\sqrt{N}\ \operatorname{vec}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{A},\mathbf{B}}^{p} - \boldsymbol{\Sigma}_{\mathbf{A},\mathbf{B}}) \right\|_2^2 + o_\mathbb{P}(1),$$

where $\boldsymbol{P}_{0,N} = (\widehat{\boldsymbol{K}}_{\mathbf{B},N}^{-1/2}\boldsymbol{W}_{0,N,\perp}) \otimes (\widehat{\boldsymbol{K}}_{\mathbf{A},N}^{-1/2}\boldsymbol{U}_{0,N,\perp})$. The associated normal subspaces converge. Bases may rotate within these subspaces; such rotations leave both squared norms and covariance eigenvalues unchanged. Slutsky's theorem therefore yields the squared norm of a centered Gaussian vector with covariance $\boldsymbol{P}^\mathsf{T}\boldsymbol{V}\boldsymbol{P}$. Diagonalizing that covariance proves Eq. (16).

The fitted trailing singular subspaces consistently estimate the same normal subspaces. Together with $\widehat{\boldsymbol{V}} \xrightarrow{\mathbb{P}} \boldsymbol{V}$, this implies convergence of the eigenvalues of $\widehat{\boldsymbol{\Omega}}$. At least one positive eigenvalue makes the limiting mixture continuous with an unambiguous positive-level quantile. Consistent quantiles and vanishing numerical error establish pointwise size. Singular $\boldsymbol{V}$ is allowed; only its nonzero projected directions contribute.

This argument requires only consistency, not a root-N rate, for the working factors: their randomness moves the reference matrix $\boldsymbol{H}_{0,N}$ *within* the rank-r manifold. It does not generate an additional first-order normal displacement. By contrast, a preprocessing change to $\widehat{\boldsymbol{\Sigma}}^p_{\mathbf{A},\mathbf{B}}$ must be $o_\mathbb{P}(N^{-1/2})$ or included in its CLT. □

Why a singular raw covariance estimate is not an obstacle here. If $\boldsymbol{V}_{\text{unique}}$ is the covariance of the distinct-pair influence vector, then $\boldsymbol{V} = \boldsymbol{J}_{\mathbf{A},\mathbf{B}}\boldsymbol{V}_{\text{unique}}\boldsymbol{J}_{\mathbf{A},\mathbf{B}}^\mathsf{T}$ may be singular when the block repeats entries. We do not invert it: we fit a positive-definite working weight and evaluate $\boldsymbol{P}^\mathsf{T}\boldsymbol{V}\boldsymbol{P}$. Zero mixture weights remain zero. If *all* projected weights are zero, however, the first-order statistic is degenerate and the nondegenerate calibration theorem does not apply. Keeping duplicate dependence is different from imposing duplicate equalities on the fitted $\boldsymbol{C}$.

G Working Likelihood and Local Power

Proof of Proposition 1. For fixed $\boldsymbol{K}$, the working log likelihood, up to constants independent of $\boldsymbol{C}$, is

$$\ell_K(\boldsymbol{C}) = -\frac{N}{2}(\mathbf{Y} - \text{vec}(\boldsymbol{C}))^\mathsf{T}\boldsymbol{K}^{-1}(\mathbf{Y} - \text{vec}(\boldsymbol{C})).$$

The unrestricted maximum occurs at $\boldsymbol{C} = \widehat{\boldsymbol{\Sigma}}^p_{\mathbf{A},\mathbf{B}}$; the constrained maximum is the minimum-distance fit. The determinant of $\boldsymbol{K}/N$ cancels because $\boldsymbol{K}$ is the same in both models. Thus twice the maximized log-likelihood difference is exactly T_K. If $\boldsymbol{K}$ is fitted from data, this is an algebraic *plug-in working-likelihood* statement, not a conditional distributional assertion given the random fitted $\boldsymbol{K}$. □

