PERMUTATION-BASED RANK TEST IN THE PRESENCE OF DISCRETIZATION AND APPLICATION IN CAUSAL DISCOVERY WITH MIXED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances have shown that statistical tests for the rank of cross-covariance matrices play an important role in causal discovery. These rank tests include partial correlation tests as special cases and provide further graphical information about latent variables. Existing rank tests typically assume that all the continuous variables can be perfectly measured, and yet, in practice many variables can only be measured after discretization. For example, in psychometric studies, the continuous level of certain personality dimensions of a person can only be measured after being discretized into order-preserving options such as disagree, neutral, and agree. Motivated by this, we propose **M**ixed data **P**ermutation-based **R**ank **T**est (MPRT), which properly controls the statistical errors even when some or all variables are discretized. Theoretically, we establish the exchangeability and estimate the asymptotic null distribution by permutations; as a consequence, MPRT can effectively control the Type I error in the presence of discretization while previous methods cannot. Empirically, our method is validated by extensive experiments on synthetic data and real-world data to demonstrate its effectiveness as well as applicability in causal discovery (our code will be available).

1 INTRODUCTION AND RELATED WORK

Recent advances have shown that the rank of a cross-covariance matrix and its statistical test play essential roles in multiple fields of statistics especially in causal discovery (Sullivant et al., 2010; Spirtes, 2013). From one perspective, Independence and Conditional Independence (CI) are crucial concepts in causal discovery and Bayesian network learning (Pearl et al., 2000; Spirtes et al., 2000; Koller & Friedman, 2009) due to its relation to d-separations (Pearl, 1988), and it has been shown that rank tests take those linear CI tests as special cases (Sullivant et al., 2010; Di, 2009; Dong et al., 2024). From another point of view, rank of a cross-covariance matrix corresponds to t-separations in a graph (Sullivant et al., 2010), which contain graphical information that can be used to identify latent variables (Huang et al., 2022; Dong et al., 2024). A more detailed discussion about related work can be found in Appendix D.

Existing statistical rank tests (Anderson, 1984) are often built upon Canonical Correlation Analysis (CCA) (Jordan, 1875; Hotelling, 1992), with a likelihood ratio based test statistics. Despite their effectiveness, existing methods rely on the strong assumption that all the variables concerned can be perfectly measured. However, in many fields, it is often the case that the best available data are just discretized approximations of some underlying continuous variable (formally defined in Eq. 1). For example, in mental health, anxiety levels are often categorized into levels such as mild, moderate, or severe, according to some latent thresholds (Johnson et al., 2019). Examples can be found in multiple fields such as finance (Changsheng & Yongfeng, 2012), psychology (Lord & Novick, 2008), biometrics (Finney, 1952) and econometrics (Nerlove & Press, 1973), where continuous variables are often assumed to be observed as discretized values.

When discretization is present, existing rank tests can hardly work. The main reason lies is that the discretized values only reflect the order of the data, leading to cross-covariance estimates that may differ significantly from the underlying cross-covariance matrix (also illustrated in Figure 1). Furthermore, even though the true underlying cross-covariance matrix can be estimated by maximum likelihood-based methods such as polychoric and polyserial correlations (Olsson et al., 1982; Olsson, 1979), they cannot be directly plugged into existing rank tests. This is because the involved discretization and maximum likelihood processes change the distribution of test statistics to a considerable extent and thus the p-values cannot be correctly calculated. As a consequence, Type I errors of existing methods cannot be effectively controlled. Both of these points are elaborated in Section 2.2.

To properly address the issue of discretization, in this paper, we propose a novel statistic rank test based on permutation, i.e., Mixed data Permutation-based Rank Test (MPRT) that can accommodate continuous, partially discretized, or fully discretized observations. Specifically, in the presence of discretization, the underlying cross-covariance can be estimated by maximum likelihood estimator, but the information loss resulting from discretization and the additional estimation steps make the derivation of the null distribution highly non-trivial. To this end, we start with the continuous case and establish exchangeability of linear projections of concerned variables (captured by Theorem 4), based on which the null distribution can be empirically estimated by permutations. When some observations are discretized, the exchangeability still holds but we do not have direct access to permutable data. Fortunately, we show that the concerned statistic distribution can still be consistently estimated by properly using permuted discretized observations (captured by Theorem 5). We summarize our major contributions as follows.

- To our best knowledge, we propose the first statistic rank test i.e., Mixed data Permutation-based Rank Test (MPRT), that properly deals with the problem of discretization. Rank test takes partial correlation CI test as a special case and thus the problem is crucial to many scientific fields such as psychology, biometrics, and econometrics, where discretizations are ubiquitous.
- Theoretically, we estimate the asymptotic null distribution by effectively making use of data permutations, and thus properly controls the Type I error. The setting considered is rather general: for the test of $\text{rank}(\Sigma_{\mathbf{X},\mathbf{Y}})$, both $\mathbf{X}$ and $\mathbf{Y}$ are allowed to be either fully continuous, partially discretized, or fully discretized. Therefore, our method also includes the fully-continuous rank test as a special case.
- Empirically, we validate our novel rank test under multiple synthetic settings where our method is shown to control Type I error properly and Type II error effectively, while existing methods cannot. We also use a real-world dataset to show the practicability of the proposed rank test and illustrate its application in causal discovery.

2 PRELIMINARIES

2.1 PROBLEM SETTING

Suppose that we have a set of M observed random variables $\mathbf{V} = \{\mathsf{V}_j\}_{j=1}^{M}$ that are jointly Gaussian. However, for some of these variables, direct observations are unavailable. We use $\mathbb{C}_{\mathbf{V}}$ and $\mathbb{D}_{\mathbf{V}}$ to denote the index set of those variables in $\mathbf{V}$ that we have direct observations and that of those we only have order-preserving discretized observations, respectively. Assume that we have N i.i.d., observations of these variables. The underlying true data matrix is $\boldsymbol{D} \in \mathbb{R}^{N\times M}$, while we only have access to $\tilde{\boldsymbol{D}}$, where some columns are discretized. Specifically, for $j \in \mathbb{C}_{\mathbf{V}}$, $\tilde{\boldsymbol{D}}_{:,j} = \boldsymbol{D}_{:,j}$, while for those $j \in \mathbb{D}_{\mathbf{V}}$, the observations are discretized in the following fashion:

$$\tilde{\boldsymbol{D}}_{i,j} = t, \text{ if } T_t^j < \boldsymbol{D}_{i,j} \leq T_{t+1}^j, \text{ for } i \in \{1, ..., N\}, t \in \{1, ..., C_j\}, \tag{1}$$

where C_j is the cardinality of the domain of the discretized observation of V_j, T_t^j refers to the t-th threshold for variable V_j, $T_1^j \triangleq -\infty$, and $T_{C_j+1}^j \triangleq \infty$.

We are interested in the rank of the population cross-covariance matrix over certain combinations of variables, e.g., $\Sigma_{\mathbf{X},\mathbf{Y}}$, where $\mathbf{X} \subseteq \mathbf{V}$ and $\mathbf{Y} \subseteq \mathbf{V}$ ($\mathbf{X}$ and $\mathbf{Y}$ are not necessarily disjoint). The rank information is crucial to causal discovery (Spirtes et al., 2000) and will be detailed in Section 2.2. Ideally, we would expect that we have infinite datapoints and there is no discretization; in this case, the sample covariance $\hat{\Sigma}_{\mathbf{X},\mathbf{Y}}$ would be exactly the same as the population covariance, and the rank can be easily calculated by linear algebra. However, in practice we only have finite datapoints and for some of the variables we only have discretized observations. Thus, it is crucial to consider the following problem: in the finite sample case and in the presence of discretization, we only have access to $\tilde{\boldsymbol{D}}$ instead of $\boldsymbol{D}$, how to build a valid statistic test that properly controls the Type I error for testing the rank of a cross-covariance matrix $\Sigma_{\mathbf{X},\mathbf{Y}}$?

2.2 Why this Problem is Important?

In this section we will briefly discuss why rank test is important in the context of causal discovery as well as why it is crucial to deal with discretization.

(i) Rank Test Takses Linear CI Test as a Special Case

In causal discovery, we aim to find the underlying causal graph among variables given observational data. The most classical approach is to use conditional independence (CI) relationships to identify d-separations in a graph; see, e.g., the PC algorithm (Spirtes et al., 2000). This idea is captured by the following theorem.

Theorem 1 (Conditional Independence and D-separation (Pearl, 1988))**.** *Under the Markov and faithfulness assumption, for disjoint sets of variables* $\mathbf{A}$*,* $\mathbf{B}$ *and* $\mathbf{C}$*,* $\mathbf{C}$ *d-separates* $\mathbf{A}$ *and* $\mathbf{B}$ *in graph* $\mathcal{G}$*, iff* $\mathbf{A} \perp\!\!\!\perp \mathbf{B}|\mathbf{C}$ *holds for every distribution in the graphical model associated to* $\mathcal{G}$*.*

In practice, we often consider linear causal models where the CI test can be done by e.g., Fisher-Z (Fisher et al., 1921). It has been shown that, for linear causal models, d-separations between variables can also be uncovered by rank tests, which is summarized in the following theorem.

