A Conditional Independence Test in the Presence of Discretization

Anonymous Authors[1]

Abstract

Testing conditional independence has many applications, such as in Bayesian network learning and causal discovery. Different test methods have been proposed. However, existing methods generally can not work when only discretized observations are available. Specifically, consider X_1, $\tilde{X}_2$ and X_3 are observed variables, where $\tilde{X}_2$ is a discretization of latent variables X_2. Applying existing test methods to the observations of X_1, $\tilde{X}_2$ and X_3 can lead to a false conclusion about the underlying conditional independence of variables X_1, X_2 and X_3. Motivated by this, we propose a conditional independence test specifically designed to accommodate the presence of such discretization. To achieve this, we design the bridge equations to estimate the underlying conditional independence. An appropriate test statistic and its asymptotic distribution under the null hypothesis of conditional independence have also been derived. Both theoretical proofs and empirical validation have been provided, demonstrating the effectiveness of our test methods.

1. Introduction

Conditional independence testing is an important task in statistical analysis and machine learning. It can determine whether two variables X_1 and X_2, are independent given a set of other variables $\boldsymbol{Z}$. This implies that once $\boldsymbol{Z}$ is known, the values of X_1 provide no additional information about X_2, and vice versa. Such testing is crucial in different domains, including causal discovery and Bayesian network learning, as it helps in understanding and modeling the relationships and dependencies within data.

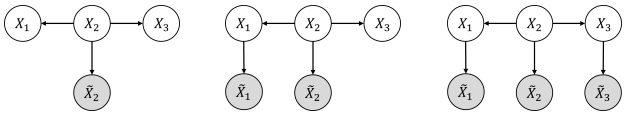
Existing conditional independence tests assume direct access to observations from all variables and have proven effectiveness. However, in many real-world scenarios, the observations of some variables can be discretized. Specifically, in this case, instead of having the observations of X, only the discretized observations from a variable $\tilde{X}$ resulting from a discretized transformation of the variable X are available. For instance, in healthcare, although the progression of cancer is a continuous process, it is discretized into stages (e.g., Stage I, II, III, IV) for clinical and treatment purposes. Similarly, in environmental studies, researchers often work with discretized classifications of continuous variables like pollution levels, categorized into ranges for analysis. In finance, variables such as asset values are binned into ranges for assessing creditworthiness or investment risks (e.g., sell, hold, and strong buy).

When in the presence of discretization, observations can not reflect the underlying conditional independence relationships. Directly applying current conditional independence tests to those observations can lead to false conclusions. For instance, consider the three data generative processes presented through *causal graphical models* (Pearl et al., 2000) in Fig 1. Let X_i represent continuous variables, and $\tilde{X}_i$ denote these variables after discretization. The gray shade of a node indicates that the variable is observable, white shade of a node indicates that the variable is latent. Under the causal Markov condition (Pearl et al., 2000; Spirtes et al., 2000) that *a variable is conditionally independent of all others (except its effects) given its direct causes*, variables X_1 and X_3 should be conditionally independent given X_2. However, this independence may not hold after discretization. For example, in Fig 1(a), conditioning on $\tilde{X}_2$, X_1 and X_3 is not conditionally independent in general. This phenomenon arises because the latent variable X_2, serving as a common cause for both X_1 and X_3, is not factored into the conditioning set. Consequently, since X_1 and X_3 are derived from X_2, they inherently contain some information about X_2, leading to a dependency between them. As a result, applying existing methods directly to observations yields erroneous conclusions about the conditional independence of X_1 and X_3 given $\tilde{X}_2$. Due to the same reason, checking the conditional independence also fails in Fig 1(b) and Fig 1(c).

To accurately test the conditional independence in the presence of discretization, we introduce a novel testing method tailored specifically for this scenario. To achieve this, we design the bridge equations to estimate the underlying con-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.



(a) (b) (c)

Figure 1: Different data generative processes presented through causal graphical models. Under causal Markov condition in the presence of discretization. The causal graphic model entails conditional independence relations. Observed variables cannot reflect the conditional independence of X_1 and X_3 given X_2.

ditional independence from observations. Appropriate test statistics and their asymptotic distribution under the null hypothesis of unconditional independence and conditional independence have also been derived, respectively. Our method is adept at assessing the conditional independence of underlying continuous variables across a range of scenarios, including cases where both observed variables are continuous, both are discretized, or one is continuous and the other is discretized.

2. Related Work

Testing for conditional independence is pivotal in the field of causal discovery (Spirtes et al., 2000), and a variety of methods exist for performing conditional independence 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 conditional independence by comparing the estimations of $p(X_1|X_2, \boldsymbol{Z})$ and $p(X_1|X_2)$.

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 discrete. 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 conditional independence 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 conditional independence 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 discrete variables are generated from transformation of latent linear Gaussian variables.

3. DCT: A Conditional Independence Test in the Presence of Discretization

Problem Setting Consider a set of independent and identically distributed (i.i.d.) p-dimensional random vectors, denoted as $\tilde{\boldsymbol{X}} = (X_1, X_2, \ldots, \tilde{X}_j, \ldots, \tilde{X}_p)^T$. In this set, some variables, indicated by a tilde ($\sim$), such as $\tilde{X}_j$, are assumed to 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. This function, $g_j : \mathcal{X} \to \tilde{\mathcal{X}}$, maps the continuous domain of X_j onto the discrete domain of $\tilde{\mathcal{X}}_j$, such that $\tilde{X}_j = g_j(X_j)$ for each observation. The cardinality of the domain after discretization is at least 2 and smaller than infinity. Given n observations $\{\tilde{\boldsymbol{x}}^1, \tilde{\boldsymbol{x}}^2, \ldots, \tilde{\boldsymbol{x}}^n\}$ randomly sampled from $\tilde{\boldsymbol{X}}$, the goal is to assess both conditional and unconditional independence among the variables of the vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_j, \ldots, X_p)^T$. In our model, we assume $\boldsymbol{X} \sim N(0, \Sigma)$, Σ 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, which is detailed in Appendix A.8.

Preliminary Framework for DCT In the design of a hypothesis test, two principal challenges typically need to be addressed. Firstly, there is the issue of establishing a connection between existing observations and the parameter of interest, a task referred to as the *estimation*. Secondly, one must determine the relationship between the estimated parameter and the true parameter to draw conclusions, a process known as the *inference* problem.

Specifically, to design a hypothesis test, one must first establish a test statistic. This statistic is calculated from observations. The next step is to derive the underlying distribution

of this test statistic under the null hypothesis (when given an infinite sample size). Within this framework, to check whether the null hypothesis holds, one starts by calculating the value of the test statistic based on observations. Then, if the value is likely to have been sampled from the derived distribution of the test statistic, it suggests that the null hypothesis is likely to hold.

More concretely, in the context of an unconditional independence test, we formulate a test statistic that is calculated from observations of two variables, where we are interested in their independence relationship. We then derive the distribution of the test statistic under the null hypothesis that the two variables are assumed to be independent. This derived distribution is leveraged to calculate the likelihood that the observed value of our test statistic t, based on the observations of the two variables, could occur under the assumption of their independence.

In the context of a conditional independence test, we formulate another test statistic that is calculated from observations of two variables, along with observations of other variables. In this case, we are interested in their independence relationship conditional on these other variables. Similarly, we then derive the distribution of the test statistic under the null hypothesis that the two variables are assumed to be conditionally independent.

Our objective is to deduce the independence and conditional independence relationships within the original linear Gaussian model, based on its discretized observations. In the context of a linear Gaussian model, this challenge is directly equivalent to constructing statistical inferences for its covariance matrix ($\mathbf{\Sigma}$) and its precision matrix ($\mathbf{\Omega}$). In the ensuing section, we illustrate the design of a bridge equation to address the estimation problem of $\mathbf{\Sigma}$. We then demonstrate how to design a proper statistical inference for $\mathbf{\Sigma}$. Following this, we connect $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ through nodewise regression. In such a way, a statistical inference for conditional independence is effectively designed.

3.1. Design *Bridge Equation* for Test Statistics

Role of Bridge Equation The bridge equation establishes a connection between the underlying covariance $\sigma_{i,j}$ of two continuous variables X_{j_1} and X_{j_2} and observations. This equation is leveraged in estimating the covariance σ_{j_1,j_2} and provides a theoretical basis for the test statistics. In general, its form is as follows.

$$\hat{\tau}_{j_1,j_2} = T(\sigma; \hat{\mathbf{\Lambda}}), \tag{1}$$

where $\hat{\tau}_{j_1,j_2}$ is a statistic which can also be estimated from observations. The set $\hat{\mathbf{\Lambda}}$ contains additional parameters that might be required by the bridge function T. This set of parameters can be estimated from observations.

When in the presence of discretization, the discrete transformations make the sample covariance matrix based on $\tilde{\boldsymbol{X}}$ inconsistent with the covariance matrix of $\boldsymbol{X}$. Then the covariance matrix of $\tilde{\boldsymbol{X}}$ should not serve as an estimation of the covariance matrix of $\boldsymbol{X}$. To obtain the estimation $\hat{\sigma}_{j_1,j_2}$ of σ_{j_1,j_2}, the bridge equation is leveraged. *Specifically, given estimated $\hat{\tau}_{j_1,j_2}$ and $\hat{\mathbf{\Lambda}}$, $\hat{\sigma}_{j_1,j_2}$ can be obtained by solving the bridge equation $\hat{\tau}_{j_1,j_2} = T(\sigma; \hat{\mathbf{\Lambda}})$.* As a result, the covariance matrix of $\boldsymbol{X}$ and the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ can be estimated. These estimated matrices contain information about both unconditional and conditional independence among the variables, respectively.

Additionally, the bridge equation provides a theoretical basis for the test statistics. Specifically, our objectives extend beyond simply estimating the covariance matrix $\mathbf{\Sigma}$ and the precision matrix $\mathbf{\Omega}$, we also aim to rigorously test both unconditional independence and conditional independence relationships among the variables of $\boldsymbol{X}$. This involves employing p-values to test the confidence level in the independence of variables, providing a quantitative measure of the strength of these relationships and a clear indication of whether the null hypothesis of independence can be rejected. To accomplish this objective, it is crucial to develop two distinct test statistics, each with its unique asymptotic distributions under the null hypotheses of unconditional and conditional independence, respectively. The design of these test statistics, along with their asymptotic distributions, are integrally linked to the bridge equation T. The details are provided in Section 3.2 and Section 3.3.

