Differentiable Causal Structure Learning with Identifiability Guarantees by NOTIME

Jeroen Berrevoets[1*] **Jakob Raymaekers**[2] **Mihaela van der Schaar**[1]
Tim Verdonck[2] **Ruicong Yao**[2,3]
[1]University of Cambridge [2]University of Antwerp [3]KU Leuven

Abstract

The introduction of the NOTEARS algorithm resulted in a wave of research on differentiable Directed Acyclic Graph (DAG) learning. Differentiable DAG learning transforms the combinatorial problem of identifying the DAG underlying a Structural Causal Model (SCM) into a constrained continuous optimization problem. Being differentiable, these problems can be solved using gradient-based tools which allow integration into other differentiable objectives. However, in contrast to classical constrained-based algorithms, the identifiability properties of differentiable algorithms are poorly understood. We illustrate that even in the well-known Linear Non-Gaussian Additive Model (LiNGAM), the current state-of-the-art methods do not identify the true underlying DAG. To address the issue, we propose NOTIME (*Non-combinatorial Optimization of Trace exponential and Independence MEasures*), the first differentiable DAG learning algorithm with *provable* identifiability guarantees under the LiNGAM by building on a measure of (joint) independence. With its identifiability guarantees, NOTIME remains invariant to normalization of the data on a population level, a property lacking in existing methods. NOTIME compares favourably against NOTEARS and other (scale-invariant) differentiable DAG learners, across different noise distributions and normalization procedures. Introducing the first identifiability guarantees to general LiNGAM is an important step towards practical adoption of differentiable DAG learners.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. *Authors listed in alphabetical order.

1 INTRODUCTION

Causality is at the heart of how we understand the world and is central to many fields of science. Causal discovery is the process of inferring causal networks from data, which are usually represented by Directed Acyclic Graphs (DAGs). For a long time, research was limited to combinatorial structural learning algorithms such as PC [Spirtes et al., 2001, Colombo et al., 2014], GES [Chickering, 2002, Hauser and Bühlmann, 2012], DirectLiNGAM [Shimizu et al., 2011] and RESIT [Peters et al., 2014] which iteratively perform (conditional) independence tests or greedy score-based selection. Separate procedures have to be applied to learn the causal structure/order and the model parameters.

More recent advances, however, formulate the NP-hard combinatorial problem as a constrained continuous program that can be solved by standard numerical optimizers. In that way, both the structure and the parameters can be learned simultaneously. For example, in the seminal work NOTEARS [Zheng et al., 2018], the author proposed the first differentiable DAG learning algorithm where the optimal weighted adjacency matrix solves a regularized least-squares problem under an exponential acyclicity constraint. GOLEM [Ng et al., 2020] focuses on the linear Gaussian model and optimizes a likelihood-based score function under a soft sparsity and DAG constraint. DAGMA [Bello et al., 2022] proposes a new acyclicity characterization based on the log-determinant which improved the performance in detecting cycles and speed. Extensions to nonlinear additive noise models are possible via score functions or neural networks and Sobolev estimators [Lachapelle et al., 2019, Zheng et al., 2020].

Despite the popularity of such differentiable DAG learning methods, their identifiability properties in the context of structural causal models (SCMs) are rather restricted. It can be shown that the above methods only identify the DAG underlying a linear SCM on the condition that all noise variables have equal variance [Loh and Bühlmann, 2014]. This is disappointing since

it is known that for linear SCMs, it is theoretically possible to identify the underlying DAG provided that the noise variables are non-Gaussian with arbitrary noise scale [Shimizu et al., 2006, Peters et al., 2017]. This is a much weaker and more practical condition than requiring the noise variables to have the same variance. A reliable method for learning DAGs should identify the true DAG in these linear non-Gaussian additive models (LiNGAMs).

The lack of identifiability under general LiNGAM limits the application of the existing differentiable DAG learning methods to many real-world problems with heteroscedastic noise. Additionally, they can lead to unexpected and false results. For example, Reisach et al. [2021] showed that the performances of NOTEARS and GOLEM deteriorate heavily when the data is first normalized. This is clearly undesirable, as the scale of the variables should never influence the direction of causality (e.g., whether a vaccine is effective shouldn't depend on whether the dose is measured in oz or mL, and changing the unit definitely shouldn't make the vaccine cause the disease it is treating.). What is happening, of course, is that the normalization procedure changes the variance of the noise variables. As a result, the condition of noise variables with equal variances is no longer fulfilled, rendering the methods unreliable.

To address the issue, we propose the first differentiable DAG learning algorithm with provable identifiability guarantees under the LiNGAM. We achieve this by building the objective function on a measure of (joint) independence. In particular, we use a loss function quantifying the dependency of the fitted residuals through the d-dimensional Hilbert Schmidt Independence Criterion (dHSIC) [Gretton et al., 2007, Pfister et al., 2018]. This approach is inspired by the fact that independent noise assumption is the key factor that makes the additive noise model identifiable. Specifically, the main contributions of this paper are:

1. A differentiable causal structure learning algorithm with a loss function based on dHSIC.
2. We show theoretically that the algorithm identifies the true LiNGAM on a population level and, as a consequence, is invariant to data scaling.
3. We show empirically that NOTIME generally outperforms the competition in the LiNGAM model across different noise distributions and under different normalizations of the data.
4. Our research highlights that the sample size in relation to the dimension of the data also plays an important role for differentiable causal learning methods with identifiable guarantees on LiNGAM, aligned with previous studies on other models. This opens an avenue for theoretical studies of the behavior.

The paper is organized as follows: In Section 2, we introduce LiNGAM, NOTEARS, and the joint independence measure dHSIC which form the foundation of the method. In Section 3, we introduce the proposed loss function which incorporates dHSIC in the learning framework of NOTEARS. We prove and illustrate the theoretical guarantee of NOTIME and discuss its initialization. Section 4 presents an empirical study comparing NOTIME with popular differentiable causal learning methods, GES, and SortNRegress [Reisach et al., 2023]. Finally, we conclude in Section 6.

2 PRELIMINARIES

Linear Structural Equation Models. In this paper, we consider the linear structural equation models (SEM). Let $\boldsymbol{X} = (X_1, \dots, X_d)^\top$ be a d-variate random vector. A linear SEM can be written as

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{B} \in \mathbb{R}^{d\times d}$ is the weight matrix which can be permuted to be upper triangular and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_d)^\top$ is the noise vector which satisfies $(\varepsilon_1, \dots, \varepsilon_d)$ jointly independent, $\varepsilon_i \perp\!\!\!\perp X_i$ for all $1 \le i, j \le d$. Shimizu et al. [2006] showed that linear non-Gaussian additive models (LiNGAM) are identifiable from the distribution, i.e., the weight matrix can be fully recovered based on observational data. The key insight is that, on a population level, any other model would violate the joint independence assumption on the noise variables. Several combinatorial algorithms, including ICA-LiNGAM [Shimizu et al., 2006], DirectLiNGAM [Shimizu et al., 2011] and pairwise LiNGAM [Hyvärinen and Smith, 2013], have been proposed which are guaranteed to identify the correct DAG on a population level.