Theorem 2 (D-separation by Rank Test (Dong et al., 2024))**.** *Suppose a linear causal model with graph* $\mathcal{G}$ *and assume rank faithfulness (Spirtes, 2013). For disjoint variable sets* $\mathbf{A}$*,* $\mathbf{B}$*, and* $\mathbf{C}$*, we have* $\mathbf{C}$ *d-separates* $\mathbf{A}$ *and* $\mathbf{B}$ *in graph* $\mathcal{G}$*, if and only if rank*$(\Sigma_{\mathbf{A}\cup\mathbf{C},\mathbf{B}\cup\mathbf{C}}) = |\mathbf{C}|$.

The above Theorem 2 says that d-separations can also be inferred from rank of a cross-covariance matrix, and thus for causal discovery of linear causal models, partial correlation test / linear CI test can be substituted by rank test.

(ii) Rank Relates to T-separation that Indicates Latent Variables

Next, we show that rank of cross-covariance informs something beyond d-separations. Specifically, t-separations (Sullivant et al., 2010) can be inferred from rank, and t-separations can be used to identify latent variables. The relation between rank and t-separations is given as follows.

Theorem 3 (Rank and T-separation (Sullivant et al., 2010))**.** *Given two sets of variables* $\mathbf{A}$ *and* $\mathbf{B}$ *from a linear model with graph* $\mathcal{G}$ *and assume rank faithfulness. We have:*

$$\textit{rank}(\Sigma_{\mathbf{A},\mathbf{B}}) = \min\{|\mathbf{C_A}| + |\mathbf{C_B}| : (\mathbf{C_A}, \mathbf{C_B}) \textit{ t-separates } \mathbf{A} \textit{ from } \mathbf{B} \textit{ in } \mathcal{G}\}, \tag{2}$$

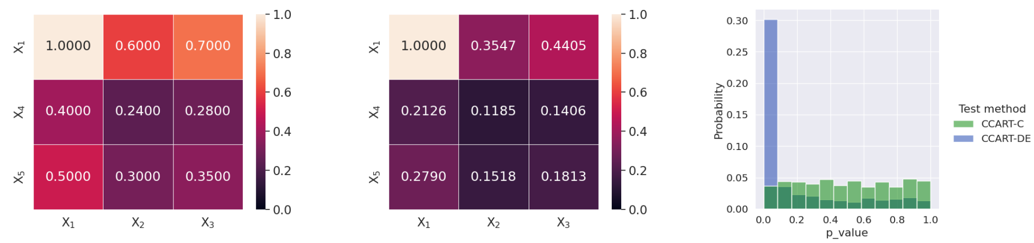
where $\Sigma_{\mathbf{A},\mathbf{B}}$ *is the cross-covariance over* $\mathbf{A}$ *and* $\mathbf{B}$*.*

The left-hand side of Equation 2 is about properties of the observational distribution, while the right-hand side describes properties of the graph. An example highlighting the greater informativeness of rank compared to CI is as follows. Consider the graph $\mathcal{G}$ in Figure 5, where $\{\mathsf{X}_1, \mathsf{X}_2\}$ and $\{\mathsf{X}_3, \mathsf{X}_4\}$ are d-separated by L_1, but we can never infer that from any CI test, i.e., we can never check whether $\{\mathsf{X}_1, \mathsf{X}_2\} \perp\!\!\!\perp \{\mathsf{X}_3, \mathsf{X}_4\}|\mathsf{L}_1$ holds, as L_1 is not observed. In contrast, using rank information, we can infer that $\text{rank}(\Sigma_{\{\mathsf{X}_1\mathsf{X}_2\},\{\mathsf{X}_3\mathsf{X}_4\}}) = 1$, which implies $\{\mathsf{X}_1, \mathsf{X}_2\}$ and $\{\mathsf{X}_3, \mathsf{X}_4\}$ are t-separated by one latent variable. The rationale behind is that the t-separation of two set of variables $\mathbf{A}, \mathbf{B}$ by $(\mathbf{C_A}, \mathbf{C_B})$ can be inferred through rank information, without actually observing any element in $(\mathbf{C_A}, \mathbf{C_B})$. A more detailed discussion can be found in (Dong et al., 2024).

(iii) Discretization is Ubiquitous and Needs to be Handled

Discretization is ubiquitous in many scientific fields. For instance, it is common to come across concepts that cannot be measured directly, such as depression, anxiety, attitude, and the observations of such variables are often the result of coarse-grained measurement of the underlying continuous ones. More examples can be found in fields like psychology (Lord & Novick, 2008), biometrics (Finney, 1952) and econometrics (Nerlove & Press, 1973), where it is widely accepted to assume a continuous variable underlies a dichotomous or polychotomous observed one.

In the context of rank test, what should we do to deal with such a ubiquitous discretization problem? One naive way is to just treat these ordinal values as continuous ones and test the rank of a cross-covariance matrix as usual, and yet it cannot work. The reason lies in that the observed values of these discretized variables just represent the ordering and the values can be rather arbitrary. For example, assume that the original continuous observations are discretized into three levels represented by $\{1, 2, 3\}$ respectively; one can alternatively uses $\{1, 2, 2.1\}$ or $\{1, 2, 10^{16}\}$ to represent the three



(a) Population cross-covariance matrix over continuous variables.

(b) Cross-covariance matrix using discretized data with $N \to \infty$.

(c) Distributions of p-values of CCART-C and CCART-DE.

Figure 1: Subfigures (a) and (b) together show that we cannot directly take the discrete values for the calculation of rank of the covariance matrix. Subfigure (c) shows that directly plugging an estimated cross-covariance matrix into a rank test does not work as Type I cannot be controlled.

levels. If we directly use the ordinal values, the resulting cross-covariance matrix can be very different from the ground truth one, leading to meaningless results. An example can be found in Figure 1, where (a) shows the population cross-covariance and (b) shows the counterpart calculated by using discretized observations. Even with infinite samples, the two matrices are totally different, and the rank of the matrix in (a) is 1 while rank of that in (b) is 3. Next, we will show that, even if we can use maximum likelihood to estimate the correlation first, the problem is still highly non-trivial.

2.3 Classical Rank Test with Estimated Correlation

We have shown that the naive solution of directly using the ordinal values cannot work. Thus, one may wonder another straightforward one - estimate the correlations first (which can be done by maximizing likelihood, detailed in Section 3.3), and then plug the estimated correlations into a standard CCA rank test. In this section we will show that this straightforward solution cannot work either; more specifically, the Type-I errors cannot be effectively controlled.

We start with a brief introduction to the classical rank test, which is based on Canonical Correlation Analysis (CCA) (Jordan, 1875; Hotelling, 1992). The key design of a test typically is to find a suitable statistic and to derive its distribution under the null hypothesis. As for rank test of cross-covariance $\Sigma_{\mathbf{X},\mathbf{Y}}$, statistics based on CCA scores between $\mathbf{X}$ and $\mathbf{Y}$ are found to be very effective. For $|\mathbf{X}| = P$, $|\mathbf{Y}| = Q$, and $K = \min(P, Q)$, the CCA problem is as follows:

$$\max_{\boldsymbol{A}\in\mathbb{R}^{P\times K},\boldsymbol{B}\in\mathbb{R}^{Q\times K}} \operatorname{tr}(\boldsymbol{A}^T\hat{\Sigma}_{\mathbf{X},\mathbf{Y}}\boldsymbol{B}), \text{ s.t., } \boldsymbol{A}^T\hat{\Sigma}_{\mathbf{X}}\boldsymbol{A} = \boldsymbol{B}^T\hat{\Sigma}_{\mathbf{Y}}\boldsymbol{B} = \boldsymbol{I}. \tag{3}$$

Assume that the solution to Eq. 3 leads to CCA scores between $\mathbf{X}$ and $\mathbf{Y}$ as $\{r_i\}_{i=1}^K$. With the null hypothesis that rank$(\Sigma_{\mathbf{X},\mathbf{Y}}) \leq k$, referred to as $\mathcal{H}_0^k$, we would expect that the top-k CCA scores are non-zero and the rest ones are all zero. This leads to a likelihood-ratio-based test statistics (Anderson, 1984) under $\mathcal{H}_0^k$ as follows.

$$\lambda_k = -\left(N - \frac{P+Q+3}{2}\right)\ln(\Pi_{i=k+1}^K(1-r_i^2)), \tag{4}$$

which has been shown to approximately follow a chi-square distribution with degree of freedom $(P-k+1)(Q-k+1)$. To perform the rank test, one only has to calculate λ_k and the related chi-square distribution to get the p-value.

