Causal Discovery with Unequal Edge Error Tolerance

Joni Shaska
University of Southern California
shaska@usc.edu

Urbashi Mitra
University of Southern California
ubli@usc.edu

***Abstract*—We introduce an algorithm for causal discovery in the setting of linear additive Gaussian noise models that controls the false positive rate of individual edges. The ability to control individual edge errors has only recently started to gain attention. However, recent methods that make guarantees on the existence of edges can only do so in the limit of infinite samples or are computationally infeasible even for modestly sized graphs. By looking at the residual sum of squares between two vertices we can detect the presence of an edge while controlling the false positive rate for a given number of samples. We call our algorithm ϵ-CUT and compare it to other state-of-the-art algorithms. In some instances, ϵ-CUT achieves a 12 percent reduction in the false negative rate for a given false positive rate over the different algorithms.**

***Index Terms*—causal inference, finite-sample results, error control**

I. Introduction

Understanding the underlying causes of phenomena affected by multiple variables can often be done via the representation of causal graphs [1]. These graphs are often assumed to be directed acyclic graphs. Applying causal graph discovery can have utility in disciplines as diverse as topology inference in wireless networks [2], gene networks in biology (e.g. [3]), the impact of medications, and optimizing the effect of advertising [4]. As a result, there is a wide range of research regarding causal graph discovery. For instance, [5] shows that greedy search algorithms based on likelihood methods, such as those introduced in [6] are optimal (though not necessarily computationally efficient). Causation entropy is introduced in [7] for causal discovery of time series data, and further analyzed in [8]. Causal discovery is posed as a matrix completion problem in [9]. However, most graph discovery algorithms do not provide finite-sample guarantees on individual edges discovered by the algorithm, which is unfavorable in many applications, especially when the existence of an edge between two vertices may lead to different decisions. Because of this, we introduce ϵ-CUT, an algorithm for *Causal discovery with Unequal edge error Tolerance*.

A common way to control the sparsity of causal graphs in discovery algorithms is through a regularization term such as ℓ_0 in BIC [6] and ℓ_1 in LASSO neighborhood regression [10] and NOTEARS [11]. A hyper-parameter determines the strength of the regularization, and a higher parameter value indicates the algorithm favors sparsely connected graphs. Unfortunately, aside from a well-known result in [10] (which deals with connected components rather than individual edges) there is no current way to determine *a priori* how the hyper-parameter value affects the edge error rates of various discovery algorithms.

Recently, researchers have begun to emphasize providing guarantees on the edge error rates and confidence bounds on the outputs of discovery algorithms. For instance, a generalized likelihood ratio test is inverted in [12] to obtain confidence bounds on the causal effects between individual vertices. However, the results in [12] rely on asymptotic results, and hence cannot provide guarantees in the finite sample setting. In [13] a causal bandit problem is studied and different edge errors are treated unequally. This paper considers the setting introduced in [14], which formulates the causal graph discovery problem as a Neyman-Pearson hypothesis testing problem, where the false negative rate is minimized subject to a false positive constraint. The optimal detector is derived in [14], but is found to be computationally infeasible even for modestly sized graphs. Hence, the algorithm introduced in this paper still satisfies the finite-sample constraint on the false positive rate, while significantly reducing computational complexity.

Our contributions are as follows:

1) We derive an algorithm that seeks to minimize the false negative rate while keeping the false positive rate below a pre-specified tolerance level ϵ. We call this method ϵ-CUT.
2) We show ϵ-CUT's false positive rate will always satisfy the user-specified tolerance. This result requires no asymptotic assumptions on the convergence of distributions or consistency conditions and hence, is a finite-sample result.
3) We investigate the performance of ϵ-CUT compared to other causal discovery methods, namely LASSO neighborhood regression [10] and NOTEARS [11]. We study two examples. The first considers a two-vertex system with varying edge weights and shows ϵ-CUT performance improves relative to the other algorithms for increasing edge weights. In the second, we fix the edge weight and vary the size of the graph, again showing a performance gain for even modestly sized graphs.

Much of our notation mimics that used in [8], [14] and [7].

