A Conditional Independence Test in the Presence of Discretization

Anonymous authors
Paper under double-blind review

Abstract

Testing conditional independence (CI) has many important applications, such as Bayesian network learning and causal discovery. Although several approaches have been developed for learning CI structures for observed variables, those existing methods generally fail to work when the variables of interest can not be directly observed and only discretized values of those variables are available. For example, if X_1, $\tilde{X}_2$ and X_3 are the observed variables, where $\tilde{X}_2$ is a discretization of the latent variable X_2, applying the existing methods to the observations of X_1, $\tilde{X}_2$ and X_3 would lead to a false conclusion about the underlying CI of variables X_1, X_2 and X_3. Motivated by this, we propose a CI test specifically designed to accommodate the presence of discretization. To achieve this, a bridge equation and nodewise regression are used to recover the precision coefficients reflecting the conditional dependence of the latent continuous variables under the nonparanormal model. An appropriate test statistic has been proposed, and its asymptotic distribution under the null hypothesis of CI has been derived. Theoretical analysis, along with empirical validation on various datasets, rigorously demonstrates the effectiveness of our testing methods.

1 Introduction

Independence and conditional independence (CI) are fundamental concepts in statistics. They are leveraged for exploring queries in statistical inference, such as sufficiency, parameter identification, and ancillarity (Dawid, 1979). They also play a central role in emerging areas such as causal discovery (Koller & Friedman, 2009), graphical model learning, and feature selection (Xing et al., 2001). Tests for CI have attracted increasing attention from both theoretical and application sides.

Formally, the problem is to test the CI of two variables X_1 and X_2 given a random vector (a set of other variables) $\boldsymbol{Z}$. In statistical notation, the null hypothesis is written as $H_0 : X_1 \perp\!\!\!\perp X_2 \mid \boldsymbol{Z}$, where $\perp\!\!\!\perp$ denotes "independent from." The alternative hypothesis is written as $H_1 : X_1 \not\perp\!\!\!\perp X_2 \mid \boldsymbol{Z}$, where $\not\perp$ denotes "dependent with." The null hypothesis implies that once $\boldsymbol{Z}$ is known, the values of X_{j_1} provide no additional information about X_{j_2}, and vice versa. Different tests have been designed to handle different scenarios, including Gaussian variables with linear dependence (Yuan & Lin, 2007; Peterson et al., 2015; Mohan et al., 2012; Ren et al., 2015) and non-linear dependence (Fukumizu et al., 2004; Zhang et al., 2012; Strobl et al., 2019; Sen et al., 2017; Aliferis et al., 2010) (*For detailed related work, please refer to App. F*).

Given observations of X_1, X_2, and $\boldsymbol{Z}$, the CI can be effectively tested with existing methods. However, in many scenarios, accurately measuring continuous variables of interest is challenging due to limitations in data collection. Sometimes the data obtained are approximations represented as discretized values. For example, in finance, variables such as asset values cannot be measured and are binned into ranges for assessing investment risks (e.g., sell, hold, and strong buy) (Changsheng & Yongfeng, 2012; Damodaran, 2012). Similarly, in mental health, anxiety levels are often assessed using scales like the GAD-7, which categorizes responses into levels such as mild, moderate, or severe (Mossman et al., 2017; Johnson et al., 2019). In the entertainment industry, the quality of movies is typically summarized through viewer ratings (Sparling & Sen, 2011; Dooms et al., 2013).

When discretization is present, existing CI tests can fail to determine the CI of underlying continuous variables. This issue arises because existing CI tests treat discretized observations as observations of continuous variables, leading to incorrect conclusions about their CI relationships. More precisely, the problem lies in the discretization process, which introduces new discrete variables. Consequently, *although the intent is to test the CI of the underlying continuous variables, what is actually being tested is the CI involving a mix of both continuous and newly introduced discrete variables*. In general, this CI relationship is inconsistent with the one among the underlying continuous variables.

As illustrated in Fig. 1, we show different data-generative processes using causal graphical models (Pearl, 2000) in the presence of discretization. A gray node indicates an observable variable, while a white node indicates a latent variable. Variables denoted by X_j (without a tilde $\sim$) represent continuous variables, which may not be observed; while variables denoted by $\tilde{X}_j$ represent observed discretized variables derived from X_j due to discretization. In Fig. 1(a), X_2 is latent, and only its discrete counterpart $\tilde{X}_2$ is observed. In this case, rather than observing X_1, X_2, and X_3, we only observe X_1, $\tilde{X}_2$, and X_3. Existing CI methods use these observations to test ***whether*** $X_1 \perp\!\!\!\perp X_3 \mid \{X_2\}$, but what is actually being tested is ***whether*** $X_1 \perp\!\!\!\perp X_3 \mid \{\tilde{X}_2\}$. In fact, according to the *causal Markov condition* (Spirtes et al., 2000), it can be inferred from Fig. 1(a) that $X_1 \perp\!\!\!\perp X_3 \mid \{X_2\}$ and $X_1 \not\perp\!\!\!\perp X_3 \mid \{\tilde{X}_2\}$. This mismatch leads to existing CI methods, that employ observations to check the CI relationships between X_1 and X_3 given X_2, to reach incorrect conclusions. Due to the same reason, checking the CI also fails in Fig 1(b) and Fig 1(c).

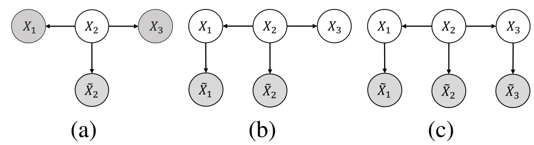


Figure 1: We illustrate different data generative processes with causal graphical models. The discretization process introduces new discrete variables which are denoted with a tilde ($\sim$).

In this paper, we design a CI test specifically for handling the presence of discretization. An appropriate test statistic for the CI of latent continuous variables, based solely on discretized observations, is derived. To develop this test, we first estimate the covariance between latent continuous variables and discretized observations. This is achieved by constructing bridge equations that enable the estimation of covariance using statistics derived from discretized observations. Subsequently, to utilize the estimated covariance of latent continuous variables for testing CI relationships, we apply a node-wise regression approach (Callot et al., 2019). This method allows us to derive test statistics for CI based on the estimated covariance. By assuming that the continuous variables follow a Gaussian distribution, we can derive the asymptotic distributions of the test statistics under the null hypothesis of CI. The major contributions of our paper include that

- We develop a CI test for ensuring accurate analysis in scenarios where data has been discretized, which are common due to limitations in data collection or measurement techniques, such as in financial analysis and healthcare.
- Our CI test can handle various scenarios including 1). Both variables X_{j_1} and X_{j_2} are discretized 2). Both variables X_{j_1} and X_{j_2} are continuous. 3). One of the variables X_{j_1} or X_{j_2} is discretized.
- We compare our test with the existing methods on both synthetic and real-world datasets, confirming that our method can effectively estimate the CI of the underlying continuous variables and outperform the existing tests applied on the discretized observations.

2 Problem Setting and Necessity of Correction

Problem Setting Consider a set of independent and identically distributed (i.i.d.) p-dimensional random vectors, denoted as $\tilde{\boldsymbol{X}} = (X_1, X_2, \dots, \tilde{X}_j, \dots, \tilde{X}_p)^T$. In this set, some variables, indicated by a tilde ($\sim$), such as $\tilde{X}_j$, follow a discrete distribution. For each such variable, there exists a corresponding latent Gaussian random variable X_j. The transformation from X_j to $\tilde{X}_j$ is governed by an unknown monotone nonlinear function g_j and a thresholding function f_j. The function $f_j \circ g_j : \mathcal{X} \to \tilde{\mathcal{X}}$ maps the continuous domain of X_j onto the discrete domain of $\tilde{\mathcal{X}}_j$. Specifically, for each variable X_j, there exists a finite constant vector $\mathbf{d}_j = (d_{j,1}, \dots, d_{j,M-1})$ characterized by strictly increasing elements such that

$$\tilde{X}_j = f_j(g_j(X_j)) = \begin{cases} 1 & 0 < g_j(X_j) < d_{j,1} \\ m & d_{j,m-1} < g_j(X_j) < d_{j,m} \\ M & g_j(X_j) > d_{j,M-1} \end{cases} \tag{1}$$

This model is also known as the nonparanormal model (Liu et al., 2009). The cardinality of the domain after discretization is at least 2 and smaller than infinity. Our goal is to assess both conditional and unconditional independence among the variables of the vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$. In our model, we assume $\boldsymbol{X} \sim N(0, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ only contain 1 among its diagonal, i.e., $\sigma_{jj} = 1$ for all $j \in [1, \ldots, p]$. One should note this assumption is *without loss of generality*. We provide a detailed discussion of our assumption in App. B.9.

Why the correction is essential? We aim to propose a CI test that serves as a correction to infer the correct CI relationships among the latent continuous variables of interest. One question that arises is whether the discretized variables exhibit the same conditional independence as their original continuous counterparts, i.e., the correction is not needed. This concern becomes more significant when the level of discretization is high. To highlight the effect of discretization, we propose the following theorem. In essence, the discretization inevitably introduces distortions, leading to potentially false conclusions. The proof can be found in Appendix B.1.

Theorem 2.1. *Let X_1, X_2 and X_3 be jointly Gaussian random variables such that $X_1 \perp\!\!\!\perp X_3 | X_2$, $\tilde{X}_2 = f_j(g_j(X_2))$ is the descretized observation as defined in equation 1. Then the conditional independence between X_1 and X_3 given $\tilde{X}_2$ doesn't hold, i.e., $X_1 \not\perp\!\!\!\perp X_3 | \tilde{X}_2$.*

3 DCT: A Discretization-Aware CI Test

Notation Throughout this work, we use X_j to denote the j-th component of the vector of variables $\boldsymbol{X}$. We denote the sample mean of X_j by $\mathbb{E}_n[X_j]$, and the expectation by $\mathbb{E}[X_j]$. The empirical probability is represented by $\mathbb{P}_n$ whereas the true probability is denoted by $\mathbb{P}$. For a matrix $\mathbf{X}$, $\mathbf{X}_{-j}$ represents all columns of $\mathbf{X}$ except the j-th column, $\mathbf{X}_{-j-j}$ denotes the submatrix obtained by removing both the j-th column and row, and $\mathbf{X}_{-jj}$ represents the j-th column of $\mathbf{X}$ with the j-th row removed. For any parameter α, we use $\hat{\alpha}$ to denote its estimation. $\mathbb{1}\{condition\}$ is 1 if the condition holds true, 0 otherwise. For a full notation table, please refer to Appendix A.

To develop a CI test, we need to design a test statistic that can reflect the conditional dependence relation and be calculated from observations. Next, it is essential to derive the underlying distribution of this statistic under the null hypothesis that the tested variables are conditionally (or unconditionally) independent. By calculating the value of the test statistic from observations and determining if this statistic is likely to be drawn from the derived distribution (i.e., calculating the *p-value* and comparing it with the significance level α), we can decide if the null hypothesis should be rejected.

Our objective is to deduce the independence and CI relationships within the original multivariate Gaussian variable $\boldsymbol{X}$, based on its discretized observations $\tilde{\boldsymbol{X}}$. In the context of a multivariate Gaussian model, this challenge is directly equivalent to constructing statistical inferences for its covariance matrix $\boldsymbol{\Sigma} = (\sigma_{j_1 j_2})$ and its precision matrix $\boldsymbol{\Omega} = (\omega_{jk}) = \boldsymbol{\Sigma}^{-1}$ (Baba et al., 2004). The covariance matrix $\boldsymbol{\Sigma}$ captures the pairwise covariances between variables, while the precision matrix $\boldsymbol{\Omega}$ provides information about the CI between variables. Specifically, the entry ω_{jk} in the precision matrix is the partial correlation coefficient between variables X_j and X_k, which can be used to test whether these variables are conditionally independent given some other variables. Technically, we are interested in two things: (1) the calculation of the covariance $\hat{\sigma}_{j_1 j_2}$ and the precision coefficient (or the partial correlation coefficient) $\hat{\omega}_{jk}$, serving as the estimation of $\sigma_{j_1 j_2}$ and ω_{jk} respectively (in this paper, a variable with a hat indicates its estimation); and (2) the derivation of the distribution of $\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2}$ and $\hat{\omega}_{jk} - \omega_{jk}$ under the null hypothesis of independence and CI.

In the remainder of section, 1). we first introduce ***bridge equations*** to address the estimation challenge of the covariance $\sigma_{j_1 j_2}$; 2). we proceed to derive the distribution of $\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2}$, demonstrating it is ***asymptotically normal***; 3). utilizing ***nodewise regression***, we establish the relationship between the covariance matrix $\boldsymbol{\Sigma}$ and the precision matrix $\boldsymbol{\Omega}$, where the regression parameter $\beta_{j,k}$ acts as an effective surrogate for ω_{jk}. Leveraging the distribution of $\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2}$, we further illustrate that $\hat{\beta}_{j,k} - \beta_{j,k}$ is also ***asymptotically normal***.

3.1 Estimating Covariance Through Observations

Our first task is to establish the connection between the underlying covariance $\sigma_{j_1 j_2}$ of the continuous pair X_{j_1} and X_{j_2} with their observed counterparts. In the presence of discretization, the sample covariance matrix computed from $\tilde{\boldsymbol{X}}$ is inconsistent with the covariance matrix of $\boldsymbol{X}$. To obtain the estimation $\hat{\sigma}_{j_1,j_2}$ consistent with σ_{j_1,j_2}, the bridge equation is leveraged. In general, its form is as follows.

$$\hat{\tau}_{j_1,j_2} = T(\hat{\sigma}_{j_1 j_2}; \hat{\boldsymbol{\Lambda}}), \tag{2}$$

where $\hat{\sigma}_{j_1 j_2}$ is the estimated covariance, $\hat{\tau}_{j_1,j_2}$ is a statistic that can also be estimated from observations, and $\hat{\boldsymbol{\Lambda}}$ is a set of additional parameters required by the function $T(\cdot)$. The specific form of the function $T(\cdot)$ will be derived later. Both $\hat{\tau}_{j_1,j_2}$ and $\hat{\boldsymbol{\Lambda}}$ should be able to be calculated purely relying on observations. *Then, given the calculated $\hat{\tau}_{j_1,j_2}$ and $\hat{\boldsymbol{\Lambda}}$, $\hat{\sigma}_{j_1,j_2}$ can be obtained by solving the bridge equation.* As a result, the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{X}$ can be estimated, which contains information about both unconditional independence and CI (which can be derived from its inverse).

To estimate the covariance of a latent multivariate Gaussian distribution, we need to design appropriate $\hat{\tau}_{j_1,j_2}$, $\hat{\boldsymbol{\Lambda}}$, and $T(\cdot)$. Notably, bridge equations have to be designed to handle the possible cases: C1. both observed variables are discretized; C2. one variable is continuous while the other is discretized. For C3. both variables remain continuous, we can easily take its sample covariance as the estimated covariance. We will show that cases C1 and C2 can be merged into a single form of bridge equation with different parameters and a binarization operation applied to the observations. Our bridge equations are presented in Def. 3.1, Def. 3.2.

3.1.1 Bridge Equations for Discretized and Mixed Pairs

Let us first address the challenging cases where both observed variables are discretized or where one variable is continuous while the other is discretized. In general, different bridge equations would need to be designed to handle each case individually. *However, in our analysis, we provide a unified bridge equation that is applicable to both cases.* This is achieved by *binarizing* the observed variables, thereby unifying both cases into a binary case. As some information may be lost in the binarization process, this unification may require more data samples compared to using tailored bridge functions for each specific case. Developing specific bridge equations for each case to improve sample efficiency is left in future work.