In Eq 3, $\hat{\Sigma}_{\mathbf{X},\mathbf{Y}}$ refers to the sample covariance $\frac{\boldsymbol{D}^{\mathbf{X}T}\boldsymbol{D}^{\mathbf{Y}}}{N-1}$. In the presence of discretization, we only have access to $\tilde{\boldsymbol{D}}^{\mathbf{X}}$ and $\tilde{\boldsymbol{D}}^{\mathbf{Y}}$, but we can still estimate the cross-correlation by maximizing the likelihood (detailed in Section 3.3), and take the estimation into Eq. 3 to calculate the CCA scores and thus the test statistics. However, due to the information loss introduced by discretization and the additional maximum likelihood steps, the distribution of the statistics is changed to a considerable extent. An example is shown in Figure 1 (c), where CCART-C refers to CCA rank test using the original continuous observations and CCART-DE refers to first estimating the correlations by maximum likelihood using discrete data and then plugging it into the CCA rank test. As shown, the p-values of CCART-C are uniformly distributed while the p-values of CCART-DE are clearly not;

most of them are near to zero and thus the test tends to reject everything, leading to unacceptably large Type I errors (also validated in Section 4.2 and Figure 2).

Ideally, we would expect to derive the updated distribution of the statistics, and yet the involved likelihood maximization steps make it very difficult. Therefore, we aim to solve this problem by estimating the empirical cdf of the null distribution using permutations, detailed in what follows.

3 Mixed Data Permutation-based Rank Test

In this section, we propose our novel Mixed data Permutation-based Rank Test (MPRT). We start with the all continuous case.

3.1 All Continuous Case

Assume that we are interested in the rank of $\Sigma_{\mathbf{X},\mathbf{Y}}$, where $|\mathbf{X}| = P$ and $|\mathbf{Y}| = Q$ and their corresponding data matrices are $\tilde{\boldsymbol{D}}^{\mathbf{X}} \in \mathbb{R}^{N\times P}$ and $\tilde{\boldsymbol{D}}^{\mathbf{Y}} \in \mathbb{R}^{N\times Q}$ respectively. The first crucial step is to solve the CCA problem defined in Eq 3, by Singular Value Decomposition (SVD) as follows.

$$\boldsymbol{USV} = \hat{\Sigma}_{\mathbf{X}}^{-\frac{1}{2}}\hat{\Sigma}_{\mathbf{X},\mathbf{Y}}\hat{\Sigma}_{\mathbf{Y}}^{-\frac{1}{2}}, \tag{5}$$

$$\boldsymbol{A} = \hat{\Sigma}_{\mathbf{X}}^{-\frac{1}{2}T}\boldsymbol{U} \text{ and } \boldsymbol{B} = \hat{\Sigma}_{\mathbf{Y}}^{-\frac{1}{2}T}\boldsymbol{V}^{T}, \tag{6}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are two linear projection matrices and the two CCA variables are $\mathbf{C}_{\mathbf{X}} = \boldsymbol{A}^T\mathbf{X}$ and $\mathbf{C}_{\mathbf{Y}} = \boldsymbol{B}^{\boldsymbol{T}}\mathbf{Y}$. $\mathbf{C}_{\mathbf{X}}$ and $\mathbf{C}_{\mathbf{Y}}$ have two good properties: (i) $\hat{\Sigma}_{\mathbf{C}_{\mathbf{X}}} = \hat{\Sigma}_{\mathbf{C}_{\mathbf{Y}}} = \boldsymbol{I}$, and $\hat{\Sigma}_{\mathbf{C}_{\mathbf{X}},\mathbf{C}_{\mathbf{Y}}}$ is a diagonal matrix; (ii) under null hypothesis $\mathcal{H}_0^k$: rank$(\Sigma_{\mathbf{X},\mathbf{Y}}) \le k$, only the top-k diagonal entries of $\Sigma_{\mathbf{C}_{\mathbf{X}},\mathbf{C}_{\mathbf{Y}}}$ are nonzero and the rest of the diagonal entries should be zero. Taking these two into consideration, we have the exchangeability between $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$, which is formalized in the following Theorem 4 (proof of which can be found in Appendix).

Theorem 4 (Exchangeability of $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$)**.** *Given a set of variables $\mathbf{V}$ that are jointly gaussian, under null hypothesis $\mathcal{H}_0^k$: rank$(\Sigma_{\mathbf{X},\mathbf{Y}}) \le k$, where $\mathbf{X},\mathbf{Y} \subseteq \mathbf{V}$, random vectors $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$ are asymptotically independent with each other.*

Based on the exchangeability between $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$, we can permute the data matrix of $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$ in order to get resampling of $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$. Specifically, given a random permutation matrix $\boldsymbol{P}$, $\boldsymbol{P}\tilde{\boldsymbol{D}}^{\mathbf{C}_{\mathbf{X}}}_{:,k:}$ and $\tilde{\boldsymbol{D}}^{\mathbf{C}_{\mathbf{Y}}}_{:,k:}$ together serve as N i.i.d. resamplings from the joint distribution of $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$. Further, the statistics in Eq. 4 only depends on the k-th to K-th CCA scores between $\mathbf{X}$ and $\mathbf{Y}$, which can be equivalently calculated by the first to $(K-k)$-th CCA scores between $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$, formally captured by the following Lemma 1.

Lemma 1 (Alternative Way to Calculate Statistic in Eq. 4)**.** *Let the CCA score between $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$ be $\{\hat{r}_i\}_1^{K-k}$. Then the statistic defined in Eq. 4 can also be formulated as:*

$$\lambda_k = -\left(N - \frac{P+Q+3}{2}\right)\ln(\Pi_{i=1}^{K-k}(1-\hat{r}_i^2)). \tag{7}$$

By Lemma 1, we know that the test statistics only depends on $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$. Further, $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$ can be resampled by permutations. Taking these two into consideration, we can make use of permutation to estimate the empirical CDF of the null distribution, and thus correctly calculate the p-value. Below we give a detailed description of the procedure to do the permutation and consequently calculate the p-value. Given $\boldsymbol{A}$ and $\boldsymbol{B}$, we have the observed data matrix of the two canonical variables as $\tilde{\boldsymbol{D}}^{\mathbf{C}_{\mathbf{X}}} = \tilde{\boldsymbol{D}}^{\mathbf{X}}\boldsymbol{A}$ and $\tilde{\boldsymbol{D}}^{\mathbf{C}_{\mathbf{Y}}} = \tilde{\boldsymbol{D}}^{\mathbf{Y}}\boldsymbol{B}$ (where $\tilde{\boldsymbol{D}}^{\mathbf{C}_{\mathbf{X}}}, \tilde{\boldsymbol{D}}^{\mathbf{C}_{\mathbf{Y}}} \in \mathbb{R}^{N\times K}$). For each random $N \times N$ permutation matrix $\boldsymbol{P}$, we use $\boldsymbol{P}\tilde{\boldsymbol{D}}^{\mathbf{C}_{\mathbf{X}}}_{:,k:}$ and $\tilde{\boldsymbol{D}}^{\mathbf{C}_{\mathbf{Y}}}_{:,k:}$ to calculate the test statistics under permutation $\boldsymbol{P}$ as $\lambda_k^{\boldsymbol{P}}$ following Eq. 7, and the p-value is obtained as:

$$p_k = \mathbb{E}\,\mathbf{1}_{[\lambda_k^{\boldsymbol{P}} \ge \lambda_k]}, \tag{8}$$

where the expectation is taken over random permutations.

3.2 MIXED CASE - IN THE PRESENCE OF DISCRETIZATION

Here we discuss the case where some columns of the data matrices $\tilde{\boldsymbol{D}}^{\mathbf{X}}$ and $\tilde{\boldsymbol{D}}^{\mathbf{Y}}$ are discretized. Under such a scenario, one can still estimate $\hat{\Sigma}_{\mathbf{X}}$, $\hat{\Sigma}_{\mathbf{X},\mathbf{Y}}$, and $\hat{\Sigma}_{\mathbf{Y}}$ by maximizing likelihood, which will be detailed in Section 3.3. After that, $\boldsymbol{A}$ and $\boldsymbol{B}$ can still be estimated following Eq. 5 and Eq. 6, and the exchangeability between $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$ still holds.