Bridge Equations in Different Cases In scenarios involving discretization, the design of bridge equations has to be tailored to accommodate different cases characterized by the characteristics of the variables involved. Specifically,

- C1. *Both Variables are Directly Observed:* In the simplest case, the observations for both continuous variables X_{j_1} and X_{j_2} are directly available. The bridge function does not involve the complexities of transformation or additional $\hat{\mathbf{\Lambda}}$ have to be estimated.
- C2. *Both Variables are Latent:* Both continuous variables X_{j_1} and X_{j_2} are latent, with their observed counterparts $\tilde{X}_{j_1}$ and $\tilde{X}_{j_2}$ being discrete transformations of these latent variables. The required bridge function here aims to establish connections from the discretized observations back to the latent covariance.
- C3. *One Variable is Latent and One is Observed:* One variable $\tilde{X}_{j_2}$ is observed as a discrete transformation of a latent variable X_{j_2} , while the other X_{j_1} is available in its continuous form. The bridge function must be adept at handling the interaction between discrete and continuous data in this scenario.

In the end, we found that cases C2 and C3 can be merged together, with the same form of bridge equation but different parameters. We present our bridge equations in definitions 3.1, 3.3 and 3.4.

Definition 3.1 (Bridge Equation for A Continuous-Variable Pair)**.** For two continuous variables X_{j_1} and X_{j_2} , the bridge equation is defined as:

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

In the simplest scenario above, where there is no discretized transformation, the sample covariance of X_{j_1} and X_{j_2} provides a consistent estimation. In this context, the bridge function T acts merely as an identity mapping. However, the situation becomes more complex with the introduction of discretized transformations. In these cases, the bridge function must be specifically designed and include additional parameters $\hat{\mathbf{\Lambda}}$. Most importantly, this function needs to be a one-to-one mapping to ensure that, given the parameters $\hat{\mathbf{\Lambda}}$ and $\hat{\tau}_{j_1,j_2}$, the $\hat{\sigma}_{j_1,j_2}$ can be uniquely identified. Now we define the additional parameter required by the bridge function T when dealing with scenarios involving discretization as follows. Let $\mathbb{P}_n Z$ denote the average of a random variable Z given n i.i.d. observation of Z.

Definition 3.2. Let Φ be the cumulative distribution function (cdf) of the standard normal distribution, $\hat{h}_j = \Phi^{-1}(1-\hat{\tau}_j)$, where $\hat{\tau}_j = \sum_{i=1}^n \mathbb{1}_{\{\tilde{x}_j^i > \mathbb{P}_n \tilde{X}_j\}}/n$. We further denote $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$.

Definition 3.3 (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:

$$\begin{aligned}\hat{\tau}_{j_1,j_2} &:= \hat{P}(\tilde{X}_{j_1} > \mathbb{P}_n \tilde{X}_{j_1}, \tilde{X}_{j_2} > \mathbb{P}_n \tilde{X}_{j_2}) \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\tilde{x}_{j_1}^i > \mathbb{P}_n \tilde{X}_{j_1}, \tilde{x}_{j_2}^i > \mathbb{P}_n \tilde{X}_{j_2}\}} = T(\sigma; \{\hat{h}_{j_1}, \hat{h}_{j_2}\}),\end{aligned}$$

and the bridge function is

$$T(\sigma; \{\hat{h}_{j_1}, \hat{h}_{j_2}\}) = \int_{x_1 > \hat{h}_{j_1}} \int_{x_2 > \hat{h}_{j_2}} \phi(x_{j_1}, x_{j_2}; \sigma) dx_{j_1} dx_{j_2},$$

where ϕ is the probability density function of a bivariate normal distribution.

Definition 3.4 (Bridge Equation for A Continuous-Discretized-Variable Pair)**.** For one continuous variable X_{j_1} and one discrete variable $\tilde{X}_{j_2}$, the bridge equation is defined as:

$$\hat{\tau}_{j_1,j_2} := \hat{P}(X_{j_1} > 0, \tilde{X}_{j_2} > \mathbb{P}_n \tilde{X}_{j_2}) = T(\sigma; \{0, \hat{h}_{j_2}\}).$$

Estimating Covariance σ To estimate σ_{j_1,j_2}, by leveraging Eq. (1), $\hat{\sigma}_{j_1,j_2} = T^{-1}(\hat{\tau}_{j_1,j_2}; \hat{\theta})$. For two continuous variables X_{j_1} and X_{j_2} , The analytic solution can be simply obtained as follows.

$$\hat{\sigma}_{j_1,j_2} = \frac{1}{n}\sum_{i=1}^n x_{j_1}^i x_{j_2}^i - \frac{1}{n}\sum_{i=1}^n x_{j_1}^i \frac{1}{n}\sum_{i=1}^n x_{j_2}^i.$$

For the cases involving the discretized variable as proposed in Definition 3.3 and Definition 3.4, obtaining an analytic solution for $\hat{\sigma}_{j_1,j_2}$ is challenging. To estimate σ_{j_1,j_2}, we optimize it such that $\hat{\sigma}_{j_1,j_2}$ is the value that satisfies Eq. (1) by solving the objective

$$\min_{\sigma} ||\hat{\tau}_{j_1,j_2} - T(\sigma; \{\hat{h}_{j_1}, \hat{h}_{j_2}\})||^2 \quad s.t. -1 < \sigma < 1. \quad (2)$$

The $\hat{\tau}_{j_1,j_2}$ will be one-to-one mapping with calculated $\hat{\sigma}_{j_1,j_2}$ for fixed $\hat{h}_{j_1}, \hat{h}_{j_2}$. The proof is provided in Appendix A.2

3.2. Unconditional Independence Test

Let $\psi_{\hat{\theta}} = [f_{\hat{\theta}}^1(\cdot), f_{\hat{\theta}}^2(\cdot), f_{\hat{\theta}}^3(\cdot)]^T$ contain a group of functions parameterized by $\hat{\theta} = (\hat{\sigma}_{j_1,j_2}, \hat{h}_{j_1}, \hat{h}_{j_2})$. Define $\mathbb{P}_n \psi_{\hat{\theta}}$ as sample mean of these functions evaluated at n sample points. Similarly, $\mathbb{P}_n \psi_{\hat{\theta}} \psi_{\hat{\theta}}^T$ is defined as sample mean of the outer product $\psi_{\hat{\theta}} \psi_{\hat{\theta}}^T$. The notation $P\psi_{\hat{\theta}} := E\mathbb{P}_n \psi_{\hat{\theta}}$ denotes the expectations of the functions in $\psi_{\hat{\theta}}$. Furthermore, let $\psi_{\hat{\theta}}'$ denote the derivative of the functions contained in $\psi_{\hat{\theta}}$. The specific form and domain of $\psi_{\hat{\theta}}$ will be defined later. In the following theorem, we formulate the general form of the distribution used to test the statistic $\hat{\sigma}_{j_1,j_2}$.

Theorem 3.5 (Independence Test)**.** *In our settings, under the null hypothesis that two observed variables indexed with j_1 and j_2 are statistically independent under our framework, the independence can be tested using the statistic*

$$\hat{\sigma}_{j_1,j_2} = T^{-1}(\hat{\tau}_{j_1,j_2}; \hat{\theta}).$$

This statistic is approximated to follow a normal distribution, as detailed below.

$$\hat{\sigma}_{j_1,j_2} \overset{approx}{\sim} N\left(0, \tfrac{1}{n}((\mathbb{P}_n \psi_{\hat{\theta}}')^{-1} \mathbb{P}_n \psi_{\hat{\theta}} \psi_{\hat{\theta}}^T (\mathbb{P}_n \psi_{\hat{\theta}}'^T)^{-1})_{1,1}\right) \quad (3)$$

The above theorem utilizes the estimated parameter $\hat{\theta}$. As a result, the statistic is approximately normally distributed. Ideally, the true parameter θ_0 should be used, which is an expectation of $\hat{\theta}$. Specifically, the distribution is given as:

$$\begin{aligned}&N\left(0, \frac{1}{n}\left((E\mathbb{P}_n \psi_{\hat{\theta}}')^{-1} E\mathbb{P}_n \psi_{\hat{\theta}} \psi_{\hat{\theta}}^T (E\mathbb{P}_n \psi_{\hat{\theta}}'^T)^{-1}\right)_{1,1}\right) \\ =&N\left(0, \frac{1}{n}\left((P\psi_{\theta_0}')^{-1} P\psi_{\theta_0} \psi_{\theta_0}^T (P\psi_{\theta_0}'^T)^{-1}\right)_{1,1}\right). \quad (4)\end{aligned}$$

It is important to note that the approximation error caused by using $\hat{\theta}$ instead of θ_0 diminishes as the sample size increases. This is because the value of $\hat{\theta}$ converges to the true

underlying parameter θ_0 with increasing sample size. We left the detailed derivation in Appendix A.3.

For conducting an unconditional independence test, the covariance $\hat{\sigma}_{j_1,j_2}$ serves as the test statistic and can be estimated from observed data as aforementioned. The missing piece is to derive the specific form and domain of $\psi_{\hat{\theta}}$ in Theorem 3.5. Once $\psi_{\hat{\theta}}$ is clearly established, we can directly calculate its derivative $\psi'_{\hat{\theta}}$, its average $\mathbb{P}_n\psi_{\hat{\theta}}$, and the average over the product $\mathbb{P}_n(\psi_{\hat{\theta}}\psi_{\hat{\theta}}^T)$. These elements determine the variance of the normal distribution, which is essential for statistical inference. In other words, assessing the probability of the observed $\hat{\sigma}_{j_1,j_2}$ being drawn from this normal distribution. A high probability would suggest that the statistic is consistent with the null hypothesis, indicating the independence of X_{j_1} and X_{j_2}. Conversely, a low probability would imply a deviation from the null hypothesis, suggesting a potential dependence between these variables.

Since the variables being tested for independence can be both discretized, only one being discretized, or neither being discretized. This results in different forms of $\psi'_{\hat{\theta}}$ consequently differs across these scenarios. Let Z_{j_1} and Z_{j_2} be any two random variables indexed by j_1 and j_2. Note that they can be either discrete or continuous. Let $\mathbb{P}_n Z_{j_1}$ denote the sample mean of Z_{j_1}. Let $\hat{\sigma}^i_{j_1,j_2} = z^i_{j_1} \cdot z^i_{j_2} - \mathbb{P}_n Z_{j_1} \cdot \mathbb{P}_n Z_{j_2}$ denote the sample covariance based on a i-th pairwise observation of the variables Z_{j_1} and Z_{j_2}. Let $\hat{\tau}^i_{j_1} = \mathbb{1}_{\{z^i_{j_1} > \mathbb{P}_n Z_{j_1}\}}$ and $\hat{\tau}^i_{j_2} = \mathbb{1}_{\{Z^i_{j_2} > \mathbb{P}_n Z_{j_2}\}}$, each calculated based on i-th observations of the variables Z_{j_1} and Z_{j_2}, respectively. Let $\hat{\tau}^i_{j_1,j_2}$ be $\hat{\tau}^i_{j_1} \cdot \hat{\tau}^i_{j_2}$. In subsequent discussions, to help distinguish between different types of variables, instead of using Z_{j_1} and Z_{j_2}, we use X_{j_1} and X_{j_2} for continuous observed variables and $\tilde{X}_{j_1}$ and $\tilde{X}_{j_2}$ for discretized observed variables. We now demonstrate the different forms of $\psi'_{\hat{\theta}}$ that arise in different cases.