Differentiable Causal Structure Learning. NOTEARS [Zheng et al., 2018] is the first algorithm for causal structure learning which can be solved entirely by continuous optimization, thereby eliminating the need to solve a combinatorial problem. The optimization procedure uses a combination of a score function (typically the MSE) and a sparsity penalty (with parameter λ) to score a DAG (represented by the weight matrix $\boldsymbol{W}$). The key ingredient is its exponential constraint which is differentiable and holds if and only if the DAG is acyclic:

$$\begin{aligned} \boldsymbol{W}^*_{\boldsymbol{X}} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \frac{1}{2n} ||\mathbf{X}\boldsymbol{W} - \mathbf{X}||_F^2 + \lambda ||\boldsymbol{W}||_1 \\ \text{s.t.} \quad \operatorname{Tr}(e^{\boldsymbol{W}\odot\boldsymbol{W}}) - d = 0. \end{aligned}$$

The optimization process employs an augmented Lagrangian method, which iteratively solves unconstrained subproblems to impose the continuous DAG constraint on the solution. In practice, the constraint may not be perfectly fulfilled due to machine precision and thus thresholding on the estimated weights is applied to ensure the acyclicity of the output. Subsequent research proposed different acyclicity constraints for training efficiency, c.f. Bello et al. [2022].

Identifiability in Linear SEMs. We have emphasized the fact that under the LiNGAM assumption, the true underlying DAG is identifiable from observational data. When the noise variables are Gaussian, identifiability of the linear SEM is impossible in general. One exception is when the variances of the noise variables are known up to a constant factor [Loh and Bühlmann, 2014]. Then no matter the noise type, the true model is the only minimizer of a weighted MSE, resulting in the identifiability. In particular, the linear Gaussian equal variance (EV) model [Peters and Bühlmann, 2014] belongs to this class, in which case the classical (unweighted) MSE is minimized by the true model. As the unweighted MSE is a popular choice of the loss function in the NOTEARS and other methods, it inherits these identifiability properties. In other words, NOTEARS is only guaranteed to identify the DAG underlying the SEM when the variances of the noise terms are equal. It should be clear that this is extremely restrictive and that the LiNGAM assumption is much more attractive in practice. After all, real datasets often have many clearly heteroscedastic non-Gaussian variables and while they may sometimes have some approximately Gaussian variables, they are rarely exactly Gaussian. Perhaps more importantly, the assumption of knowing the variances of the error terms is very unrealistic and impractical.

This issue was not obvious initially, as the original empirical study of NOTEARS generated data from a linear SEM in which each independent noise term had equal variance. As this happens to be the specific setting in which NOTEARS (or the MSE loss more generally) does work, the shortcomings of NOTEARS were not discovered. Gradually, evidence for these restricted identifiability properties has appeared. For example, Reisach et al. [2021] illustrated that NOTEARS fails on normalized data and attributed it to low varsortability but did not give a theoretical explanation. In light of the previous discussion, this fact is easily explained: the noise variances are no longer the same after normalizing the data, and hence the identifiability condition is violated , which is a fundamental problem due to the inconsistency of the loss. Note that Ng et al. [2024] also empirically validated that varsortability is essentially *not* related to the performance of differentiable causal learning methods in general.

One may wonder how well or poorly NOTEARS (or the MSE loss) may perform in case the identifiability condition is violated. After all, it could be possible that on a population level, the underlying DAG is still identified with a high probability that is somewhat smaller than one. Unfortunately, this is not the case. The following two-dimensional example inspired by Loh and Bühlmann [2014] illustrates that NOTEARS will identify the *wrong* DAG (the reverse causal order) with probability one!

Example 1. *Let* $\boldsymbol{X} = (X_1, X_2)$ *be a random vector defined by*

$$\begin{aligned} X_1 &:= \varepsilon_1 \\ X_2 &:= -0.5X_1 + \varepsilon_2 \end{aligned}$$

where ε_1 *and* ε_2 *are arbitrary non-Gaussian noise variables with* $\mathbb{E}[\varepsilon_1] = \mathbb{E}[\varepsilon_2] = 0$, $Var[\varepsilon_1] = 1$, *and* $Var[\varepsilon_2] = 0.25$. *We then compute the expected NOTEARS loss under the weight matrix* $\boldsymbol{W}_1 = \begin{bmatrix} 0 & w_1 \\ 0 & 0 \end{bmatrix}$, *and* $\boldsymbol{W}_2 = \begin{bmatrix} 0 & 0 \\ w_2 & 0 \end{bmatrix}$. *Obviously,* $\boldsymbol{W}_1$ *gives the correct DAG which encodes the causation from* X_1 *to* X_2 *via* $\boldsymbol{X} = \boldsymbol{X}\boldsymbol{W} + \boldsymbol{\varepsilon}$. *We will show that NOTEARS would wrongly identify* $\boldsymbol{W}_2$ *as the underlying DAG on a population level.*

First, we can assume without loss of generality that $w_1, w_2 \leq 0$ *due to the true data generation mechanism and the loss function. Denote the objective function with* $\boldsymbol{\Theta}(\boldsymbol{W}, \boldsymbol{X})$. *We have,*

$$\begin{aligned} \boldsymbol{\Theta}(\boldsymbol{W}_1, \boldsymbol{X}) =& Var(X_1) + Var(-0.5X_1 - w_1X_1) \\ &+ Var(\varepsilon_2) - \lambda w_1 \\ =& 1.25 + 0.25(1 + 2w_1)^2 - \lambda w_1 \\ \boldsymbol{\Theta}(\boldsymbol{W}_2, \boldsymbol{X}) =& Var(X_2) + Var(X_1 + 0.5w_2X_1) \\ &+ Var(w_2\varepsilon_2) - \lambda w_2 \\ =& 0.5 + 0.25(2 + w_2)^2 + 0.25w_2^2 - \lambda w_2. \end{aligned}$$

Minimizing these loss functions w.r.t. w_1 *and* w_2 *respectively yields* $w_1 = \min(0, 0.5\lambda - 0.5))$ *and* $w_2 = \min(0, \lambda - 1)$, *with corresponding losses*

$$\begin{aligned} \boldsymbol{\Theta}(\boldsymbol{W}_1^*, \boldsymbol{X}) =& \frac{5}{4} - \frac{\lambda^2}{4} + \frac{\lambda}{2} \\ \boldsymbol{\Theta}(\boldsymbol{W}_2^*, \boldsymbol{X}) =& 1 - \frac{1}{2}\lambda^2 + \lambda \end{aligned}$$

Now it is clear that for $0 \leq \lambda < 1$, *we have* $\boldsymbol{\Theta}(\boldsymbol{W}_1^*, \boldsymbol{X}) > \boldsymbol{\Theta}(\boldsymbol{W}_2^*, \boldsymbol{X})$ *and hence NOTEARS reverses the causal order. When* $\lambda \geq 1$, *the optimal weight matrix is the zero matrix. To conclude, the MSE loss always fails to find the true causal direction in this simple example. Therefore, the training loss should be carefully selected to avoid its reliance on the variable scale and ensure consistency.*