This work has been funded in part by one or more of the following grants: ARO W911NF1910269, ARO W911NF2410094, DOE DE-SC0021417, Swedish Research Council 2018-04359, NSF CCF-2008927, NSF RINGS-2148313, NSF CCF-2200221, NSF CCF-2311653, ONR 503400-78050, ONR N00014-22-1-2363, NSF A22-2666-S003.

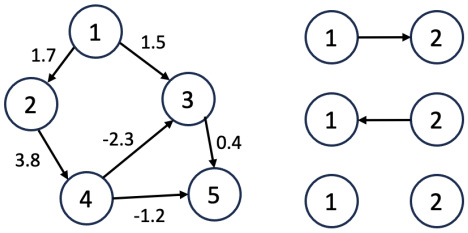


Fig. 1: (L) An example, five-node, directed, acyclic graph with edge weights. (R) Two-node graphs corresponding to the matrices $\boldsymbol{A}_0$, $\boldsymbol{A}_1$, and $\boldsymbol{A}_2$, respectively.

II. Problem Formulation

Consider a *directed acyclic graph* $\mathcal{G}$ with edge set $\mathcal{E}$, vertex set $\mathcal{V}$ with $|\mathcal{V}| = d$, and corresponding weighted adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$. We assume a prior on the adjacency matrix, denoted by $\pi_{\boldsymbol{A}}$. We have a series of n observations $\{X_k\}_{k=1}^n$ generated by the underlying distribution $P_{\boldsymbol{A}}$ conditioned on the realization of $\boldsymbol{A}$. Given a vertex $i \in \mathcal{V}$, let $X_k(i)$ be the k-th measurement $k = 1, .., n$ of vertex i. Moreover, let $\mathcal{Z}(i)$ denote the parent set of vertex i. Assuming $\mathcal{Z}(i)$ is non-empty, let i_j be the j-th parent of vertex i, $j = 1, ..., |\mathcal{Z}(i)|$. Define the vector

$$Z_k(i) = [X_k(i_1), X_k(i_2), ..., X_k(i_{|\mathcal{Z}(i)|})]^\top, \tag{1}$$

i.e., the vector comprised of the k-th measurements of the parents of i. Moreover, we assume a linear, additive, exogenous inputs model, i.e.,

$$X_k \quad = \quad \boldsymbol{A}X_k + W_k, \tag{2}$$

where the exogeneous input terms $\{W_k\}_{k=1}^n$ are Gaussian with equal covariance. This assumption is commonly made in the literature [6], [8], [10], [12], and it is also worth underscoring that if one removes the equivariance assumption, recoverability of $\boldsymbol{A}$ is no longer guaranteed even in the asymptotic case [6]. Given the underlying adjacency matrix $\boldsymbol{A}$, the goal is to recover the *support* $\boldsymbol{\chi}$ of $\boldsymbol{A}$, where $\chi_{i,j} = 1 \iff \boldsymbol{A}_{i,j} \neq 0$ *or* $\boldsymbol{A}_{j,i} \neq 0$.

A. Error Metrics

We define the following error rates which were first introduced in [7] and further studied in [8] and [14],

$$\textit{false positive}: \quad \epsilon^+ = \frac{\mathbb{E}\sum_{i,j} \mathbb{1}\{\hat{\boldsymbol{\chi}}_{i,j} = 1, \boldsymbol{A}_{i,j} = 0\}}{\mathbb{E}\sum_{i,j} \mathbb{1}\{\boldsymbol{A}_{i,j} = 0\}}, \tag{3}$$

$$\textit{false negative}: \quad \epsilon^- = \frac{\mathbb{E}\sum_{i,j} \mathbb{1}\{\hat{\boldsymbol{\chi}}_{i,j} = 0, \boldsymbol{A}_{i,j} \neq 0\}}{\mathbb{E}\sum_{i,j} \mathbb{1}\{\boldsymbol{A}_{i,j} \neq 0\}}, \tag{4}$$

where the expectations are taken with respect to detector $\hat{\boldsymbol{\chi}}$ as well as the prior distribution on the matrix $\boldsymbol{A}$, $\pi_{\boldsymbol{A}}$. We wish to find the detector $\hat{\boldsymbol{\chi}}$ that solves the following optimization problem.

$$\inf_{\hat{\boldsymbol{\chi}}} \epsilon^- \quad \text{s.t. } \epsilon^+ \leq \epsilon, \tag{5}$$

where $0 < \epsilon < 1$. We underscore that the detector $\hat{\chi}$ captures the detection of **all** edges in the causal graph; we shall provide a strategy for the detection of each individual edge, but our performance analysis will be over all edges.