Lemma 3.6 ($\psi_{\hat{\theta}}$ for A Continuous-Variable Pair)**.** *For two continuous variables X_{j_1} and X_{j_2}, $\psi_{\hat{\theta}}$ is defined as*

$$\psi_{\hat{\theta}} = \hat{\sigma}^i_{j_1,j_2} - \hat{\sigma}_{j_1,j_2}. \tag{5}$$

Proposition 3.7 ($\psi_{\hat{\theta}}$ for A Discretized-Variable Pair)**.** *For discretized variables $\tilde{X}_{j_1}$ and $\tilde{X}_{j_2}$, $\psi_{\hat{\theta}}$ is defined as*

$$\psi_{\hat{\theta}} = \begin{pmatrix} \hat{\tau}^i_{j_1,j_2} - T(\hat{\sigma}_{j_1,j_2}; \{\hat{h}_{j_1}, \hat{h}_{j_2}\}) \\ \hat{\tau}^i_{j_1} - \bar{\Phi}(\hat{h}_{j_1}) \\ \hat{\tau}^i_{j_2} - \bar{\Phi}(\hat{h}_{j_2}) \end{pmatrix}. \tag{6}$$

Corollary 3.8 ($\psi_{\hat{\theta}}$ for A Continuous-Discretized-Variable Pair)**.** *For one discretized variable $\tilde{X}_{j_2}$ and one continuous variable X_{j_1}, $\psi_{\hat{\theta}}$ is defined as*

$$\psi_{\hat{\theta}} = \begin{pmatrix} \hat{\tau}^i_{j_1,j_2} - T(\hat{\sigma}_{j_1,j_2}; \{0, \hat{h}_{j_2})\} \\ \hat{\tau}^i_{j_1} - \bar{\Phi}(\hat{h}_{j_2}) \end{pmatrix}. \tag{7}$$

Up to this point, our discussion has been confined to the case of covariance σ. In the next section, we will present the results of our proposed conditional independence test.

3.3. Conditional Independence Testing

The real challenge lies in the inference of conditional independence, i.e., the estimation of the precision matrix $\hat{\omega}_{j_1,j_2}$ and the corresponding distribution $\hat{\omega}_{j_1,j_2} - \omega_{j_1,j_2}$. In the following, we first use β_{j_1,j_2} which is gotten using nodewise regression as a substitution of testing for $\omega_{j_1 j_2} = 0$, we then construct the influence function of $\hat{\beta}_{j_1,j_2} - \beta_{j_1,j_2}$ as the combination of influence function of $\hat{\sigma}_{j_1,j_2} - \sigma_{j_1,j_2}$ and show it will also be asymptotically normal.

Preliminary of nodewise regression To utilize covariance for testing conditional independence, it is necessary to establish a relationship between the estimated covariance and a metric capable of reflecting conditional independence. To achieve this, we employ the nodewise regression which effectively builds the connection between covariance and precision matrix. Suppose we can access observations $\{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^n\}$ from latent continuous variables $\boldsymbol{X} = (X_1, \ldots, X_p) \sim N(0, \boldsymbol{\Sigma})$, nodewise regression will do regression on every dimension with all other dimensions as predictors.

$$x^i_{j_1} = \sum_{j_1 \neq j_2} x^i_{j_2} \beta_j + \epsilon^i_{j_1}. \tag{8}$$

It can be shown that there are deterministic relationships between the regression coefficients with the covariance and precision matrix of $\boldsymbol{X}$:

$$\begin{aligned} \beta_j &= \boldsymbol{\Sigma}^{-1}_{-j-j} \boldsymbol{\Sigma}_{-jj} \in \mathbb{R}^{p-1}, \\ \beta_{j,k} &= -\frac{\omega_{jk}}{\omega_{jj}}, \quad j \neq k, \end{aligned} \tag{9}$$

where $\boldsymbol{\Sigma}_{-j-j}$ is the submatrix of $\boldsymbol{\Sigma}$ without jth column and jth row, and the $\boldsymbol{\Sigma}_{-jj}$ is the vector of jth column without jth row. $\beta_{j,k} \in \mathbb{R}$ is the surrogate of ω_{jk} to capture the independence relationship of X_j with X_k conditioning on other variables. We can use the definitions 3.1, 3.3 and 3.4 to get the estimation $\hat{\boldsymbol{\Sigma}}_{-j-j}$ and $\hat{\boldsymbol{\Sigma}}_{-jj}$ and thus get the estimation $\hat{\beta}_j$. The proof of Eq (9) can be found in Appendix A.7.1.

Statistical Inference for ω_{j_1,j_2} Nodewise regression offers a robust 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. The underlying relationship between these variables is as follows:

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

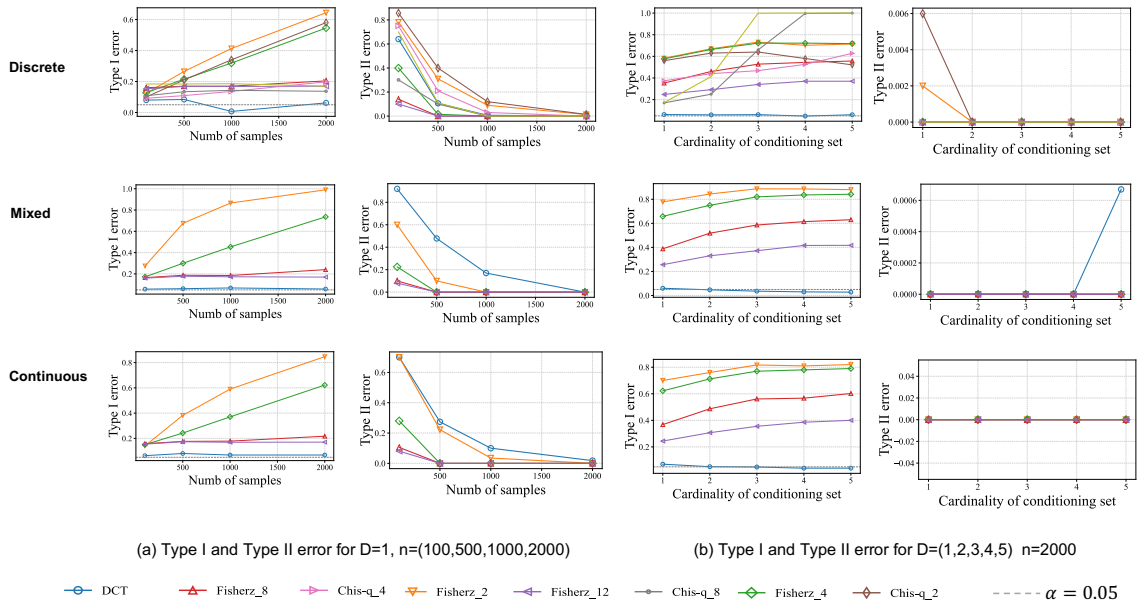


Figure 2: Comparison of results of Type I and Type II error (1 minus power) for all three types of tested data (continuous, mixed, discrete) and different number of samples and cardinality of conditioning set. The suffix attached to a test's name denotes the cardinality of discretization; for example, "Fsherz_4" signifies the application of the Fsher-z test to data discretized into four levels. Chi-square test is only applicable for the discrete case.

For ease of notation, we express the distribution of the difference between the estimated covariance with the true covariance as

$$\hat{\sigma}_{j_1,j_2} - \sigma_{j_1,j_2} = \frac{1}{n}\sum_{i=1}^{n} \xi^i_{j_1,j_2} \tag{10}$$

according to Eq. (4), which is asymptotically normal. The whole vector of $\hat{\beta}_j - \beta_j$ is nothing but a linear combination of dependent Gaussian variables. For notational convenience, we express $\hat{\mathbf{\Sigma}}_{-j-j} - \mathbf{\Sigma}_{-j-j} = \frac{1}{n}\sum_{i=1}^{n} \Xi^i_{-j,-j}$ and $\hat{\mathbf{\Sigma}}_{-jj} - \mathbf{\Sigma}_{-jj} = \frac{1}{n}\sum_{i=1}^{n} \Xi^i_{-j,j}$ as the matrix form of difference between estimated covariance with the true one. Consequently, we have

Theorem 3.9 (Conditional Independence test)**.** *In our settings, under the null hypothesis that X_j and X_k are conditional statistically independent given a set of variables $\boldsymbol{Z}$, the statistic*

$$\hat{\beta}_{j,k} = (\hat{\mathbf{\Sigma}}^{-1}_{-j-j}\hat{\mathbf{\Sigma}}_{-jj})_{[k]}, \tag{11}$$

where $[k]$ denote the corresponding elements of variable X_k in $\hat{\mathbf{\Sigma}}^{-1}_{-j-j}\hat{\mathbf{\Sigma}}_{-jj}$. $\hat{\beta}_{j,k}$ has the asymptotic distribution:

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

where

$$B^i = \begin{bmatrix} \Xi^i_{-j,-j} \\ \Xi^i_{-j,j} \end{bmatrix} \text{ and}$$

$$a^{[k]}_l = \begin{cases} \left(\hat{\mathbf{\Sigma}}^{-1}_{-j-j}\right)_{[k],l}, & \text{for } l \in \{1,\dots,p-1\} \\ \sum_{q=1}^{n}\left(\hat{\mathbf{\Sigma}}^{-1}_{-j-j}\right)_{[k],l}\left(\tilde{\beta}_j\right)_q, & \text{for } l \in \{p,\dots,p^2-p\} \end{cases}$$

and $\tilde{\beta}_j$ is β_j whose $\beta_{j,k} = 0$.

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 conditional independence test. We leave the detailed derivation in Appendix A.7.2,

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 Appendix B.