Measuring Dependence. The independent noise assumption is the key to the identifiablity of LiNGAM and more general additive noise models (ANM) Peters et al. [2014]. The true causal model would minimize the score DM($\mathbf{res_1}, \mathbf{res_2}, .., \mathbf{res_d}$) where, $\mathbf{res_i}$ is an n-dimensional vector of the residuals of the i-th variable from the model and DM is a joint dependence measure (which is small when the components are jointly independent). The development of measures of (joint) independence is an active area of research, and several proposals have appeared in recent years. Prime examples include the Hilbert–Schmidt Independence Criterion (HSIC, dHSIC) [Gretton et al., 2007, Pfister et al., 2018], Distance multivariance [Böttcher et al., 2019], and measures based on the popular distance correlation [Jin and Matteson, 2018]. On a population level, these functionals map a d-variate distribution to a non-negative number, which is zero if and only if the d components are mutually independent. Among them, the dHSIC often performs very well. The dHSIC estimator on finite samples is defined as

$$\begin{aligned}\widehat{\mathrm{dHSIC}}(x_1,\dots,x_d) = &\frac{1}{n}\sum_{M_2(n)}\prod_{j=1}^{d} k^j(x^j_{i_1},x^j_{i_2}) \\ &+\frac{1}{n^{2d}}\sum_{M_{2d}(n)}\prod_{j=1}^{d} k^j(x^j_{i_{2j-1}},x^j_{i_{2j}}) \\ &-\frac{2}{n^{d+1}}\sum_{M_{d+1}(n)}\prod_{j=1}^{d} k^j(x^j_{i_1},x^j_{i_{j+1}})\end{aligned}$$

where d is the number of variables, n is the number of data samples, $k(\cdot,\cdot)$ is a characteristic kernel, x_i^j is the i-th element of the j-th column, and $M_q(n)$ is the q-fold Cartesian product over the set $\{1,\dots n\}$. Pfister et al. [2018] provide an implementation with a time complexity of $\mathcal{O}(d\cdot n^2)$. We note that the algorithm for estimating dHSIC consists solely of tensor operations, and thus if the kernel functions $k(\cdot,\cdot)$ are differentiable, it is simple to take the gradient of $\widehat{\mathrm{dHSIC}}$ with contemporary automatic differentiation software. In the literature, the Gaussian kernel is widely used for kernel-based (joint) independence test [Pfister et al., 2018, Zhang et al., 2011], and proved to have good performance. The bandwidth is usually chosen by convention [Pfister et al., 2018].

We note that although the HSIC and dHSIC have been widely applied in the causal discovery literature, they are mainly used as independence tests. For example, the RESIT algorithm [Mooij et al., 2009, Peters et al., 2014] iteratively applies HSIC-based tests for order searching and variable pruning to identify the underlying additive noise models. Immer et al. [2023] considered the bivariate location-scale noise model, and provided a learning algorithm based on a RESIT-type estimation with the HSIC tests. Monti et al. [2020] considered general bivariate nonlinear causal discovery and developed an algorithm based on HSIC and nonlinear ICA. Recently, Huang et al. [2020] considered causal discovery from heterogeneous/nonstationary data and developed an algorithm based on an extended version of the HSIC to measure the dependence between causal modules. Extension of HSIC and dHSIC for causal structure learning on functional data was also developed by Laumann et al. [2023].

3 METHODOLOGY

Identifiable Differential DAG Learning. In this section, we propose a differentiable DAG learning algorithm with identifiability guarantees under the LiNGAM model. As a consequence, the method is scale-invariant on a population level, which aligns with the idea that the causal ordering of the underlying DAG should not be dependent on the scale of the variables. Therefore, it overcomes the shortcomings of the existing methods. In order to obtain such a method, we propose to replace the MSE in the loss of NOTEARS by a measure of joint dependence of the residuals. This yields the proposed method NOTIME: *Non-combinatorial Optimization of Trace exponential and Independence MEasures*. NOTIME optimizes a weight matrix $\boldsymbol{W}$ to have jointly independent residuals while being sparse and enforcing acyclicity:

$$\begin{aligned}\boldsymbol{W}^*_{\boldsymbol{X}} = &\operatorname*{argmin}_{\boldsymbol{W}} \mathrm{dHSIC}(\boldsymbol{X}-\boldsymbol{X}\boldsymbol{W}) + \lambda\,||\boldsymbol{W}\odot\boldsymbol{\Sigma}||_1 \\ &\text{s.t.}\quad \mathrm{Tr}(e^{\boldsymbol{W}\odot\boldsymbol{W}}) - d = 0.\end{aligned} \tag{1}$$

The first term quantifies the joint (in)dependence of the residuals of the model. The second term penalizes the matrix of weights $\boldsymbol{W}$ with an ℓ_1-norm, such that on finite samples we can estimate DAGs that are sparse (i.e. not necessarily fully connected). Here, $\boldsymbol{\Sigma}_{ij} = \frac{\sigma_{X_j}}{\sigma_{X_i}}$ is a de-standardization matrix, which converts a normalized weight matrix $\boldsymbol{W}$ back to fit the original data scale by taking the Hadamard product, which we denote as $\odot$. This ensures that the severity of the sparsity penalty does not depend on the scale of the variables, and is similar in spirit to the practice of scaling the variables before doing regularized regression like the lasso. Finally, the third term enforces the acyclicity of the graph corresponding with the estimated weight matrix $\boldsymbol{W}^*$ in a similar way as the original NOTEARS.

Remark 1. *We note that the terms in the objective function of Equation 1 can be replaced by alternatives serving a similar purpose. In particular, the first term can be replaced by any measure of joint independence that is minimized when the input variables are jointly in-*

dependent such as the multivariance criterion [Böttcher et al., 2019]. The term second can be replaced by any penalty-inducing sparsity in the estimated weight matrix. Finally, the third term can be replaced by any penalty enforcing acyclicity in $\boldsymbol{W}$ such as the log-determinant function used in Bello et al. [2022].

Theoretical Properties. It turns out that, on a population level, this continuous optimization problem defined in Equation (1) does have the property of identifying the true DAG underlying the data generating process in case it is a LiNGAM structure [Shimizu et al., 2006], as stated in Theorem 1 below. For a proof, we refer to Appendix B. The idea is that by the identifiability of the linear independent component analysis (ICA) model, all the minimizers to the problem have the form $I - PS(I - \boldsymbol{B})$, where P is the permutation matrix, S is a diagonal scaling matrix, and I is the identity. By using the acyclicity constraint, we conclude that there is a unique solution with $P = S = I$.