However, to get the resampling of $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$ by permutation, one has to apply linear transformation $\boldsymbol{A}$ and $\boldsymbol{B}$ to get $\tilde{\boldsymbol{D}}^{\mathbf{C_X}} = \tilde{\boldsymbol{D}}^{\mathbf{X}}\boldsymbol{A}$ and $\tilde{\boldsymbol{D}}^{\mathbf{C_Y}} = \tilde{\boldsymbol{D}}^{\mathbf{Y}}\boldsymbol{B}$, respectively. In the all continuous case, it is straightforward, but in the presence of discretization, it makes no sense to apply a linear transformation $\boldsymbol{A}$ to $\tilde{\boldsymbol{D}}^{\mathbf{X}}$, when some columns of $\tilde{\boldsymbol{D}}^{\mathbf{X}}$ are just ordinal values. As a consequence, we cannot make use of Lemma 4 to get a resampling of $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$ to calculate the statistic λ_k and estimate the p-value anymore.

Fortunately, it can be shown that to calculate $\lambda_k^{\boldsymbol{P}}$, one does not have to really get the exact resampling from $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$. Instead, for each random permutation $\boldsymbol{P}$, we can get a consistent estimation of $\{\hat{r}_i\}_1^{K-k}$ and consequently calculate $\lambda_k^{\boldsymbol{P}}$. This is formalized by the following Theorem 5.

Theorem 5 (Consistent Estimation of $\{\hat{r}_i\}_1^{K-k}$ under Permutation $\boldsymbol{P}$)**.** *Under permutation $\boldsymbol{P}$, the empirical CCA scores between $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$, i.e., $\{\hat{r}_i\}_1^{K-k}$, are the singular values of $\hat{\Sigma}^{-\frac{1}{2}}_{\mathbf{C}_{\mathbf{X}k:}}\hat{\Sigma}_{\mathbf{C}_{\mathbf{X}k:},\mathbf{C}_{\mathbf{Y}k:}}\hat{\Sigma}^{-\frac{1}{2}}_{\mathbf{C}_{\mathbf{Y}k:}}$, which can be consistently estimated by:*

$$((\boldsymbol{A}^T\hat{\Sigma}_{\mathbf{X}}\boldsymbol{A})_{k:,k:})^{-\frac{1}{2}}\left((\boldsymbol{A}^T\frac{\boldsymbol{D}^{\mathbf{X}^T}\boldsymbol{P}^T\boldsymbol{D}^{\mathbf{Y}}}{N-1}\boldsymbol{B})_{k:,k:}\right)((\boldsymbol{B}^T\hat{\Sigma}_{\mathbf{Y}}\boldsymbol{B})_{k:,k:})^{-\frac{1}{2}}, \tag{9}$$

where $\frac{\boldsymbol{D}^{\mathbf{X}^T}\boldsymbol{P}^T\boldsymbol{D}^{\mathbf{Y}}}{N-1}$ can be consistently estimated by using $\tilde{\boldsymbol{D}}^{\mathbf{X}}$ and $\boldsymbol{P}^T\tilde{\boldsymbol{D}}^{\mathbf{Y}}$ and assuming unit variance of variables.

Remark 1 (Remark on Theorem 5)**.** *Theorem 5 implies that we can consistently estimate $\lambda_k^{\boldsymbol{P}}$ by making use of randomly permuted data $\tilde{\boldsymbol{D}}^{\mathbf{X}}$ and $\boldsymbol{P}^T\tilde{\boldsymbol{D}}^{\mathbf{Y}}$. Note that although here the transpose of permutation applies to $\tilde{\boldsymbol{D}}^{\mathbf{Y}}$, the correctness of the process still relies on the exchangeability between $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$, and does not need the exchangeability between $\mathbf{X}$ and $\mathbf{Y}$. In words, doing permutation on $\tilde{\boldsymbol{D}}^{\mathbf{X}}\boldsymbol{A}$ will meet the problem of applying linear transformation to data that might contain ordinal values, and Theorem 5 provides a way to bypass the problem by permuting $\tilde{\boldsymbol{D}}^{\mathbf{Y}}$ instead.*

Till now, the remaining problem is how to consistently estimate cross-covariance matrices in the presence of discretization, and it will be detailed in what follows.

3.3 CORRELATION ESTIMATION IN THE PRESENCE OF DISCRETIZATION

Assume that we concern the rank of $\Sigma_{\mathbf{X},\mathbf{Y}}$, where some of the variables are discretized and $\mathbf{X}$ and $\mathbf{Y}$ are not necessarily disjoint. As mentioned, for those variables that we only have discretized observations, their variance can never be determined. Further, the rank of a cross-covariance matrix is equivalent to the rank of the corresponding cross-correlation matrix. Without loss of generality, we can assume all variables to have unit variance and zero mean. Thus, we sometimes use correlation and covariance interchangeably. The remaining crucial step is to estimate the correlation matrix for $\mathbf{V} = \mathbf{X} \cup \mathbf{Y}$, i.e., $\hat{\boldsymbol{R}}$, by data $\tilde{\boldsymbol{D}} \in \mathbb{R}^{N\times|\boldsymbol{V}|}$. As some elements of $\mathbf{V}$ are discrete, we use $\mathbb{C}_{\mathbf{V}}$ and $\mathbb{D}_{\mathbf{V}}$ to denote the index set of continuous variables and discrete variables in $\mathbf{V}$ respectively.

We first introduce the overall objective function for correlation estimation as follows.

$$\hat{\boldsymbol{R}} = \arg\min_{\boldsymbol{R}\in\mathbb{R}^{M\times M}} \mathcal{L}(\tilde{\boldsymbol{D}}, \boldsymbol{R}), \tag{10}$$

$$\mathcal{L}(\tilde{\boldsymbol{D}}, \boldsymbol{R}) = -\sum_{1\le i<j\le M} \log p_{ij}(\tilde{\boldsymbol{D}}_{:,ij}; \boldsymbol{R}_{i,j}), \tag{11}$$

where the optimization objective is minimizing pair-wise negative log-likelihood, also referred to as pseudo likelihood, instead of the real joint log-likelihood over all variables. The reason lies in that optimizing over the joint log-likelihood is very computationally expensive and the pseudo likelihood is tractable while also serves as a consistent estimator (Besag, 1974).

Next, we will specify the pair-wise log-likelihood in three different scenarios - between two continuous variables, between a continuous and a discrete variable, and between two discrete variables.

(i) Likelihood for Two Continuous Variables

If both $i \in \mathbb{C}_{\mathbf{V}}$ and $j \in \mathbb{C}_{\mathbf{V}}$, the likelihood function is just the joint gaussian pdf parametrized by $\boldsymbol{R}_{i,j}$ given as follows:

$$\log p_{ij}(\tilde{\boldsymbol{D}}_{:,ij}; \boldsymbol{R}_{i,j}) = (1/2)\left(\operatorname{tr}\left(\begin{bmatrix}1, \boldsymbol{R}_{i,j}\\ \boldsymbol{R}_{i,j}, 1\end{bmatrix}^{-1}\begin{bmatrix}1, \hat{\boldsymbol{R}}_{i,j}\\ \hat{\boldsymbol{R}}_{i,j}, 1\end{bmatrix}\right) + \log\det\begin{bmatrix}1, \boldsymbol{R}_{i,j}\\ \boldsymbol{R}_{i,j}, 1\end{bmatrix}\right), \tag{12}$$

where $\hat{\boldsymbol{R}}_{i,j}$ is the empirical correlation matrix that can be directly calculated from data $\tilde{\boldsymbol{D}}_{:,ij}$.

(ii) Likelihood for a Continuous and a Discrete Variable

If $i \in \mathbb{C}_{\mathbf{V}}$ and $j \in \mathbb{D}_{\mathbf{V}}$, then the log-likelihood (also known as polyserial correlation estimation (Olsson et al., 1982)) can be factorized as follows.

$$\log p_{ij}(\tilde{\boldsymbol{D}}_{:,ij}; \boldsymbol{R}_{i,j}) = \frac{1}{N}\sum_{k=1}^{N} \log p(\mathsf{V_i} = \tilde{\boldsymbol{D}}_{k,i}) p(\mathsf{V_j} = \tilde{\boldsymbol{D}}_{k,j} | \mathsf{V_i} = \tilde{\boldsymbol{D}}_{k,i}, \boldsymbol{R}_{i,j}), \tag{13}$$

where $p(\mathsf{V_i} = \tilde{\boldsymbol{D}}_{k,i})$ is a standard gaussian pdf. For a specific value of $\tilde{\boldsymbol{D}}_{k,j}$, say, t, we have that:

$$p(\mathsf{V_j} = \tilde{\boldsymbol{D}}_{k,j} | \mathsf{V_i} = \tilde{\boldsymbol{D}}_{k,i}, \boldsymbol{R}_{i,j}) = p(T^j_t < \mathsf{V_j} \leq T^j_{t+1} | \mathsf{V_i} = \tilde{\boldsymbol{D}}_{k,i}, \boldsymbol{R}_{i,j}), \tag{14}$$

$$= \Phi\left(\frac{T^j_{t+1} - \boldsymbol{R}_{i,j}\tilde{\boldsymbol{D}}_{k,i}}{(1 - \boldsymbol{R}^2_{i,j})^{1/2}}\right) - \Phi\left(\frac{T^j_{t} - \boldsymbol{R}_{i,j}\tilde{\boldsymbol{D}}_{k,i}}{(1 - \boldsymbol{R}^2_{i,j})^{1/2}}\right), \tag{15}$$

where Φ is the standard gaussian cdf. We note that the thresholds T are unknown, thus it could be taken as free parameters during optimization. In practice, it is more efficient to estimate the thresholds first by using inverse gaussian cdf as follows:

$$\hat{T}^j_{t+1} = \Phi^{-1}\left(\frac{\sum_{k=1}^{N} \mathbf{1}_{[\tilde{\boldsymbol{D}}_{k,j} \leq t]}}{N}\right). \tag{16}$$