Theoretically, continuous variables and discrete variables can be further discretized into binary variables. Imagine we have the observed variable $\tilde{X}_{j_1}$ with the possible values "low", "medium", "high", we can create a dividing point: everything above becomes "very high", everything below becomes "very low". This binarization process is also applicable to the continuous variable. Note that $\tilde{X}_j$ is just the discretized version of its corresponding continuous variable X_j, this dividing point directly responds to a specific value in the original continuous domain, which we denote as the boundary h_j. Multiple choices of h_j are possible. In this paper, we define h_j as the boundary in the continuous domain that corresponds to the mean of its discretized counterpart $\tilde{X}_j$. Mathematically, we define h_j as follows: for any single discretized variable $\tilde{X}_j$, there exists a constant c_j such that $h_j = g_j^{-1}(c_j)$ satisfying

$$\mathbb{1}\{\tilde{x}_j^i > \mathbb{E}[\tilde{X}_j]\} = \mathbb{1}\{g_j(x_j^i) > c_j\} = \mathbb{1}\{x_j^i > h_j\}.$$

Estimating the boundary Since the continuous variable X_j follows a normal distribution according to our assumption, we can thus construct the relation $\mathbb{P}(\tilde{X}_j > \mathbb{E}[\tilde{X}_j]) = 1 - \Phi(h_j)$, where Φ is the cumulative distribution function (cdf) of a standard normal distribution. Apparently, we do not have access to the true probability. However, we can easily obtain its estimation by counting how many samples drop in the region larger than its sample mean. Specifically,

$$\hat{h}_j = \Phi^{-1}(1 - \hat{\tau}_j), \tag{3}$$

where $\hat{\tau}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\tilde{x}_j^i > \mathbb{E}_n[\tilde{X}_j]\}$, serving as the estimation of $\mathbb{P}(\tilde{X}_j > \mathbb{E}[\tilde{X}_j])$. We further denote $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$.

Intuition of estimating covariance The question now is to estimate the latent covariance $\sigma_{j_1 j_2}$ for the observed discrete pair $(\tilde{X}_{j_1}, \tilde{X}_{j_2})$ or mixed pair $(\tilde{X}_{j_1}, X_{j_2})$. Leveraging the binarization process, there exists boundaries h_{j_1}, h_{j_2} that partition the continuous variables pair X_{j_1} and X_{j_2} to a 2×2 contingency table. The area of each cell in this table represents the joint probability of the pair (X_{j_1}, X_{j_2}) falling with a specific region defined by those boundaries. In this paper, we focus on the top-right cell of the contingency table, which represents the joint probability of both variables exceeding their respective boundaries.

Mathematically, we denote $\bar{\Phi}(z_1, z_2; \rho) = \mathbb{P}(Z_1 > z_1, Z_2 > z_2)$, where (Z_1, Z_2) follows a bivariate normal distribution with mean zero, variance one and covariance ρ. For a discretized pair of observed variables $(\tilde{X}_{j_1}, \tilde{X}_{j_2})$, We define

$$\tau_{j_1,j_2} := \mathbb{P}(\tilde{X}_{j_1} > \mathbb{E}[\tilde{X}_{j_1}], \tilde{X}_{j_2} > \mathbb{E}[\tilde{X}_{j_2}]) = \bar{\Phi}(h_{j_1}, h_{j_2}; \sigma_{j_1 j_2}).$$

That is, the probability of discretized variables larger than their mean can be expressed as a function of underlying covariance. This equation serves as the key to estimating latent covariance based on the discretized observations. Similarly, we can estimate this probability by counting samples dropped into the region of both variables exceeding their sample means. Mathematically,

$$\hat{\tau}_{j_1,j_2} := \mathbb{P}_n(\tilde{X}_{j_1} > \mathbb{E}_n[\tilde{X}_{j_1}], \tilde{X}_{j_2} > \mathbb{E}_n[\tilde{X}_{j_2}]) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\tilde{x}^i_{j_1} > \mathbb{E}_n[\tilde{X}_{j_1}], \tilde{x}^i_{j_2} > \mathbb{E}_n[\tilde{X}_{j_2}]\}. \quad (4)$$

Since $\bar{\Phi}(h_{j_1}, h_{j_2}; \sigma_{j_1 j_2})$ is a function of $\sigma_{j_1 j_2}$, by substituting the parameters $\tau_{j_1,j_2}, h_{j_1}, h_{j_2}$ as their estimation, we can construct the bridge equation as follows:

Definition 3.1 (Bridge Equation for A Discretized-Variable Pair)**.** For discretized variables $\tilde{X}_{j_1}$ and $\tilde{X}_{j_2}$, the bridge equation is defined as:

$$\hat{\tau}_{j_1,j_2} = T(\hat{\sigma}_{j_1 j_2}; \{\hat{h}_{j_1}, \hat{h}_{j_2}\}),$$

where $T(\hat{\sigma}_{j_1 j_2}; \{\hat{h}_{j_1}, \hat{h}_{j_2}\}) = \int_{z_1 > \hat{h}_{j_1}} \int_{z_2 > \hat{h}_{j_2}} \phi(z_1, z_2; \hat{\sigma}_{j_1 j_2})\, dz_1\, dz_2$, and ϕ is the probability density function of a bivariate normal distribution with mean zero and covariance $\hat{\sigma}_{j_1 j_2}$, we note that $\hat{h}_{j_1}, \hat{h}_{j_2}$ can be simply calculated using equation 3 and $\hat{\tau}_{j_1,j_2}$ can be calculated using equation 4.

Following the same intuition, we can directly apply the same bridge equation to estimate the covariance of mixed pairs. The only difference is there is no need to estimate the boundary $\hat{h}_j$ for the continuous variable. Instead, we can incorporate its true mean of zero into the equation.

Definition 3.2 (Bridge Equation for A Continuous-Discretized-Variable Pair)**.** For one continuous variable X_{j_1} and one discretized variable $\tilde{X}_{j_2}$, the bridge function is defined as follows:

$$\hat{\tau}_{j_1,j_2} = \mathbb{P}_n(X_{j_1} > 0, \tilde{X}_{j_2} > \mathbb{E}_n[\tilde{X}_{j_2}]) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x^i_{j_1} > 0, \tilde{x}^i_{j_2} > \mathbb{E}_n[\tilde{X}_{j_2}]\} = T(\sigma_{j_1 j_2}; \{0, \hat{h}_{j_2}\}),$$

and the function $T(\cdot)$ has the same form of Def. 3.1.

3.1.2 Calculation of Estimated Covariance

For the continuous case where there is no discretization transformation, the sample covariance provides a consistent estimation of the true one. That is, for an observable pair of continuous variables (X_{j_1}, X_{j_2}), we can simply obtain the analytic solution of estimated covariance:

$$\hat{\sigma}_{j_1 j_2} = \frac{1}{n}\sum_{i=1}^{n} x^i_{j_1} x^i_{j_2} - \frac{1}{n}\sum_{i=1}^{n} x^i_{j_1} \frac{1}{n}\sum_{i=1}^{n} x^i_{j_2} \quad (5)$$

For the cases involving the discretized variable as proposed in Def. 3.1 and Def. 3.2, we can rely on the property that variance $\mathbf{\Sigma}$ only contains 1 among the diagonal, which implies the covariance $\sigma_{j_1 j_2}$ should vary from -1 to 1. Thus, we can calculate the estimated covariance by solving the objective

$$\hat{\sigma}_{j_1 j_2} = \arg\min_{\sigma'_{j_1 j_2}} ||\hat{\tau}_{j_1,j_2} - T(\sigma'_{j_1 j_2}; \{\hat{h}_{j_1}, \hat{h}_{j_2}\})||^2 \quad s.t. -1 < \sigma'_{j_1 j_2} < 1. \quad (6)$$

The $\hat{\tau}_{j_1,j_2}$ is a one-to-one mapping with calculated $\hat{\sigma}_{j_1 j_2}$, $\hat{h}_{j_1}$ and $\hat{h}_{j_2}$, which is proved in App. B.3

3.2 Unconditional Independence Test

The estimation of covariance $\hat{\sigma}_{j_1 j_2}$ can be effectively solved using the designed bridge equation. Now, we focus on deriving the distribution of $\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2}$. These results is used as an unconditional independence test in the presence of the discretization. Moreover, Thm. 3.3, Lem. 3.4, Lem. 3.5 and Lem. 3.6 will be leveraged in the derivation process of the CI test in Section 3.3. The detailed derivation steps for both unconditional independence test and CI test are relatively complicated, therefore, we will provide a general intuition. For a complete derivation, please refer to the App. B.4.

Assume we are interested in the true parameter θ_0, e.g., for discretized pairs, $\theta = (\hat{\sigma}_{j_1 j_2}, h_{j_1}, h_{j_2})$. We denote $\hat{\theta}$ as its estimation which is close to θ_0, and $f(\theta)$ is a continuous function. By leveraging Taylor expansion, we have

$$f(\hat{\theta}) = f(\theta_0) + f'(\theta_0)(\hat{\theta} - \theta_0) + \dots, \tag{7}$$

where the second order terms and more are omitted, which directly constructs the relationship between the estimated parameter with the true one. Rearrange the term, we get $\hat{\theta} - \theta_0 = (f(\hat{\theta}) - f(\theta_0))/f'(\theta_0)$. If the denominator is a constant and the numerator can be expressed as a sum of i.i.d samples, we can see $\hat{\theta} - \theta_0$ will be asymptotically normal according to the central limit theorem (Van der Vaart, 2000).

Let $\psi_{\hat{\theta}} = [f^1_{\hat{\theta}}(\cdot), \dots]^T$ contains a group of functions parameterized by $\hat{\theta}$. We define the functions evaluated at one sample as $\psi^i_{\hat{\theta}} = \psi_{\hat{\theta}}(\mathbf{z}^i)$, where $\mathbf{z}^i$ denotes the i-th sample point. We define the sample mean of these functions evaluated at n points as $\mathbb{E}_n[\psi_{\hat{\theta}}] = \frac{1}{n}\sum_{i=1}^n \psi^i_{\hat{\theta}}$, similarly, $\mathbb{E}_n[\psi_{\hat{\theta}}\psi^T_{\hat{\theta}}] = \frac{1}{n}\sum_{i=1}^n \psi^i_{\hat{\theta}}{\psi^i_{\hat{\theta}}}^T$ and $\psi'_{\hat{\theta}}$ denotes the Jacobian matrix $\frac{\partial \psi_{\hat{\theta}}}{\partial \hat{\theta}}$. We now provide the main result of derived distribution $\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2}$ under the hull hypothesis that tested pairs are independent.

Theorem 3.3 (Independence Test). *Under the null hypothesis that the Gaussian variables (X_{j_1}, X_{j_2}) are statistically independent $\sigma_{j_1 j_2} = 0$, the test statistics $\hat{\sigma}_{j_1 j_2}$ obtained according to Def. 3.1 for discretized pairs $(\tilde{X}_{j_1}, \tilde{X}_{j_2})$, Def. 3.2 for mixed pairs $(X_{j_1}, \tilde{X}_{j_2})$ and equation 5 for continuous pairs, is asymptotically normal:*

$$\sqrt{n}(\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2}) \xrightarrow{d} N\left(0, ((\mathbb{E}_n[\psi'_{\hat{\theta}}])^{-1}\mathbb{E}_n[\psi_{\hat{\theta}}\psi^T_{\hat{\theta}}](\mathbb{E}_n[\psi'^T_{\hat{\theta}}])^{-1})_{1,1}\right), \tag{8}$$

where the specific form of $\psi^i_{\hat{\theta}}$ are presented in Lem. 3.4,Lem. 3.5 and Lem. 3.6.

We now provide the specific forms of $\psi^i_{\hat{\theta}}$. Since the variables being tested for independence can be both discretized, only one being discretized, or neither being discretized—-the form of $\psi_{\hat{\theta}}$ varies accordingly. The specific forms of $\psi_{\hat{\theta}}$ in these scenarios are defined as follows:

Lemma 3.4. *($\psi^i_{\hat{\theta}}$ for A Continuous-Variable Pair). For two continuous variables X_{j_1} and X_{j_2} with their corresponding i-th samples $x^i_{j_1}, x^i_{j_2}$:*

$$\psi^i_{\hat{\theta}} := x^i_{j_1}x^i_{j_2} - \mathbb{E}_n[X_{j_1}]\mathbb{E}_n[X_{j_2}] - \hat{\sigma}_{j_1 j_2}, \tag{9}$$

Lemma 3.5 ($\psi^i_{\hat{\theta}}$ for A Discretized-Variable Pair). *For discretized variables $\tilde{X}_{j_1}$ and $\tilde{X}_{j_2}$, with their corresponding i-th samples $\tilde{x}^i_{j_1}, \tilde{x}^i_{j_2}$:*

$$\psi^i_{\hat{\theta}} := \begin{pmatrix} \hat{\tau}^i_{j_1, j_2} - T(\hat{\sigma}_{j_1 j_2}; \{\hat{h}_{j_1}, \hat{h}_{j_2}\}) \\ \hat{\tau}^i_{j_1} - \bar{\Phi}(\hat{h}_{j_1}) \\ \hat{\tau}^i_{j_2} - \bar{\Phi}(\hat{h}_{j_2}) \end{pmatrix}, \tag{10}$$

where $\hat{\tau}^i_{j_1 j_2} = \mathbb{1}\{\tilde{x}^i_{j_1} > \mathbb{E}_n[\tilde{X}_{j_1}], \tilde{x}^i_{j_2} > \mathbb{E}_n[\tilde{X}_{j_2}]\}, \hat{\tau}^i_{j_1} = \mathbb{1}\{\tilde{x}^i_j > \mathbb{E}_n[\tilde{X}_{j_1}]\}$, and similarly for $\hat{\tau}^i_{j_2}$.

Lemma 3.6 ($\psi^i_{\hat{\theta}}$ for A Continuous-Discretized-Variable Pair). *For one discretized variable $\tilde{X}_{j_2}$ and one continuous variable X_{j_1}, with their corresponding i-th sample point $\tilde{x}^i_{j_2}, x^i_{j_1}$:*

$$\psi^i_{\hat{\theta}} := \begin{pmatrix} \hat{\tau}^i_{j_1, j_2} - T(\hat{\sigma}_{j_1 j_2}; \{0, \hat{h}_{j_2}\}) \\ \hat{\tau}^i_{j_2} - \bar{\Phi}(\hat{h}_{j_2}) \end{pmatrix}, \tag{11}$$

where $\hat{\tau}^i_{j_1 j_2} = \mathbb{1}\{x^i_{j_1} > 0, \tilde{x}^i_{j_2} > \mathbb{E}_n[\tilde{X}_{j_2}]\}, \hat{\tau}^i_{j_2} = \mathbb{1}\{\tilde{x}^i_j > \mathbb{E}_n[\tilde{X}_{j_2}]\}$.

Derivation of forms of $\psi_{\hat{\theta}}$ for different cases and their corresponding distribution defined in Eq equation 8 can be found in App. B.5, App. B.6, App. B.7. Up to this point, our discussion has been confined to the case of covariance $\sigma_{j_1 j_2}$, the indicator of unconditional independence. In the next section, we will present the results of our CI test.

3.3 Conditional Independence (CI) Test

To construct a CI test of our model, we are interested at two things: calculation of the estimated precision coefficient $\hat{\omega}_{jk}$ and the derivation of the corresponding distribution $\hat{\omega}_{jk} - \omega_{jk}$. While obtaining $\hat{\omega}_{jk}$ from the $\hat{\mathbf{\Sigma}}$ is straightforward, it leaves the inference problem unresolved. Thus, we leverage nodewise regression and show the regression parameter $\beta_{j,k}$ serving as a surrogate of testing for $\omega_{jk} = 0$, we then construct the formulation of $\hat{\beta}_{j,k} - \beta_{j,k}$ as the combination of formulation of $\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2}$ and show it will also be asymptotically normal.