Theorem 1 (Identification of LiNGAM structures)**.** *Let $\boldsymbol{X}$ be a d-variate random vector generated by the LiNGAM model*

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{B}$ can be permuted to a strictly lower triangular matrix and the components of $\boldsymbol{\varepsilon}$ are independent and non-Gaussian. Then $\boldsymbol{B}$ is identified by NOTIME, assuming an oracle dHSIC, i.e. $\boldsymbol{W}^* = \boldsymbol{B}$ where

$$\begin{aligned} \boldsymbol{W}^* &:= \operatorname*{argmin}_{\boldsymbol{W}} dHSIC(\boldsymbol{X} - \boldsymbol{W}\boldsymbol{X}) \\ &s.t. \qquad \operatorname{Tr}(e^{\boldsymbol{W}\odot\boldsymbol{W}}) - d = 0 \end{aligned}$$

Remark 2. *The result is stated on a population level, in which case $\lambda = 0$ guarantees the identification (without further information about the size of the absolute values of the entries in $\boldsymbol{B}$). In finite samples, however, $\lambda > 0$ can often help to trade some bias for a reduced variance. In order to get asymptotic identification, λ should vanish with growing n (assuming d stays fixed). This is aligned with the theory in high-dimensional statistics [Bühlmann and Van De Geer, 2011].*

Remark 3. *The result of Theorem 1 holds for any oracle measure of joint independence that is non-negative and zero if and only if the components are jointly independent, and for any penalty enforcing acyclicity in the weight matrix.*

A comparison of the methods is provided in Table 1. So far, NOTIME is the only differentiable DAG learning method with scale-invariance and guarantee to a broad class of linear models.

Example 2. *To further illustrate the identifiability properties of NOTIME, we present a small numerical experiment. We generate samples from the distribution induced by the DAG $X_1 \to X_2 \to X_3$ and $X_1 \to X_3$ with structural equations*

$$\begin{aligned} X_1 &:= \varepsilon_1 \\ X_2 &:= 0.25X_1 + \varepsilon_2 \\ X_3 &:= 0.25X_1 + 0.25X_2 + \varepsilon_3 \end{aligned}$$

and $\varepsilon_1, \varepsilon_2, \varepsilon_3$ are proportional to a t_5, exponential and t_{10} distribution respectively. We let the sample size vary from $n = 20$ to $n = 500$, and generate 100 datasets for each sample size. On each of these 100 datasets, we run NOTEARS, GES, SortNRegress, and NOTIME. We calculate the average structural Hamming distance of the four methods. The result of this experiment is shown in Figure 1. As the sample size increases, the structural Hamming distance of NOTIME goes down, indicating that the probability of indentifying the correct DAG converges to 1. In contrast, NOTEARS and GES quickly converge to an average structural Hamming distance of 3, which is the same performance as the zero graph and as good as randomly guessing the connections. NOTEARS performs particularly poorly here, because the marginal variances of the variables decrease for the variables with lower causal order in this data. SortNRegress hovers around an average structural Hamming distance of 4, meaning that it performs worse than the zero graph here. The result also provides empirical evidence that the gradient-based optimization method used in NOTIME can identify the global optimizer of (1). Theoretical guarantees of such optimization schemes are limited to the bivariate case [Deng et al., 2023]. In addition, for high dimensional space, finding the

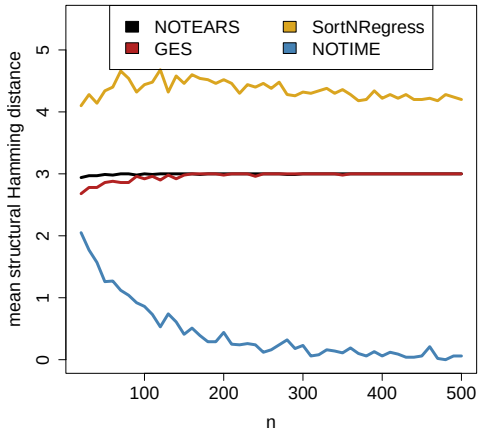


Figure 1: **Identification in n.** Average structural Hamming distance for increasing sample size. As the sample sizes increases, the probability of learning the true underlying DAG converges to 1 for NOTIME, unlike for the other methods.

Table 1: **Comparison of the methods.** In particular, GOLEM can identify the equivalent class to the linear Gaussian model, and all the differentiable methods output DAGs after thresholding. On a population level, scaling will not affect the graph identified by ICA-LiNGAM.

Method	Differentiable	Criterion	Guarantee	Scale-invariant	DAG
ICA-LiNGAM	✗	ICA	LiNGAM	✗*	✓
DirectLiNGAM	✗	DM	LiNGAM	✓	✓
GES	✗	BIC	Markov blanket	✓	✗
NOTEARS	✓	MSE	Linear EV	✗	✓*
GOLEM	✓	Likelihood	Linear Gaussian*	✗	✓*
NOTIME	✓	DM	LiNGAM	✓	✓*

globally optimal DAG based on any score function is a problem that cannot be solved exactly in polynomial time given the super-exponential scaling of the search space.

Optimization. We adopt the same optimization scheme as NOTEARS where the constrained problem (1) is solved by the augmented Lagrangian methods, c.f. Equations (11) to (14) in Zheng et al. [2018]. Since the dependence score does not have a simple closed-form derivative, we use standard gradient solvers to compute it. Note that each term of the objective function in Equation (1) is scale-equivariant. As a result, the estimated matrix $\boldsymbol{W}$ is also scale-equivariant and the sparsity pattern of $\boldsymbol{W}$ is scale-invariant. Therefore, we can start by scaling the variables before running the optimization, and de-standardize the estimated matrix $\boldsymbol{W}$ at the end. This approach has the added advantage that the search space of the elements of $\boldsymbol{W}$ is restricted to $[-1, 1]$. The following proposition makes the interaction of NOTIME with scaling the variables precise (see Appendix C for a proof, which essentially exploits the bijection between the problem on the original and scaled data). Note that $\widehat{\text{dHSIC}}$ is biased by $\mathcal{O}(\frac{1}{n})$, and is less sensitive for sparser graphs [Zhang et al., 2023], we thus de-bias and rescale it for different d such that the zero matrices have the same loss magnitude to fix the range of λ in the experiments, i.e., we optimize

$$\gamma(n \cdot \widehat{\text{dHSIC}}(\boldsymbol{X} - \boldsymbol{W}\boldsymbol{X}) - 1) \tag{2}$$

with a scaling parameter γ in practice.

Proposition 1. *Suppose we have a dataset $\boldsymbol{X}$ and let $\boldsymbol{W}^*_{\boldsymbol{X}}$ be the solution to the optimization of Equation (1) on $\boldsymbol{X}$. Now consider the scaled data $\boldsymbol{Z} := \boldsymbol{X}\boldsymbol{S}$ where $\boldsymbol{S}$ is a diagonal matrix with $\boldsymbol{S}_{jj} := s_j > 0$ for all $j \in \{1, \dots, d\}$. Then the solution to the optimization of Equation (1) on $\boldsymbol{Z}$ satisfies*

$$\boldsymbol{W}^*_{\boldsymbol{Z}} = \boldsymbol{C} \odot \boldsymbol{W}^*_{\boldsymbol{X}},$$

where $\boldsymbol{C}_{ij} = \boldsymbol{s}_i / \boldsymbol{s}_j$.