(iii) Likelihood for Two Discrete Variables

If both $i \in \mathbb{D}_{\mathbf{V}}$ and $j \in \mathbb{D}_{\mathbf{V}}$, then the log-likelihood is given as follows (also known as polychoric correlation estimation (Olsson, 1979; Jöreskog, 1994)).

$$\log p_{ij}(\tilde{\boldsymbol{D}}_{:,ij}; \boldsymbol{R}_{i,j}) = \frac{1}{N}\sum_{k=1}^{N} \log p(\mathsf{V_i} = \tilde{\boldsymbol{D}}_{k,i}, \mathsf{V_j} = \tilde{\boldsymbol{D}}_{k,j} | \boldsymbol{R}_{i,j}) \tag{17}$$

$$= \frac{1}{N}\sum_{k=1}^{N} \log(\Phi_2(T^i_{\tilde{\boldsymbol{D}}_{k,i}+1}, T^j_{\tilde{\boldsymbol{D}}_{k,j}+1}; \boldsymbol{R}_{i,j}) + \Phi_2(T^i_{\tilde{\boldsymbol{D}}_{k,i}}, T^j_{\tilde{\boldsymbol{D}}_{k,j}}; \boldsymbol{R}_{i,j}) \tag{18}$$

$$- \Phi_2(T^i_{\tilde{\boldsymbol{D}}_{k,i}+1}, T^j_{\tilde{\boldsymbol{D}}_{k,j}}; \boldsymbol{R}_{i,j}) - \Phi_2(T^i_{\tilde{\boldsymbol{D}}_{k,i}}, T^j_{\tilde{\boldsymbol{D}}_{k,j}+1}; \boldsymbol{R}_{i,j})), \tag{19}$$

where $\Phi_2(.,.,r)$ is the joint cdf of two standard gaussian variables with correlation r and the thresholds for each variable can also be estimated by using Eq. 16.

3.4 Parameterization Trick for Rank Test

We note that the optimization problem defined in Eq. 10 does not constrain the space to be a pseudo-correlation matrix - a matrix that is PSD with unit diagonal elements. If we only care about the maximum likelihood estimator, the pseudo-correlation requirement might be unnecessary. However, as we rely on SVD for CCA and rank test, the requirement of being pseudo-correlation matrix is crucial. A classical way to solve this problem is by projected gradient descent: we project the current solution to the space of pseudo-correlation matrices after each step of gradient descent. Yet, in practice we found this solution less effective, due to that the projection itself cannot be analytically solved and thus an additional optimization step to solve projection is required.

Algorithm 1: MPRT: Mixed data Permutation-based Rank Test

Input : Sample $\tilde{\boldsymbol{D}}^{\mathbf{X}}$, $\tilde{\boldsymbol{D}}^{\mathbf{Y}}$, indexes of discretized columns, null hypothesis $\mathcal{H}_0^k$: rank$(\Sigma_{\mathbf{X},\mathbf{Y}}) \leq k$, and significant level α;

Output : True (fail to reject $\mathcal{H}_0^k$) or False (reject $\mathcal{H}_0^k$);

$P = |\mathbf{X}|$, $Q = |\mathbf{Y}|$, and $K = \min(P, Q)$;

Get $\hat{\Sigma}_{\mathbf{X}}$, $\hat{\Sigma}_{\mathbf{X},\mathbf{Y}}$, and $\hat{\Sigma}_{\mathbf{Y}}$ as submatrices of $\hat{\boldsymbol{R}}$ by Eq. 22 (unit variance assumed);

Calculate $\boldsymbol{A}$ and $\boldsymbol{B}$ following Eqs. 5 and 6.;

Let $\boldsymbol{P} = \boldsymbol{I}$ (no permutation), calculate $\{\hat{r}_i\}_1^{K-k}$ following Eq. 9 and then the statistic λ_k following Eq. 7;

for *each random permutation* $\boldsymbol{P}$ **do**

    Calculate $\{\hat{r}_i\}_1^{K-k}$ under $\boldsymbol{P}$ following Eq. 9 and then the statistic under $\boldsymbol{P}$, i.e., $\lambda_k^{\boldsymbol{P}}$, following Eq. 4;

Calculate p-value p_k by Eq. 8;

return $p_k \geq \alpha$

To this end, we directly parameterize the space of pseudo-correlation matrices in a geometric way following (Rousseeuw & Molenberghs, 1993), given as follows.

$$\boldsymbol{R} = \boldsymbol{U}^T\boldsymbol{U}, \tag{20}$$

$$\boldsymbol{U}_{j,i} = \begin{cases} \cos\boldsymbol{\theta}_{i-j+1,i}\Pi_{k=1}^{i-j}\sin\boldsymbol{\theta}_{k,i}, & j \leq i \\ 0, & j > i \end{cases}, \text{ s.t., } \boldsymbol{\theta}_{i,i} = 0, \ \forall i. \tag{21}$$

Therefore, we have an alternative way to parameterize the correlation matrix, which gives rise to the following new formulation of our objective function (instead of Eq. 10):

$$\hat{\boldsymbol{R}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{D}, \boldsymbol{R}). \tag{22}$$

We summarize the overall testing procedure of our proposed MPRT in Algorithm 1.

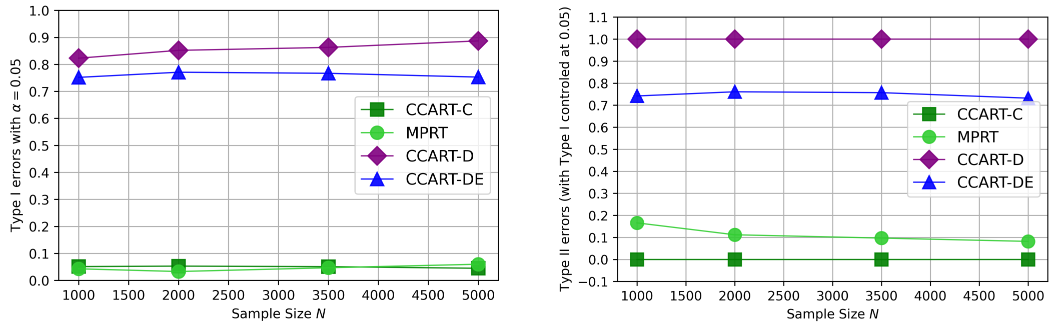
4 Experiments

4.1 Experimental Setting

To empirically validate the proposed Mixed data Permutation-based Rank Test (MPRT), we apply our method to synthetic data and compare it with the following methods. (i) CCART-C: CCA-based Rank Test (Anderson, 1984) that use the original continuous observation as input; as it has access to the original observations, its performance is taken as the best possible performance that we can achieve. (ii) CCART-D: CCA-based Rank Test with Discrete data; it directly takes the ordinal values as input. (iii) CCART-DE: CCA-based Rank Test with Discrete data Estimating covariance; it takes the estimated correlation matrix as input (following Eq. 22).

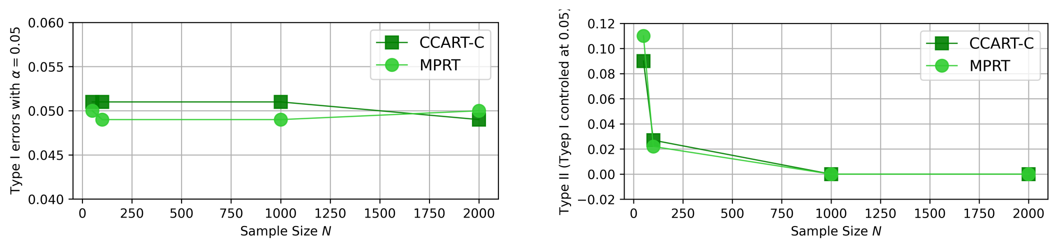
We consider two scenarios: mixed data scenario where data are partially discretized, and all continuous scenario where all the original observations are available. The first scenario is to illustrate how well can we handle discretization while the second is to show that our method can serve as a general rank test method as we also work well when there is no discretization. In terms of performance, we concern both Type I errors and Type II errors. Specifically, we expect a good test can properly control the Type I errors given a significance level α, while the Type II errors should be as small as possible. We consider different sample sizes, and for each comparison, we consider 3000 random trials. For MPRT, we randomly generated 200 permutations to calculate the p-value. The ground truth covariance matrices are randomly generated. For the mixed scenario, we uniformly generate two thresholds from $[-1.5, 1.5]$ for each variable that should be discretized, and use the thresholds together with $-\infty$ and ∞ to discretize the continuous observations into three categories $\{1, 2, 3\}$.