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 condi-

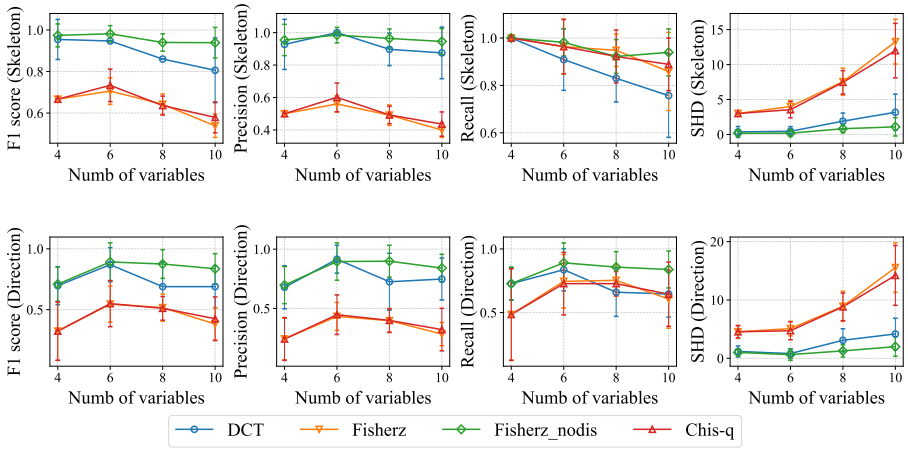


Figure 3: Experimental comparison of causal discovery on synthetic datasets with changing number of nodes p=(4,6,8,10) and fixed samples n=5000. "Fisherz_nodis" is the Fisher-z test applied to original continuous data. We evaluate F_1 (↑), Precision (↑), Recall (↑) and SHD (↓) on both skeleton and DAG.

tions. 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 $D = (1, 2, \ldots, 5)$. It is assumed that every variable within this conditioning set is effective, i.e., they influence the conditional independence of the variables under test. 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. For this, we first generate the true continuous linear Gaussian data Y, W from Z_i ($i = 1$ for the first case, summed over n for the second); they were constructed as the form $a_i Z_i + E$ ($\sum_{i=1}^{n} a_i Z_i + E$ respectively), where a_i is randomly sampled from $U(0.5, 1.5)$, E follows a standard normal distribution independent with all other variables. Hence, $Y \perp\!\!\!\perp W|Z$. These data then undergo multi-level discretization into $K = (2, 4, 8, 12)$ distinct values, with discretization boundaries randomly generated based on the variable range. The first column in Figure 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 linear 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. Hence, we construct the relationship $Y \not\perp\!\!\!\perp W|Z$. The same discretization approach is applied here. The second column in Figure 2 (a) (b) shows the Type II error with the changing number of samples and cardinality of the conditioning set compared with other methods.

We can see that with derived null distribution, the Type I errors of quantity 0.05 are approximated very well for all three data types of both scenarios from Figure 2 (a), while other approaches exhibit significantly higher Type I error rates, these rates escalate with the increase in both the number of samples and the size of the conditioning set. In essence, such methods are more prone to erroneously conclude that tested variables are conditionally dependent.

Another notable observation is that compared to alternative tests that demonstrate considerable power even with smaller sample sizes—a reasonable outcome given these tests' propensity to judge tested variables as conditionally dependent—our approach does not exhibit satisfactory power until the sample size reaches 2000. This is especially true for mixed and continuous cases. An explanation for this phenomenon is that our method binarizes discretized data, which may not effectively utilize all observations. This aspect could be considered for future research. Moreover, our test displays 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

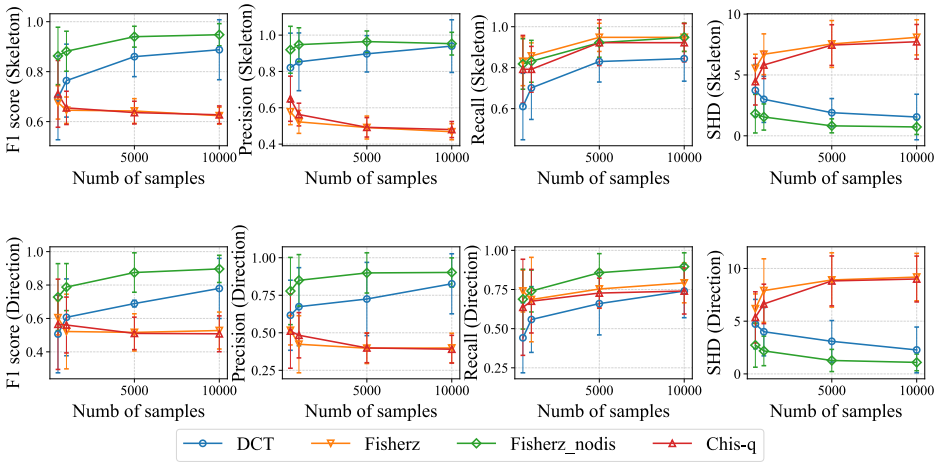


Figure 4: Experimental comparison of causal discovery on synthetic datasets with fixed nodes p=8 and changing sample size n=(500, 1000, 5000, 10000). "Fisherz_nodis" is the Fisher-z test applied to original continuous data. We evaluate F_1 (↑), Precision (↑), Recall (↑) and SHD (↓) on both skeleton and DAG.

condition that the causal graph of variables $X_1, \ldots, X_p$ 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 linear 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.

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 $n = (500, 1000, 5000, 10000)$.

Comparative analysis is conducted using the PC algorithm integrated with various testing methods. Specifically, we compare DCT against the Fsher-z test applied to discretized data, the chi-square test, and the Fsher-z test on original continuous data, the latter serving as a theoretical upper bound for comparison. As long as the PC can only return a completed partially directed acyclic graph (CPDAG). To ease the comparison, we use the same orientation rules (Dor & Tarsi, 1992) implemented by Causal-DAG to convert a CPDAG (Chandler Squires, 2018) into a DAG. We evaluate both the undirected skeleton and the directed graph with the criterion of structural Hamming distance (SHD), the F1 score, the precision and the recall. For each setting, we run 10 graph instances with different seeds and report the mean and standard deviation in Figure 3 and Figure 4.

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 provide a novel testing method to uncover the latent conditional independence given only the discretized observations of a linear Gaussian model. Compared with existing CI tests, our testing effectively serves as a correction by avoiding the erroneous judgment caused by discretization. Both theoretical proofs and experiments demonstrate the validity of our test methods.

Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. 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.

Asratian, A. S., Denley, T. M., and Häggkvist, R. *Bipartite graphs and their applications*, volume 131. Cambridge university press, 1998.

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

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

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

Doran, G., Muandet, K., Zhang, K., and Schölkopf, B. A permutation-based kernel conditional independence test. In *UAI*, pp. 132–141, 2014.

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

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

F.R.S., K. P. 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.

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

Li, L., Ng, I., Luo, G., Huang, B., Chen, G., Liu, T., Gu, B., and Zhang, K. Federated causal discovery from heterogeneous data, 2024.

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

Mohan, K., Chung, M., Han, S., Witten, D., Lee, S.-I., and Fazel, M. Structured learning of gaussian graphical models. *Advances in neural information processing systems*, 25, 2012.

Pearl, J. et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000.

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

Ren, Z., Sun, T., Zhang, C.-H., and Zhou, H. H. Asymptotic normality and optimalities in estimation of large gaussian graphical models. 2015.

Sen, R., Suresh, A. T., Shanmugam, K., Dimakis, A. G., and Shakkottai, S. Model-powered conditional independence test. *Advances in neural information processing systems*, 30, 2017.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvarinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. Directlingam: A direct method for learning a linear non-gaussian structural equation model, 2011.

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

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

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

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

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

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

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

Zhang, Y., Zhang, Z., Liu, K., and Qian, G. An improved iamb algorithm for markov blanket discovery. *J. Comput.*, 5(11):1755–1761, 2010.

A. Proof of Things

A.1. Proof of $\hat{\theta} \xrightarrow{p} \theta_0$

Lemma A.1. *For the estimation $\hat{\theta}$ which is calculated using bridge equation 3.1 3.3 3.4, as a zero of Ψ_n defined in Eq (25),(31), (34) , 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

$$n \to \infty, \quad \hat{\tau}_j = \frac{1}{n}\sum_{i=1}^{n} X_{i,j} \xrightarrow{p} E(X_j) = P(X_j = 1) = \tau_j. \tag{12}$$

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

$$n \to \infty, \quad \hat{\tau}_{j_1,j_2} = \frac{1}{n}\sum_{i=1}^{n} X_{i,j_1} X_{i,j_2} \xrightarrow{p} E(X_{j_1} X_{j_2}) = P(X_{j_1} = 1, X_{j_2} = 1) = \tau_{j_1,j_2}. \tag{13}$$

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

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 $(\hat{\sigma}_{j_1,j_2}, \hat{h}_{j_2}) \xrightarrow{p} (\sigma_{j_1,j_2}, h_{j_2})$.

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

$$n \to \infty, \quad \hat{\sigma}_{j_1,j_2} = \frac{1}{n}\sum_{i=1}^{n} x_{j_1}^i x_{j_2}^i - \frac{1}{n}\sum_{i=1}^{n} x_{j_1}^i \frac{1}{n}\sum_{i=1}^{n} x_{j_2}^i \xrightarrow{p} E(X_{j_1} X_{j_2}) - E(X_{j_1})E(X_{j_2}) = \sigma_{j_1,j_2}. \tag{14}$$

A.2. Proof of one-to-one mapping between $\hat{\tau}_{j_1,j_2}$ with $\hat{\sigma}_{j_1,j_2}$

Lemma A.2. *For any fixed $\hat{h}_{j_1}$ and $\hat{h}_{j_2}$, $T(\sigma; \{\hat{h}_{j_1}, \hat{h}_{j_2}\}) = \int_{x_1 > \hat{h}_{j_1}} \int_{x_2 > \hat{h}_{j_2}} \phi(x_{j_1}, x_{j_2}; \sigma) dx_{j_1} dx_{j_2}$, is a strictly monotonically increasing function on $\sigma \in (-1, 1)$.*

Proof To prove the Lemma, we just need to show the gradient $\frac{\partial T(\sigma; \{\hat{h}_{j_1}, \hat{h}_{j_2}\}}{\partial \sigma} > 0$ for $\sigma \in (-1, 1)$.

$$\frac{\partial T(\sigma; \{\hat{h}_{j_1}, \hat{h}_{j_2}\}}{\partial \sigma} == \frac{1}{2\pi\sqrt{(1-\sigma^2)}} \exp\left(-\frac{(\hat{h}_{j_1}^2 - 2\sigma \hat{h}_{j_1}\hat{h}_{j_2} + \hat{h}_{j_2}^2)}{2(1-\sigma^2)}\right), \tag{15}$$

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

A.3. Proof of Theorem 3.5

In this section, we provide the proof of Theorem 3.5, 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 true distribution P, 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

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \psi_\theta(\boldsymbol{x}_i) = \mathbb{P}_n \psi_\theta. \tag{16}$$

We further specify the form

$$\Psi(\theta) = \int \psi_\theta(\boldsymbol{x}) d\boldsymbol{x} = P\psi_\theta. \tag{17}$$

Assume the estimator $\hat{\theta}$ is a zero of Ψ_n, i.e., $\Psi_n(\hat{\theta}) = 0$ and will converge 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

$$0 = \Psi_n(\hat{\theta}) = \Psi_n(\theta_0) + (\hat{\theta} - \theta_0)\Psi_n'(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)\Psi_n''(\theta_0). \tag{18}$$