B. Definitions

Definition 1. *We define the following probabilities and weights:*

$$P_{i,j}^+ = \mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{\chi}_{i,j} = 0), \tag{6}$$

$$Q_{i,j}^- = \mathbb{P}(\hat{\boldsymbol{\chi}}_{i,j} = 0 | \boldsymbol{\chi}_{i,j} = 1). \tag{7}$$

$$w_{i,j}^+ = \frac{\mathbb{P}(\boldsymbol{A}_{i,j} = 0)}{\sum_{k,l} \mathbb{P}(\boldsymbol{A}_{k,l} = 0)}, \tag{8}$$

$$w_{i,j}^- = \frac{\mathbb{P}(\boldsymbol{A}_{i,j} \neq 0)}{\sum_{k,l} \mathbb{P}(\boldsymbol{A}_{k,l} \neq 0)}. \tag{9}$$

□

With definitions 1, we can straightforwardly write out our error rates as,

$$\epsilon^+ = \sum_{i,j} w_{i,j}^+ P_{i,j}^+, \qquad \epsilon^- = \sum_{i,j} w_{i,j}^- Q_{i,j}^-. \tag{10}$$

III. Main Results

We present ϵ-CUT along with our main results. Namely, ϵ-CUT satisfies the false positive constraint in (5). Due to space constraints, the proofs are omitted, however, we give the basic intuition of our algorithm, which is best explained through a simple example.

Consider a system with two vertices, see Figure 1. We receive observation $\{X_k\}_1^n$ with $X_k = [X_k(1), X_k(2)]^\top$, $k = 1, 2, ..., n$. As stated before, we assume a linear system,

$$\begin{bmatrix} X_k(1) \\ X_k(2) \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} X_k(1) \\ X_k(2) \end{bmatrix} + \begin{bmatrix} W_k(1) \\ W_k(2) \end{bmatrix}, \tag{11}$$

where the $[W_k(1), W_k(2)]^\top$ vectors are *i.i.d.* Gaussian vectors with zero mean and covariance matrix $\sigma^2 I$. Since we restrict ourselves to directed acyclic graphs, the adjacency matrix $\boldsymbol{A}$ can only take one of three possible forms, which we denote as follows,

$$\boldsymbol{A}_0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \boldsymbol{A}_1 = \begin{bmatrix} 0 & 0 \\ a & 0 \end{bmatrix}, \qquad \boldsymbol{A}_2 = \begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix}, \tag{12}$$

where $a \in \mathbb{R} \setminus \{0\}$. If $\boldsymbol{A} = \boldsymbol{A}_0$, we have that for all k,

$$\mathbb{E}[X_k(1)^2] = \mathbb{E}[W_k(1)^2] = \sigma^2, \tag{13}$$

$$\mathbb{E}[X_k(2)^2] = \mathbb{E}[W_k(2)^2] = \sigma^2. \tag{14}$$

Alternatively, if $\boldsymbol{A} = \boldsymbol{A}_1$, we still have $\mathbb{E}[X_k(1)^2] = \sigma^2$, but

$$\mathbb{E}[X_k(2)^2] = \mathbb{E}[\big(aX_k(1) + W_k(2)\big)^2] = (a^2 + 1)\sigma^2. \tag{15}$$

Similarly, if $\boldsymbol{A} = \boldsymbol{A}_2$, then we have

$$\mathbb{E}[X_k(1)^2] = (a^2 + 1)\sigma^2, \qquad \mathbb{E}[X_k(2)^2] = \sigma^2. \tag{16}$$