The following lemma formalizes the properties of nodewise regression that enable this approach:

Lemma 3.7. *[Nodewise Regression Properties] For a p-dimensional multivariate normal variable $\mathbf{X} = (X_1, \dots, X_p) \sim N(0, \mathbf{\Sigma})$ with covariance matrix $\mathbf{\Sigma}$ and precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} = (\omega_{jk})_{1 \le j,k \le p}$. For any $j \in \{1, \dots, p\}$, consider the nodewise regression where each X_j is regressed on all other variables:*

$$X_j = \sum_{k \neq j} X_k \beta_{j,k} + \epsilon_j,$$

where $\beta_{j,k}$ is the regression coefficient of X_k in predicting X_j, $\beta_j = (\beta_{j,k})_{k \neq j} \in \mathbb{R}^{p-1}$ is the vector of all coefficients, and ϵ_j is the residual term. Then the following relationships hold:

$$\begin{aligned} \beta_j &= \mathbf{\Sigma}_{-j-j}^{-1} \mathbf{\Sigma}_{-jj} \in \mathbb{R}^{p-1}, \\ \beta_{j,k} &= -\frac{\omega_{jk}}{\omega_{jj}}, \quad j \neq k. \end{aligned} \tag{12}$$

The derivation can be found in Appendix B.8.1. The lemma establishes the deterministic relationships between the regression coefficient $\beta_{j,k}$ and the entry of precision matrix ω_{jk}. Since ω_{jj} will never be zero (due to the positive definiteness $\mathbf{\Omega}$), we can conclude $\beta_{j,k}$ serves as an effective surrogate of ω_{jk}. Moreover, β_j can be expressed in terms of the submatrices of the covariance matrix $\mathbf{\Sigma}$. We can further conduct its estimation $\hat{\beta}_j = (\hat{\beta}_{j,k})_{k \neq j} = \hat{\mathbf{\Sigma}}_{-j-j}^{-1} \hat{\mathbf{\Sigma}}_{-jj}$, where the estimated covariance terms can be obtained using Def. 3.1, 3.2 and equation 5.

Statistical Inference for $\beta_{j,k}$ Nodewise regression offers a direct solution for the estimation problem. A pertinent inquiry pertains to the construction of the distribution of $\hat{\beta}_j - \beta_j$. It is crucial to recognize that the distribution of $\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2}$ is already established. Therefore, if we can conceptualize $\hat{\beta}_j - \beta_j$ as a linear combination of $\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2}$, the problem is directly solved, i.e., the $\hat{\beta}_j - \beta_j$ is linear combination of dependent Gaussian variables. The underlying relationship between these variables is as follows:

$$\hat{\beta}_j - \beta_j = -\hat{\mathbf{\Sigma}}_{-j-j}^{-1} \left((\hat{\mathbf{\Sigma}}_{-j-j} - \mathbf{\Sigma}_{-j-j})\beta_j - (\hat{\mathbf{\Sigma}}_{-jj} - \mathbf{\Sigma}_{-jj}) \right).$$

The derivation is provided in App. B.8.2. For ease of notation, we further express the distribution of the difference between the estimated covariance and the true covariance as

$$\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2} = \frac{1}{n} \sum_{i=1}^{n} \xi_{j_1, j_2}^i. \tag{13}$$

The specific form of ξ_{j_1, j_2}^i is given in App. B.5, B.6, B.7 respectively for different cases. For notational convenience, we express $\hat{\mathbf{\Sigma}}_{-j-j} - \mathbf{\Sigma}_{-j-j} = \frac{1}{n} \sum_{i=1}^{n} \Xi_{-j,-j}^i$ and $\hat{\mathbf{\Sigma}}_{-jj} - \mathbf{\Sigma}_{-jj} = \frac{1}{n} \sum_{i=1}^{n} \Xi_{-j,j}^i$, where ξ_{j_1, j_2} is the element of the matrix Ξ at the position indexed by (j_1, j_2). We now propose the statistic and its asymptotic distribution for the CI test in the following theorem.

Theorem 3.8 (Conditional Independence test). *Under the null hypothesis that Gaussian variables X_j and X_k are conditional statistically independent given all other variables $\mathbf{X}_{-\{jk\}}$, i.e., $\beta_{j,k} = 0$, the testing statistic*

$$\hat{\beta}_{j,k} = (\hat{\mathbf{\Sigma}}_{-j-j}^{-1} \hat{\mathbf{\Sigma}}_{-jj})_{[k]}, \tag{14}$$

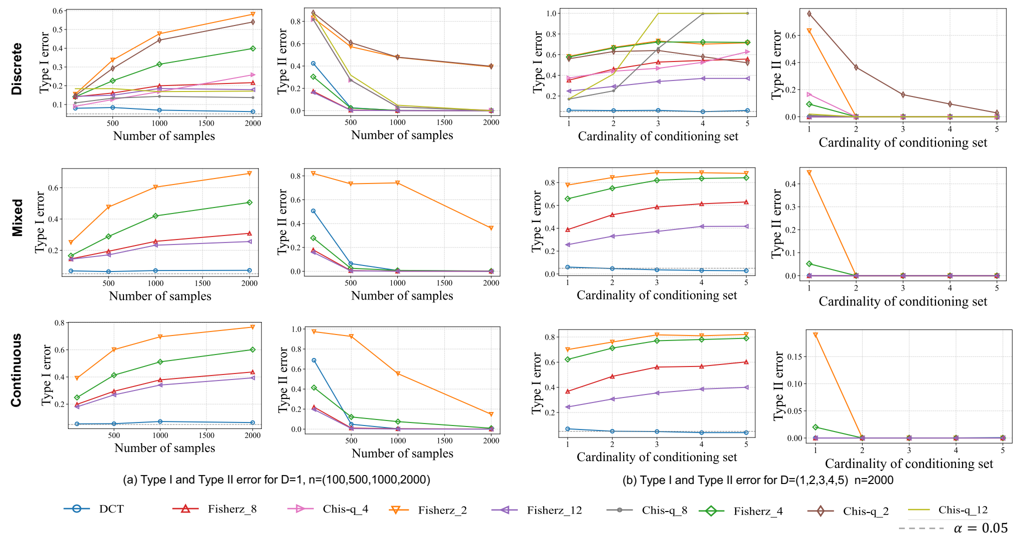


Figure 2: Comparison of results of Type I and calibrated Type II error ($1 -$ power) for all three types of tested data (continuous, mixed, discrete) and different number of samples and cardinality of conditioning set. The suffix attached to a test's name denotes the cardinality of discretization; for example, "Fisherz_4" signifies the application of the Fisher-z test to data discretized into four levels. Chi-square test is only applicable for the discrete case.

where $[k]$ denotes the element corresponding to the variable X_k in $\hat{\mathbf{\Sigma}}_{-j-j}^{-1}\hat{\mathbf{\Sigma}}_{-jj}$, has the asymptotic distribution:

$$\sqrt{n}(\hat{\beta}_{j,k} - \beta_{j,k}) \sim N(0, a^{[k]T}\frac{1}{n}\sum_{i=1}^{n} vec(B_{-j}^{i})vec(B_{-j}^{i})^{T}a^{[k]}),$$

$$\text{where } B^{i} = \begin{bmatrix} \Xi_{-j,j}^{i} & ^{T} \\ \Xi_{-j,-j}^{i} & \end{bmatrix}, \quad a_{l}^{[k]} = \begin{cases} \left(\hat{\mathbf{\Sigma}}_{-j-j}^{-1}\right)_{[k],l}, & \text{for } l \in \{1, \ldots, p-1\} \\ \sum_{q=1}^{p-1}\left(\hat{\mathbf{\Sigma}}_{-j-j}^{-1}\right)_{[k],l}\left(\tilde{\beta}_{j}\right)_{q}, & \text{for } l \in \{p, \ldots, p^{2}-p\} \end{cases}$$

and $\tilde{\beta}_j$ is β_j whose $\beta_{j,k} = 0$; vec is row-wise vectorization of a matrix.

In practice, we can plug in the estimation of regression parameter $\hat{\beta}_j$ and set $\hat{\beta}_{j,k} = 0$ as the substitution of $\tilde{\beta}_j$ to calculate the variance and do the CI test. Specifically, we can obtain the $\hat{\beta}_{j,k}$ using equation 14 where the estimated covariance terms can be calculated by solving the bridge equation Eq. 2. Under the null hypothesis that $\beta_{j,k} = 0$ (conditional independence), we can take the calculated $\hat{\beta}_{j,k}$ into the distribution defined in Thm. 3.8 and obtain the p-value. If the p-value is smaller than the predefined significance level α (normally set at 0.05), we will infer the tested pairs are conditionally dependent; otherwise, we do not. The detailed derivation of the Thm. 3.8 can be found in App. B.8.2. The pseudo code of DCT is provided in Appendix D.

4 Experiments

We applied the proposed method DCT to synthetic data to evaluate its practical performance and compare it with Fisher-Z test (Fisher, 1921) (for all three data types) and Chi-Square test (F.R.S., 2009) (for discrete data only) as baselines. Specifically, we investigated its Type I and Type II error and its application in causal discovery. The experiments investigating its robustness, performance in denser graphs and effectiveness in a real-world dataset can be found in App. E.

4.1 On the Effect of the Cardinality of Conditioning Set and the Sample Size

Our experiment investigates the variations in Type I and Type II error (1 minus power) probabilities under two conditions. In the first scenario, we focus on the effects of modifying the sample size, denoted as $n = (100, 500, 1000, 2000)$, while conditioning on a single variable. In the second, the sample size is held constant at 2000, and we vary the cardinality of the conditioning set, represented as $D = (1, 2, \dots, 5)$. It is assumed that every variable within this conditioning set is effective, i.e., they influence the CI of the tested pairs. We repeat each test 1500 times.

We use Y, W to denote the variables being tested and use Z to denote the variables being conditioned on. The discretized versions of the variables are denoted with a tilde symbol (e.g., $\tilde{Z}$). For both conditions, we evaluate three distinct types of observations of tested variables: continuous observations for both variables (Y, W), discrete observations for both variables $(\tilde{Y}, \tilde{W})$ and a mixed type $(\tilde{Y}, W)$. The variables in the conditioning set will always be discretized observations $(\tilde{Z})$.

To see how well the derived asymptotic null distribution approximates the true one, we verify if the probability of Type I error aligns with the significance level α preset in advance. We generate true continuous multivariate Gaussian data Y, W from Z_i (single $i = 1$ for the first scenario, and summed over n for the second), structured as $a_i Z_i + E$ and $\sum_{i=1}^{n} a_i Z_i + E$, where a_i is sampled from $U(0.5, 1.5)$ and E follows a standard normal distribution, independent of all other variables. This ensures $Y \perp\!\!\!\perp W | Z$. The data are then discretized into $K = (2, 4, 8, 12)$ levels, with boundaries randomly set based on the variable range. The first column in Fig. 2 (a) (b) shows the resulting probability of Type I errors at the significance level $\alpha = 0.05$ compared with other methods.

A good test should have as small a probability of Type II error as possible, i.e., a larger power. To test the power of our DCT, we generate the continuous multivariate Gaussian data Z_i from Y, W; constructed as $Z_i = a_i Y + b_i W + E$, where a_i, b_i are sampled from $U(0.5, 1.5)$ and E follows a standard normal distribution independent with all others, i.e., $Y \not\perp\!\!\!\perp W | Z$. The same discretization approach is applied here. One should note that directly comparing the p-value with a common predefined significance level is unfair since all baselines tend to produce very small p-values. Therefore, all tests are calibrated[1] in this experiment. The second column in Fig. 2 (a) and (b) correspondingly shows the calibrated Type II error as the number of samples and the cardinality of the conditioning set change, compared to other methods.

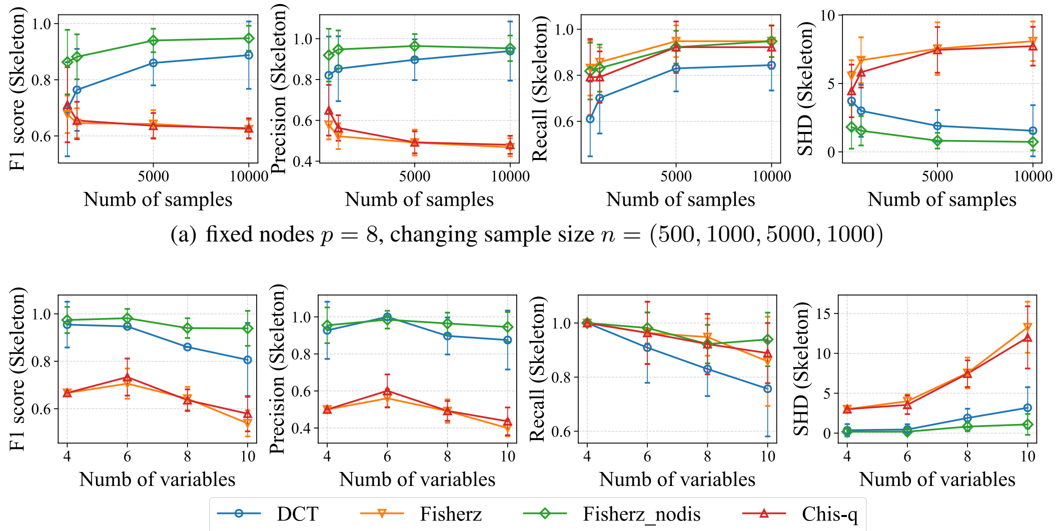
From Fig. 2 (a), we note that the Type I error rates with our derived null distribution are well-approximated at 0.05 across all three data types in both scenarios. In contrast, other testing methods show significantly higher Type I error rates, increasing with the number of samples and the size of the conditioning set. This indicates that such methods are more prone to erroneously concluding that tested variables are conditionally dependent. Additionally, while alternative tests demonstrate considerable power with smaller sample sizes, our approach requires a sample size of 1000 to achieve satisfactory power, particularly in mixed and continuous cases. A possible explanation for this phenomenon is that our method binarizes discretized data, which may not effectively utilize all observations. This aspect warrants further investigation in future research. Moreover, our test shows remarkable stability in response to changes in the number of conditioning sets.

4.2 Application in Causal Discovery

Causal discovery aims at looking for the true causal structure from the data. Under the assumption of causal Markov condition that the causal relationships among variables can be expressed by a Directed Acyclic Graph (DAG) $\mathcal{G}$ and its statistical independence is entailed in this graphic model, faithfulness ensures that the statistical independencies observed in the data can be reliably used to infer the causal structure. Given both assumptions, constraint-based causal discovery, e.g., PC algorithm (Spirtes et al., 2000) recovers the graph structure relying on testing the conditional independence of observation. Apparently, in the presence of discretization, the failures of testing conditional independence will seriously impair the resulting DAG.

To evaluate the efficacy of the DCT, we construct the true DAG $\mathcal{G}$ utilizing the Bipartite Pairing (BP) model as detailed in (Asratian et al., 1998), with the number of edges being one fewer than the number of nodes. The subsequent generation of true multivariate Gaussian data involves assigning causal weights drawn from a uniform distribution $U \sim (0.5, 2)$ and incorporating noise via samples from a standard normal distribution for each variable.

[1] Calibration is the process of empirically finding the decision threshold to match the desired significance level, ensuring accurate control of Type I error.



(a) fixed nodes $p = 8$, changing sample size $n = (500, 1000, 5000, 1000)$

(b) fixed sample size $n = 5000$, changing node $p = (4, 6, 8, 10)$

Figure 3: Experimental result of skeleton discovery on synthetic data for changing sample size (a) and changing number of nodes (b). Fisherz_nodis is the Fisher-z test applied to original continuous data. We evaluate F_1 (↑), Precision (↑), Recall (↑) and SHD (↓).

Following this, we binarize the data, setting the threshold randomly based on each variable's range. Our experiment is divided into two scenarios: In the first, we set the number of samples $n = 5000$, with the number of nodes p varying across 4, 6, 8, and 10. In the second scenario, we fix the number of nodes at $p = 8$ and explore sample sizes $n = (500, 1000, 5000, 10000)$.

A comparative analysis is performed using the PC algorithm integrated with various testing methods. Specifically, we compare DCT against the Fisher-z test applied to discretized data, the Chi-Square test, and the Fisher-z test on the original continuous data, the latter serving as a theoretical upper bound. Since the PC algorithm only returns a completed partially directed acyclic graph (CPDAG), we apply the same orientation rules from Dor & Tarsi (1992), as implemented by Causal-DAG (Chandler Squires, 2018), to convert a CPDAG into a DAG for easier comparison. We evaluate both the undirected skeleton and the directed graph using structural Hamming distance (SHD), F1 score, precision, and recall as evaluation metrics. For each setting, we run 10 graph instances with different seeds and report the mean and standard deviation for skeleton discovery in Figure 3 and DAG discovery in Figure 4 in Appendix C.