We also apply the proposed MPRT method with mixed data to the classical causal discovery method PC algorithm (Spirtes et al., 2000) and see whether our test method can better test CI relations compared to the classical Fisher-Z CI test (Fisher et al., 1921), in the presence of discretization. Fisher-Z is only compared by the result of PC and cannot be not compared in the previous setting, as linear CI relations can only correspond to a part of the rank information. Finally, we employ a real-life dataset to illustrate the applicability of the proposed method in real-life scenarios.



(a) The probability of Type I errors with $\alpha = 0.05$. (b) Type II errors (effective Type I controlled at 0.05).

Figure 2: The probability of Type I and Type II errors with **mixed data**, by different rank test methods, under different sample sizes $N = 1000, 2000, 5000$.



(a) The probability of Type I errors with $\alpha = 0.05$. (b) Type II errors (effective Type I controlled at 0.05).

Figure 3: The probability of Type I and Type II errors with **continuous data**, by different rank test methods, under different sample sizes $N = 50, 100, 1000, 2000$.

4.2 Analysis on Type I and Type II Errors under Different Sample Sizes

In this section we analyze the performance of each method in terms of Type I and Type II errors under different sample sizes. For the mixed data scenario, the result is shown in Figure 2. Specifically, one can see that both our proposed MPRT and CCART-C can properly control the Type I errors as the Type I errors of them are both very close to the significance level $\alpha = 0.05$; in contrast, CCART-D and CCART-DE totally failed to control the Type I errors. As for Type II errors, it can be found that the Type II errors of MPRT are quite small, and decreases with the increase of sample size N, while CCART-D and CCART-DE cannot benefit from the increase in sample size. We note that it is very natural that MPRT cannot beat CCART-C as CCART-C takes the original continuous observation as input while MPRT takes mixed data as input. We show the performance of CCART-C just in order to show the minimal possible Type II errors that one can achieve in the presence of discretization.

We also show the performance when both CCART-C and MPRT have access to the original continuous observations, as in Figure 3. Specifically, both methods properly control the Type I errors as in the subfigure 3(a). For the Type II errors, the performance of CCART-C and MPRT is almost the same. This is as expected, as in this scenario both methods use exactly the same test statistics except that CCART-C uses the analytically derived null distribution to get the p-value while MPRT uses the empirical CDF to calculate the p-value; the two results are expected to be exactly the same asymptotically.

Taking the performance under these two scenarios together into consideration it can be argued that MPRT is a very general and valid rank test as it can handle all continous data, partially discretized data, and all discretized data and the Type I are properly controlled while the power is also good.

4.3 Application in Causal Discovery

In this section we validate our test using the PC algorithm (Spirtes et al., 2000). Specifically, we consider linear causal models with gaussian noises $\mathsf{V}_i = \sum_{\mathsf{V}_j \in \mathrm{Pa}(\mathsf{V}_i)} a_{ij} \mathsf{V}_j + \varepsilon_{\mathsf{V}_i}$, where the edge

(a) Discovered personality substructure for Openness. (b) Discovered substructure for Neuroticism.

Figure 4: Application of MPRT in causal discovery using real-life Big Five human personality data.

Table 1: F1 score and SHD of the PC algorithm, with different CI test methods (↑ means the bigger the better while ↓ the smaller the better).

	F1 score for skeleton ↑			SHD for skeleton ↓		
CI test method	$N = 500$	$N = 1000$	$N = 2000$	$N = 500$	$N = 1000$	$N = 2000$
MPRT	**0.84**	**0.9**	**0.96**	**0.80**	**0.60**	**0.20**
Fisher-Z	0.81	0.80	0.78	1.20	1.20	1.40
CCART-D	0.75	0.79	0.77	1.60	1.60	1.80
CCART-DE	0.80	0.85	0.83	1.40	1.30	1.60

coefficients and the variance of the noises are randomly generated. We consider the scenario where data are partially discretized and compare MPRT with Fisher-Z to see which one works better with PC. We employ F1 score $\text{F1} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall}+\text{Precision}}$ for skeleton (the bigger the better) and Structural Hamming Distance (SHD) for skeleton (the smaller the better) to evaluate the performance. As shown in Table 1, MPRT achieves the best performance in terms of both F1 and SHD, under all sample sizes. This validates the claim that MPRT can serve as a powerful CI test for causal discovery in the presence of discretization.

4.4 Real-world Causal Discovery Application

In this section, we further validate our proposed MPRT method using a real-world Big Five Personality dataset `https://openpsychometrics.org/`. It consists of 50 personality indicators and close to 20,000 data points. Each Big Five personality dimension, namely, Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (O-C-E-A-N), are designed to be measured with their own 10 indicators and the values of each variable are ordinal: Disagree, slightly disagree, Neutral, Slightly agree, and Agree. We employ RLCD (Dong et al., 2024), a recently proposed rank based causal discovery method with our MPRT method. We choose 7 items from openness and 6 items from neuroticism to verify our method.

The results are shown in Figure 4. Specifically, for openness we discovered two latent variables. L2 corresponds to whether a person has a lot of ideas while L1 corresponds to the general concept of openness. As for neuroticism, we also discovered two latent variables. L1 relates more to one's emotions while L2 relates to one's stress level. In contrast, if we directly use the ordinal values to do the rank test, i.e., using CCART-D, all the p-values tend to be very small, and thus we have to use very small significance level (around 1e-10) in order to have some structures discovered; yet using such an extremely small alpha value will induce a lot of Type II errors. This result illustrates the superiority of using MPRT in the presence of discretizations in real-life scenarios, and again empirically validate the proposed method.

5 Conclusion

In this paper, we propose a novel permutation-based rank test that works in the presence of discretization. It is rather general as it can accommodate fully continuous data, partially discretized data, or fully discretized data as input, and it can effectively control the Type I errors while the Type II is also small. Extensive empirical studies validate our method.

REFERENCES

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.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

Hu Changsheng and Wang Yongfeng. Investor sentiment and assets valuation. *Systems Engineering Procedia*, 3:166–171, 2012.

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.

Herbert A David. The beginnings of randomization tests. *The American Statistician*, 62(1):70–72, 2008.

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.

Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pp. 132–141, 2014.

David John Finney. Probit analysis: a statistical treatment of the sigmoid response curve. 1952.

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

Ronald Aylmer Fisher et al. 014: On the" probable error" of a coefficient of correlation deduced from a small sample. 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.

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.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*, pp. 162–190. Springer, 1992.

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.

Sverre Urnes Johnson, Pål Gunnar Ulvenes, Tuva Øktedalen, and Asle Hoffart. Psychometric properties of the general anxiety disorder 7-item (gad-7) scale in a heterogeneous psychiatric sample. *Frontiers in psychology*, 10:1713, 2019.

Camille Jordan. Essai sur la géométrie à n dimensions. *Bulletin de la Société mathématique de France*, 3:103–174, 1875.

Karl G Jöreskog. On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, 59(3):381–389, 1994.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Frederic M Lord and Melvin R Novick. *Statistical theories of mental test scores*. IAP, 2008.

Marc Nerlove and S James Press. *Univariate and multivariate log-linear and logistic models*, volume 1306. Rand Corporation, 1973.

Ulf Olsson. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460, 1979.

Ulf Olsson, Fritz Drasgow, and Neil J Dorans. The polyserial correlation coefficient. *Psychometrika*, 47:337–347, 1982.

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

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000.

Fortunato Pesarin and Luigi Salmaso. The permutation testing approach: a review. *Statistica*, 70(4): 481–509, 2010.

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.

Peter J Rousseeuw and Geert Molenberghs. Transformation of non positive semidefinite correlation matrices. *Communications in Statistics–Theory and Methods*, 22(4):965–984, 1993.

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, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

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

William J Welch. Construction of permutation tests. *Journal of the American Statistical Association*, 85(411):693–698, 1990.

Anderson M Winkler, Olivier Renaud, Stephen M Smith, and Thomas E Nichols. Permutation inference for canonical correlation analysis. *Neuroimage*, 220:117065, 2020.

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.