Initialization. Even though NOTIME, like NOTEARS, is a gradient-based continuous optimization problem, it is not convex and thus not guaranteed to return the global optimum, which is a common challenge for differentiable learning methods [Ng et al., 2024]. Therefore, the starting values of $\boldsymbol{W}$ given to the solver can influence the final result. As we will illustrate later, we found that in many cases, the method benefits from initializing $\boldsymbol{W}$ with the output of a combinatorial LiNGAM algorithm. This is somewhat expected given the relationship between the two. There is an additional reason why nudging the solver in the right direction can be beneficial over the default of intializing the matrix $\boldsymbol{W}$ by the zero-matrix. Consider the simple example of a bivariate DAG where $X_1 \to X_2$, then both the true structure $\boldsymbol{W}_1 = \begin{bmatrix} 0 & 0 \\ \alpha & 0 \end{bmatrix}$ and the alternative $\boldsymbol{W}_2 = \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix}$ will typically reduce the dependence between the resulting residual matrix (even though only the first leads to actual independence on a population level). This suggests that initializing with the zero matrix is like initializing on a local maximum of the objective function, and an inaccuracy in the first gradient calculation could thus lead the optimizer astray.

Computational Complexity The complexity to compute the residuals and the dHSIC statistic is both $\mathcal{O}(nd^2)$. This is also the complexity of calculating the loss. The calculation of the loss gradient requires evaluations of each coordinate. Therefore, its complexity is $\mathcal{O}(nd^2 \cdot d^2) = \mathcal{O}(nd^4)$. The method can be sped up by using faster measures of independence. There are many such options, including distance covariance and multivariance, however, we found dHSIC to be the best performer in our initial experiments, c.f. Appendix D, and therefore we put it forward as our primary proposal. We note that this is a reasonable price to pay for the general identifiability. We provide more details in Appendix E on the complexity and running time.

4 EXPERIMENTS ON SIMULATED DATA

In this section, we present an empirical study to complement the theoretical results and illustrate the favorable performance of NOTIME over existing alternatives.

Computing. All experiments were conducted on 24 Intel Xeon Platinum 8360Y processors with 36 cores, 2.40 GHz frequency, and 8GB of memory for each core. The wall time in total is 48 hours.

Simulation Setup. Our setup is inspired by the conventions in the literature of differentiable DAG learning methods [Zheng et al., 2018, Ng et al., 2020, Bello et al., 2022]. The datasets of sizes $n = 1000$ are generated according to a distribution induced by random linear SEMs with $d \in \{10, 20, 30, 50, 100\}$ variables. The graphs are randomly chosen from the Erdös-Renyi (ER) graphs with d edges. The connection strengths are sampled randomly from a uniform distribution on $[-2, -0.5] \cup [0.5, 2]$. For the noise variables, we consider three different error distributions to inspect the robustness of NOTIME to different regimes. In particular, we generate the noise variables from the lognormal, exponential and $t(3)$-distributions. Finally, we fit the different methods on the original data, the data scaled to equal variances, and the data scaled so that the marginal variances of the variables are in the reverse order(the variable which originally had the highest variance receives the lowest variance etc.). This should highlight that NOTIME is not sensitive to the scale of the variables, whereas many alternatives are.

Methods Under Comparison. For the competing methods, we include NOTEARS [Zheng et al., 2018], TOPO [Deng et al., 2023], DAGMA [Bello et al., 2022], GOLEM-NV [Ng et al., 2020], GES [Kalisch et al., 2012, Hauser and Bühlmann, 2012] and ICA-LiNGAM [Shimizu et al., 2006]. We removed sortNRegress as the results are much worse. We compare them with NOTIME initialized with a combinatorial LiNGAM algorithm as discussed in the previous section (by default NOTIME-ICA stands for NOTIME with dHSIC loss and ICA-LiNGAM initialization). For each method, we compute the average structural hamming distance (ASHD) over 10 replications per combination of the simulation parameters. Each time, we run the algorithm for $\lambda \in \{10^{-4}, 10^{-3}, 0.005, 0.01, 0.1, 0.5, 1, 2\}$, a cutoff threshold in $\{0.05, 0.1, 0.3\}$ (for GOLEM-NV and TOPO we used the recommended parameters from the authors, and for NOTIME we choose γ in Equation (2) as Table 14) and report the best result. This allows us to purely compare the objective functions without them being influenced by the selection of λ.

Results. Table 2 compares ASHD between NOTIME and the competitors. The best method and the best differentiable method are highlighted. Table 3 compares all differentiable methods initialized by ICA-LiNGAM. There are several key points to make here.

1. Stable performance. NOTIME has a stable performance over the different simulation setups. It strongly outperforms the competition in almost all situations where the variances do not carry information about the causal order. In other situations, it is competitive.

2. Scale matters for benchmarks. NOTEARS, DAGMA, and GOLEM-NV (with LiNGAM initialization) perform well as long as the variables are in the original scale. In that case, the noise variances are equal and the MSE loss is consistent. Thus these methods can identify the causal structure. As soon as this information disappears by standardizing the variables to equal or reverse scales, they fail. In particular, for reversely rescaled data, they essentially perform no better than the zero graph. Although GES is stable after rescaling, it does not identify the correct causal graph. In addition, its performance was only marginally better on the exponential noise.

3. LiNGAM initialization works well for NOTIME. Initializing NOTIME by ICA-LiNGAM generally improves the performance. In addition, NOTIME improves on the starting value. We also notice that NOTEARS and DAGMA do not gain much from LiNGAM initialization. While GOLEM-NV-ICA performs competitively on data with the original or equal scale, it still fails on reversely rescaled data. This again highlights the importance of a consistent loss.

4. Role of thresholding and sparsity parameters. We also investegated the importance of thresholding and sparsity parameters in Appendix D.3.2 as Ng et al. [2024]. In general, we found that the optimal choice of the parameters is data-driven. Small thresholds perform well for reversely scaled data, while large thresholds are better for data scaled by the other methods.

5. Identifiability shows that large samples help. Perhaps most interestingly, the identifiability properties allow us to meaningfully analyze the importance of large sample sizes in causal discovery. Therefore, we additionally tested the sample size $n = 100$ and $n = 1000$ for relatively small $d \in \{3, 4, 5, 10\}$ with each node having a 50% chance of being connected (so it can be as dense as ER4 for $d = 10$). We choose relatively small d as the results on them are more sensitive to the change of the sample size. Table 4 in Appendix D.1 shows the comparison of ASHD between NOTIME and the competition for $n = 100$, whereas Table 5 there presents the same results for $n = 1000$. NOTIME-zero only differs from NOTIME-ICA by zero initialization, whereas NOTIME(m)-zero additionally uses multivari-

Table 2: **Accuracy results.** ASHD of NOTIME versus benchmarks on lognormal, exponential, and t_3 noise with $n = 1000$. Each panel in the table represents one noise type. The best method and the best differentiable method are highlighted.