Rearrange the equation above, we have

$$\begin{aligned} \hat{\theta} - \theta_0 &= -\frac{\Psi_n(\theta_0)}{\Psi_n'(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)\Psi_n''(\theta_0)} \\ &= -\frac{\frac{1}{n}\sum_{i=1}^n \psi_\theta(\boldsymbol{x}_i)}{\Psi_n'(\theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)\Psi_n''(\theta_0)}. \end{aligned} \tag{19}$$

According to the central limit theorem, the numerator will be asymptotic normal with variance $P\psi_{\theta_0}^2/n$ 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 $\hat{\theta} - \theta_0 = o_P(1)$. [1] As long as the denominator converges in probability and the numerator converges in distribution, according to Slusky's lemma, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow N\left(0, \frac{P\psi_{\theta_0}^2}{(P\psi_{\theta_0}')^2}\right). \tag{20}$$

Extend into the high-dimensional case we should have

$$\hat{\theta} - \theta_0 = -(\Psi_n'(\theta_0))^{-1}\Psi_n(\theta_0), \tag{21}$$

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

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow N\left(0, (P\psi_{\theta_0}')^{-1} P\psi_{\theta_0}\psi_{\theta_0}^T (P\psi_{\theta_0}'^T)^{-1}\right), \tag{22}$$

Specifically, in our case $\theta = (\sigma_{j_1,j_2}, \Lambda)$, where Λ is another parameter set influencing the estimation of σ_{j_1,j_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

$$\hat{\sigma}_{j_1,j_2} \overset{\text{approx}}{\sim} N\left(0, ((\mathbb{P}_n\psi_{\hat{\theta}}')^{-1}\mathbb{P}_n\psi_{\hat{\theta}}\psi_{\hat{\theta}}^T(\mathbb{P}_n\psi_{\hat{\theta}}'^T)^{-1})_{1,1}\right), \tag{23}$$

which should converge to Eq. (22) when $n \to \infty$.

A.4. Proof of Proposition 3.7

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 true distribution P, 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

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n \psi_\theta(\boldsymbol{x}_i) = \mathbb{P}_n\psi_\theta. \tag{24}$$

The Eq. (22) states that under certain conditions, the difference between the estimation $\hat{\theta}$ with the true parameters θ_0 will be asymptotically normal with specific variance only related to the true parameter. Let's first focus on the most challenging case where both variables are discretized observations and our interested parameter will include $\hat{\theta} = (\hat{\sigma}_{j_1,j_2}, \hat{h}_{j_1}, \hat{h}_{j_2})$ (Although we only care about the distribution of $\hat{\sigma}_{j_1,j_2} - \sigma_{j_1,j_2}$, the estimation of boundary $\hat{h}_{j_1}$ and $\hat{h}_{j_2}$ will influence the estimation of $\hat{\sigma}_{j_1,j_2}$, thus we need to consider all of them).

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

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

$$\Psi_n(\hat{\theta}) = \begin{pmatrix} \Psi_n(\hat{\sigma}_{j_1,j_2}) \\ \Psi_n(\hat{h}_{j_1}) \\ \Psi_n(\hat{h}_{j_2}) \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \psi_{\hat{\theta}}(\boldsymbol{x}^i) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \hat{\tau}^i_{j_1,j_2} - T(\hat{\sigma}_{j_1,j_2}; \{\hat{h}_{j_1}, \hat{h}_{j_2}\}) \\ \hat{\tau}^i_{j_1} - \bar{\Phi}(\hat{h}_{j_1}) \\ \hat{\tau}^i_{j_2} - \bar{\Phi}(\hat{h}_{j_2}) \end{pmatrix} = \mathbf{0}. \tag{25}$$

$$\Psi_n(\theta_0) = \begin{pmatrix} \Psi_n(\sigma_{j_1,j_2}) \\ \Psi_n(h_{j_1}) \\ \Psi_n(h_{j_2}) \end{pmatrix} = \frac{1}{n}\sum_{i=1}^{n} \psi_{\theta_0}(\boldsymbol{x}^i) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \hat{\tau}^i_{j_1,j_2} - T(\sigma_{j_1,j_2}; \{h_{j_1}, h_{j_2}\}) \\ \hat{\tau}^i_{j_1} - \bar{\Phi}(h_{j_1}) \\ \hat{\tau}^i_{j_2} - \bar{\Phi}(h_{j_2}) \end{pmatrix}. \tag{26}$$

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

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

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

$$\begin{aligned}
\frac{\partial \Psi_n(\sigma_{j_1,j_2})}{\partial \sigma_{j_1,j_2}} &= -\frac{1}{2\pi\sqrt{(1-\sigma^2_{j_1,j_2})}} \exp\left(-\frac{(h^2_{j_1} - 2\sigma_{j_1,j_2}h_{j_1}h_{j_2} + h^2_{j_2})}{2(1-\sigma^2_{j_1,j_2})}\right); \\
\frac{\partial \Psi_n(\sigma_{j_1,j_2})}{\partial h_{j_1}} &= \int_{h_{j_2}}^{\infty} \frac{1}{2\pi\sqrt{1-\sigma^2_{j_1,j_2}}} \exp\left(-\frac{h^2_{j_1} - 2\sigma_{j_1,j_2}h_{j_1}x_2 + x^2_2}{2(1-\sigma^2_{j_1,j_2})}\right) dx_2; \\
\frac{\partial \Psi_n(\sigma_{j_1,j_2})}{\partial h_{j_2}} &= \int_{h_{j_1}}^{\infty} \frac{1}{2\pi\sqrt{1-\sigma^2_{j_1,j_2}}} \exp\left(-\frac{h^2_{2} - 2\sigma_{j_1,j_2}h_{j_2}x_1 + x^2_1}{2(1-\sigma^2_{j_1,j_2})}\right) dx_1; \\
\frac{\partial \Psi_n(h_{j_1})}{\partial \sigma_{j_1,j_2}} &= 0; \\
\frac{\partial \Psi_n(h_{j_1})}{\partial h_{j_1}} &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{h^2_{j_1}}{2}\right); \\
\frac{\partial \Psi_n(h_{j_1})}{\partial h_{j_2}} &= 0; \\
\frac{\partial \Psi_n(h_{j_2})}{\partial \sigma_{j_1,j_2}} &= 0; \\
\frac{\partial \Psi_n(h_{j_2})}{\partial h_{j_1}} &= 0; \\
\frac{\partial \Psi_n(h_{j_2})}{\partial h_{j_2}} &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{h^2_{j_2}}{2}\right).
\end{aligned} \tag{27}$$

For simplicity of notation, we define

$$\hat{\sigma}_{j_1,j_2} - \sigma_{j_1,j_2} = \frac{1}{n}\sum_{i=1}^{n} \xi^i_{j_1,j_2}, \tag{28}$$

where the specific form is of $\{\xi^i_{j_1,j_2}\}$ is defined in Eq. (A.4). We should note that $\{\xi^i_{j_1,j_2}\}$ are i.i.d random variables with mean zero (this property will be the key to the derivation of inference of conditional independence). As long as our estimation $\hat{\theta}$ converge in probability to θ_0 as proved in A.1, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow N\left(0, ((P\psi'_{\theta_0})^{-1} P\psi_{\theta_0}\psi^T_{\theta_0}(P\psi'^T_{\theta_0})^{-1})_{1,1}\right), \tag{29}$$

where ψ_{θ_0} is defined in Eq. (26). However, in practice, we don't have access to either P or θ_0. 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

$$\frac{1}{n}\left((\mathbb{P}_n\psi'_{\hat{\theta}})^{-1}\mathbb{P}_n\psi_{\hat{\theta}}\psi_{\hat{\theta}}^T(\mathbb{P}_n\psi_{\hat{\theta}}'^T)^{-1}\right)_{1,1}. \tag{30}$$

A.5. Proof of Corollary 3.8

Use the same line of procedure as in proof of Proposition 3.7, for mixed pair of observations where X_{j_1} is continuous and X_{j_2} is discrete, we can construct the criterion function

$$\Psi_n(\hat{\theta}) = \begin{pmatrix}\Psi_n(\hat{\sigma}_{j_1,j_2})\\ \Psi_n(\hat{h}_{j_2})\end{pmatrix} = \frac{1}{n}\sum_{i=1}^n \psi_{\hat{\theta}}(\boldsymbol{x}^i) = \frac{1}{n}\sum_{i=1}^n \begin{pmatrix}\hat{\tau}^i_{j_1,j_2} - T(\hat{\sigma}_{j_1,j_2};\{0,\hat{h}_{j_2}\})\\ \hat{\tau}^i_{j_2} - \bar{\Phi}(\hat{h}_{j_2})\end{pmatrix} = \mathbf{0}. \tag{31}$$

$$\Psi_n(\theta_0) = \begin{pmatrix}\Psi_n(\sigma_{j_1,j_2})\\ \Psi_n(h_{j_2})\end{pmatrix} = \frac{1}{n}\sum_{i=1}^n \psi_{\theta_0}(\boldsymbol{x}^i) = \frac{1}{n}\sum_{i=1}^n \begin{pmatrix}\hat{\tau}^i_{j_1,j_2} - T(\sigma_{j_1,j_2};\{0,h_{j_2}\})\\ \hat{\tau}^i_{j_2} - \bar{\Phi}(h_{j_2})\end{pmatrix}. \tag{32}$$

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

$$\hat{\theta} - \theta_0 = \begin{pmatrix}\hat{\sigma}_{j_1,j_2} - \sigma_{j_1,j_2}\\ \hat{h}_{j_2} - h_{j_2}\end{pmatrix} = -\frac{1}{n}\sum_{i=1}^n \begin{pmatrix}\frac{\partial\Psi_n(\sigma_{j_1,j_2})}{\partial\sigma_{j_1,j_2}} & \frac{\partial\Psi_n(\sigma_{j_1,j_2})}{\partial h_{j_2}}\\ \frac{\partial\Psi_n(h_{j_2})}{\partial\sigma_{j_1,j_2}} & \frac{\partial\Psi_n(h_{j_2})}{\partial h_{j_2}}\end{pmatrix}^{-1}\begin{pmatrix}\hat{\tau}^i_{j_1,j_2} - T(\sigma_{j_1,j_2};\{0,h_{j_2}\})\\ \hat{\tau}^i_{j_2} - \bar{\Phi}(h_{j_2}).\end{pmatrix}, \tag{33}$$

where the specific form of each entry of the gradient matrix can be found in Eq. (27). Using exactly the same procedure, we should have the same formation of the variance calculated as Eq. (30) with a different definition of ψ_{θ_0} and $\psi_{\hat{\theta}}$ defined in Eq. (32) (31).