A PROOFS

A.1 PROOF OF THEOREM 4

Theorem 4 (Exchangeability of $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$)**.** *Given a set of variables $\mathbf{V}$ that are jointly gaussian, under null hypothesis $\mathcal{H}_0^k$: rank$(\Sigma_{\mathbf{X},\mathbf{Y}}) \leq k$, where $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$, random vectors $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$ are asymptotically independent with each other.*

Proof of Theorem 4. Asymptotically $\hat{\Sigma}_{\mathbf{X}}$, $\hat{\Sigma}_{\mathbf{Y}}$, and $\hat{\Sigma}_{\mathbf{X},\mathbf{Y}}$ are the same as $\Sigma_{\mathbf{X}}$, $\Sigma_{\mathbf{Y}}$, and $\Sigma_{\mathbf{X},\mathbf{Y}}$, respectively. Under the null hypo that $\mathcal{H}_0^k$: rank$(\Sigma_{\mathbf{X},\mathbf{Y}}) \leq k$, we have that the population covariance between $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$ are all zeros. Given that all variables are jointly gaussian, $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$ are also jointly gaussian. Thus $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$ are asymptotically independent. $\square$

A.2 PROOF OF LEMMA 1

Lemma 1 (Alternative Way to Calculate Statistic in Eq. 4)**.** *Let the CCA score between $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$ be $\{\hat{r}_i\}_1^{K-k}$. Then the statistic defined in Eq. 4 can also be formulated as:*

$$\lambda_k = -\left(N - \frac{P+Q+3}{2}\right)\ln(\Pi_{i=1}^{K-k}(1-\hat{r}_i^2)). \tag{7}$$

Proof of Lemma 1. The CCA scores between $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$ are just the diagonal entries of their cross-covariance matrix, which corresponds to the k to K CCA scores between $\mathbf{X}$ and $\mathbf{Y}$. Thus we have $\hat{r}_i = r_{i+k}$ for $i = \{1, ..., K-k\}$, and thus $\lambda_k = -(N - \frac{P+Q+3}{2})\ln(\Pi_{i=k+1}^{K}(1-r_i^2))$. $\square$

A.3 PROOF OF THEOREM 5

Theorem 5 (Consistent Estimation of $\{\hat{r}_i\}_1^{K-k}$ under Permutation $\boldsymbol{P}$)**.** *Under permutation $\boldsymbol{P}$, the empirical CCA scores between $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$, i.e., $\{\hat{r}_i\}_1^{K-k}$, are the singular values of $\hat{\Sigma}_{\mathbf{C}_{\mathbf{X}k:}}^{-\frac{1}{2}}\hat{\Sigma}_{\mathbf{C}_{\mathbf{X}k:},\mathbf{C}_{\mathbf{Y}k:}}\hat{\Sigma}_{\mathbf{C}_{\mathbf{Y}k:}}^{-\frac{1}{2}}$, which can be consistently estimated by:*

$$((\boldsymbol{A}^T\hat{\Sigma}_{\mathbf{X}}\boldsymbol{A})_{k:,k:})^{-\frac{1}{2}}\left((\boldsymbol{A}^T\frac{\boldsymbol{D}^{\mathbf{X}^T}\boldsymbol{P}^T\boldsymbol{D}^{\mathbf{Y}}}{N-1}\boldsymbol{B})_{k:,k:}\right)((\boldsymbol{B}^T\hat{\Sigma}_{\mathbf{Y}}\boldsymbol{B})_{k:,k:})^{-\frac{1}{2}}, \tag{9}$$

where $\frac{\boldsymbol{D}^{\mathbf{X}^T}\boldsymbol{P}^T\boldsymbol{D}^{\mathbf{Y}}}{N-1}$ can be consistently estimated by using $\tilde{\boldsymbol{D}}^{\mathbf{X}}$ and $\boldsymbol{P}^T\tilde{\boldsymbol{D}}^{\mathbf{Y}}$ and assuming unit variance of variables.

Proof of Theorem 5. We are interested in $\hat{\Sigma}_{\mathbf{C}_{\mathbf{X}k:}}^{-\frac{1}{2}}\hat{\Sigma}_{\mathbf{C}_{\mathbf{X}k:},\mathbf{C}_{\mathbf{Y}k:}}\hat{\Sigma}_{\mathbf{C}_{\mathbf{Y}k:}}^{-\frac{1}{2}}$. Assume that we have access to the original data $\boldsymbol{D}^{\mathbf{X}}$ and $\boldsymbol{D}^{\mathbf{Y}}$. By the exchangeability, for each random $\boldsymbol{P}$, we have $(\boldsymbol{P}\boldsymbol{D}^{\mathbf{X}}\boldsymbol{A})_{:,k:}$ and $(\boldsymbol{D}^{\mathbf{Y}}\boldsymbol{B})_{:,k:}$ are the N samples from joint distribution of $\mathbf{C}_{\mathbf{X}k:}$ and $\mathbf{C}_{\mathbf{Y}k:}$. Then the $\hat{\Sigma}_{\mathbf{C}_{\mathbf{X}k:}}^{-\frac{1}{2}}$, $\hat{\Sigma}_{\mathbf{C}_{\mathbf{X}k:},\mathbf{C}_{\mathbf{Y}k:}}$, and $\hat{\Sigma}_{\mathbf{C}_{\mathbf{Y}k:}}^{-\frac{1}{2}}$ are as follows:

$$\begin{aligned}
\hat{\Sigma}_{\mathbf{C}_{\mathbf{X}k:}}^{-\frac{1}{2}} &= (\frac{((\boldsymbol{P}\boldsymbol{D}^{\mathbf{X}}\boldsymbol{A})_{:,k:})^T(\boldsymbol{P}\boldsymbol{D}^{\mathbf{X}}\boldsymbol{A})_{:,k:}}{N-1})^{-\frac{1}{2}}, && (23)\\
&= (\frac{((\boldsymbol{P}\boldsymbol{D}^{\mathbf{X}}\boldsymbol{A})^T(\boldsymbol{P}\boldsymbol{D}^{\mathbf{X}}\boldsymbol{A}))_{k:,k:}}{N-1})^{-\frac{1}{2}}, && (24)\\
&= (\frac{(\boldsymbol{A}^T\boldsymbol{D}^{\mathbf{X}^T}\boldsymbol{D}^{\mathbf{X}}\boldsymbol{A})_{k:,k:}}{N-1})^{-\frac{1}{2}}, && (25)\\
&= ((\boldsymbol{A}^T\hat{\Sigma}_{\mathbf{X}}\boldsymbol{A})_{k:,k:})^{-\frac{1}{2}}. && (26)
\end{aligned}$$

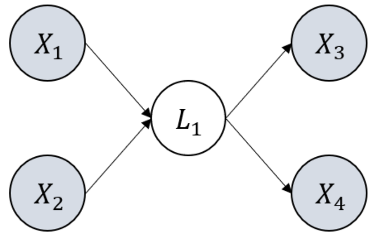


Figure 5: An illustrative example to show that rank contains more graphical information than CI. When using CI, we cannot deduce that $\{\mathbf{X}_1, \mathbf{X}_2\}$ and $\{\mathbf{X}_3, \mathbf{X}_4\}$ are d-separated by $\mathbf{L_1}$ as $\mathbf{L_1}$ is latent, while by using rank we can.

$$\hat{\Sigma}_{\mathbf{C_Y}_{k:}}^{-\frac{1}{2}} = (\frac{((\boldsymbol{D}^{\mathbf{Y}}\boldsymbol{B})_{:,k:})^T(\boldsymbol{D}^{\mathbf{Y}}\boldsymbol{B})_{:,k:}}{N-1})^{-\frac{1}{2}}, \tag{27}$$

$$= (\frac{((\boldsymbol{D}^{\mathbf{Y}}\boldsymbol{B})^T(\boldsymbol{D}^{\mathbf{Y}}\boldsymbol{B}))_{k:,k:}}{N-1})^{-\frac{1}{2}}, \tag{28}$$

$$= (\frac{(\boldsymbol{B}^T\boldsymbol{D}^{\mathbf{Y}^T}\boldsymbol{D}^{\mathbf{Y}}\boldsymbol{B})_{k:,k:}}{N-1})^{-\frac{1}{2}}, \tag{29}$$

$$= ((\boldsymbol{B}^T\hat{\Sigma}_{\mathbf{Y}}\boldsymbol{B})_{k:,k:})^{-\frac{1}{2}}. \tag{30}$$

$$\hat{\Sigma}_{\mathbf{C_X}_{k:},\mathbf{C_Y}_{k:}} = \frac{((\boldsymbol{P}\boldsymbol{D}^{\mathbf{X}}\boldsymbol{A})_{:,k:})^T(\boldsymbol{D}^{\mathbf{Y}}\boldsymbol{B})_{:,k:}}{N-1}, \tag{31}$$

$$= \frac{((\boldsymbol{P}\boldsymbol{D}^{\mathbf{X}}\boldsymbol{A})^T\boldsymbol{D}^{\mathbf{Y}}\boldsymbol{B})_{k:,k:}}{N-1}, \tag{32}$$

$$= (\frac{(\boldsymbol{A}^T\boldsymbol{D}^{\mathbf{X}^T}\boldsymbol{P}^T\boldsymbol{D}^{\mathbf{Y}}\boldsymbol{B})_{k:,k:}}{N-1}), \tag{33}$$

$$= (\boldsymbol{A}^T\frac{\boldsymbol{D}^{\mathbf{X}^T}\boldsymbol{P}^T\boldsymbol{D}^{\mathbf{Y}}}{N-1}\boldsymbol{B})_{k:,k:}. \tag{34}$$

Further, $\tilde{\boldsymbol{D}}^{\mathbf{X}}$ and $\boldsymbol{P}^T\tilde{\boldsymbol{D}}^{\mathbf{Y}}$ can be taken as sampled from the joint distribution of two independent gaussian random vectors. As each of them are marginally gaussian, they are also jointly gaussian. Thus, $\frac{\boldsymbol{D}^{\mathbf{X}^T}\boldsymbol{P}^T\boldsymbol{D}^{\mathbf{Y}}}{N-1}$ can be consistently estimated by maximizing likeilhood as in Eq. 22. $\square$

B Other Definitions

B.1 T-separation

The definitions of trek and t-separation are as follows.