For completeness, the full tangent-space calculation identifies which part of an error covariance matters for local power. Given an error vector $\mathbf{e}$, the normal restriction on a residual $\mathbf{v}$ is $\boldsymbol{N}_0^\mathsf{T}\mathbf{v} = \boldsymbol{N}_0^\mathsf{T}\mathbf{e}$. For any $\boldsymbol{K}_0 \succ 0$, a Lagrange multiplier calculation gives

$$\min_{\mathbf{v}:\boldsymbol{N}_0^\mathsf{T}\mathbf{v}=\boldsymbol{N}_0^\mathsf{T}\mathbf{e}} \mathbf{v}^\mathsf{T}\boldsymbol{K}_0^{-1}\mathbf{v} = (\boldsymbol{N}_0^\mathsf{T}\mathbf{e})^\mathsf{T}(\boldsymbol{N}_0^\mathsf{T}\boldsymbol{K}_0\boldsymbol{N}_0)^{-1}(\boldsymbol{N}_0^\mathsf{T}\mathbf{e}), \tag{22}$$

$$\mathbf{v}_{\min} = \boldsymbol{K}_0\boldsymbol{N}_0(\boldsymbol{N}_0^\mathsf{T}\boldsymbol{K}_0\boldsymbol{N}_0)^{-1}\boldsymbol{N}_0^\mathsf{T}\mathbf{e}.$$

Under local alternatives, the same distance expansion gives a Gaussian normal error with quadratic weight $(\boldsymbol{N}_0^\mathsf{T}\boldsymbol{K}_0\boldsymbol{N}_0)^{-1}$. The next proof specializes this calculation to a one-dimensional normal space.

Fixed alternatives and information claims. If $\text{rank}(\boldsymbol{\Sigma}^{\text{alt}}_{\mathbf{A},\mathbf{B}}) > r$, $\widehat{\boldsymbol{\Sigma}}^p_{\mathbf{A},\mathbf{B}} \xrightarrow{\mathbb{P}} \boldsymbol{\Sigma}^{\text{alt}}_{\mathbf{A},\mathbf{B}}$, and stable factors and a bounded consistent covariance estimate persist, the limiting weighted distance from $\boldsymbol{\Sigma}^{\text{alt}}_{\mathbf{A},\mathbf{B}}$ to the closed rank-null set is positive. Thus T_K diverges at rate N while its plug-in reference quantile stays bounded, giving consistency against that fixed alternative. This is not a statement that pairwise moments exhaust observed-data information. Even under MCAR, a correctly specified joint model can leverage relationships with other variables to estimate covariances more efficiently. Proving semiparametric efficiency would require a model-specific efficient influence function and an explicit information bound; neither is claimed here.

G.1 Proof of the Strict Complete-Case Power Comparison

Proof of Theorem 6. Let $\mathbf{e}$ be the last coordinate vector in $\mathbb{R}^{r+1}$ and $\mathbf{n}_0 = \mathbf{e} \otimes \mathbf{e}$. At $\tau = 0$, the cross-block has rank r with a fixed positive singular-value gap. Its one-dimensional normal space is spanned by $\mathbf{n}_0$; the normal local shift is τ. Let $\mathsf{R}_{*,t}$ indicate joint observation of the two residual variables, and let J_t indicate a test-wise complete row. The corresponding normal influence contributions under the null are

$$\mathsf{h}_{*,t} = \frac{\mathsf{R}_{*,t}}{q_*}\mathsf{X}_{*,t}\mathsf{Y}_{*,t}, \qquad \mathsf{h}_{\mathrm{CC},t} = \frac{\mathsf{J}_t}{q_{\mathrm{all}}}\mathsf{X}_{*,t}\mathsf{Y}_{*,t}.$$

Their means are zero at $\tau = 0$. MCAR and independence of the unit-variance residual variables at this null give $\mathbb{E}[\mathsf{X}_*^2\mathsf{Y}_*^2] = 1$, hence

$$\Omega_{\mathrm{PW}} = \mathbb{E}[\mathsf{h}_{*,t}^2] = q_*^{-1}, \qquad \Omega_{\mathrm{CC}} = \mathbb{E}[\mathsf{h}_{\mathrm{CC},t}^2] = q_{\mathrm{all}}^{-1}.$$

Both variances use the original $\sqrt{N}$ scaling. The normal marginal limits under the stated local Gaussian alternatives have means τ and these variances. The centered triangular-array CLTs hold because the Gaussian covariances approach a fixed positive-definite limit, their higher moments are uniformly bounded, and the mask probabilities are fixed. Random pair counts and estimated means have the same first-order expansions as under the null. The covariance estimate has the same probability limit along this local sequence.