According to the result, DCT exhibits performance nearly on par with the theoretical upper bound across metrics such as F1 score, precision, and Structural Hamming Distance (SHD) when the number of variables (p) is small and the sample size (n) is large. Despite a decline in performance as the number of variables increases with a smaller sample size, DCT significantly outperforms both the Fisher-Z test and the Chi-square test. Notably, in almost all settings, the recall of DCT is lower than that of the baseline tests, which is a reasonable outcome *since these tests tend to infer conditional dependencies, thereby retaining all edges given the discretized observations.* For instance, a fully connected graph, would achieve a recall of 1.

5 Conclusion

In this paper, we present a new testing method tailored for scenarios commonly encountered in real-world applications, where variables, though inherently continuous, are only observable in their discretized forms. Our method distinguishes itself from existing CI tests by effectively mitigating the misjudgment introduced by discretization and accurately recovering the CI relationships of latent continuous variables. We substantiate our approach with theoretical results and empirical validation, underscoring the effectiveness of our testing methods.

REFERENCES

Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(1), 2010.

Armen S Asratian, Tristan MJ Denley, and Roland Häggkvist. *Bipartite graphs and their applications*, volume 131. Cambridge university press, 1998.

Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4): 657–664, 2004.

Laurent Callot, Mehmet Caner, Esra Ulasan, and A. Özlem Önder. A nodewise regression approach to estimating large portfolios, 2019.

Chandler Squires. `causaldag`*: creation, manipulation, and learning of causal models*, 2018. URL `https://github.com/uhlerlab/causaldag`.

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

Aswath Damodaran. *Investment valuation: Tools and techniques for determining the value of any asset*, volume 666. John Wiley & Sons, 2012.

A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(1):1–15, 1979.

Simon Dooms, Toon De Pessemier, and Luc Martens. Movietweetings: a movie rating dataset collected from twitter. In *Workshop on Crowdsourcing and human computation for recommender systems, CrowdRec at RecSys*, volume 2013, pp. 43, 2013.

Dorit Dor and Michael Tarsi. A simple algorithm to construct a consistent extension of a partially oriented graph. 1992. URL `https://api.semanticscholar.org/CorpusID:122949140`.

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

Jianqing Fan, Han Liu, Yang Ning, and Hui Zou. High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(2):405–421, 2017.

Ronald Aylmer Fisher. On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. *Metron*, 1:3–32, 1921.

Karl Pearson F.R.S. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 1*, 50:157–175, 2009. URL `https://api.semanticscholar.org/CorpusID:121472089`.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan): 73–99, 2004.

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*, 10:449461, 2019.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press, 2009. ISBN 9780262013192. URL `https://books.google.co.in/books?id=7dzpHCHzNQ4C`.

Loka Li, Ignavier Ng, Gongxu Luo, Biwei Huang, Guangyi Chen, Tongliang Liu, Bin Gu, and Kun Zhang. Federated causal discovery from heterogeneous data, 2024.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(10), 2009.

Dimitris Margaritis. Distribution-free learning of bayesian network structure in continuous domains. In *AAAI*, volume 5, pp. 825–830, 2005.

Karthik Mohan, Mike Chung, Seungyeop Han, Daniela Witten, Su-In Lee, and Maryam Fazel. Structured learning of gaussian graphical models. *Advances in neural information processing systems*, 25, 2012.

Sarah A Mossman, Marissa J Luft, Heidi K Schroeder, Sara T Varney, David E Fleck, Drew H Barzman, Richard Gilman, Melissa P DelBello, and Jeffrey R Strawn. The generalized anxiety disorder 7-item (gad-7) scale in adolescents with generalized anxiety disorder: signal detection and validation. *Annals of clinical psychiatry: official journal of the American Academy of Clinical Psychiatrists*, 29(4):227, 2017.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000. ISBN 0521773628. URL `http://www.worldcat.org/isbn/0521773628`.

Christine Peterson, Francesco C Stingo, and Marina Vannucci. Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174, 2015.

Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical models. 2015.

Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. *Advances in neural information processing systems*, 30, 2017.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model, 2011.

E Isaac Sparling and Shilad Sen. Rating: how difficult is it? In *Proceedings of the fifth ACM conference on Recommender systems*, pp. 149–156, 2011.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1):20180017, 2019.

Liangjun Su and Halbert White. A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, 24(4):829–864, 2008.

A. W. van der Vaart. *Stochastic Convergence*, pp. 5–24. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998a.

A. W. van der Vaart. *M–and Z-Estimators*, pp. 41–84. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998b. doi: 10.1017/CBO9780511802256.006.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Eric P Xing, Michael I Jordan, Richard M Karp, et al. Feature selection for high-dimensional genomic microarray data. In *Icml*, volume 1, pp. 601–608. Citeseer, 2001.

Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

Yishi Zhang, Zigang Zhang, Kaijun Liu, and Gangyi Qian. An improved iamb algorithm for markov blanket discovery. *J. Comput.*, 5(11):1755–1761, 2010.

Appendix for

"A Conditional Independence Test in the Presence of Discretization"

Appendix organization:

A NOTATION TABLE

Category	Description
Number and Indices	
n	Number of samples
p	Number of variables
j_1, j_2, j, k	Index of a variable $j_1, j_2, j, k \in (1, \ldots, p)$
Random Variables	
$\boldsymbol{X}$	A vector of Gaussian variables
$\tilde{\boldsymbol{X}}$	A vector of variables whose partial variables are discretized versions of $\boldsymbol{X}$
$\boldsymbol{\Sigma}$	Covariance of $\boldsymbol{X}$
$\boldsymbol{\Sigma}_{-j-j}$	Submatrix of $\boldsymbol{\Sigma}$ with j-th row and j-th column removed
$\boldsymbol{\Sigma}_{-jj}$	j-th column of $\mathbf{X}$ with j-th row removed
$\boldsymbol{\Omega}$	Precision matrix of $\boldsymbol{X}$, equals to $\boldsymbol{\Sigma}^{-1}$
X_j	j-th component of the $\boldsymbol{X}$
$\boldsymbol{X}_{-\{jk\}}$	All other variables of $\boldsymbol{X}$ with X_j and X_k removed
$\sigma_{j_1 j_2}$	Covariance between X_{j_1} and X_{j_2}
ω_{jk}	Precision coefficient ω_{jk}
x_j^i	i-th sample of X_j
$\tilde{x}_j^i$	i-th sample of $\tilde{X}_j$
h_j	The boundary in the continuous domain that corresponds to the mean of $\tilde{X}_j$
τ_j	Probability of $\tilde{X}_j$ larger than its mean: $\mathbb{P}(\tilde{X}_j > \mathbb{E}[\tilde{X}_j])$
$\beta_{j,k}$	Regression coefficient of X_k in predicting X_j
β_j	Vector of all coefficients regressing X_j
$\xi^i_{j_1 j_2}$	Influence function component, it represents the influence of the i-th observation on the covariance estimation error
Ξ^i	Matrix form of ξ^i
Estimation of Variables	
$\hat{\sigma}_{j_1 j_2}$	Estimation of $\sigma_{j_1 j_2}$, calculated using equation 6, equation 5
$\hat{\boldsymbol{\Sigma}}$	Estimation of $\boldsymbol{\Sigma}$, matrix form of $\hat{\sigma}_{j_1 j_2}$
$\hat{\omega}_{jk}$	Estimation of ω_{jk}
$\hat{h}_j$	Estimation of h_j, calculated using equation 3
$\hat{\tau}_{j_1}$	Estimation of τ_{j_1}, calculated as $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\tilde{x}_j^i > \mathbb{E}_n(\tilde{X}_j)\}$
$\hat{\beta}_j$	Estimation of $\hat{\beta}_j$, calculated as $\hat{\boldsymbol{\Sigma}}_{-j-j}^{-1}\hat{\boldsymbol{\Sigma}}_{-jj}$
Functions and Operators	
$\mathbb{P}$	True probability
$\mathbb{P}_n$	Sample probability
$\mathbb{E}[Z]$	Expectation of a random variable Z
$\mathbb{E}_n[Z]$	Sample mean of a random variable Z over n samples
$\mathbb{1}$	1 condition: is 1 if the condition is true, 0 otherwise
$\Phi(z)$	Cumulative distribution function of a standard normal distribution
$\bar{\Phi}(z)$	$1 - \Phi(z)$, corresponding to the $\mathbb{P}(Z > z)$
$\bar{\Phi}(z_1, z_2; \rho)$	$\mathbb{P}(Z_1 > z_1, Z_2 > z_2)$, where (Z_1, Z_2) follows a bivariate normal distribution with mean zero, variance one and covariance ρ.
$\psi_{\hat{\theta}}$	A group of functions parametrized by $\hat{\theta}$
$\psi^i_{\hat{\theta}}$	$\psi_{\hat{\theta}}$ evaluated at sample i
$\psi'_{\hat{\theta}}$	Jacobian matrix of $\frac{\partial \psi_{\hat{\theta}}}{\partial \hat{\theta}}$
For Discretized Pair $\tilde{X}_{j_1}, \tilde{X}_{j_2}$	
$\tau_{j_1 j_2}$	Probability of both $\tilde{X}_{j_1}$ and $\tilde{X}_{j_2}$ larger than their mean: $\mathbb{P}(\tilde{X}_{j_1} > \mathbb{E}[\tilde{X}_{j_1}], \tilde{X}_{j_2} > \mathbb{E}[\tilde{X}_{j_2}])$
$\hat{\tau}_{j_1 j_2}$	Estimation of $\tau_{j_1 j_2}$: $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\tilde{x}^i_{j_1} > \mathbb{E}_n[\tilde{X}_{j_1}], \tilde{x}^i_{j_2} > \mathbb{E}_n[\tilde{X}_{j_2}]\}$

Category	Description
$\hat{\tau}^i_{j_1 j_2}$	A sample of $\hat{\tau}_{j_1 j_2}$: $\mathbb{1}\{\tilde{x}^i_{j_1} > \mathbb{E}_n[\tilde{X}_{j_1}], \tilde{x}^i_{j_2} > \mathbb{E}_n[\tilde{X}_{j_2}]\}$
For Mixed Pair $X_{j_1}, \tilde{X}_{j_2}$	
$\tau_{j_1 j_2}$	Probability of both X_{j_1} and $\tilde{X}_{j_2}$ larger than their mean: $\mathbb{P}(X_{j_1} > 0, \tilde{X}_{j_2} > \mathbb{E}[\tilde{X}_{j_2}])$
$\hat{\tau}_{j_1 j_2}$	Estimation of $\tau_{j_1 j_2}$: $\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{x^i_{j_1} > 0, \tilde{x}^i_{j_2} > \mathbb{E}_n[\tilde{X}_{j_2}]\}$
$\hat{\tau}^i_{j_1 j_2}$	A sample of $\hat{\tau}_{j_1 j_2}$: $\mathbb{1}\{\tilde{x}^i_{j_1} > 0, \tilde{x}^i_{j_2} > \mathbb{E}_n[\tilde{X}_{j_2}]\}$

B Proof and Derivations

B.1 Proof of Thm.2.1

If the X_1, X_2 and X_2 are jointly Gaussian and $X_1 \perp\!\!\!\perp X_3|X_3$, we have

$$Cov(X_1, X_3|X_2) = 0. \tag{15}$$

To test if X_1, X_3 are conditional independent given $\tilde{X}_2$, we are interested if $Cov(X_1, X_3|\tilde{X}_2)$ equals zero. Using the law of total covariance, we have

$$Cov(X_1, X_3|\tilde{X}_2) = \mathbb{E}[Cov(X_1, X_3|X_2, \tilde{X}_2)|\tilde{X}_2] + Cov(\mathbb{E}[X_1|X_2, \tilde{X}_2], \mathbb{E}[X_3|X_2, \tilde{X}_2]|\tilde{X}_2). \tag{16}$$

Since $\tilde{X}_2$ is the deterministic function of X_2, $\tilde{X}_2$ will be conditional independent with X_1 and X_3 given X_2. Therefore,

$$Cov(X_1, X_3|X_2, \tilde{X}_2) = Cov(X_1, X_3|X_2) = 0. \tag{17}$$

The first term of equation 16 is zero. We now focus on the second term. Similarly, we have

$$\mathbb{E}[X_1|X_2, \tilde{X}_2] = \mathbb{E}[X_1|X_2], \quad \mathbb{E}[X_3|X_2, \tilde{X}_2] = \mathbb{E}[X_3|X_2], \tag{18}$$

due to the conditional independence. One can see

$$Cov(X_1, X_3|X_2, \tilde{X}_2) = Cov(E[X_1|X_2], \mathbb{E}[X_3|X_2]|\tilde{X}_2). \tag{19}$$

Without loss of generality, we assume the mean of X_1, X_2 and X_3 are zero. Then $\mathbb{E}[X_1|X_2]$ and $\mathbb{E}[X_3|X_2]$ are scaled versions of X_2. The original equation becomes

$$Cov(X_1, X_3|X_2, \tilde{X}_2) = c \cdot Var(X_2|\tilde{X}_2), \tag{20}$$

where c is a constant. We know that

$$Var(X_2|\tilde{X}_2) = \mathbb{E}[(X_2 - \mathbb{E}[X_2|\tilde{X}_2])^2|\tilde{X}_2], \tag{21}$$

which will be zero if and only if X_2 is almost surely a function of $\tilde{X}_2$. That means given $\tilde{X}_2$, the value of X_2 is determined exactly without any randomness, which clearly doesn't hold true in our discretization framework. Thus, $X_1 \not\perp\!\!\!\perp X_3|\tilde{X}_2$, which completes the proof.

B.2 Proof of $\hat{\theta} \xrightarrow{p} \theta_0$

Lemma B.1. *For the estimation $\hat{\theta}$ which is calculated using bridge equation3.1 3.2 and equation 5, as a zero of Ψ_n defined in equation 34,equation 41, equation 44 , will converge in probability to $\theta_0 = (\sigma_{j_1 j_2}, h_{j_1}, h_{j_2}), (\sigma_{j_1 j_2}, h_{j_2}), (\sigma_{j_1 j_2})$ respectively.*

Proof We first focus on the most challenging one where both variables are discrete. According to the law of large numbers, for the estimated boundary $\hat{h}_{j_1}$ and $\hat{h}_{j_2}$ whose calculations are defined as $\hat{h}_j = \Phi^{-1}(1 - \hat{\tau}_j)$, we should have

$$n \to \infty, \quad \hat{\tau}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\tilde{x}_j^i > \mathbb{E}_n[\tilde{X}_j]\} \xrightarrow{p} \mathbb{P}(\tilde{X}_j > \mathbb{E}[\tilde{X}_j]). \tag{22}$$

Recall the definition $\mathbb{P}(\tilde{X}_j > E[\tilde{X}_j]) = 1 - \Phi(h_j)$, according to continuous mapping theorem (Vaart, 1998a), as long as the function $\Phi^{-1}(1 - \cdot)$ is continuous, we should have $\hat{h}_j \xrightarrow{p} h_j$. And thus $\hat{h}_{j_1} \xrightarrow{p} h_{j_1}, \hat{h}_{j_2} \xrightarrow{p} h_{j_2}$.

We have $\hat{\tau}_{j_1,j_2} = \bar{\Phi}(\hat{h}_{j_1}, \hat{h}_{j_2}, \hat{\sigma}_{j_1 j_2})$ and the estimation $\hat{\sigma}_{j_1 j_2}$ can be obtained through solving the function. Similarly, we also have

$$\begin{aligned} n \to \infty, \quad \hat{\tau}_{j_1,j_2} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\tilde{x}_{j_1}^i > \mathbb{E}_n[\tilde{X}_{j_1}]\}\mathbb{1}\{\tilde{x}_{j_2}^i > \mathbb{E}_n[\tilde{X}_{j_2}]\} &\xrightarrow{p} \mathbb{P}(\tilde{x}_{j_1}^i > E[\tilde{X}_{j_1}], \tilde{x}_{j_2}^i > E[\tilde{X}_{j_2}]) \\ &= \tau_{j_1,j_2}. \end{aligned} \tag{23}$$

Similarly, according to the continuous mapping theorem, we have $\hat{\sigma}_{j_1j_2} \xrightarrow{p} \sigma_{j_1j_2}$. Thus, the parameter $(\hat{\sigma}_{j_1j_2}, \hat{h}_{j_1}, \hat{h}_{j_2}) \xrightarrow{p} (\sigma_{j_1j_2}, h_{j_1}, h_{j_2})$.