Hence, we study the following hypothesis testing problem,

$$H_0 : \mathbb{E}[X_k(1)^2] = \mathbb{E}[X_k(2)^2], \tag{17}$$

$$H_1 : \mathbb{E}[X_k(1)^2] \neq \mathbb{E}[X_k(2)^2]. \tag{18}$$

Our hypothesis test is given as follows. If H_0 is true, then the empirical estimates of the variances should be roughly

Algorithm 1: ϵ-CUT

Specify user false alarm tolerance $0 < \epsilon \leq 1$. Let $2^{\mathcal{V}}$ denote the power set of $\mathcal{V}$ (the vertex set). Let $\mathcal{B}$ denote the set of all unique pairs of edges, i.e., the set of all pairs (i, j), such that $i \neq j$, $i, j \in \mathcal{V}$. The cumulative distribution function (cdf) of a chi-squared distribution with l degrees of freedom is denoted by $F_l(x)$.

For each $(i, j) \in \mathcal{B}$:

1) For each $\hat{\mathcal{Z}}(i) \in 2^{\mathcal{V}}$ and $\hat{\mathcal{Z}}(j) \in 2^{\mathcal{V}}$ with $j \notin \hat{\mathcal{Z}}(i)$ and $i \notin \hat{\mathcal{Z}}(j)$:
   a) Perform linear least squares regression for i and j on $\hat{\mathcal{Z}}(i)$ and $\hat{\mathcal{Z}}(j)$, respectively, obtaining the vectors of coefficients $\hat{\alpha}$ and $\hat{\beta}$.
   b) Compute
$$\hat{\sigma}_i^2 = \sum_{k=1}^{n} \big(X_k(i) - \hat{\alpha}^\top \hat{Z}_k(i)\big)^2, \quad \hat{\sigma}_j^2 = \sum_{k=1}^{n} \big(X_k(j) - \hat{\beta}^\top \hat{Z}_k(j)\big)^2,$$
   i.e., the residual sum of squares for vertices i and j.
   c) Compute $\tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)}$ such that
$$\begin{aligned} 2 - F_{n-p}\Big(\frac{\tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} - |q-p|\sigma^2}{2\sigma^2} + n - p\Big) + F_{n-p}\Big(-\frac{\tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} - |q-p|\sigma^2}{2\sigma^2} + n - p\Big) \\ - F_{n-q}\Big(\frac{\tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} - |q-p|\sigma^2}{2\sigma^2} + n - q\Big) + F_{n-q}\Big(-\frac{\tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} - |q-p|\sigma^2}{2\sigma^2} + n - q\Big) = \epsilon, \end{aligned}$$
   where $p = |\hat{\mathcal{Z}}(i)|$ and $q = |\hat{\mathcal{Z}}(j)|$.
   d) If $|\hat{\sigma}_i^2 - \hat{\sigma}_j^2| \leq \tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)}$, declare $\hat{\chi}_{i,j} = 0$.
2) If all potential parent sets $\hat{\mathcal{Z}}(i) \in 2^{\mathcal{V}}$ and $\hat{\mathcal{Z}}(j) \in 2^{\mathcal{V}}$ have been tested without declaring $\hat{\chi}_{i,j} = 0$, declare $\hat{\chi}_{i,j} = 1$.

equal with high probability. That is, for a properly specified threshold τ, the event

$$|\sigma_1^2 - \sigma_2^2| \leq \tau, \tag{19}$$

where

$$\hat{\sigma}_1^2 = \sum_{k=1}^{n} X_k(1)^2, \quad \hat{\sigma}_2^2 = \sum_{k=1}^{n} X_k(2)^2, \tag{20}$$

should occur with high probability. To extend the intuition to the case of more than two vertices, notice that if we condition on the parents of vertex i, σ_i^2 is now given as

$$\hat{\sigma}_i^2 = \sum_{k=1}^{n} \big(X_k(i) - \alpha_i^\top Z_k(i)\big)^2, \tag{21}$$

where α_i is the vector of non-zero edge weights in the ith row of $\boldsymbol{A}$. With the intuition of the algorithm explained, the only challenge left is to determine how to select the threshold τ to satisfy the false positive constraint. For this, the following lemma is useful.