scaletype	**original**					**equal**					**reverse**				
d	10	20	30	50	100	10	20	30	50	100	10	20	30	50	100
lognormal noise															
ICA-LiNGAM	0.4	1.7	3.9	6.8	23.7	2.4	7.2	12.3	30.3	80.9	10.8	30.5	64.5	110.7	297.8
GES	4.2	7.0	15.3	54.5	102.5	4.2	7.0	15.3	54.5	102.5	4.2	7.0	15.3	54.5	102.5
DAGMA	**0.3**	**0.6**	**0.0**	**1.2**	**1.8**	7.4	15.9	23.1	35.1	71.5	13.0	25.5	36.5	71.3	149.8
NOTEARS	0.6	0.9	2.8	2.5	2.9	10.0	17.0	26.2	40.0	82.5	13.3	26.1	39.8	72.5	150.5
GOLEM-NV	16.5	33.7	52.5	88.4	175.6	16.1	26.1	42.2	69.5	147.8	9.2	19.1	32.9	49.8	102.0
TOPO	1.4	2.2	2.7	15.3	35.4	20.2	45.4	65.1	95.2	173.5	30.8	83.6	133.5	188.3	481.2
NOTIME-ICA	1.8	7.1	10.9	10.0	23.8	**1.8**	**6.9**	**10.9**	**21.4**	**55.3**	**1.7**	**6.8**	**10.9**	**32.5**	**71.6**
exponential noise															
ICA-LiNGAM	**0.1**	0.4	2.0	3.0	38.7	1.5	11.9	25.8	54.8	256.2	12.6	41.1	70.5	128.0	367.6
GES	6.0	7.3	9.8	48.5	92.3	6.0	**7.3**	**9.8**	48.5	92.3	6.0	**7.3**	**9.8**	**48.5**	**92.3**
DAGMA	**0.2**	**0.0**	**0.9**	**0.9**	**0.2**	7.1	16.9	**23.4**	**33.1**	**71.6**	10.2	20.5	30.7	55.3	114.1
NOTEARS	0.3	0.3	1.7	1.5	0.4	8.8	18.4	25.4	38.1	81.0	10.3	20.6	31.0	55.2	113.9
GOLEM-NV	16.3	33.7	53.4	87.6	175.8	16.6	25.6	41.8	72.7	145.5	9.5	20.7	32.8	51.7	108.4
TOPO	2.9	2.4	2.6	14.3	15.5	21.3	46.1	63.9	90.4	172.6	33.2	79.2	132.9	176.7	468.3
NOTIME-ICA	0.6	8.4	17.5	20.8	71.8	**0.7**	**10.9**	**23.6**	39.9	100	**2.0**	**11.7**	**24.3**	**49.1**	100
t_3 noise															
ICA-LiNGAM	**0.0**	0.8	1.0	2.6	11.7	1.1	13.7	12.9	26.1	117.7	12.6	35.7	69.3	115.0	318.8
GES	5.0	7.3	12.6	48.7	92.5	5.0	**7.3**	12.6	48.7	92.5	5.0	**7.3**	12.6	48.7	**92.5**
DAGMA	**0.0**	**0.7**	**0.0**	**0.8**	**0.4**	8.0	16.8	22.7	33.6	**68.7**	12.1	23.4	34.6	65.5	136.4
NOTEARS	0.4	1.2	0.9	1.9	3.0	9.8	18.1	26.9	37.8	79.5	12.7	24.4	35.7	65.8	136.4
GOLEM-NV	17.3	34.1	52.8	88.9	177.6	17.1	25.3	42.2	72.1	148.0	7.9	20.4	34.0	54.2	103.3
TOPO	0.2	1.1	0.6	14.2	17.3	14.8	35.2	51.1	82.6	161.4	24.4	58.9	95.8	155.7	457.4
NOTIME-ICA	0.4	14.9	10.2	11.2	40.8	**0.3**	**16.0**	**10.8**	**22.6**	86.5	**1.7**	**15.3**	**11.2**	**37.0**	100

ance as the independence criterion instead of dHSIC. We observe that even for $d = 5$ variables, almost all methods struggle to perform better than the zero graph (which would have an average structural hamming distance of 5) when using $n = 100$ samples. This is in stark contrast with many of the previous empirical studies which would often consider higher dimensions d with limited sample sizes but fail to recognize that these settings are only feasible for a select subset of distributions (i.e., those where the variances carry information about the causal order). For $n = 1000$, the scenario with $d = 5$ is tackled very well by NOTIME, indicating that a sample size of $n = 1000$ is enough to recover the true DAG relatively reliably in this case. For $d = 10$, we see that the performance is slightly better for $n = 1000$ than for $n = 100$, but still practically quite poor. This indicates that for $d = 10$, even larger sample sizes than $n = 1000$ are needed to reliably learn the causal DAG in practice. Our results are aligned with recent research [Gao et al., 2022, Zhao et al., 2022, Ng et al., 2024] and provide new insights on the sample complexity of *consistent and differentiable* causal learning method for LiNGAM which was previously not available.

5 REAL DATA EXPERIMENT

In this section, we compare NOTEARS with NOTIME on a real-world dataset Airfoil self noise[1]. There are 6 variables and 1504 observations. The known tiers for the causal ordering[2] are 1. Chord, Attack, Velocity. 2. Frequency, Displacement. 3. Pressure. Edges within the first tier or from later tiers to former tiers are forbidden. We use the default parameters from NOTEARS for both methods, with $\lambda = 0.1$ and a

[1] https://archive.ics.uci.edu/dataset/291/airfoil+self+noise (CC BY 4.0 license)

[2] https://github.com/cmu-phil/example-causal-datasets/tree/main/real/airfoil-self-noise (CC0-1.0 license)

Table 3: **Accuracy results.** ASHD of NOTIME versus differentiable methods initialized by ICA-LiNGAM on the same setting as Table 2. The best method is highlighted.

scaletype	**original**					**equal**					**reverse**				
d	10	20	30	50	100	10	20	30	50	100	10	20	30	50	100
lognormal noise															
NOTEARS-ICA	0.5	0.8	1.6	2.7	2.8	9.4	17.3	25.7	41.5	89.2	15.8	30.7	42.9	72.5	150.4
DAGMA-ICA	0.3	0.6	**0.0**	15.8	1.8	7.2	15.9	22.4	40.3	62.2	15.6	29.9	43.1	82.0	173.8
GOLEM-NV-ICA	**0.0**	**0.0**	2.1	**0.1**	**0.8**	**1.8**	**5.6**	**10.7**	**19.9**	**49.0**	10.5	25.2	42.8	69.7	159.1
NOTIME-ICA	1.8	7.1	10.9	10.0	23.8	**1.8**	6.9	10.9	21.4	55.3	**1.7**	**6.8**	**10.9**	**32.5**	**71.6**
exponential noise															
NOTEARS-ICA	0.2	0.2	0.5	1.4	0.7	8.3	18.7	25.5	**38.9**	84.6	11.8	22.5	33.1	55.2	113.9
DAGMA-ICA	0.2	**0.0**	0.5	10.1	**0.2**	7.1	17.0	22.6	43.4	**69.4**	11.4	22.4	36.9	68.5	168.5
GOLEM-NV-ICA	**0.0**	**0.0**	**0.3**	**0.0**	19.4	2.2	**10.4**	**20.5**	39.1	123.7	12.6	32.8	48.9	85.5	191.3
NOTIME-ICA	0.6	8.4	17.5	20.8	71.8	**0.7**	10.9	23.6	39.9	100	**2.0**	**11.7**	**24.3**	**49.1**	100
t_3 noise															
NOTEARS-ICA	0.5	0.7	0.8	1.9	2.9	8.9	18.4	26.1	37.9	87.2	14.9	28.5	40.2	65.8	136.3
DAGMA-ICA	**0.0**	0.7	0.0	4.8	**0.4**	8.0	16.0	22.0	40.8	**59.7**	14.8	28.1	40.3	73.7	185.1
GOLEM-NV-ICA	**0.0**	**0.0**	**0.0**	**0.1**	0.5	2.3	**10.9**	12.2	**21.2**	69.1	11.4	30.2	45.4	70.5	167.9
NOTIME-ICA	0.4	14.9	10.2	11.2	40.8	**0.3**	16.0	**10.8**	22.6	86.5	**1.7**	**15.3**	**11.2**	**37.0**	100