Apparently, the result above could easily extend to the mixed case where we fix $\hat{h}_1 = h_1 = 0$. Using the same procedure, we should have $(\hat{\sigma}_{j_1j_2}, \hat{h}_{j_2}) \xrightarrow{p} (\sigma_{j_1j_2}, h_{j_2})$.

For the continuous case whose estimated variance is calculated as $\hat{\sigma}_{j_1j_2} = \frac{1}{n}\sum_{i=1}^n x_{j_1}^i x_{j_2}^i - \frac{1}{n}\sum_{i=1}^n x_{j_1}^i \frac{1}{n}\sum_{i=1}^n x_{j_2}^i$., according to law of large numbers, we should have

$$n \to \infty, \quad \hat{\sigma}_{j_1j_2} = \frac{1}{n}\sum_{i=1}^n x_{j_1}^i x_{j_2}^i - \frac{1}{n}\sum_{i=1}^n x_{j_1}^i \frac{1}{n}\sum_{i=1}^n x_{j_2}^i \xrightarrow{p} \mathbb{E}(X_{j_1}X_{j_2}) - \mathbb{E}(X_{j_1})\mathbb{E}(X_{j_2}) = \sigma_{j_1j_2}. \tag{24}$$

B.3 Proof of one-to-one mapping between $\hat{\tau}_{j_1,j_2}$ with $\hat{\sigma}_{j_1j_2}$

Lemma B.2. *For any fixed $\hat{h}_{j_1}$ and $\hat{h}_{j_2}$, $T(\sigma'_{j_1j_2}; \{\hat{h}_{j_1}, \hat{h}_{j_2}\}) = \int_{x_1>\hat{h}_{j_1}}\int_{x_2>\hat{h}_{j_2}} \phi(x_{j_1}, x_{j_2}; \sigma)dx_{j_1}dx_{j_2}$, is a strictly monotonically increasing function on $\sigma \in (-1, 1)$.*

Proof To prove the lemma, we just need to show the gradient $\frac{\partial T(\sigma'_{j_1j_2}; \{\hat{h}_{j_1}, \hat{h}_{j_2}\}}{\partial \sigma} > 0$ for $\sigma'_{j_1j_2} \in (-1, 1)$.

$$\frac{\partial T(\sigma_{j_1j_2}); \{\hat{h}_{j_1}, \hat{h}_{j_2}\}}{\partial \sigma'_{j_1j_2}} == \frac{1}{2\pi\sqrt{(1-\sigma'^2_{j_1j_2})}} \exp\left(-\frac{(\hat{h}_{j_1}^2 - 2\sigma'_{j_1j_2}\hat{h}_{j_1}\hat{h}_{j_2} + \hat{h}_{j_2}^2)}{2(1-\sigma'^2_{j_1j_2})}\right), \tag{25}$$

which is obviously positive for $\sigma'_{j_1j_2} \in (-1, 1)$. Thus, we have one-to-one mapping between $\hat{\tau}_{j_1j_2}$ with the calculated $\hat{\sigma}_{j_1j_2}$ for fixed $\hat{h}_{j_1}$ and $\hat{h}_{j_2}$.

B.4 Proof of Thm. 3.3

In this section, we provide the proof of Thm. 3.3, which utilizes a regular statistical tool: Z-estimator (Vaart, 1998b). Specifically, we are interested in the parameter θ and we have it estimation $\hat{\theta}$. Let $\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}$ are sampled from some distribution, we can construct the function characterized by the parameter θ related the $\boldsymbol{x}$ as $\psi_\theta(\boldsymbol{x})$. As long as we have n observations, we can construct the function as follows

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n \psi_\theta(\boldsymbol{x}_i) = \mathbb{E}_n[\psi_\theta]. \tag{26}$$

We further specify the form

$$\Psi(\theta) = \int \psi_\theta(\boldsymbol{x})d\boldsymbol{x} = \mathbb{E}[\psi_\theta]. \tag{27}$$

Assume the estimator $\hat{\theta}$ is a zero of Ψ_n, i.e., $\Psi_n(\hat{\theta}) = 0$ and will converge in probability to θ_0, which is a zero of Ψ, i.e., $\Psi(\theta_0) = 0$. Expand $\Psi_n(\hat{\theta})$ in a Taylor series around θ_0, we should have

$$0 = \Psi_n(\hat{\theta}) = \Psi_n(\theta_0) + (\hat{\theta} - \theta_0)\Psi'_n(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^2\Psi''_n(\theta_0). \tag{28}$$

Rearrange the equation above, we have

$$\begin{aligned}\hat{\theta} - \theta_0 &= -\frac{\Psi_n(\theta_0)}{\Psi'_n(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^2\Psi''_n(\theta_0)} \\ &= -\frac{\frac{1}{n}\sum_{i=1}^n \psi_\theta(\boldsymbol{x}_i)}{\Psi'_n(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^2\Psi''_n(\theta_0)}.\end{aligned} \tag{29}$$

According to the central limit theorem, the numerator will be asymptotic normal with variance $\mathbb{E}[\psi_{\theta_0}^2]/n$ as the mean $\Psi(\theta_0) = 0$ is zero. The first term of denominator $\Psi_n'(\theta_0)$ will converge in probability to $\Psi'(\theta_0)$ according to the law of large numbers. The second term $\hat{\theta} - \theta_0 = o_P(1)$. [2] As long as the denominator converges in probability and the numerator converges in distribution, according to Slusky's lemma, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow N\left(0, \frac{\mathbb{E}[\psi_{\theta_0}^2]}{\mathbb{E}[\psi_{\theta_0}']^2}\right). \tag{30}$$

Extend into the high-dimensional case we should have

$$\hat{\theta} - \theta_0 = -\Psi_n'(\theta_0)^{-1}\Psi_n(\theta_0) \tag{31}$$

where the second order term is omitted, further assume the matrix $\mathbb{E}[\psi_{\theta_0}']$ is invertible, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow N\left(0, (\mathbb{E}[\psi_{\theta_0}'])^{-1}\mathbb{E}[\psi_{\theta_0}\psi_{\theta_0}^T](\mathbb{E}[\psi_{\theta_0}'^T])^{-1}\right), \tag{32}$$

Specifically, in our case $\theta_0 = (\sigma_{j_1j_2}, \mathbf{\Lambda})$, where $\mathbf{\Lambda}$ is another parameter set influencing the estimation of $\sigma_{j_1j_2}$ (will discuss case in case in later proof). In the practical scenario, we only have access to the estimated parameter $\hat{\theta}$ and the empirical distribution $\mathbb{P}_n$, thus we have

$$\hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2} \overset{\text{approx}}{\sim} N\left(0, ((\mathbb{E}_n[\psi_{\hat{\theta}}'])^{-1}\mathbb{E}_n[\psi_{\hat{\theta}}\psi_{\hat{\theta}}^T](\mathbb{E}_n[\psi_{\hat{\theta}}'^T])^{-1})_{1,1}\right). \tag{33}$$

Under the null hypothesis of independent, $\sigma_{j_1j_2} = 0$. We provide the proof that $\hat{\theta} \xrightarrow{p} \theta_0$ of our case in App. B.2. Thus, $\mathbb{E}_n[\psi_{\hat{\theta}}]$, the function parameterized by $\hat{\theta}$, should also converge in $\mathbb{E}_n[\psi_{\theta_0}]$ when $n \to \infty$. Besides, by the law of large numbers, $\mathbb{E}_n[\psi_{\hat{\theta}_0}]$ will converge to $E[\psi_{\hat{\theta}_0}]$. Thus, the equation above will converge to equation 32 when $n \to \infty$.

B.5 Derivation of Lem. 3.5

Let's first focus on the most challenging case where both variables are discretized observations and our interested parameter will include $\hat{\theta} = (\hat{\sigma}_{j_1j_2}, \hat{h}_{j_1}, \hat{h}_{j_2})$ (Although we only care about the distribution of $\hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}$, the estimation of boundary $\hat{h}_{j_1}$ and $\hat{h}_{j_2}$ will influence the estimation of $\hat{\sigma}_{j_1j_2}$, thus we need to consider all of them).

The next step will be to *construct an appropriate criterion function ψ such that $\Psi_n(\hat{\theta}) = \mathbf{0}$*. Given n observations $\{\tilde{\boldsymbol{x}}^1, \tilde{\boldsymbol{x}}^2, \dots, \tilde{\boldsymbol{x}}^n\}$, which are discretized version of $\{\boldsymbol{x}^1, \boldsymbol{x}^2, \dots, \boldsymbol{x}^n\}$ we should have

$$\Psi_n(\hat{\theta}) = \begin{pmatrix} \Psi_n(\hat{\sigma}_{j_1j_2}) \\ \Psi_n(\hat{h}_{j_1}) \\ \Psi_n(\hat{h}_{j_2}) \end{pmatrix} = \frac{1}{n}\sum_{i=1}^n \psi_{\hat{\theta}}(\tilde{\boldsymbol{x}}^i) = \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} \hat{\tau}_{j_1,j_2}^i - T(\hat{\sigma}_{j_1j_2}; \{\hat{h}_{j_1}, \hat{h}_{j_2}\}) \\ \hat{\tau}_{j_1}^i - \bar{\Phi}(\hat{h}_{j_1}) \\ \hat{\tau}_{j_2}^i - \bar{\Phi}(\hat{h}_{j_2}) \end{pmatrix} = \mathbf{0}. \tag{34}$$

$$\Psi_n(\theta_0) = \begin{pmatrix} \Psi_n(\sigma_{j_1j_2}) \\ \Psi_n(h_{j_1}) \\ \Psi_n(h_{j_2}) \end{pmatrix} = \frac{1}{n}\sum_{i=1}^n \psi_{\theta_0}(\tilde{\boldsymbol{x}}^i) = \frac{1}{n}\sum_{i=1}^n \begin{pmatrix} \hat{\tau}_{j_1,j_2}^i - T(\sigma_{j_1j_2}; \{h_{j_1}, h_{j_2}\}) \\ \hat{\tau}_{j_1}^i - \bar{\Phi}(h_{j_1}) \\ \hat{\tau}_{j_2}^i - \bar{\Phi}(h_{j_2}) \end{pmatrix}. \tag{35}$$

The difference between the estimated parameter with the true parameter can be expressed as

$$\hat{\theta} - \theta_0 = \begin{pmatrix} \hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2} \\ \hat{h}_{j_1} - h_{j_1} \\ \hat{h}_{j_2} - h_{j_2} \end{pmatrix} = -\frac{1}{n}\sum_{i=1}^n \begin{pmatrix} \frac{\partial\Psi_n(\sigma_{j_1j_2})}{\partial\sigma_{j_1j_2}} & \frac{\partial\Psi_n(\sigma_{j_1j_2})}{\partial h_{j_1}} & \frac{\partial\Psi_n(\sigma_{j_1j_2})}{\partial h_{j_2}} \\ \frac{\partial\Psi_n(h_{j_1})}{\partial\sigma_{j_1j_2}} & \frac{\partial\Psi_n(h_{j_1})}{\partial h_{j_1}} & \frac{\partial\Psi_n(h_{j_1})}{\partial h_{j_2}} \\ \frac{\partial\Psi_n(h_{j_2})}{\partial\sigma_{j_1j_2}} & \frac{\partial\Psi_n(h_{j_2})}{\partial h_{j_1}} & \frac{\partial\Psi_n(h_{j_2})}{\partial h_{j_2}} \end{pmatrix}^{-1} \cdot \begin{pmatrix} \hat{\tau}_{j_1,j_2}^i - T(\sigma_{j_1j_2}; \{h_{j_1}, h_{j_2}\}) \\ \hat{\tau}_{j_1}^i - \bar{\Phi}(h_{j_1}) \\ \hat{\tau}_{j_2}^i - \bar{\Phi}(h_{j_2}) \end{pmatrix}, \tag{36}$$

[2]We will not provide proof of this in this paper; however, interested readers may refer to (Vaart, 1998b)

where the specific form of each entry of the gradient matrix is expressed as

$$\begin{aligned}
\frac{\partial \Psi_n(\sigma_{j_1 j_2})}{\partial \sigma_{j_1 j_2}} &= -\frac{1}{2\pi\sqrt{(1-\sigma_{j_1,j_2}^2)}}\exp\left(-\frac{(h_{j_1}^2 - 2\sigma_{j_1,j_2}h_{j_1}h_{j_2} + h_{j_2}^2)}{2(1-\sigma_{j_1,j_2}^2)}\right);\\
\frac{\partial \Psi_n(\sigma_{j_1 j_2})}{\partial h_{j_1}} &= \int_{h_{j_2}}^{\infty} \frac{1}{2\pi\sqrt{1-\sigma_{j_1 j_2}{}^2}} \exp\left(-\frac{h_{j_1}^2 - 2\sigma_{j_1 j_2}h_{j_1}x_2 + x_2^2}{2(1-\sigma_{j_1,j_2}^2)}\right) dx_2;\\
\frac{\partial \Psi_n(\sigma_{j_1 j_2})}{\partial h_{j_2}} &= \int_{h_{j_1}}^{\infty} \frac{1}{2\pi\sqrt{1-\sigma_{j_1 j_2}{}^2}} \exp\left(-\frac{h_{j_2}^2 - 2\sigma_{j_1 j_2}h_{j_2}x_1 + x_1^2}{2(1-\sigma_{j_1 j_2}{}^2)}\right) dx_1;\\
\frac{\partial \Psi_n(h_{j_1})}{\partial \sigma_{j_1 j_2}} &= 0;\\
\frac{\partial \Psi_n(h_{j_1})}{\partial h_{j_1}} &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{h_{j_1}^2}{2}\right);\\
\frac{\partial \Psi_n(h_{j_1})}{\partial h_{j_2}} &= 0;\\
\frac{\partial \Psi_n(h_{j_2})}{\partial \sigma_{j_1 j_2}} &= 0;\\
\frac{\partial \Psi_n(h_{j_2})}{\partial h_{j_1}} &= 0;\\
\frac{\partial \Psi_n(h_{j_2})}{\partial h_{j_2}} &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{h_{j_2}^2}{2}\right).
\end{aligned} \tag{37}$$

For simplicity of notation, we define

$$\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2} = \frac{1}{n}\sum_{i=1}^{n} \xi_{j_1,j_2}^i, \tag{38}$$

where the specific form is of $\{\xi_{j_1,j_2}^i\}$ is defined in equation 36. We should note that $\{\xi_{j_1,j_2}^i\}$ are i.i.d random variables with mean zero (this property will be the key to the derivation of inference of CI). As long as our estimation $\hat{\theta}$ converge in probability to θ_0 as proved in B.2, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow N\left(0, ((\mathbb{E}[\psi'_{\theta_0}])^{-1}\mathbb{E}[\psi_{\theta_0}\psi_{\theta_0}^T](\mathbb{E}[\psi'^T_{\theta_0}])^{-1})\right), \tag{39}$$

where ψ_{θ_0} is defined in equation 35. However, in practice, we don't have access to either θ_0 or the true expectation. In this scenario, we can plug in the empirical distribution of $\mathbb{P}_n\psi_{\hat{\theta}}$ to get the estimated variance, i.e., the actual variance used in the calculation of $\hat{\sigma}_{j_1 j_2} - \sigma_{j_1 j_2}$ is

$$\frac{1}{n}\left((\mathbb{E}_n[\psi'_{\hat{\theta}}])^{-1}\mathbb{E}_n[\psi_{\hat{\theta}}\psi_{\hat{\theta}}^T](\mathbb{E}_n[\psi'^T_{\hat{\theta}}])^{-1}\right)_{1,1}. \tag{40}$$