Lemma 1. *For any n and $i \in \mathcal{V}$ let $\mathcal{Z}(i)$ and $Z_k(i)$ be defined as in the top of Section II. Define*

$$\hat{\sigma}_i^{*2} = \sum_{k=1}^{n} \big(X_k(i) - \hat{\alpha}_i^\top Z_k(i)\big)^2, \tag{22}$$

where $\hat{\alpha}_i$ is the vector of coefficients resulting from performing least squares for vertex i on $\{Z_k(i)\}_{k=1}^n$. Then, conditioned on $\{Z_k(i)\}_{k=1}^n$, $\hat{\sigma}_i^{*2}/\sigma^2$ follows a chi-squared distribution with $n - p$ degrees of freedom, where $p = |\mathcal{Z}(i)|$.

With Lemma 1, we can prove the following result.

Theorem 1. *Assume we have a linear system as described in* (2)*. Then, for any number of samples n and any given false positive tolerance ϵ, the false positive rate of ϵ-CUT, denoted by ϵ_1^+, satisfies $\epsilon_1^+ \leq \epsilon$.*

Proof Sketch. Since $\epsilon^+ = \sum_{i,j} w_{i,j}^+ P_{i,j}^+$, it suffices to show that for all i, j, $P_{i,j}^+ \leq \epsilon$. Note that

$$\begin{aligned} P_{i,j}^+ = \int_{\boldsymbol{A}} \int_{\boldsymbol{X}^{\setminus i,j}} \mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{X}^{\setminus i,j}) \mathbb{P}_{\boldsymbol{A}}(\boldsymbol{X}^{\setminus i,j}) \\ \frac{\mathbb{1}\{\boldsymbol{A}_{i,j} = 0 \cup \boldsymbol{A}_{j,i} = 0\} \pi_{\boldsymbol{A}}}{\mathbb{P}(\boldsymbol{\chi}_{i,j} = 0)}, \end{aligned} \tag{23}$$

where $\boldsymbol{X}^{\setminus i,j}$ denotes the set of all measurements except those from the ith and jth vertex. That is, $\boldsymbol{X}^{\setminus i,j} = \{X_k^{\setminus i,j}\}_{k=1}^n$ where $X_k^{\setminus i,j} = [X_k(1), ..., X_k(i-1), X_k(i+1), ..., X_k(j-1), X_k(j+1), ..., X_k(d)]^\top$. Then, it suffices to show that for any $\boldsymbol{A}$ and $\boldsymbol{X}^{\setminus i,j}$ we have $\mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{X}^{\setminus i,j}) \leq \epsilon$. This is done by first noticing that for $\hat{\sigma}_i^2$ and $\hat{\sigma}_j^2$ as defined in Algorithm 1,

$$\mathbb{P}_{\boldsymbol{A}}(\hat{\boldsymbol{\chi}}_{i,j} = 1 | \boldsymbol{X}^{\setminus i,j}) \tag{24}$$

$$= \mathbb{P}_{\boldsymbol{A}}\Big(\bigcap_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} |\hat{\sigma}_i^2 - \hat{\sigma}_j^2| > \tau_{\hat{\mathcal{Z}}(i),\hat{\mathcal{Z}}(j)} | \boldsymbol{X}^{\setminus i,j} \Big) \tag{25}$$

$$\leq \mathbb{P}_{\boldsymbol{A}}\big(|\hat{\sigma}_i^{*2} - \hat{\sigma}_j^{*2}| > \tau_{\mathcal{Z}(i),\mathcal{Z}}(j) | \boldsymbol{X}^{\setminus i,j}\big). \tag{26}$$

Using Lemma 1 together with a series of inequalities and algebraic manipulations, (26) is upper bounded by the expression given in Algorithm 1 c) which completes the proof. □

Some important notes and comparisons are to be made regarding ϵ-CUT and Theorem 1.