cutoff threshold 0.3, which were proved to work well in previous studies. For NOTIME, we use the dHSIC measure and initialize either by the zero matrix or ICA-LiNGAM. The results are shown in Appendix F. We clearly see that NOTIME provides sparser estimations. In particular, we only have one or two edges which are supposed to be forbidden when initialized by the zero matrix and ICA-LiNGAM respectively. On the contrary, the graph from NOTEARS is rather dense. There are in total 7 edges which should not exist. In addition, loops are observed. Even for the larger $\lambda = 5$, NOTEARS still identifies 5 wrong edges. To conclude, NOTIME shows better performance in this example.

6 CONCLUSION

We summarize the key takeaways and potential avenues for future research. The key takeaways are

- Differentiable causal discovery with identifiability guarantees is possible.
- NOTIME achieves this and as a result outperforms the competition in the LiNGAM family.
- A large sample size is crucial for reliable causal discovery, even in small dimensions.

This work opens up several interesting avenues for feature research. For example, with identifiability guarantees, it now makes sense to thoroughly study the sample sizes required for reliable causal discovery in practice and how this scales with the number of variables. Additionally, there are many variants of NOTIME which could be studied and whose performances could be compared. Finally, one may think about how to extend these results to nonlinear SEMs. A relatively straightforward extension beyond strictly linear models is to additive models with known link functions. In that case, the algorithm can essentially be used as-is. The extension to nonlinear additive models with unknown link functions requires specific treatment if one wants to preserve the identifiability properties. In contrary to the linear case, merely assuming joint independence of the residuals may not guarantee identifying the true DAG, which is similar to the case in the nonlinear ICA [Hyvärinen and Pajunen, 1999], therefore, additional assumptions need to be made, which becomes a potential future direction.

Limitations The LiNGAM assumption is central to the theoretical and empirical results in this paper. The LiNGAM family is only a subset of the known conditions under which identifiability is possible. A more flexible family is that of the nonlinear additive noise models, which we consider a future challenge for differentiable DAG learning with identifiability guarantees.

Acknowledgements

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen" programme (JR, TV, RY).

References

Bello, K., Aragam, B., and Ravikumar, P. (2022). Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. In *Advances in Neural Information Processing Systems*, volume 35, pages 8226–8239.

Böttcher, B., Keller-Ressel, M., and Schilling, R. (2019). Distance multivariance: New dependence measures for random vectors. *The Annals of Statistics*, 47(5):2757–2789.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554.

Colombo, D., Maathuis, M. H., et al. (2014). Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314.

Deng, C., Bello, K., Aragam, B., and Ravikumar, P. K. (2023). Optimizing notears objectives via topological swaps. In *International Conference on Machine Learning*, pages 7563–7595. PMLR.

Gao, M., Tai, W. M., and Aragam, B. (2022). Optimal estimation of gaussian dag models. In *International Conference on Artificial Intelligence and Statistics*, pages 8738–8757. PMLR.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20.

Hauser, A. and Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13:2409–2464.

Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., and Schölkopf, B. (2020). Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53.

Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439.

Hyvärinen, A. and Smith, S. M. (2013). Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14(1):111–152.

Immer, A., Schultheiss, C., Vogt, J. E., Schölkopf, B., Bühlmann, P., and Marx, A. (2023). On the identifiability and estimation of causal location-scale noise models. In *International Conference on Machine Learning*, pages 14316–14332. PMLR.

Jin, Z. and Matteson, D. S. (2018). Generalizing distance covariance to measure and test multivariate mutual dependence via complete and incomplete v-statistics. *Journal of Multivariate Analysis*, 168:304–322.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26.

Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. (2019). Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*.

Laumann, F., Von Kügelgen, J., Park, J., Schölkopf, B., and Barahona, M. (2023). Kernel-based independence tests for causal structure learning on functional data. *Entropy*, 25(12):1597.

Loh, P.-L. and Bühlmann, P. (2014). High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(1):3065–3105.

Monti, R. P., Zhang, K., and Hyvärinen, A. (2020). Causal discovery with general non-linear relationships using non-linear ica. In *Uncertainty in Artificial Intelligence*, pages 186–195. PMLR.

Mooij, J., Janzing, D., Peters, J., and Schölkopf, B. (2009). Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th annual International Conference on Machine Learning*, pages 745–752.

Ng, I., Ghassami, A., and Zhang, K. (2020). On the role of sparsity and dag constraints for learning linear dags. In *Advances in Neural Information Processing Systems*, volume 33, pages 17943–17954.

Ng, I., Huang, B., and Zhang, K. (2024). Structure learning with continuous optimization: A sober look and beyond. In Locatello, F. and Didelez, V., editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 71–105. PMLR.

Peters, J. and Bühlmann, P. (2014). Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. MIT press.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053.

Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2018). Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31.

Reisach, A. G., Seiler, C., and Weichwald, S. (2021). Beware of the simulated dag! varsortability in additive noise models. *Advances in Neural Information Processing Systems*, 34.

Reisach, A. G., Tami, M., Seiler, C., Chambaz, A., and Weichwald, S. (2023). A scale-invariant sorting criterion to find a causal order in additive noise models. In *Advances in Neural Information Processing Systems*, pages 785–807.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10).

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvarinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., Bollen, K., and Hoyer, P. (2011). Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12(Apr):1225–1248.

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, prediction, and search*. MIT press.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 804–813. AUAI Press.

Zhang, T., Zhang, Y., and Zhou, T. (2023). Statistical insights into hsic in high dimensions. *Advances in Neural Information Processing Systems*, 36:19145–19156.

Zhao, R., He, X., and Wang, J. (2022). Learning linear non-gaussian directed acyclic graph with diverging number of nodes. *Journal of Machine Learning Research*, 23(269):1–34.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pages 9472–9483.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. (2020). Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] See Sections 2 and 3 and Appendices B and C.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] See Sections 3 and 4.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] See Appendix A.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] See Section 3.
   (b) Complete proofs of all theoretical results. [Yes] See Sections 2 and 3.
   (c) Clear explanations of any assumptions. [Yes] See Section 3.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] See Appendix A.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See Section 4.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] See Section 4.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] See Section 4.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] See Sections 4 and 5.
   (b) The license information of the assets, if applicable. [Yes] See Sections 4 and 5
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] See Appendix A.
   (d) Information about consent from data providers/curators. [Not Applicable] We use public datasets.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Differentiable Causal Structure Learning with Identifiability Guarantees: Supplementary Materials

A CODE

We provide the code for NOTIME at https://github.com/STAN-UAntwerp/NOTIME.