B.6 Derivation of Lem. 3.6

Use the same procedure as in the derivation of Lem. 3.5, for mixed pair of observations where X_{j_1} is continuous and $\tilde{X}_{j_2}$ is discrete, we can construct the criterion function

$$\Psi_n(\hat{\theta}) = \begin{pmatrix}\Psi_n(\hat{\sigma}_{j_1 j_2})\\ \Psi_n(\hat{h}_{j_2})\end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n}\psi_{\hat{\theta}}(\tilde{\boldsymbol{x}}^i) = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\hat{\tau}_{j_1,j_2}^i - T(\hat{\sigma}_{j_1 j_2}; \{0, \hat{h}_{j_2}\})\\ \hat{\tau}_{j_2}^i - \bar{\Phi}(\hat{h}_{j_2})\end{pmatrix} = \mathbf{0}. \tag{41}$$

$$\Psi_n(\theta_0) = \begin{pmatrix}\Psi_n(\sigma_{j_1 j_2})\\ \Psi_n(h_{j_2})\end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n}\psi_{\theta_0}(\tilde{\boldsymbol{x}}^i) = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\hat{\tau}_{j_1,j_2}^i - T(\sigma_{j_1 j_2}; \{0, h_{j_2}\})\\ \hat{\tau}_{j_2}^i - \bar{\Phi}(h_{j_2})\end{pmatrix}. \tag{42}$$

The difference between the estimated parameter with the true parameter can be expressed as

$$\hat{\theta}-\theta_0 = \begin{pmatrix} \hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2} \\ \hat{h}_{j_2} - h_{j_2} \end{pmatrix} = -\frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \frac{\partial \Psi_n(\sigma_{j_1j_2})}{\partial \sigma_{j_1j_2}} & \frac{\partial \Psi_n(\sigma_{j_1j_2})}{\partial h_{j_2}} \\ \frac{\partial \Psi_n(h_{j_2})}{\partial \sigma_{j_1j_2}} & \frac{\partial \Psi_n(h_{j_2})}{\partial h_{j_2}} \end{pmatrix}^{-1} \begin{pmatrix} \hat{\tau}^i_{j_1,j_2} - T(\sigma_{j_1j_2}; \{0, h_{j_2}\}) \\ \hat{\tau}^i_{j_2} - \bar{\Phi}(h_{j_2}). \end{pmatrix}, \tag{43}$$

where the specific form of each entry of the gradient matrix can be found in equation 37. Using exactly the same procedure, we should have the same formation of the variance calculated as equation 40 with a different definition of ψ_{θ_0} and $\psi_{\hat{\theta}}$ defined in equation 42 equation 41.

B.7 Derivation of Lem. 3.4

Use the same line of procedure as in the derivation of Lem. 3.5, for a continuous pair of variables, we can construct the criterion function

$$\Psi_n(\hat{\theta}) = \Psi_n(\hat{\sigma}_{j_1j_2}) = \frac{1}{n}\sum_{i=1}^{n} x^i_{j_1}x^i_{j_2} - \frac{1}{n}\sum_{i=1}^{n} x^i_{j_1}\frac{1}{n}\sum_{i=1}^{n} x^i_{j_2} - \hat{\sigma}_{j_1j_2} = 0. \tag{44}$$

$$\Psi_n(\theta_0) = \Psi_n(\sigma_{j_1j_2}) = \frac{1}{n}\sum_{i=1}^{n} x^i_{j_1}x^i_{j_2} - \frac{1}{n}\sum_{i=1}^{n} x^i_{j_1}\frac{1}{n}\sum_{i=1}^{n} x^i_{j_2} - \sigma_{j_1j_2}. \tag{45}$$

Denote $\frac{1}{n}\sum_{i=1}^{n} x^i_{j_1}$ as $\bar{x}_{j_1}$ and $\frac{1}{n}\sum_{i=1}^{n} x^i_{j_2}$ as $\bar{x}_{j_2}$. We should have

$$\hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2} = \frac{1}{n}\sum_{i=1}^{n} x^i_{j_1}x^i_{j_2} - \bar{x}_{j_1}\bar{x}_{j_2} - \sigma_{j_1j_2}. \tag{46}$$

According to equation 30, we have

$$\sqrt{n}(\hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}) \rightsquigarrow N\left(0, \frac{\mathbb{E}[\psi^2_{\theta_0}]}{(\mathbb{E}[\psi'_{\theta_0}])^2}\right). \tag{47}$$

where $(E[\psi'_{\theta_0}])^2 = 1$. In practical calculation, we have the variance

$$\frac{1}{n}\mathbb{E}_n[\psi^2_{\hat{\theta}}]/(\mathbb{E}_n[\psi'_{\hat{\theta}}])^2 = \frac{1}{n^2}\sum_{i=1}^{n}(x^i_{j_1}x^i_{j_2} - \bar{x}_{j_1}\bar{x}_{j_2} - \hat{\sigma}_{j_1j_2})^2. \tag{48}$$

B.8 Proof of Thm. 3.8

B.8.1 Proof of Lem. 3.7

Consider our latent continuous variables $\boldsymbol{X} = (X_1, \ldots, X_p) \sim N(0, \boldsymbol{\Sigma})$ and do nodewise regression

$$X_j = \boldsymbol{X}_{-j}\beta_j + \epsilon_j, \tag{49}$$

where $\boldsymbol{X}_{-j}$ is the submatrix of $\boldsymbol{X}$ with X_j removed. We can divide its covariance $\boldsymbol{\Sigma}$ and its precision matrix $\Omega = \boldsymbol{\Sigma}^{-1}$ into the predictor $\mathbf{X}_{-j}$ and outcome variable X_j in our regression:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{jj} & \boldsymbol{\Sigma}_{j-j} \\ \boldsymbol{\Sigma}_{-jj} & \boldsymbol{\Sigma}_{-j-j} \end{pmatrix} \quad \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{jj} & \boldsymbol{\Omega}_{j-j} \\ \boldsymbol{\Omega}_{-jj} & \boldsymbol{\Omega}_{-j-j} \end{pmatrix}. \tag{50}$$

Just like regular linear regression, we can get

$$n \to \infty, \quad \beta_j = \boldsymbol{\Sigma}^{-1}_{-j-j}\boldsymbol{\Sigma}_{-jj}. \tag{51}$$

From the invertibility of a block matrix

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A-BD^{-1}C)^{-1} & -(A-BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A-BD^{-1}C)^{-1} & D^{-1}+D^{-1}C(A-BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}. \tag{52}$$

If A and D is invertible, we will have

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A-BD^{-1}C)^{-1} & 0 \\ 0 & (D-CA^{-1}B)^{-1} \end{bmatrix} \begin{bmatrix} I & -BD^{-1} \\ -CA^{-1} & I \end{bmatrix}. \tag{53}$$

Thus, we can get:

$$\begin{aligned} \mathbf{\Omega}_{jj} &= (\mathbf{\Sigma}_{jj} - \mathbf{\Sigma}_{j-j}\mathbf{\Sigma}_{-j-j}^{-1}\mathbf{\Sigma}_{-jj})^{-1}; \\ \mathbf{\Omega}_{j-j} &= -\left(\mathbf{\Sigma}_{jj} - \mathbf{\Sigma}_{j-j}\mathbf{\Sigma}_{-j-j}^{-1}\mathbf{\Sigma}_{-jj}\right)^{-1}\mathbf{\Sigma}_{j-j}(\mathbf{\Sigma}_{-j-j})^{-1}. \end{aligned} \tag{54}$$

Move one step forward:

$$-\mathbf{\Omega}_{jj}^{-1}\mathbf{\Omega}_{j-j} = \mathbf{\Sigma}_{j-j}(\mathbf{\Sigma}_{-j-j})^{-1}. \tag{55}$$

Take transpose for both sides, as long as $\mathbf{\Omega}$ is a symmetric matrix and $\mathbf{\Omega}_{-jj} = \mathbf{\Omega}_{j-j}^T$, we will have

$$-\mathbf{\Omega}_{jj}^{-1}\mathbf{\Omega}_{-jj} = \mathbf{\Sigma}_{-j-j}^{-1}\mathbf{\Sigma}_{-jj} = \beta_j. \tag{56}$$

We should note testing $\mathbf{\Omega}_{-jj} = 0$ is equivalent to testing $\beta_j = 0$ as the $\mathbf{\Omega}_{jj}$ will always be nonzero. The variable $\mathbf{\Omega}_{-jj}$ captures the CI of X_j with other variables. As long as the variable $\mathbf{\Omega}_{jj}$ is just one scalar, we can get

$$\beta_{j,k} = -\frac{\omega_{jk}}{\omega_{jj}} \tag{57}$$

capturing the CI relationship between variable X_j with X_k conditioning on all other variables.

B.8.2 Detailed derivation of inference for β_j

Nodewise regression allows us to use the regression parameter β_j as the surrogate of Ω_{-jj}. The problem now transfers to constructing the inference for β_j, specifically, the derivation of distribution of $\hat{\beta}_j - \beta_j$. The overarching concept is that we are already aware of the distribution of $\hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}$ and we know that there exists a deterministic relationship between β_j with $\mathbf{\Sigma}$. Consequently, we can express $\hat{\beta}_j - \beta_j$ as a composite of $\hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}$ to establish such an inference. Specifically, we have

$$\begin{aligned} \hat{\beta}_j - \beta_j &= \hat{\mathbf{\Sigma}}_{-j-j}^{-1}\hat{\mathbf{\Sigma}}_{-jj} - \mathbf{\Sigma}_{-j-j}^{-1}\mathbf{\Sigma}_{-jj} \\ &= \hat{\mathbf{\Sigma}}_{-j-j}^{-1}\left(\hat{\mathbf{\Sigma}}_{-jj} - \hat{\mathbf{\Sigma}}_{-j-j}\mathbf{\Sigma}_{-j-j}^{-1}\mathbf{\Sigma}_{-jj}\right) \\ &= -\hat{\mathbf{\Sigma}}_{-j-j}^{-1}\left(\hat{\mathbf{\Sigma}}_{-j-j}\beta_j - \mathbf{\Sigma}_{-j-j}\beta_j + \mathbf{\Sigma}_{-j-j}\beta_j - \hat{\mathbf{\Sigma}}_{-jj}\right) \\ &= -\hat{\mathbf{\Sigma}}_{-j-j}^{-1}\left((\hat{\mathbf{\Sigma}}_{-j-j} - \mathbf{\Sigma}_{-j-j})\beta_j - (\hat{\mathbf{\Sigma}}_{-jj} - \mathbf{\Sigma}_{-jj})\right), \end{aligned} \tag{58}$$

where each entry in matrix $(\hat{\mathbf{\Sigma}}_{-j-j} - \mathbf{\Sigma}_{-j-j})$ and $(\hat{\mathbf{\Sigma}}_{-jj} - \mathbf{\Sigma}_{-jj})$ denotes the difference between estimated covariance with true covariance.

Suppose that we want to test the CI of the variable X_1 with other variables, $j = 1$. then

$$\hat{\mathbf{\Sigma}}_{-1-1} - \mathbf{\Sigma}_{-1-1} = \begin{bmatrix} \hat{\sigma}_{2,2} \dots \hat{\sigma}_{2,p} \\ \dots \\ \hat{\sigma}_{p,2} \dots \hat{\sigma}_{p,p} \end{bmatrix} - \begin{bmatrix} \sigma_{2,2} \dots \sigma_{2,p} \\ \dots \\ \sigma_{p,2} \dots \sigma_{p,p} \end{bmatrix} \tag{59}$$

$$:= \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} \xi_{2,2}^i \dots \xi_{2,p}^i \\ \dots \\ \xi_{p,2}^i \dots \xi_{p,p}^i \end{bmatrix}, \tag{60}$$

where $\{\xi^i_{j_1,j_2}\}$ are i.i.d random variables with specific form defined in equation 36 for discrete case, equation 43 for mixed case and equation 46 in continuous case. Put them together:

$$\hat{\beta}_1 - \beta_1 = \begin{bmatrix} \hat{\beta}_{1,2} - \beta_{1,2} \\ \hat{\beta}_{1,3} - \beta_{1,3} \\ \cdots \\ \hat{\beta}_{1,p} - \beta_{1,p} \end{bmatrix} = -\hat{\mathbf{\Sigma}}^{-1}_{-1-1} \frac{1}{n} \sum_{i=1}^{n} \left(\begin{bmatrix} \xi^i_{2,2} & \xi^i_{2,3} & \cdots & \xi^i_{2,p} \\ \xi^i_{3,2} & \xi^i_{3,3} & \cdots & \xi^i_{3,p} \\ \cdots & \cdots & \cdots & \cdots \\ \xi^i_{p,2} & \xi^i_{p,3} & \cdots & \xi^i_{p,p} \end{bmatrix} \begin{bmatrix} \beta_{1,2} \\ \beta_{1,3} \\ \cdots \\ \beta_{1,p} \end{bmatrix} - \begin{bmatrix} \xi^i_{2,1} \\ \xi^i_{3,1} \\ \cdots \\ \xi^i_{p,1} \end{bmatrix} \right). \tag{61}$$

As $\frac{1}{n}\sum_{i=1}^n \xi^i_{j_1,j_2}$ is asymptotically normal, the who vector of $\hat{\beta}_1 - \beta_1$ is a linear combination of Gaussian distribution. However, We cannot merely engage in a linear combination of its variance as they are dependent with each other. For example, if Y_1, Y_2 are dependent and we are trying to find out $Var(aY_1 + bY_2)$, we should have

$$Var(aY_1 + bY_2) = [a \quad b] \begin{bmatrix} Var(Y_1) & Cov(Y_1, Y_2) \\ Cov(Y_1, Y_2) & Var(Y_2) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}. \tag{62}$$

Now, suppose we are interested in the distribution of $\hat{\beta}_{1,2} - \beta_{1,2}$, we have

$$\hat{\beta}_{1,2} - \beta_{1,2} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\mathbf{\Sigma}}^{-1}_{-1-1})_{[2],:} \left(\begin{bmatrix} \xi^i_{2,2} & \xi^i_{2,3} & \cdots & \xi^i_{2,p} \\ \xi^i_{3,2} & \xi^i_{3,3} & \cdots & \xi^i_{3,p} \\ \cdots & \cdots & \cdots & \cdots \\ \xi^i_{p,2} & \xi^i_{p,3} & \cdots & \xi^i_{p,p} \end{bmatrix} \begin{bmatrix} \beta_{1,2} \\ \beta_{1,3} \\ \cdots \\ \beta_{1,p} \end{bmatrix} - \begin{bmatrix} \xi^i_{2,1} \\ \xi^i_{3,1} \\ \cdots \\ \xi^i_{p,1} \end{bmatrix} \right), \tag{63}$$

where $(\hat{\mathbf{\Sigma}}^{-1}_{-1-1})_{[2],:}$ is the row of index of X_2 of $\hat{\mathbf{\Sigma}}^{-1}_{-1-1}$ ([2] denotes the index of the variable). For ease of notation, let

$$\Xi^i_{-1,-1} = \begin{bmatrix} \xi^i_{2,2} & \xi^i_{2,3} & \cdots & \xi^i_{2,p} \\ \xi^i_{3,2} & \xi^i_{3,3} & \cdots & \xi^i_{3,p} \\ \cdots & \cdots & \cdots & \cdots \\ \xi^i_{p,2} & \xi^i_{p,3} & \cdots & \xi^i_{p,p} \end{bmatrix}, \qquad \Xi^i_{-1,1} = \begin{bmatrix} \xi^i_{2,1} \\ \xi^i_{3,1} \\ \cdots \\ \xi^i_{p,1} \end{bmatrix}, \tag{64}$$

and let

$$B^i_{-1} = \begin{pmatrix} \xi^i_{2,1} & \xi^i_{3,1} & \cdots & \xi^i_{p,1} \\ \xi^i_{2,2} & \xi^i_{2,3} & \cdots & \xi^i_{2,p} \\ \xi^i_{3,2} & \xi^i_{3,3} & \cdots & \xi^i_{3,p} \\ \cdots & \cdots & \cdots & \cdots \\ \xi^i_{p,2} & \xi^i_{p,3} & \cdots & \xi^i_{p,p} \end{pmatrix} = \begin{bmatrix} {\Xi^i_{-j,j}}^T \\ \Xi^i_{-j,-j} \end{bmatrix} \tag{65}$$

as the concatenation of those two matrices. Similarly as equation 62, The variance is calculated as

$$Var\left(\sqrt{n}(\hat{\beta}_{1,2} - \beta_{1,2})\right) = {a^{[2]}}^T \frac{1}{n} \sum_{i=1}^{n} vec(B^i_{-1}) vec(B^i_{-1})^T a^{[2]}, \tag{66}$$

where

$$a^{[2]}_l = \begin{cases} \left(\hat{\mathbf{\Sigma}}^{-1}_{-1-1}\right)_{[2],l}, & \text{for } l \in \{1, \ldots, p-1\} \\ \sum_{q=1}^{p-1} \left(\hat{\mathbf{\Sigma}}^{-1}_{-1-1}\right)_{[2],l} (\beta_1)_q, & \text{for } l \in \{p, \ldots, p^2 - p\} \end{cases} \tag{67}$$

$vec(B^i_{-1})$ is the row-wise vectorization of matrix B^i_{-1}, i.e.,

$$vec(B^i_{-1}) = \begin{pmatrix} \xi^i_{2,1} \\ \xi^i_{3,1} \\ \vdots \\ \xi^i_{p,p} \end{pmatrix} \in \mathbb{R}^{p^2-p}. \tag{68}$$

Thus, the distribution of $\hat{\beta}_{j,k} - \beta_{j,k}$ is

$$\hat{\beta}_{j,k} - \beta_{j,k} \sim N(0, {a^{[k]}}^T \frac{1}{n^2} \sum_{i=1}^{n} vec(B^i_{-j}) vec(B^i_{-j})^T) a^{[k]}). \tag{69}$$

In practice, we can plug in the estimates of β_j to estimate the interested distribution and do the CI test by hypothesizing $\beta_{j,k} = 0$.