1) Theorem 1 is a finite-sample result. This differs from much of the current literature which focuses on asymptotic recoverability guarantees of causal discovery [1], [5], [6], [9], [12]. Finite-sample results have appeared in the literature, with a notable result appearing in [10] (Theorem 3). However, the result in [10] deals with connected components instead of individual edges, which is our main objective.
2) There are many ways to solve the hypothesis-testing problem between H_0 and H_1, leading to different causal discovery algorithms. For instance, one may construct the likelihood ratio test for each edge pair (i, j). This approach is taken in [14] and is shown to be optimal, but computationally infeasible, even for modestly sized graphs and relatively simple priors $\pi_{\boldsymbol{A}}$. A generalized likelihood ratio test (GLRT) is used in [12] to obtain confidence intervals on the causal effect between two vertices, rather than directly declaring the presence or absence on an edge. Unfortunately, the results in [12] are asymptotic. Hence, ϵ-CUT circumvents the computational issues associated with [14] while providing finite-sample results.
3) Unlike several popular algorithms such as BIC [6], LASSO [10], and NOTEARS [11], ϵ-CUT requires no hyper-parameter tuning. That is, in the algorithms mentioned, a sparsity constraint is added through a regularization term (ℓ_0 in [6] and ℓ_1 in [10], [11]). The constant λ controls the sparsity of the resulting graph. Unfortunately, there is no current way to determine *a priori* how the constant λ affects the error rates, and if these rates satisfy the constraint in (5). Hence, one needs to experiment with different regularization constants. To contrast ϵ-CUT, once ϵ is given, everything in ϵ-CUT is completely specified.

IV. Numerical Results

We consider a numerical example to compare ϵ-CUT to other popular algorithms, such as NOTEARS [11] and LASSO neighborhood selection [10]. We examine some interesting phenomena and show the performance gains of ϵ-CUT over the algorithms mentioned above.

A. Summary of Prior Algorithms

1) *NOTEARS*: The NOTEARS algorithm seeks to minimize the penalized squared loss,

$$\frac{1}{2n}\sum_{k=1}^{n}\|X_k - \boldsymbol{A}X_k\|_2^2 + \lambda\|\boldsymbol{A}\|_1, \tag{27}$$

where λ is a regularizer term, $\|X_k - \boldsymbol{A}X_k\|_2$ is the l_2 norm of the vector $X_k - \boldsymbol{A}X_k$ and $\|\boldsymbol{A}\|_1$ is the l_1 norm of the matrix $\boldsymbol{A}$. It is shown in [11] that a weighted adjacency matrix $\boldsymbol{A}$ represents a directed acyclic graph if and only if

$$\text{Trace}(e^{\boldsymbol{A}\circ\boldsymbol{A}}) = d, \tag{28}$$

where $\circ$ denotes the Hadamard product (element-wise multiplication). The elements of the continuous-valued estimate $\hat{\boldsymbol{A}}$ are thresholded to determine the inactive edges.

2) *LASSO*: For a given observed node $X_k(i)$, let $Y_k^i = [X_k(1), ..., X_k(i-1), X_k(i+1), ..., X_k(d)]^\top$ for $k = 1, 2, ..., n$. Then, assuming a linear model for the observations, $\hat{X}_k(i) = A_i^\top Y_{i,k}$, where A_i is a vector of coefficients, LASSO [10] minimizes the following sparsity penalized loss over the coefficients of A_i,

$$\frac{1}{2n}\sum_{k=1}^{n}\|X_k(i) - A_i^\top Y_k^i\|_2^2 + \lambda\|A_i\|_1, \tag{29}$$

for each node i, where λ is a regularizer term.

B. Definition of Graph Prior

The prior $\pi_{\boldsymbol{A}}$ is defined as follows:

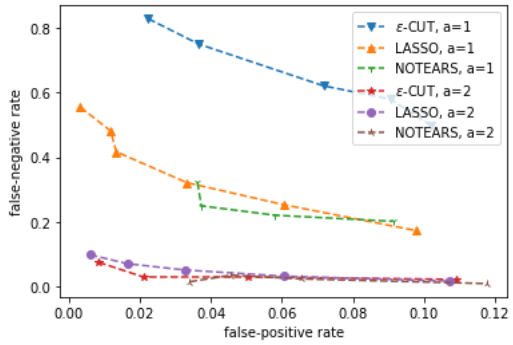
1) For a given number of vertices d, let $\mathcal{D}$ denote the set of all matrices with entries equal to either 0 or 1 corresponding to a directed acyclic graph.
2) Select a matrix from $\mathcal{D}$ uniformly at random. Denote this matrix as D. Let R be a $d \times d$ matrix of Rademacher random variables.
3) Let a be the desired weight of all the edges. Then, the final graph is given as $A = a(D \circ R)$.