A.6. Proof of Lemma 3.6

Use the same line of procedure as in Proof of Proposition 3.7, for a continuous pair of variables, we can construct the criterion function

$$\Psi_n(\hat{\theta}) = \Psi_n(\hat{\sigma}_{j_1,j_2}) = \frac{1}{n}\sum_{i=1}^n x^i_{j_1}x^i_{j_2} - \frac{1}{n}\sum_{i=1}^n x^i_{j_1}\frac{1}{n}\sum_{i=1}^n x^i_{j_2} - \hat{\sigma}_{j_1,j_2} = 0. \tag{34}$$

$$\Psi_n(\theta_0) = \Psi_n(\sigma_{j_1,j_2}) = \frac{1}{n}\sum_{i=1}^n x^i_{j_1}x^i_{j_2} - \frac{1}{n}\sum_{i=1}^n x^i_{j_1}\frac{1}{n}\sum_{i=1}^n x^i_{j_2} - \sigma_{j_1,j_2}. \tag{35}$$

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

$$\hat{\sigma}_{j_1,j_2} - \sigma_{j_1,j_2} = \frac{1}{n}\sum_{i=1}^n x^i_{j_1}x^i_{j_2} - \bar{x}_{j_1}\bar{x}_{j_2} - \sigma_{j_1,j_2}. \tag{36}$$

According to Eq. (20), we have

$$\sqrt{n}(\hat{\sigma}_{j_1,j_2} - \sigma_{j_1,j_2}) \rightsquigarrow N\left(0, \frac{P\psi^2_{\theta_0}}{(P\psi'_{\theta_0})^2}\right). \tag{37}$$

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

$$\frac{1}{n}\mathbb{P}_n\psi^2_{\hat{\theta}}/(\mathbb{P}_n\psi'_{\hat{\theta}})^2 = \frac{1}{n^2}\sum_{i=1}^n(x^i_{j_1}x^i_{j_2} - \bar{x}_{j_1}\bar{x}_{j_2} - \hat{\sigma}_{j_1,j_2})^2. \tag{38}$$

A.7. Proof of Theorem 3.9

A.7.1. Proof of Relation between β, $\mathbf{\Sigma}$ with Ω

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

$$X_j = X_{-j}\beta_j + \epsilon_j. \tag{39}$$

We can divide its covariance $\mathbf{\Sigma}$ and its precision matrix $\Omega = \mathbf{\Sigma}^{-1}$ into X and Y part in our regression:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{jj} & \mathbf{\Sigma}_{j-j} \\ \mathbf{\Sigma}_{-jj} & \mathbf{\Sigma}_{-j-j} \end{pmatrix} \quad \mathbf{\Omega} = \begin{pmatrix} \mathbf{\Omega}_{jj} & \mathbf{\Omega}_{j-j} \\ \mathbf{\Omega}_{-jj} & \mathbf{\Omega}_{-j-j} \end{pmatrix} . \tag{40}$$

Just like regular linear regression, we can get

$$n \to \infty, \quad \beta_j = \mathbf{\Sigma}_{-j-j}^{-1} \mathbf{\Sigma}_{-jj}. \tag{41}$$

From the invertibility of a block matrix

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}. \tag{42}$$

If A and D is invertible, we will have

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & (D - CA^{-1}B)^{-1} \end{bmatrix} \begin{bmatrix} I & -BD^{-1} \\ -CA^{-1} & I \end{bmatrix}. \tag{43}$$

Thus, we can get:

$$\begin{aligned} \mathbf{\Omega}_{jj} &= \mathbf{\Sigma}_{jj} - (\mathbf{\Sigma}_{j-j}\mathbf{\Sigma}_{-j-j}^{-1}\mathbf{\Sigma}_{-jj})^{-1}; \\ \mathbf{\Omega}_{j-j} &= -\left(\mathbf{\Sigma}_{jj} - (\mathbf{\Sigma}_{j-j}\mathbf{\Sigma}_{-j-j}^{-1}\mathbf{\Sigma}_{-jj})^{-1}\right)\mathbf{\Sigma}_{j-j}(\mathbf{\Sigma}_{-j-j})^{-1}. \end{aligned} \tag{44}$$

Move one step forward:

$$-\mathbf{\Omega}_{jj}^{-1}\mathbf{\Omega}_{j-j} = \mathbf{\Sigma}_{j-j}(\mathbf{\Sigma}_{-j-j})^{-1}. \tag{45}$$

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

$$-\mathbf{\Omega}_{jj}^{-1}\mathbf{\Omega}_{-jj} = \mathbf{\Sigma}_{-j-j}^{-1}\mathbf{\Sigma}_{-jj} = \beta_j. \tag{46}$$

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 conditional independence of X_j with other variables. As long as the variable $\mathbf{\Omega}_{jj}$ is just one scalar, we can get

$$\beta_{j,k} = -\frac{\omega_{jk}}{\omega_{jj}} \tag{47}$$

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

A.7.2. 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 $\hat{\sigma}_{j_1,j_2} - \sigma_{j_1,j_2}$ 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 $\hat{\sigma}_{j_1,j_2} - \sigma_{j_1,j_2}$ to establish such an inference. Specifically, we have

$$\begin{aligned} \hat{\beta}_j - \beta_j &= \hat{\mathbf{\Sigma}}_{-j-j}^{-1}\hat{\mathbf{\Sigma}}_{-jj} - \mathbf{\Sigma}_{-j-j}^{-1}\mathbf{\Sigma}_{-jj} \\ &= \hat{\mathbf{\Sigma}}_{-j-j}^{-1}\left(\hat{\mathbf{\Sigma}}_{-jj} - \hat{\mathbf{\Sigma}}_{-j-j}\mathbf{\Sigma}_{-j-j}^{-1}\mathbf{\Sigma}_{-jj}\right) \\ &= -\hat{\mathbf{\Sigma}}_{-j-j}^{-1}\left(\hat{\mathbf{\Sigma}}_{-j-j}\beta_j - \mathbf{\Sigma}_{-j-j}\beta_j + \mathbf{\Sigma}_{-j-j}\beta_j - \hat{\mathbf{\Sigma}}_{-jj}\right) \\ &= -\hat{\mathbf{\Sigma}}_{-j-j}^{-1}\left((\hat{\mathbf{\Sigma}}_{-j-j} - \mathbf{\Sigma}_{-j-j})\beta_j - (\hat{\mathbf{\Sigma}}_{-jj} - \mathbf{\Sigma}_{-jj})\right), \end{aligned} \tag{48}$$

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 conditional independence of the variable X_1 with other variables, $j = 1$, then

$$\hat{\mathbf{\Sigma}}_{-j-j} - \mathbf{\Sigma}_{-j-j} = \begin{bmatrix} \hat{\sigma}_{1,1} \dots \hat{\sigma}_{1,j-1}, \hat{\sigma}_{1,j+1} \dots \hat{\sigma}_{1,p} \\ \dots \\ \hat{\sigma}_{j-1,1} \dots \hat{\sigma}_{j-1,j-1}, \hat{\sigma}_{j-1,j+1} \dots \hat{\sigma}_{j-1,p} \\ \dots \\ \hat{\sigma}_{p,1} \dots \hat{\sigma}_{p,j-1}, \hat{\sigma}_{p,j+1} \dots \hat{\sigma}_{p,p} \end{bmatrix} - \begin{bmatrix} \sigma_{1,1} \dots \sigma_{1,j-1}, \sigma_{1,j+1} \dots \sigma_{1,p} \\ \dots \\ \sigma_{j-1,1} \dots \sigma_{j-1,j-1}, \sigma_{j-1,j+1} \dots \sigma_{j-1,p} \\ \dots \\ \sigma_{p,1} \dots \sigma_{p,j-1}, \sigma_{p,j+1} \dots \sigma_{p,p}. \end{bmatrix}. \quad (49)$$

Suppose that we want to test the conditional independence of the variable x_1 with other variables, $j = 1$. then

$$\hat{\mathbf{\Sigma}}_{-1-1} - \mathbf{\Sigma}_{-1-1} = \begin{bmatrix} \hat{\sigma}_{2,2} \dots \hat{\sigma}_{2,p} \\ \dots \\ \hat{\sigma}_{p,2} \dots \hat{\sigma}_{p,p} \end{bmatrix} - \begin{bmatrix} \sigma_{2,2} \dots \sigma_{2,p} \\ \dots \\ \sigma_{p,2} \dots \sigma_{p,p} \end{bmatrix} \quad (50)$$

$$:= \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} \xi^i_{2,2} \dots \xi^i_{2,p} \\ \dots \\ \xi^i_{p,2} \dots \xi^i_{p,p} \end{bmatrix}, \quad (51)$$

where $\{\xi^i_{j_1,j_2}\}$ are i.i.d random variables with specific form defined in Eq. (A.4) for discrete case, Eq. (33) for mixed case and Eq. (36) in continuous case. Put them together:

$$\begin{bmatrix} \hat{\beta}_{1,2} - \beta_{1,2} \\ \hat{\beta}_{1,3} - \beta_{1,3} \\ \dots \\ \hat{\beta}_{1,p} - \beta_{1,p} \end{bmatrix} = -\hat{\mathbf{\Sigma}}^{-1}_{-1-1} \frac{1}{n} \sum_{i=1}^{n} \left(\begin{bmatrix} \xi^i_{2,2} & \xi^i_{2,3} & \dots & \xi^i_{2,p} \\ \xi^i_{3,2} & \xi^i_{3,3} & \dots & \xi^i_{3,p} \\ \dots & \dots & \dots & \dots \\ \xi^i_{p,2} & \xi^i_{p,3} & \dots & \xi^i_{p,p} \end{bmatrix} \begin{bmatrix} \beta_{1,2} \\ \beta_{1,3} \\ \dots \\ \beta_{1,p} \end{bmatrix} - \begin{bmatrix} \xi^i_{2,1} \\ \xi^i_{3,1} \\ \dots \\ \xi^i_{p,1} \end{bmatrix} \right). \quad (52)$$

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 nothing but 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

$$Var(aY_1 + bY_2) = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} Var(Y_1) & Cov(Y_1, Y_2) \\ Cov(Y_1, Y_2) & Var(Y_2) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}. \quad (53)$$

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

$$\hat{\beta}_{1,2} - \beta_{1,2} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\mathbf{\Sigma}}^{-1}_{-1-1})_{[2],:} \left(\begin{bmatrix} \xi^i_{2,2} & \xi^i_{2,3} & \dots & \xi^i_{2,p} \\ \xi^i_{3,2} & \xi^i_{3,3} & \dots & \xi^i_{3,p} \\ \dots & \dots & \dots & \dots \\ \xi^i_{p,2} & \xi^i_{p,3} & \dots & \xi^i_{p,p} \end{bmatrix} \begin{bmatrix} \beta_{1,2} \\ \beta_{1,3} \\ \dots \\ \beta_{1,p} \end{bmatrix} - \begin{bmatrix} \xi^i_{2,1} \\ \xi^i_{3,1} \\ \dots \\ \xi^i_{p,1} \end{bmatrix} \right), \quad (54)$$