B.9 Discussion of assumption

In this section, we first justify why the assumption of zero mean and identity variance can be made without loss of generality. Then, we explain the rationale behind the linear Gaussian assumption.

B.9.1 zero mean and identity variance

In this section, we engage in a more thorough discussion regarding our assumptions about $\boldsymbol{X}$. Specifically, we demonstrate that this assumption of mean and variance does not compromise the generality. In other words, the true model may possess different mean and variance values, but we proceed by treating it as having a mean of zero and identity variance.

The key ingredient allowing us to assume such a model is, the discretization function g_j is an unknown nonlinear monotonic function. Suppose the g'_j maps the continuous domain to a binary variable, and we have the "groundtruth" variable, denoted X'_j, with mean a and variance b. Assume the cardinality of the discretized domain is only 2, i.e., our observation $\tilde{X}_j$ can only be 0 or 1. We further have the constant d'_j as the discretization boundary such that we have the observation

$$\tilde{X}_j = \mathbb{1}(g'_j(X'_j) > d'_j) = \mathbb{1}(X'_j > g'^{-1}_j(d_j))$$

We can always produce our assumed variable X_j with mean 0 and variance 1, such that $X_j = \frac{1}{\sqrt{b}}X'_j - \frac{a}{\sqrt{b}}$ and the same observation with a different nonlinear transformation g_j and decision boundary d_j, such that

$$\tilde{X}_j = \mathbb{1}(g_j(X_j) > d_j) = \mathbf{1}(X_j > g_j^{-1}(d_j)) = \mathbb{1}(X'_j > \sqrt{b}g_j^{-1}(d_j) + a)$$

As long as the observation $\tilde{X}_j$ is the same, we should have $\sqrt{b}g_j^{-1}(d_j) + a = g'^{-1}_j(d_j)$. Our assumed model X_j clearly mimics the "groundtruth" X'_j. Besides, according to Lem. B.3, we have one-to-one mapping between $\hat{\tau}_{j_1j_2}$ with the estimated covariance for fixed $\hat{h}_{j_1}, \hat{h}_{j_2}$. Thus, as long as the observation is the same, the estimation of covariance $\hat{\sigma}_{j_1j_2}$ remains unaffected by our assumptions regarding the mean and variance of $\boldsymbol{X}$, so do the following inference.

We further conduct casual discovery experiments to empirically validate our statement, which is shown in App. E.3.

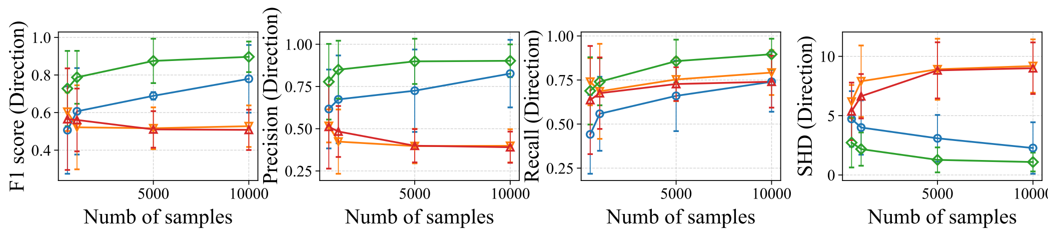
B.9.2 Discussion of Linear Gaussian Assumption

Discretization of continuous variables inevitably leads to information loss in the original data. Compared to the original distributional information, the recovered covariance matrix is naturally less accurate. Given this, constructing a valid statistical inference procedure, rather than solely relying on estimated covariance values for drawing conditional independence conclusions, is desirable.

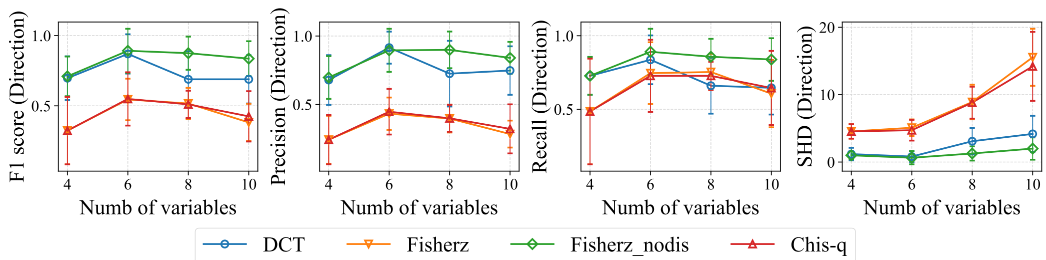
One major limitation of DCT is its reliance on the assumption that latent continuous variables follow a multivariate normal distribution. Violations of this assumption can lead to erroneous conclusions. For instance, consider a scenario where the relationship between latent variables is nonlinear, such as $X_{j_1} = X_{j_2}^2$. In this case, the covariance $\sigma_{j_1j_2}$ equals zero despite a deterministic dependency between X_{j_1} and X_{j_2}. Consequently, even if the correlation is perfectly estimated, the model fails to capture the true underlying relationship, leading to incorrect inferences.

Nevertheless, although the theoretical framework of DCT requires latent continuous variables to follow a multivariate Gaussian distribution, experimental results in various settings, even in situations in which this assumption is violated, demonstrate the usefulness and robustness of DCT, suggesting the development of this technique is essential to causal discovery from discretized continuous data. Further details of the empirical validations are provided in Appendix E.

C FIGURE OF MAIN EXPERIMENTS: CAUSAL DISCOVERY



(a) fixed nodes $p = 8$, changing sample size $n = (500, 1000, 5000, 1000)$



(b) fixed sample size $n = 5000$, changing node $p = (4, 6, 8, 10)$

Figure 4: Experiment result of DAG discovery on synthetic data for changing sample size (a) and changing number of nodes (b). Fisherz_nodis is the Fisher-z test applied to original continuous data. We evaluate F_1 (↑), Precision (↑), Recall (↑) and SHD (↓).

D PSEUDO CODE

Algorithm 1 DCT: Discretization-Aware CI Test

Require:

- Observed data matrix $\tilde{\boldsymbol{X}}' \in \mathbb{R}^{n \times d}$
- Tested pair indices j_1, j_2 with $j_1 \neq j_2$
- Conditioning set $\mathbf{S} \subseteq \{1, \dots, d\} \setminus \{j_1, j_2\}$
- Significance level α

Rearrange Data Matrix

$$\tilde{\boldsymbol{X}} = \left[\tilde{\boldsymbol{X}}'[:, j_1], \tilde{\boldsymbol{X}}'[:, j_2], \tilde{\boldsymbol{X}}'[:, \mathbf{S}]\right] \in \mathbb{R}^{n \times p}, \quad \text{where } p = 2 + |\mathbf{S}|$$

Initialize Covariance Matrix

$$\hat{\boldsymbol{\Sigma}} \leftarrow \mathbf{I}_p \quad \text{(identity matrix of size } p \times p\text{)}$$

for $q \leftarrow 1$ **to** p **do**
  for $k \leftarrow q + 1$ **to** p **do**
    if both $\tilde{X}[:, q]$ and $\tilde{X}[:, k]$ are continuous **then**
      Compute sample covariance $\hat{\sigma}_{qk}$ using equation 5
    else
      Compute covariance $\hat{\sigma}_{qk}$ using Equation equation 6
    end if
    Update covariance matrix:

$$\hat{\boldsymbol{\Sigma}}[q, k] \leftarrow \hat{\sigma}_{qk}$$

$$\hat{\boldsymbol{\Sigma}}[k, q] \leftarrow \hat{\sigma}_{qk} \quad \text{(ensuring symmetry)}$$

  end for
end for

Extract Submatrices (j_1 and j_2 correspond the first and second column of $\tilde{X}$ due to the regroup)

- Let $\hat{\boldsymbol{\Sigma}}_{-1-1} \in \mathbb{R}^{p-1 \times p-1} \leftarrow$ the submatrix of $\hat{\boldsymbol{\Sigma}}$ without 1st column and 1st row
- Let $\hat{\boldsymbol{\Sigma}}_{-11} \in \mathbb{R}^{p-1}$ be the 1st column of $\hat{\boldsymbol{\Sigma}}$ with first row removed

Compute Test Statistics

$$\hat{\beta}_{1,2} \leftarrow \hat{\boldsymbol{\Sigma}}_{-1-1}^{-1} \hat{\boldsymbol{\Sigma}}_{-11}$$

Formulate Null Distribution

$\Phi(z) \leftarrow$ Cumulative distribution function of the Normal Distribution defined in Thm. 3.8

Calculate P-value

$$p\text{-value} \leftarrow 2 \cdot \left(1 - \Phi\left(|\hat{\beta}_{1,2}|\right)\right)$$

Make Decision

if p-value $> \alpha$ **then**
  Conclude: $X_{j_1} \perp\!\!\!\perp X_{j_2} \mid X_{\mathbf{S}}$
else
  Conclude: $X_{j_1} \not\!\perp\!\!\!\perp X_{j_2} \mid X_{\mathbf{S}}$
end if
return The conditional independence decision

E Additional experiments

E.1 Linear non-Gaussian and nonlinear

Our model requires that the original data must adhere to the hypothesis of following a multivariate normal distribution, which appears to potentially limit the generalizability. Therefore, it is worthwhile to explore its robustness when such assumptions are violated. In this regard, we conducted several experiments, including scenarios involving linear non-Gaussian and nonlinear Gaussian.

For both cases, we follow the setting of our experiment where there are $p = 8$ nodes and $p - 1$ edges. We explore the effect of changing sample size $n = (100, 500, 2000, 5000)$. Specifically for linear non-Gaussian case, we adhere to some of the settings outlined by (Shimizu et al., 2011), conducting experiments where the original continuous data followed: (1) a Student's t-distribution with 3 degrees of freedom, (2) a uniform distribution, and (3) an exponential distribution. Each variable is generated as $X_i = f(PA_i) + noise$, where $noise$ follows the distribution in (1), (2), (3) correspondingly and f is an arbitrary linear function. The first three rows of Fig. 5 and Fig. 6 show the result of the linear non-Gaussian case.

For the nonlinear cases, we follow setting in (Li et al., 2024), where every variable X_i is generated as $X_i = f(WPA_i + noise)$, $noise \sim N(0, 1)$ and f is a function randomly chosen from (a) $f(x) = sin(x)$, (b) $f(x) = x^3$, (c) $f(x) = tanh(x)$, and (d) $f(x) = ReLU(x)$. W is a linear function. Similarly, we set the number of nodes at $p = 8$ and change the number of samples $n = (500, 2000, 5000)$. For both cases, we run 10 graph instances with different seeds and report the result of skeleton discovery in Fig. 5 and DAG in Fig. 6 (The same orientation rules (Dor & Tarsi, 1992) used in the main experiment are employed to convert a CPDAG (Chandler Squires, 2018) into a DAG). The last row of Fig. 5 and Fig. 6 shows the result of the nonlinear case.

Based on the experimental outcomes, DCT demonstrates marginally superior or comparable efficacy in terms of the F1-score, precision, and SHD relative to both the Fisher-Z test and the Chi-square test when dealing with small sample sizes. Nevertheless, as the sample size increases, DCT's performance clearly surpasses that of the aforementioned tests across all three evaluated metrics, especially in the linear case. Consistent with observations from the main experiment, DCT exhibits a lower recall in comparison to the baseline tests. This discrepancy can be attributed to the baseline tests being prone to incorrectly infer conditional dependence and connect a large proportion of nodes. According to the results, our test shows notable robustness under the case assumptions are violated, confirming its practical effectiveness.

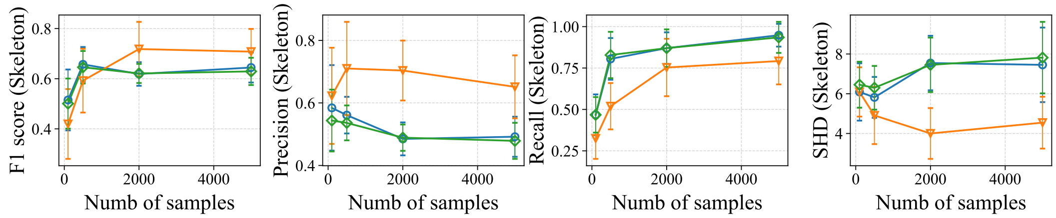
E.2 Denser graph

DCT primarily works on cases where CI is mistakenly judged as conditional dependence due to discretization. Consequently, its efficacy is more pronounced in scenarios characterized by a relatively sparse graph, as numerous instances are truly conditionally independent. Nevertheless, the investigation of causal discovery with a dense latent graph is essential for evaluating the power of a test, i.e., its ability to successfully reject the null hypothesis when the tested pairs are conditionally dependent. Thus, we conduct the experiment where $p = 8, n = 10000$ and changing edges $(p + 2, p + 4, p + 6)$. Similarly, the latent continuous data follows a multivariate Gaussian model and the true DAG $\mathcal{G}$ is constructed using BP model. We run 10 graph instances with different seeds and report the result of the skeleton discovery and DAG in Fig. 7.

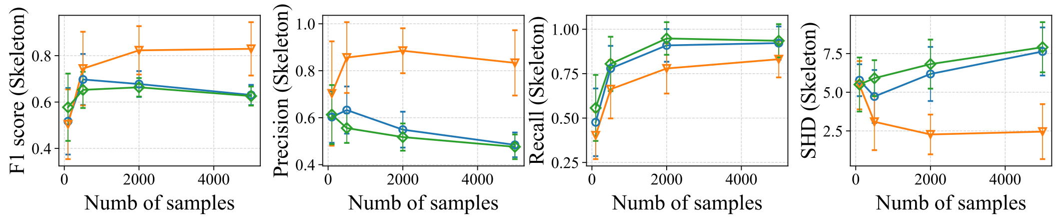
According to the experiment results, DCT exhibits better performance in terms of the F1-score, precision, and SHD relative to both the Fisher-Z test and the Chi-square test. As the graph becomes progressively denser, the superiority of the DCT correspondingly diminishes as there are few conditional independent cases in the true DAG. Due to the same reason, The recall remains lower than that of other baseline methods.