Intuitively, $\pi_{\boldsymbol{A}}$ is the prior that selects a directed acyclic graph uniformly at random, and the weights of each edge are equally likely to be either a or $-a$.

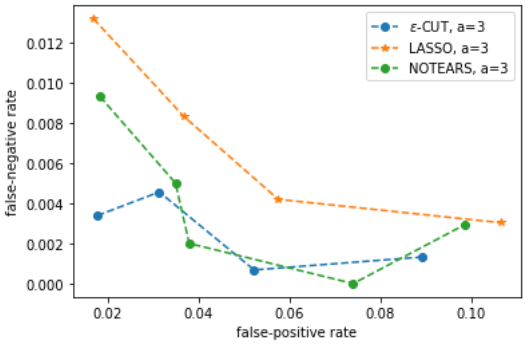
C. Results

Some numerical comparisons for a low number of samples ($n = 10$) are given in Figure 2. Figure 2 (a), considers a two-vertex system for $a \in \{1, 2\}$. The two-node system is often used in causal discovery as a base case for algorithms (such as in [15]). Interestingly, when $a = 1$, LASSO and NOTEARS considerably outperform ϵ-CUT. However, as a increases, the performance gap begins to shrink, as seen in Figure 2 (a) when $a = 2$, and as a keeps increasing, ϵ-CUT eventually begins to outperform LASSO and NOTEARS, as seen in Figure 2 (b). Interestingly, ϵ-CUT also outperforms LASSO and NOTEARS on larger graphs for a fixed a. In Figure 2 (c), we consider a graph with $d = 5$ and $a = 2$, and in Figure 2 (d), we have $d = 7$ and $a = 2$.

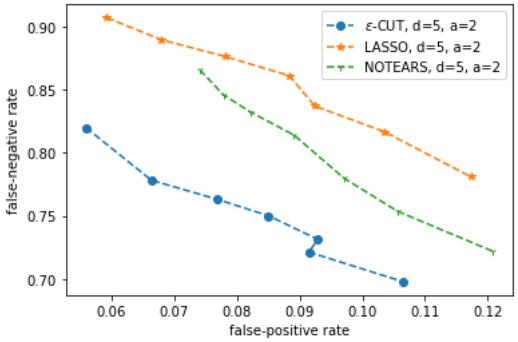
There is an intuitive explanation for the first phenomenon. First, note that ϵ-CUT seeks to control the false positive rate and selects the thresholds independently of the edge weights. Hence, if an edge has a sufficiently small weight ϵ-CUT will be more inclined to declare no edge, increasing the false negative rate. As the edge weight a increases in magnitude, the ability to distinguish between an edge and no edge increases, decreasing this trade-off. As for why the performance improves for larger graphs further analysis is needed.



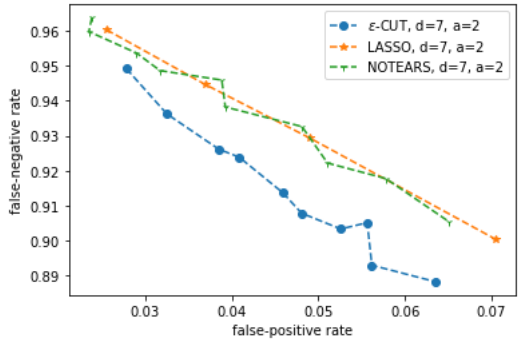
(a) Performance curves for $d = 2$ when $a \in \{1, 2\}$.



(b) Performance curves for $d = 2$ when $a = 3$.



(c) Performance curves for $d = 5$ and $a = 2$.



(d) Performance curves for $d = 7$ and $a = 2$.