Set $S = \mathbf{n}_0^{\mathsf{T}}\boldsymbol{K}_0\mathbf{n}_0 > 0$. The local profile in Eq. (22) gives

$$\mathsf{T}_K \xrightarrow{d} \mathsf{G}_{\mathrm{PW}}^2/S, \qquad \mathsf{G}_{\mathrm{PW}} \sim \mathcal{N}(\tau, \Omega_{\mathrm{PW}}).$$

In one normal dimension the whitened normal map equals $\mathbf{n}_0/\sqrt{S}$ up to sign: it spans the same normal space and satisfies $\boldsymbol{P}^{\mathsf{T}}\boldsymbol{K}_0\boldsymbol{P} = 1$. Therefore the sole mixture weight is Ω_{PW}/S. Consistent plug-in calibration gives a critical value converging to $(\Omega_{\mathrm{PW}}/S)c_\alpha$. Continuity of the limiting law implies

$$\beta_{\mathsf{T}_K}(\tau) = \mathbb{P}\{\mathsf{G}_{\mathrm{PW}}^2/\Omega_{\mathrm{PW}} > c_\alpha\} = \mathbb{P}\{\chi_1^2(q_*\tau^2) > c_\alpha\}.$$

For the complete-case sample, write N_{all} for its size and use N_{all} in place of N in the CCA statistic of Eq. (2). Then $N_{\mathrm{all}}/N \xrightarrow{\mathbb{P}} q_{\mathrm{all}}$. MCAR leaves the population distribution unchanged. Gaussianity and independence of the residual pair from the signal subvectors make the trailing canonical correlation equal to the standardized residual covariance to first order (up to sign before taking the singular value). The other canonical correlations have a fixed gap from zero. The Gaussian CCA log statistic, including any asymptotically negligible Bartlett correction, consequently has local limit

$$\mathsf{T}_{\mathrm{CCA}} \xrightarrow{d} \mathsf{G}_{\mathrm{CC}}^2/\Omega_{\mathrm{CC}} \sim \chi_1^2(q_{\mathrm{all}}\tau^2), \qquad \mathsf{G}_{\mathrm{CC}} \sim \mathcal{N}(\tau, \Omega_{\mathrm{CC}}).$$

This proves the second power formula. Finally, for $c_\alpha > 0$, $\mathbb{P}\{(\mathsf{Z}+\mu)^2 > c_\alpha\}$ is strictly increasing in $\mu > 0$: its derivative is $\phi(\sqrt{c_\alpha} - \mu) - \phi(\sqrt{c_\alpha} + \mu) > 0$, where ϕ is the standard normal density. Taking $\mu = \sqrt{q}\,|\tau|$, with $q = q_*$ or q_{all}, shows that $q_* > q_{\mathrm{all}}$ yields a strict improvement for either sign of $\tau \neq 0$. At $\tau = 0$ both limits are central χ_1^2, giving rejection probability α. □

H Relation to Complete-Data CCA

Without missingness, $\widehat{\boldsymbol{\Sigma}}^p = \widehat{\boldsymbol{\Sigma}}$ under the same $1/N$ normalization. For Gaussian observations the fourth-moment formula gives

$$\boldsymbol{V}_{(ij),(k\ell)} = \Sigma_{ik}\Sigma_{j\ell} + \Sigma_{i\ell}\Sigma_{jk}. \tag{23}$$

The second term is a cross-covariance contribution. It does not generally vanish from the full raw covariance-error matrix when $\mathrm{rank}(\boldsymbol{\Sigma}_{\mathbf{A},\mathbf{B}}) = r > 0$. Consequently, the best single-Kronecker fit to that full matrix need not equal $\boldsymbol{\Sigma}_{\mathbf{B},\mathbf{B}} \otimes \boldsymbol{\Sigma}_{\mathbf{A},\mathbf{A}}$ even without missingness.

There is a narrower, useful connection. For disjoint blocks, *if* the working factors are specifically chosen as $\widehat{\boldsymbol{K}}_{\mathbf{A}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{A},\mathbf{A}}$ and $\widehat{\boldsymbol{K}}_{\mathbf{B}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{B},\mathbf{B}}$, the whitened singular values in Eq. (14) are the canonical correlations. The statistic becomes $N\sum_{j>r}\widehat{\mathsf{c}}_j^2$, rather than the finite-sample log statistic. At a fixed regular Gaussian null, each discarded $\widehat{\mathsf{c}}_j = O_{\mathbb{P}}(N^{-1/2})$, so

$$\mathsf{T}_{\mathrm{CCA}} - N\sum_{j>r}\widehat{\mathsf{c}}_j^2 = N\sum_{j>r}O_{\mathbb{P}}(\widehat{\mathsf{c}}_j^4) = O_{\mathbb{P}}(N^{-1}). \tag{24}$$