E.3 Multivariate Gaussian with nonzero mean and non-unit variance

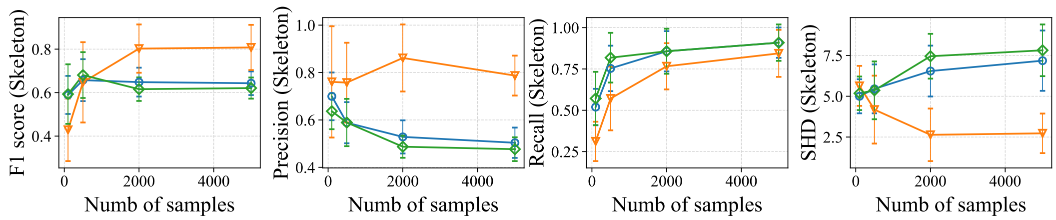
We employed a setting nearly identical to the main experiment, with the only difference being the alteration in data generation: instead of using a standard normal distribution, we used a Gaussian distribution with mean sampled from $U(-2, 2)$ and variance sampled from $U(0, 3)$. We fix the number of variables as $p = 8$ and change the number of samples $n = (100, 500, 2000, 5000)$. The Fig. 8 shows the result and demonstrates the effectiveness of our method.



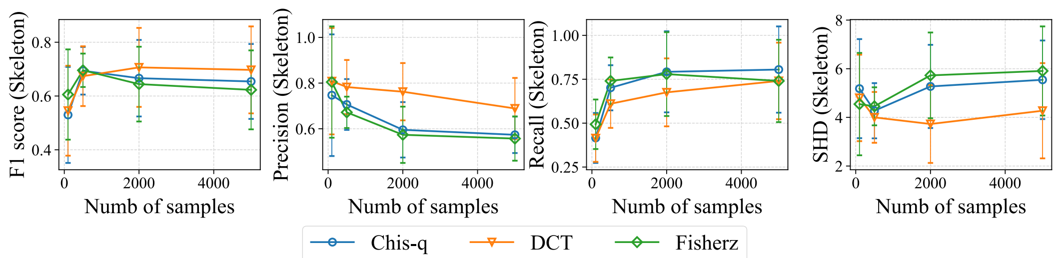
(a) Linear Exponential.



(b) Linear Student.



(c) Linear Uniform.



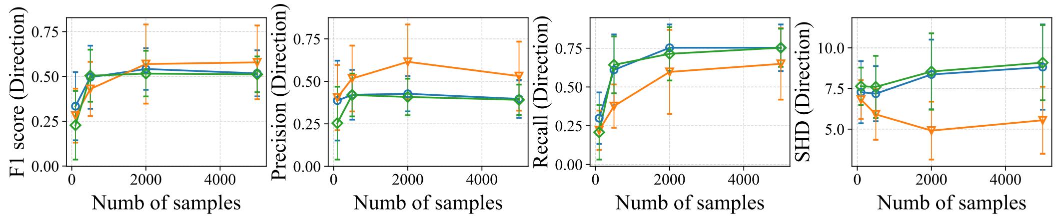
(d) Nonlinear Gaussian.

Figure 5: Experiment result of causal discovery on synthetic data with $p = 8$, $n = (100, 500, 2000, 5000)$ where the data generation process violates our assumptions. The data are generated with either nongaussian distributed (a), (b), (c) or the relations are not linear (d). The figure reports F_1 (↑), Precision (↑), Recall (↑) and SHD (↓) on skeleton.

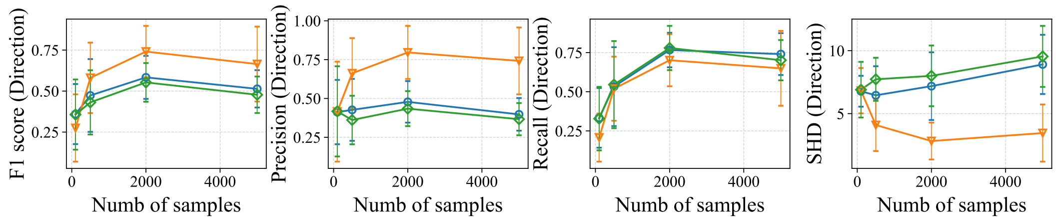
E.4 REAL-WORLD DATASET

To further validate DCT, we employ it on a real-world dataset: Big Five Personality https://openpsychometrics.org/, which includes 50 personality indicators and over 19000 data samples. Each variable contains 5 possible discrete values to represent the scale of the corresponding questions, where 1=Disagree, 2=Weakly disagree, 3=Neutral, 4=Weakly agree and 5=Agree, e.g., "N3=1" means "I agree that I worry about things". This scenario clearly suits DCT, where the degree of agreement with a certain question must be a continuous variable while we can only observe the result after categorization. We choose three variables respectively: [N3: I worry about things], [N10: I often feel blue], [N4: I seldom feel blue]. We then do the casual discovery using PC algorithm with DCT and compare it with the Chi-square test and Fisher-Z test. The result can be found in Fig. 9.

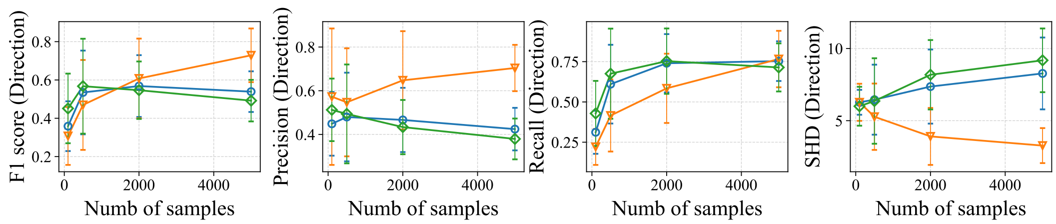
Based on the experimental outcomes, despite the absence of a groundtruth for reference, we observe that the results obtained via DCT appear more plausible than those derived from Fisher-Z and Chi-square tests. Specifically, DCT suggests the relationship $N_3 \perp\!\!\!\perp N4|N10$, which is reasonable as intuitively, the answer of 'I often feel blue' already captures the information of 'I seldom feel blue'.



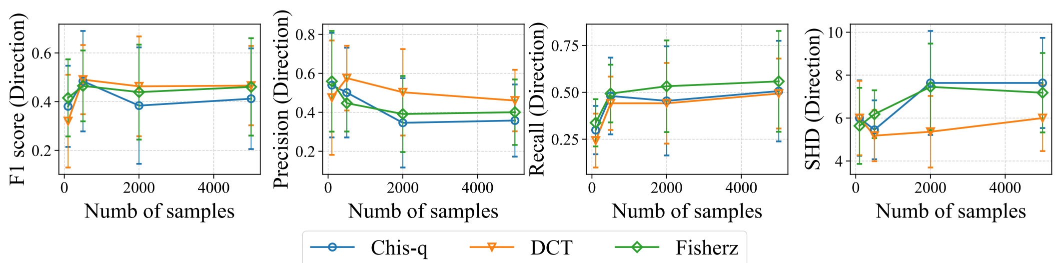
(a) Linear Exponential.



(b) Linear Student.



(c) Linear Uniform.



(d) Nonlinear Gaussian.

Figure 6: Experiment result of causal discovery on synthetic data with $p = 8$, $n = (100, 500, 2000, 5000)$ where the data generation process violates our assumptions. The data are generated with either nongaussian distributed (a), (b), (c) or the relations are not linear (d). The figure reports F_1 (↑), Precision (↑), Recall (↑) and SHD (↓) on DAG.

As a comparison, both Fisher-Z and Chi-square return a fully connected graph. The results directly correspond to our illustrative example shown in Fig. 1, substantiating the necessity of our proposed test.

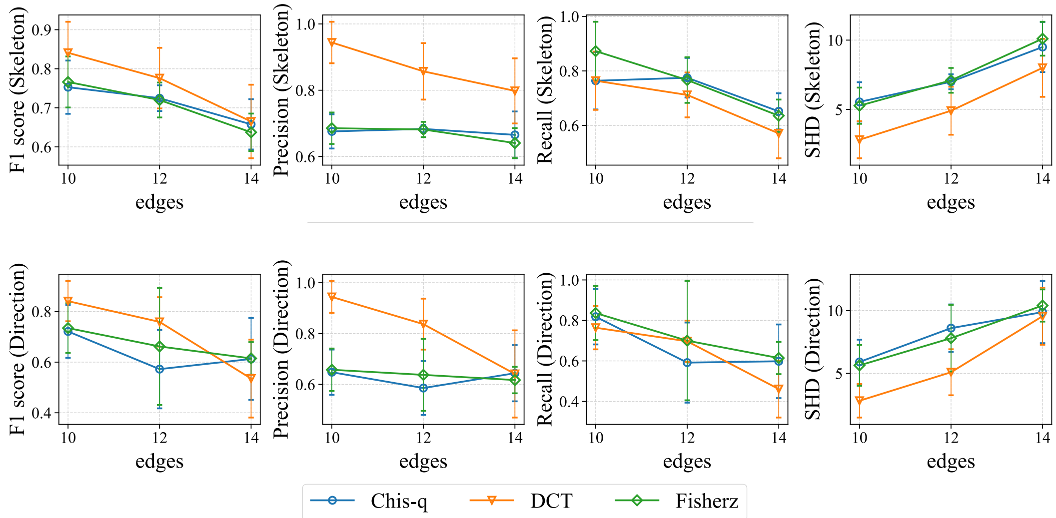


Figure 7: Experimental comparison of causal discovery on synthetic datasets for denser graphs with $p = 8, n = 10000$ and edges varying $p + 2, p + 4, p + 6$. We evaluate F_1 (↑), Precision (↑), Recall (↑) and SHD (↓) on both skeleton and DAG.

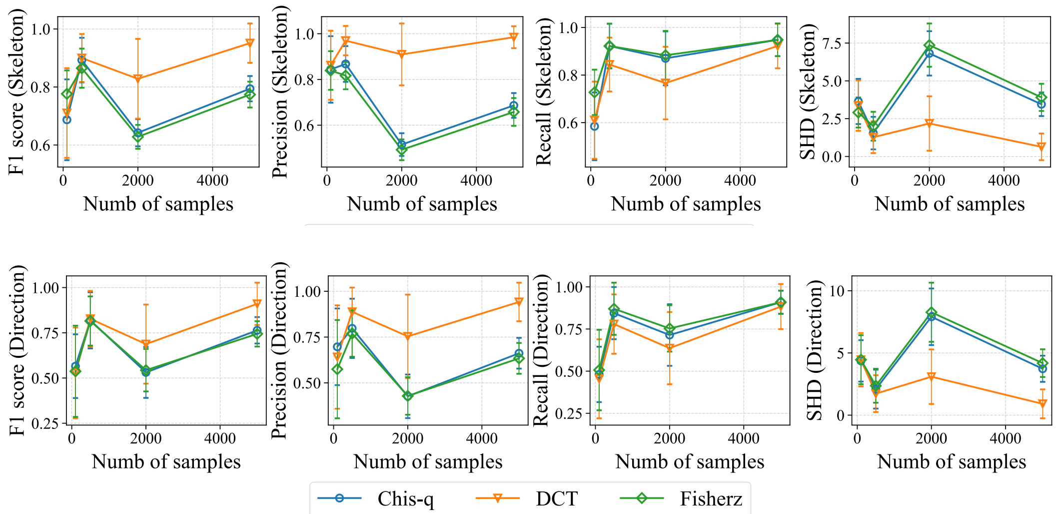
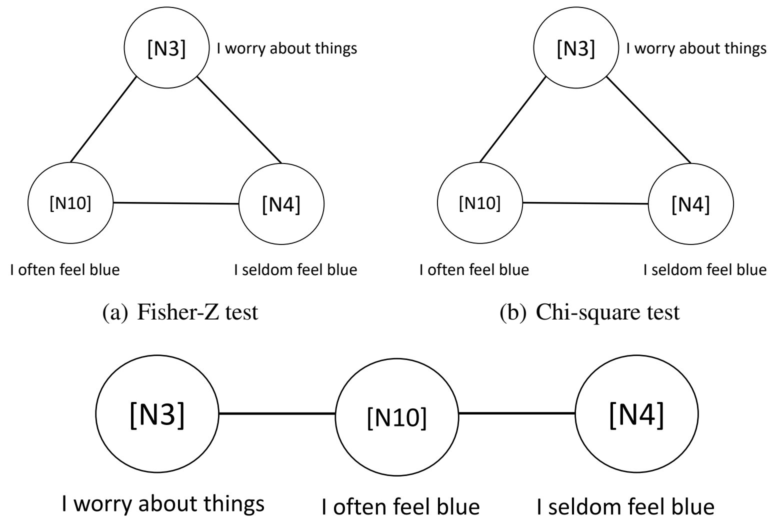


Figure 8: Experimental comparison of causal discovery on synthetic datasets for multivariate Gaussian model with $p = 8, n = (100, 500, 2000, 5000)$ and where mean is not zero. We evaluate F_1 (↑), Precision (↑), Recall (↑) and SHD (↓) on both skeleton and DAG.

F Related Work

Testing for CI is pivotal in the field of causal discovery (Spirtes et al., 2000), and a variety of methods exist for performing CI tests (CI tests). An important group of CI test methods involves the assumption of Gaussian variables with linear dependencies. For example, under this assumption, Gaussian graphical models are extensively studied (Yuan & Lin, 2007; Peterson et al., 2015; Mohan et al., 2012; Ren et al., 2015). To address CI test under Gaussian assumption, partial correlation serves as a viable method for CI testing (Baba et al., 2004). To evaluate the independence of variables X_1 and X_2 conditional on $\boldsymbol{Z}$, The technique proposed by (Su & White, 2008) determines CI by comparing the estimations of $p(X_1|X_2, \boldsymbol{Z})$ and $p(X_1|X_2)$.



(a) Fisher-Z test (b) Chi-square test

(c) DCT

Figure 9: Experimental comparison of causal discovery on the real-world dataset.

Another approach involves discretizing $\boldsymbol{Z}$ and performing independent tests within each resulting bin (Margaritis, 2005). Our work, however, diverges from these existing methods in two significant ways. Firstly, we are equipped to handle data, where partial variables are discretized. Additionally, we postulate that discrete variables are derived from the transformation of continuous variables in a latent Gaussian model. With the same assumption, the most closely related study is by (Fan et al., 2017), where the authors developed a novel rank-based estimator for the precision matrix of mixed data. However, their work stops short of providing a CI test for this method. Our research fills this gap, offering the ability to estimate the precision matrix for both discrete and mixed data and providing a rigorous CI test for our methodology.

Recent advancements in CI testing have utilized kernel methods for continuous variables influenced by nonlinear relationships. (Fukumizu et al., 2004) describes non-parametric CI relationships using covariance operators in reproducing kernel Hilbert spaces (RKHS). KCI test (Zhang et al., 2012) assesses the partial associations of regression functions linking x, y, and z, while RCI test (Strobl et al., 2019) aims to enhance the KCI test's efficiency. In KCIP test (Doran et al., 2014) employs permutations of samples to emulate CI scenarios. CCI test (Sen et al., 2017) further reformulates testing into a process that leverages the capabilities of supervised learning models. For discrete variable analysis, the G^2 test (Aliferis et al., 2010) and conditional mutual information (Zhang et al., 2010) are commonly employed. However, their method cannot deal with our setting where only discretized version of latent variables can be observed.

G Resource Usage

All the experiments are run using Intel(R) Xeon(R) CPU E5-2680 v4 with 55 processors. It costs 4 hours to run experiments in Section 3.1.

H Limiation and Broader Impacts

Limitation So far, the largest limitation of our method is to treat discretized variables as binary, which wastes the available information. Besides that, the parametric assumption limits its generalizability. However, we need to point out this is pretty normal in CI test fields.

Broader Impacts The goal of our proposed method is to test the conditional independence relationship given discretized observation. This task is essential and has broad applications. We are confident that our method will be beneficial and will not result in negative societal impacts.