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

$$\Xi^i_{-1,-1} = \begin{bmatrix} \xi^i_{2,2} & \xi^i_{2,3} & \dots & \xi^i_{2,p} \\ \xi^i_{3,2} & \xi^i_{3,3} & \dots & \xi^i_{3,p} \\ \dots & \dots & \dots & \dots \\ \xi^i_{p,2} & \xi^i_{p,3} & \dots & \xi^i_{p,p} \end{bmatrix}, \qquad \Xi^i_{-1,1} = \begin{bmatrix} \xi^i_{2,1} \\ \xi^i_{3,1} \\ \dots \\ \xi^i_{p,1} \end{bmatrix}, \quad (55)$$

and let

$$B^i_{-1} = \begin{pmatrix} \xi^i_{2,1} & \xi^i_{3,1} & \dots & \xi^i_{p,1} \\ \xi^i_{2,2} & \xi^i_{2,3} & \dots & \xi^i_{2,p} \\ \xi^i_{3,2} & \xi^i_{3,3} & \dots & \xi^i_{3,p} \\ \dots & \dots & \dots & \dots \\ \xi^i_{p,2} & \xi^i_{p,3} & \dots & \xi^i_{p,p} \end{pmatrix} \quad (56)$$

as the concatenation of those two matrices. The variance is calculated as

$$Var\left(\sqrt{n}(\hat{\beta}_{j_1,j_2} - \beta_{j_1,j_2})\right) = {a^{[2]}}^T \frac{1}{n} \sum_{i=1}^{n} vec(B^i_{-1}) vec(B^i_{-1})^T a^{[2]}, \quad (57)$$

where

$$a_l^{[2]} = \begin{cases} \left(\hat{\mathbf{\Sigma}}_{-1-1}^{-1}\right)_{[2],l}, & \text{for } l \in \{1, \dots, p-1\} \\ \sum_{q=1}^{n} \left(\hat{\mathbf{\Sigma}}_{-1-1}^{-1}\right)_{[2],l} (\beta_1)_q, & \text{for } l \in \{p, \dots, p^2-p\} \end{cases} \tag{58}$$

$vec(B_{-1}^i)$ is the squeezed vector form of matrix $vec(B_{-1}^i) \in \mathbb{R}^{p \times p-1}$, i.e.,

$$vec(B_{-1}^i) = \begin{pmatrix} \xi_{2,1}^i \\ \xi_{3,1}^i \\ \vdots \\ \xi_{p,p}^i \end{pmatrix}. \tag{59}$$

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

$$\hat{\beta}_{j,k} - \beta_{j,k} \sim N(0, a^{[k]^T} \frac{1}{n^2} \sum_{i=1}^{n} vec(B_{-j}^i) vec(B_{-j}^i)^T) a^{[k]}). \tag{60}$$

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

A.8. Discussion of assumption of $\boldsymbol{X}$

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. We further have the constant d_j' as the discretization boundary such that we have the observation

$$\tilde{X}_j = \mathbf{1}(g_j'(X_j') > d_j') = \mathbf{1}(X_j' > g_j'^{-1}(d_j))$$

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

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

As long as the observation $\tilde{X}_j$ is the same, we should have $\sqrt{b} g_j^{-1}(d_j) + a = g_j'^{-1}(d_j)$. Our assumed model X_j clearly mimics the "groundtruth" X_j'. Besides, according to Lemma A.2, we have one-to-one mapping between $\hat{\tau}_{j_1 j_2}$ 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 $\hat{\sigma}_{j_1, j_2}$ 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 Appendix B.3.

B. Additional experiments

B.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 $X_i = WX + noise$, where $noise$ follows the distribution in (1), (2), (3) correspondingly. The first three rows of Figure 5 and Figure 6 show the result.

For the nonlinear cases, we follow setting in (Li et al., 2024) , where every variable X_i is generated as $X_i = f(WX_{par} + noise)$, $noise \sim N(0, 1)$ 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) $f(x) = ReLU(x)$. 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 Figure 5 and DAG in Figure 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 Figure 5 and Figure 6 show the result.

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.

B.2. Denser graph

DCT primarily works on cases where conditional independence 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 linear 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 Figure 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 Discrete Causality Test (DCT) correspondingly diminishes. Due to the same reason, The recall remains lower than that of other baseline methods.

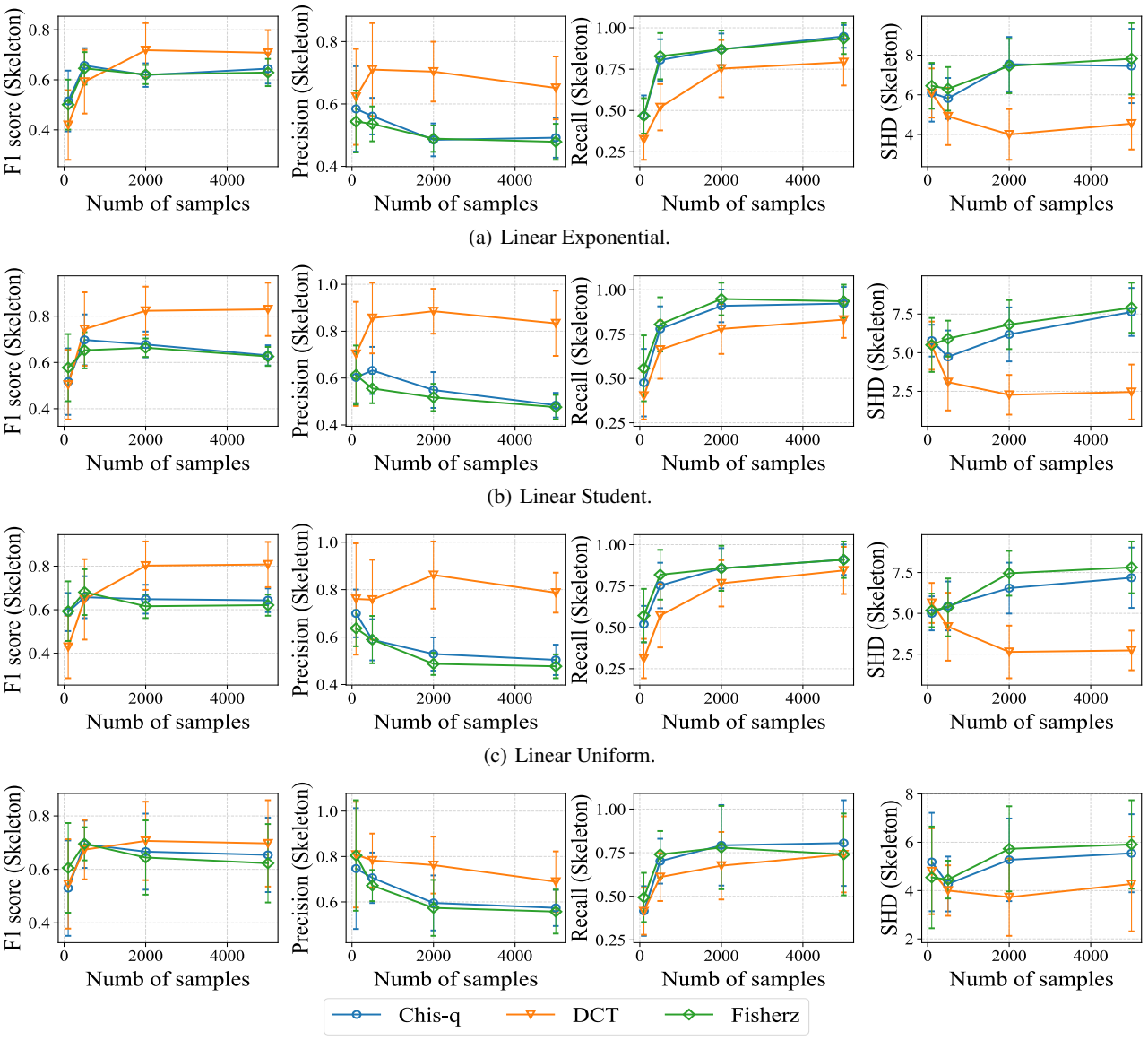
B.3. Linear Gaussian with nonzero mean and non-unit variance

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 Figure 8 shows the result and demonstrates the effectiveness of our method.

B.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 Figure 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



(a) Linear Exponential.

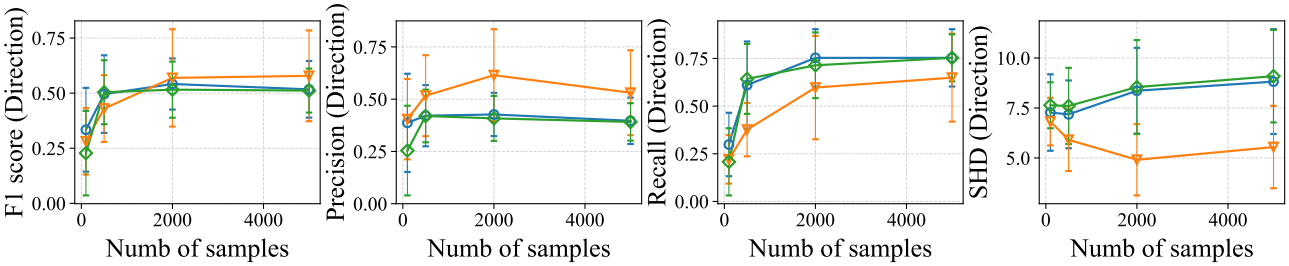
(b) Linear Student.

(c) Linear Uniform.

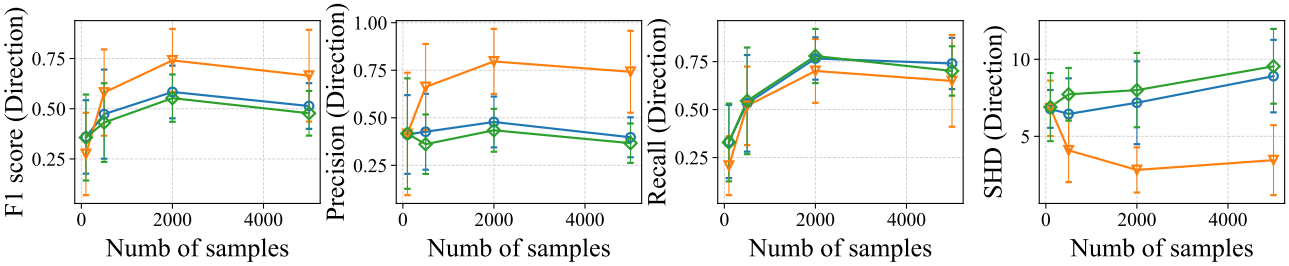
(d) Nonlinear Gaussian.