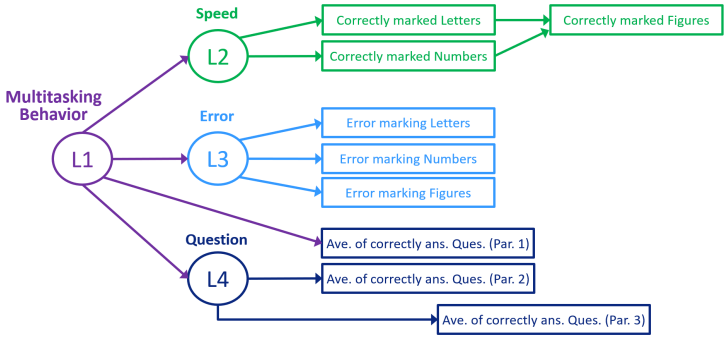


Figure 8: Given complete Multi-tasking behavior dataset, the causal graph produce by RLCD.

A bounded additive correction to N also has an $o_{\mathbb{P}}(1)$ effect in this regime. In canonical coordinates the trailing population cross-covariances vanish, removing the cross term from the relevant Gaussian normal covariance and yielding the usual chi-square reference. This calculation explains local equivalence for that particular correctly standardized choice. It does not prove that the empirical full-covariance Kronecker fit used by B-1-uncond selects those factors, and it does not establish an exact finite-sample CCA identity. For non-Gaussian data, fourth cumulants can remain even with no missingness; covariance-based calibration then has a different target from the classical Gaussian reference.

I Implementation Correspondence and Reuse

The draft describes the corrected minimum-distance version, not earlier prototypes with a second inverse weighting of the normal residual. The following details are important for reproducing its equations.

Vectorization and statistic. The Python implementation uses row-major vectorization. Its product is $\boldsymbol{K}_{\mathbf{A}} \otimes \boldsymbol{K}_{\mathbf{B}}$ and its normal map is $(\widehat{\boldsymbol{K}}_{\mathbf{A}}^{-1/2}\boldsymbol{U}_{\perp}) \otimes (\widehat{\boldsymbol{K}}_{\mathbf{B}}^{-1/2}\boldsymbol{W}_{\perp})$. These are permutation-equivalent to the column-major convention in this paper. The same effective factor roots are used to whiten, unwhiten the fitted matrix, and form the calibration map. The statistic is exactly $N \sum_{j>r} \mathsf{s}_j^2$. The extra regularized dense product returned by the generic Kronecker fitting helper is not used as a different weight for this statistic.

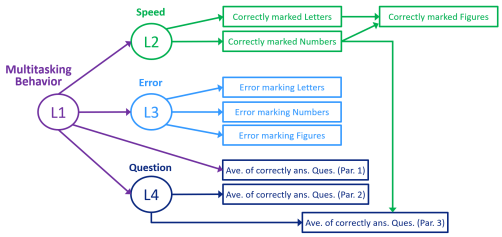
Caching. The shared cache is tied to an immutable input dataset and stores unique-pair moments once. Changing data or its missingness mask requires a new cache; changing only $\mathbf{A}, \mathbf{B}, r$ does not. The current query cache includes the method variant as well as the selected component variables and rank, preventing accidental reuse across conditional and unconditional targets. Exact repeated queries reuse the final statistic and p-value. Factor caching across *different* ranks is a possible further optimization, not assumed to be implemented here. Current wrappers also copy selected raw-data columns when preparing a query, which can introduce $O(N|\mathbf{A} \cup \mathbf{B}|)$ overhead even though cached mathematical moment operations need not revisit the rows. This overhead should be included in measured post-reuse time.

Reproducible inference. In practice, a simulation of $\mathsf{Q}^{(b)} = \sum_j \widehat{\mathsf{w}}_j \mathsf{Z}_{bj}^2$, with independent standard normal Z_{bj}, gives

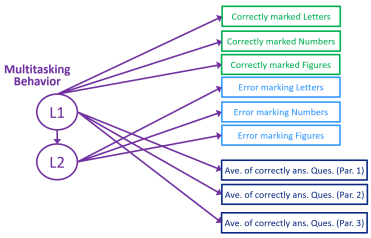
$$\widehat{\mathsf{p}}_{\mathrm{MC}} = \frac{1 + \sum_{b=1}^{B_{\mathrm{MC}}} 1\{\mathsf{Q}^{(b)} \geq \mathsf{T}_K\}}{B_{\mathrm{MC}} + 1}. \tag{25}$$

This samples the estimated reference law, not datasets, and is not a bootstrap or a finite-sample exact permutation test. Its numerical error is of order $B_{\mathrm{MC}}^{-1/2}$; fixed B_{MC} leaves that error present.