Fig. 2: Numerical results of the performance of various detectors. For all the cases, $\sigma^2 = 1$ and $n = 10$ and the regularization constant λ is varied for LASSO and NOTEARS to control the error rates. Performance curves are averaged over 2000 iterations.

V. Conclusions

This paper introduced ϵ-CUT, a causal discovery algorithm that controls the false positive rate by keeping it below a pre-specified threshold. We show the intuition of the algorithm, which revolves around the observation that if two vertices have an edge between them, their variances (when conditioned on their respective parent sets) are different and that the empirical estimates of the variances follow a chi-squared distribution. We compare ϵ-CUT to popular state-of-the-art algorithms and show the performance gains. We also show interesting phenomena, such as the performance dependence on the edge weights, and the size of the graph considered.

References

[1] J. Pearl, *Causality*. Cambridge University Press, 2009.

[2] E. Testi and A. Giorgetti, "Blind wireless network topology inference," *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 1109–1120, 2020.

[3] G. L. Lozano, J. I. Bravo, M. F. G. Diago, H. B. Park, A. Hurley, S. B. Peterson, E. V. Stabb, J. M. Crawford, N. A. Broderick, and J. Handelsman, "Introducing THOR, a model microbiome for genetic dissection of community behavior," *mBio*, vol. 10, no. 2, pp. e02 846–18, 2019. [Online]. Available: https://journals.asm.org/doi/abs/10.1128/mBio.02846-18

[4] E. Van den Broeck, K. Poels, and M. Walrave, "An experimental study on the effect of ad placement, product involvement and motives on facebook ad avoidance," *Telematics and Informatics*, vol. 35, no. 2, pp. 470–479, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736585317307645

[5] D. M. Chickering, "Learning equivalence classes of Bayesian-network structures," *The Journal of Machine Learning Research*, vol. 2, pp. 445–498, 2002.

[6] J. Peters and P. Bühlmann, "Identifiability of Gaussian structural equation models with equal error variances," *Biometrika*, vol. 101, no. 1, pp. 219–228, 2014.

[7] J. Sun, D. Taylor, and E. M. Bollt, "Causal network inference by optimal causation entropy," *SIAM Journal on Applied Dynamical Systems*, vol. 14, no. 1, pp. 73–106, 2015.

[8] X. Kang and B. Hajek, "Lower bounds on information requirements for causal network inference," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 754–759.

[9] A. Agarwal, M. Dahleh, D. Shah, and D. Shen, "Causal matrix completion," in *Proceedings of Thirty Sixth Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, G. Neu and L. Rosasco, Eds., vol. 195. PMLR, 12–15 Jul 2023, pp. 3821–3826. [Online]. Available: https://proceedings.mlr.press/v195/agarwal23c.html

[10] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *The Annals of Statistics*, vol. 34, no. 3, pp. 1436 – 1462, 2006. [Online]. Available: https://doi.org/10.1214/009053606000000281

[11] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "DAGS with no tears: Continuous optimization for structure learning," *Advances in neural information processing systems*, vol. 31, 2018.

[12] D. Strieder and M. Drton, "Confidence in causal inference under structure uncertainty in linear causal models with equal variances," *Journal of Causal Inference*, vol. 11, no. 1, p. 20230030, 2023. [Online]. Available: https://doi.org/10.1515/jci-2023-0030

[13] C. Peng, D. Zhang, and U. Mitra, "Graph identification and upper confidence evaluation for causal bandits with linear models," in *Interational Conference on Acoustics, Speech and Signal Processing*, 2024.

[14] J. Shaska and U. Mitra, "Neyman-Pearson causal inference," *To be presented at the Internation Symposium on Information Theory (2024), Athens, Greece*, 2024.

[15] D. Strieder, T. Freidling, S. Haffner, and M. Drton, "Confidence in causal discovery with linear causal models," in *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, ser. Proceedings of Machine Learning Research, C. de Campos and M. H. Maathuis, Eds., vol. 161. PMLR, 27–30 Jul 2021, pp. 1217–1226. [Online]. Available: https://proceedings.mlr.press/v161/strieder21a.html