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.

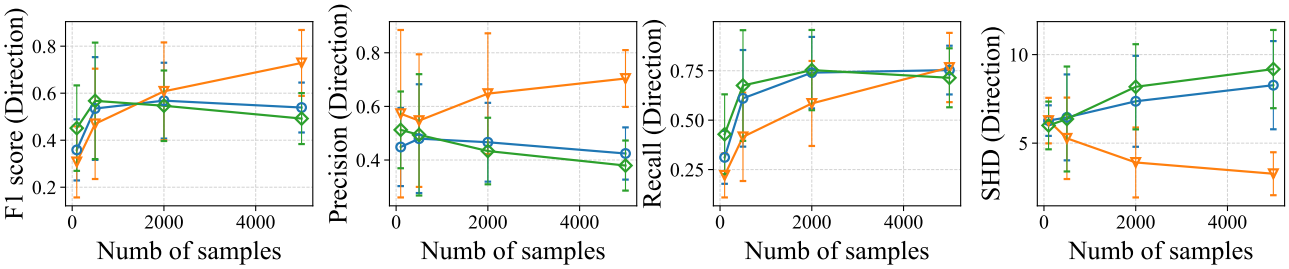
information of 'I seldom feel blue'. As a comparison, both Fisher-Z and Chi-square return a fully connected graph. The results directly correspond to our illustrative example shown in Figure 1, substantiating the necessity of our proposed test.



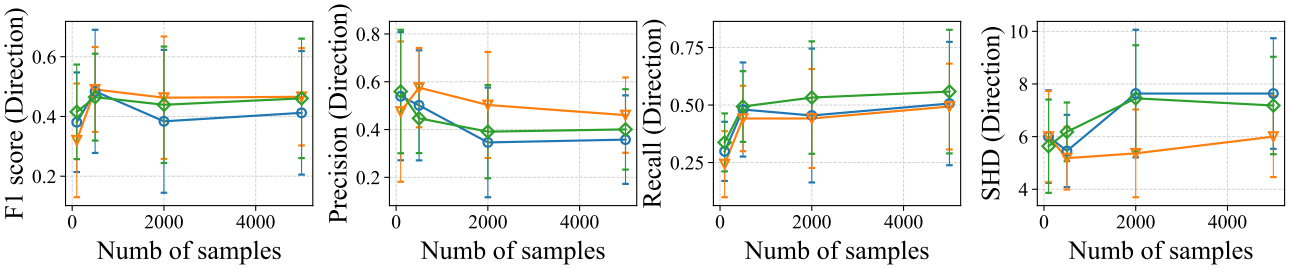
(a) Linear Exponential.



(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.

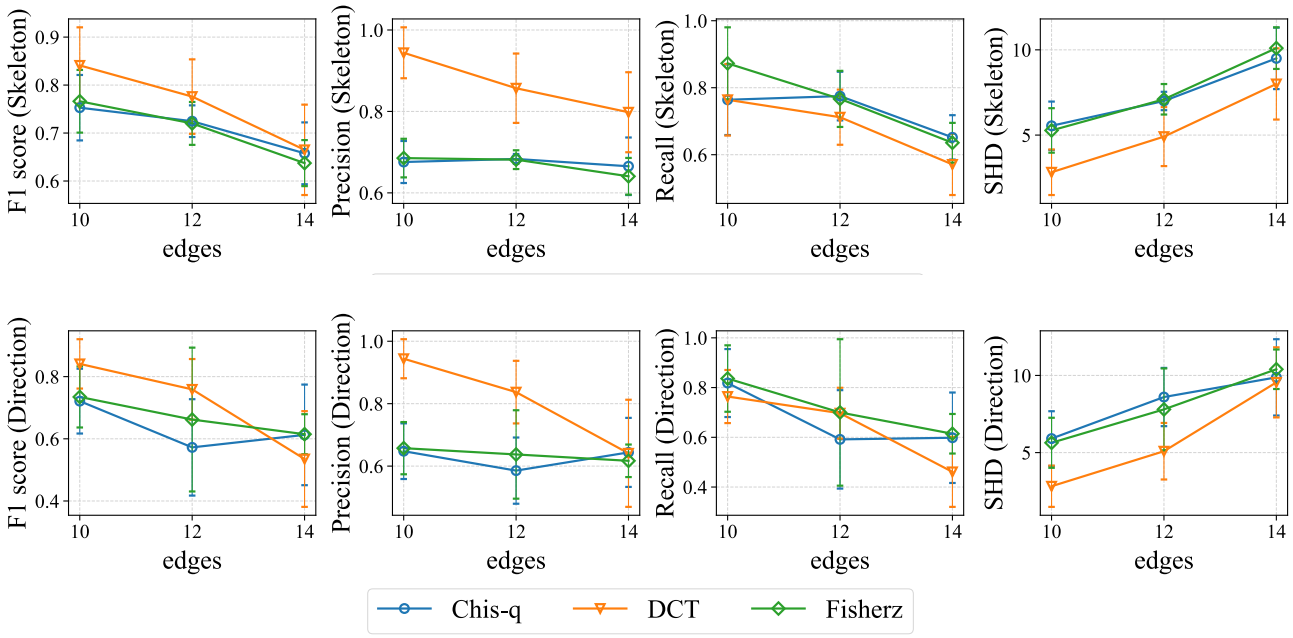


Figure 7: Experimental comparison of causal discovery on synthetic datasets for denser graphs with $p = 8, n = 10000$ and edges varying $p + 2, p + 4, p + 6$. We evaluate F_1 (↑), Precision (↑), Recall (↑) and SHD (↓) on both skeleton and DAG.

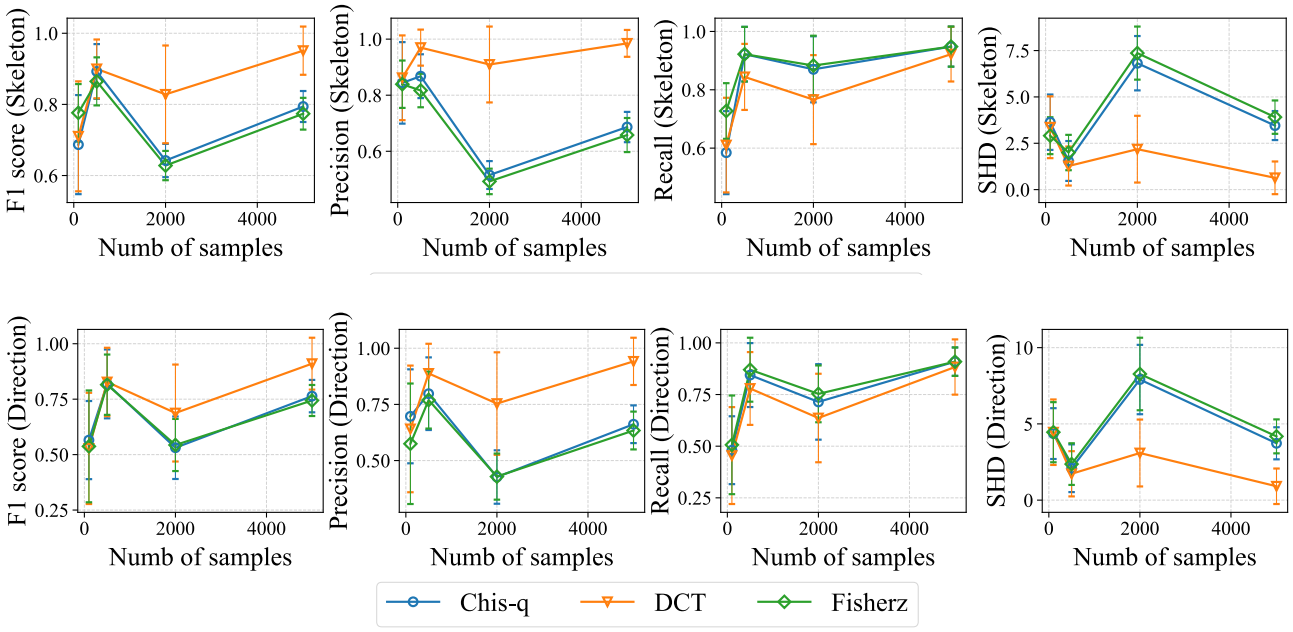
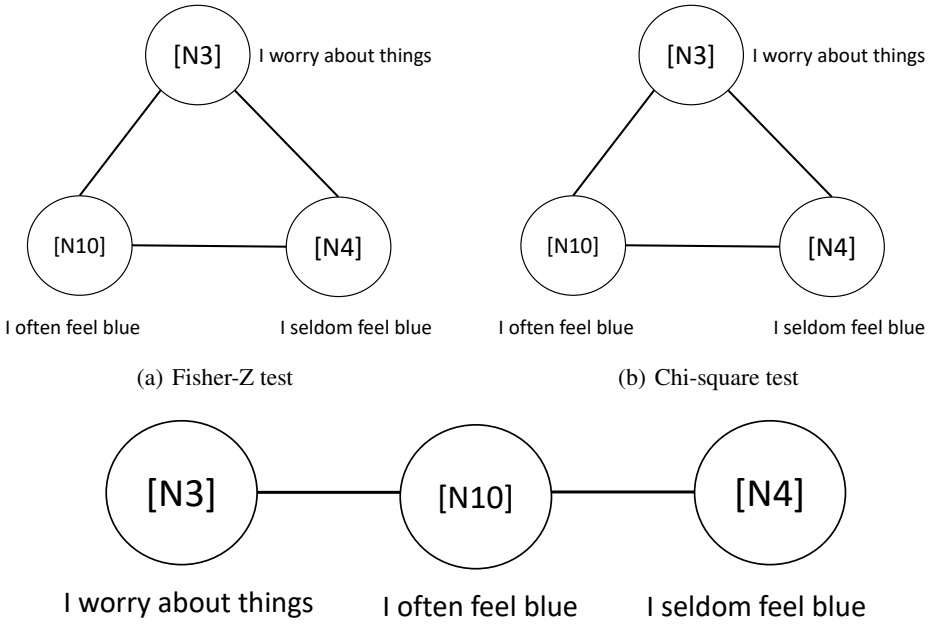


Figure 8: Experimental comparison of causal discovery on synthetic datasets for linear Gaussian model with $p = 8, n = (100, 500, 2000, 5000)$ and where mean is not zero. We evaluate F_1 (↑), Precision (↑), Recall (↑) and SHD (↓) on both skeleton and DAG.



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

(c) DCT

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