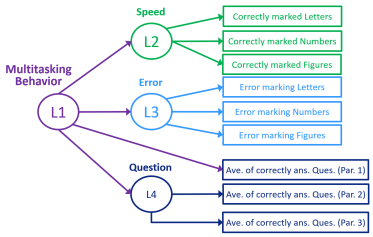
The default reference simulation uses $B_{\mathrm{MC}} = 20{,}000$ with an explicit seed. Its standard Monte Carlo error near a tail probability of 0.05 is about 0.0015. A Satterthwaite option in the code is only a moment-matched approximation to the mixture, not the reference law in Theorem 5. Report the



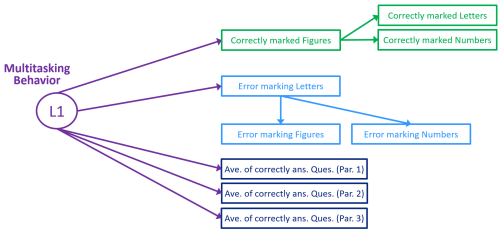
(a) The result produced by MART with RLCD.



(b) The result by pair-wise average with RLCD



(c) The result by EM imputation with RLCD



(d) The result by test-wise deletion with RLCD

Figure 9: Given Multi-tasking behavior dataset with missing values, the causal graph produced by different rank test methods.

tail-evaluation method, floor, seed, query sizes, rank, and cache policy with runtime and calibration results.

J Related Work

Conditional independence and covariance-rank tests. For Gaussian variables, conditional independence is equivalent to a zero partial correlation, which can be assessed with Fisher's z-test (Fisher, 1924; Baba et al., 2004). Beyond this setting, conditional independence can be tested through independence of regression residuals under suitable assumptions (Ramsey, 2014), or through kernel methods (Fukumizu et al., 2007; Gretton et al., 2007; Zhang et al., 2012). The general problem is difficult without restrictions on the underlying distributions; the generalized covariance measure instead tests a regression-based implication of conditional independence (Shah & Peters, 2018). Covariance-rank hypotheses extend zero partial correlation to larger variable sets and are commonly assessed through canonical correlations or likelihood-ratio statistics under complete-data models (Anderson, 1984; Robin & Smith, 2000). Related work has addressed discretized measurements through tests of partial correlation (Sun et al., 2025b;a) and covariance rank (Dong et al., 2025). These developments concern different observation models from the incomplete, continuous measurements studied here.

Rank constraints in causal discovery. Constraint-based algorithms infer graphical structure from statistical restrictions on the observed distribution. PC uses conditional independence to recover a Markov equivalence class under its assumptions (Spirtes & Glymour, 1991); FCI accommodates latent confounding and selection (Spirtes et al., 1995; Colombo et al., 2012), while CCD addresses cyclic structures (Richardson, 1996). Rank constraints contain zero partial correlations as special cases and also encode information about latent structure that conditional independence alone may miss. They have therefore been used for discovery from heterogeneous or nonstationary data (Huang et al., 2020), for latent-variable models (Silva et al., 2006; Huang et al., 2022; Dong et al., 2024), and for latent-variable discovery under selection bias (Dai et al., 2025). Such procedures may evaluate many rank hypotheses on the same dataset, making both calibration and reuse of incomplete observations consequential.

Missing data in causal discovery. Incomplete records are often handled by discarding rows globally, retaining only rows complete for each test, or imputing missing entries. Test-wise deletion retains

more data than global complete-case analysis, but may still discard rows that contain observed pairs relevant to a covariance-rank query. Strobl et al. (2018) established conditions under which test-wise deletion remains sound for FCI-type procedures with nonrandom missingness; Witte et al. (2022) studied its recoverability conditions and compared it with multiple imputation. Other approaches explicitly model the missingness process: Gain & Shpitser (2018) considered structure learning with missing-data models, and Tu et al. (2019) used missingness graphs to identify when the conditional independence relations needed by PC require correction under MAR or MNAR. For Gaussian-copula structure learning, Cui et al. (2019) estimated correlations from pairwise complete observations and used effective sample sizes in partial-correlation tests under MCAR, with a sampling-based extension for MAR. Score-based methods have also learned graphs directly from incomplete data, including an optimal-transport formulation (Vo et al., 2024). These works address graph recovery or missingness-induced selection bias. MART develops a covariance-rank test under MCAR that combines pairwise information across incomplete rows and accounts for the joint sampling uncertainty of those estimates.