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.*

Definition 2 (T-separation (Sullivant et al., 2010))**.** *Let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C_A}$, and $\mathbf{C_B}$ be four subsets of $\mathbf{V}_{\mathcal{G}}$ in graph $\mathcal{G}$ (not necessarilly disjoint). ($\mathbf{C_A}$,$\mathbf{C_B}$) t-separates $\mathbf{A}$ from $\mathbf{B}$ if for every trek (P_1,P_2) from a vertex in $\mathbf{A}$ to a vertex in $\mathbf{B}$, either P_1 contains a vertex in $\mathbf{C_A}$ or P_2 contains a vertex in $\mathbf{C_B}$.*

Example 1. *In Figure 5, there are multiple treks. For example, $\mathsf{X_4} \leftarrow \mathsf{L_1} \rightarrow \mathsf{X_3}$ is a trek between $\mathsf{X_4}$ and $\mathsf{X_3}$, $\mathsf{X_4} \leftarrow \mathsf{L_1}$ is a trek between $\mathsf{X_4}$ and $\mathsf{L_1}$, and $\mathsf{L_1} \rightarrow \mathsf{X_3}$ is a trek between $\mathsf{L_1}$ and $\mathsf{X_3}$. As for t-separations, we have $\{\mathsf{X_1}, \mathsf{X_2}\}$ and $\{\mathsf{X_3}, \mathsf{X_4}\}$ are t-separated by $(\emptyset, \{\mathsf{L_1}\})$.*

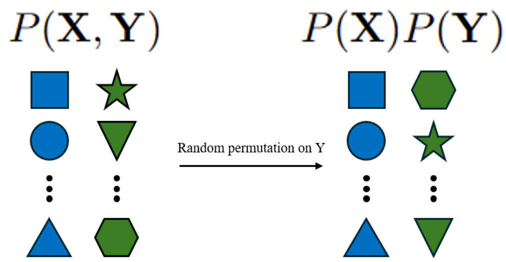


Figure 6: An illustration of exchangeability and permutation test. The left figure refer to N i.i.d. samples from $P(\mathbf{X}, \mathbf{Y})$. After random permutation on $\mathbf{Y}$, the permutated data can be considered as random i.i.d. samples from $P(\mathbf{X})$ and $P(\mathbf{Y})$. If the exchangeability holds, i.e., random vectors $\mathbf{X}$ and $\mathbf{Y}$ are independent, then we have $P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{X})P(\mathbf{Y})$, and thus the permuted data can serve as another N i.i.d. samples from $P(\mathbf{X}, \mathbf{Y})$.

C DISCUSSION

C.1 BRIEF INTRODUCTION TO PERMUTATION TEST

Permutation tests aim to empirically estimate the CDF of the null distribution of a test statistic. The core of such an CDF estimation is the exchangeability, under which we can make use of permuted data to serve as additional samples from the same distribution.

Take Figure 6 as an example. The left figure in Figure 6 refer to N i.i.d. samples from $P(\mathbf{X}, \mathbf{Y})$. After random permutation on $\mathbf{Y}$, the permutated data can be considered as random i.i.d. samples from $P(\mathbf{X})$ and $P(\mathbf{Y})$. If the exchangeability holds under the null hypothesis, i.e., random vectors $\mathbf{X}$ and $\mathbf{Y}$ are independent, then we have $P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{X})P(\mathbf{Y})$, and thus the permuted data can serve as another N i.i.d. samples from $P(\mathbf{X}, \mathbf{Y})$. Now we know how to generate additional N i.i.d. samples. As a test statistic is just a deterministic function of the N i.i.d., samples. For each randomly permuted data, we can calculate the value of the test statistic, and thus all these calculated test statistics can be considered as sampled from the distribution of the test statistic. Given these samples, we can construct the empirical CDF of the null distribution, and consequently correctly calculate the p-value.

C.2 FOR THE LINEAR NON-GAUSSIAN CASE

If we assume that the underlying continuous variables follow a linear SCM, but the joint distribution are not necessarily gaussian anymore, the proposed method can still work, as long as the parametric form is given. To be specific, we only need to modify the likelihood function in Section 3.3 according to the corresponding parametric form for correlation estimation and the proposed method can still work. As a comparison, the traditional CCA rank test must assume normality to infer the chi-square null distribution. On the other hand, if the parametric form is not given, which means we do not have any information about the shape of the distribution, it may not be possible to consistently recover the underlying correlation (due to insufficient information), and thus the problem cannot be solved.

C.3 NUMBER OF CATEGORIES AND ANALYSIS OF TYPE-I ERROR AND POWER

The proposed method can handle any level of discretization, as long as it is greater than 1, with Type-I errors properly controlled. At the same time, more levels are always beneficial, because it leads to less information loss during the discretization process, and thus the correlation matrix can be more efficiently estimated for building the test.

Regarding Type-I errors, as we establish the exchangeability even in the discretized scenario, the asymptotic null distribution can be estimated by random permutations. Consequently, Type-I errors can be properly controlled at any significance level. At the same time, we do not have theoretical result on the analysis of the power yet. To be specific, even without considering discretization, the analysis of power involves tools from advanced random matrix theories and is highly nontrivial. Furthermore, in our setting with discretized variables, the involved maximum likelihood step makes such an analysis even more challenging. To our best knowledge, there is not any existing result available for the analytic form of the power in our setting, and we plan to leave it for future exploration.

D RELATED WORK

Conditional independence and rank test. A line of conditional independence tests imposes simplifying assumptions on the distributions. For instance, when the variables have linear relations with additive Gaussian noise, the Fisher's classical z-test based on partial correlations can be used (Fisher, 1924; Baba et al., 2004). Ramsey (2014) developed an approach that separately regresses X and Y on Z, and further perform independence test on the corresponding residuals. Fukumizu et al. (2007) proposed a conditional independence test method based on Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2007). Zhang et al. (2012) further provided a kernel-based conditional test that yields pointwise asymptotic level control. Shah & Peters (2018) investigated the hardness of conditional independence test, and developed a method based on kernel-ridge regression and generalised covariance measure. On the other hand, existing statistical tests for rank of a cross-covariance matrix (Anderson, 1984) often rely on CCA (Jordan, 1875; Hotelling, 1992), with a likelihood ratio based test statistics.

Permutation test. Research and applications related to permutation tests have addressed increased attention in recent years (David, 2008; Pesarin & Salmaso, 2010; Welch, 1990). These tests lead to valid inferences while requiring weak assumptions that are commonly satisfied, base on the exchangeability of observations under the null hypothesis. Recently, a permutation-based CI test was proposed (Doran et al., 2014) and more recently a permutation-based rank test (Winkler et al., 2020). However, they cannot deal with the discretization problem. In contrast, our MPRT can take all continuous, partially discretized, or all discretized data as input, and our Type I errors can be properly controlled.

Constraint-based causal discovery. Constraint-based methods leverage statistical tests, such as conditional independence tests, to estimate the causal structure. Spirtes & Glymour (1991) proposed the PC algorithm that estimates the skeleton and orient certain edges to identify the Markov equivalence class. FCI (Spirtes et al., 1995; Colombo et al., 2012) was developed to allow for latent and selection variables, while the CCD algorithm (Richardson, 1996) can accommodate cycles. Furthermore, Huang et al. (2020) developed a constraint-based method that allows for heterogeneity or non-stationarity in the data distribution, while Silva et al. (2006); Huang et al. (2022); Dong et al. (2024) proposed algorithms based on rank test that recover the causal structure involving latent confounders.