B PROOF OF THEOREM 1

Proof. Since on a population level, the parameter for the regularization would vanish, it suffices to study the case where $\lambda = 0$. Denote the objective function with $\Theta(\boldsymbol{W}) \coloneqq \text{dHSIC}(\boldsymbol{X} - \boldsymbol{W}\boldsymbol{X})$. Note that $\Theta(\boldsymbol{W}) \geq 0$ for all matrices $\boldsymbol{W}$. First note that $\boldsymbol{B}$ is a minimizer of the objective function. We have

$$\begin{aligned}\Theta(\boldsymbol{B}) =& \text{dHSIC}(\boldsymbol{X} - \boldsymbol{B}\boldsymbol{X}) \\ =& \text{dHSIC}(\boldsymbol{\varepsilon}) \\ =& 0,\end{aligned}$$

where we have used the fact that $\boldsymbol{X}$ satisfies the LiNGAM model. Furthermore, as $\boldsymbol{B}$ can be permuted to a strictly lower triangular matrix, we have $\text{Tr}(e^{\boldsymbol{B}\odot\boldsymbol{B}}) = d$. Hence, $\boldsymbol{B}$ minimizes the objective function and satisfies the constraint.

Next, we show that $\boldsymbol{B}$ is the unique minimizer of Θ within the class of matrices satisfying the trace constraint. By the identifiability of independent component analysis (ICA) model [Comon, 1994], we have that the minimizer for $\boldsymbol{W}$ to $\text{dHSIC}(\boldsymbol{X} - \boldsymbol{W}\boldsymbol{X}) = 0$ is identified up to permutation and scale of $(I - \boldsymbol{W})$. Given that $\boldsymbol{B}$ is a minimizer of Θ, we know that all the solutions to for $\boldsymbol{W}$ to $\Theta(\boldsymbol{W})$ are of the form

$$I - PS(I - \boldsymbol{B}),$$

where P is a $d \times d$ permutation matrix and S is a $d \times d$ diagonal matrix with non-zero elements on the diagonal. Now consider any candidate $\tilde{\boldsymbol{B}} = I - \tilde{P}\tilde{S}(I - \boldsymbol{B})$. For $\tilde{\boldsymbol{B}}$ to be a permissible solution, it must be acyclic since otherwise $\text{Tr}(e^{\tilde{\boldsymbol{B}}\odot\tilde{\boldsymbol{B}}}) \neq d$. We will show that, unless $\tilde{P} = \tilde{S} = I$, $\tilde{\boldsymbol{B}}$ has at least one cycle.

Note first that we must have that $(\tilde{\boldsymbol{B}})_{ii} = 0$ for all $i \in \{1, \dots, d\}$ (though this is not a sufficient condition for acyclicity when $\boldsymbol{B}$ is not necessarily non-negative). This implies that the diagonal elements of $\tilde{\boldsymbol{B}} = I - \tilde{P}\tilde{S}(I - \boldsymbol{B}) = I - \tilde{P}\tilde{S} + \tilde{P}\tilde{S}\boldsymbol{B}$ should be zero.
Now consider the permutation matrix $\tilde{P}$ and suppose that it is not the identity matrix. The corresponding permutation can be written as a product of disjoint cycles. Additionally, there is at least one non-trivial cycle (longer than length 1), since we assume $\tilde{P} \neq I$. Suppose w.l.o.g. that this cycle involves the indices $1, \dots, J$ with $J \geq 2$. Then we have that $(I - \tilde{P}\tilde{S})_{jj} = 1$ for $j \in \{1, \dots, J\}$. Hence, we must have that $(\tilde{P}\tilde{S}\boldsymbol{B})_{jj} = -1$ for $j \in \{1, \dots, J\}$. Now this means that $\boldsymbol{B}$ itself has at least one cycle formed by the indices $1, \dots, J$, which is a contradiction. Hence, we conclude that $\tilde{P}$ must be the identity matrix.
Similarly, suppose that $\tilde{S}$ is not the identity. Then $I - \tilde{S}$ has at least one diagonal element, say $(I - \tilde{S})_{jj}$, which is non-zero. To satisfy the necessary condition for acyclicity, we must have that $(\tilde{S}\boldsymbol{B})_{jj} = -(I - \tilde{S})_{jj} \neq 0$, but this means $\boldsymbol{B}$ itself has a non-zero diagonal element and is thus cyclic, which is again a contradiction. We thus conclude that $\tilde{S} = I$.
Since, $\tilde{P} = \tilde{\boldsymbol{B}} = I$, we obtain that the only permissible solution is $\boldsymbol{B}$ itself.

□

C PROOF OF PROPOSITION 1

Proof. Denote the objective function of Equation (1) in the main paper as $\boldsymbol{\Theta}(\boldsymbol{X}, \boldsymbol{W})$. By definition of $\boldsymbol{W}_{\boldsymbol{X}}^*$, we have $\boldsymbol{\Theta}(\boldsymbol{X}, \boldsymbol{W}_{\boldsymbol{X}}^*) < \boldsymbol{\Theta}(\boldsymbol{X}, \boldsymbol{W})$ for all $\boldsymbol{W} \neq \boldsymbol{W}_{\boldsymbol{X}}^*$. We first show that $\boldsymbol{\Theta}(\boldsymbol{Z}, \boldsymbol{C} \odot \boldsymbol{W}_{\boldsymbol{X}}^*) = \boldsymbol{\Theta}(\boldsymbol{X}, \boldsymbol{W}_{\boldsymbol{X}}^*)$.

Consider the first term of the objective. We have that $\boldsymbol{Z} - (\boldsymbol{C} \odot \boldsymbol{W}_{\boldsymbol{X}}^*)\boldsymbol{Z} = (\boldsymbol{X} - \boldsymbol{W}_{\boldsymbol{X}}^* \boldsymbol{X})\, S$. Since dHSIC is scale-invariant, we have $\text{dHSIC}\left(\boldsymbol{Z} - (\boldsymbol{C} \odot \boldsymbol{W}_{\boldsymbol{X}}^*)\boldsymbol{Z}\right) = \text{dHSIC}\left((\boldsymbol{X} - \boldsymbol{W}_{\boldsymbol{X}}^* \boldsymbol{X})\, S\right) = \text{dHSIC}\left(\boldsymbol{X} - \boldsymbol{W}_{\boldsymbol{X}}^* \boldsymbol{X}\right)$.

Now consider the second term. We have $\boldsymbol{\Sigma}_{\boldsymbol{Z},ij} = \frac{\sigma_{\boldsymbol{Z}_j}}{\sigma_{\boldsymbol{Z}_i}} = \frac{\sigma_{\boldsymbol{X}_j}}{\sigma_{\boldsymbol{X}_i}} \frac{\boldsymbol{s}_j}{\boldsymbol{s}_i}$, and so $\boldsymbol{\Sigma}_{\boldsymbol{Z}} = \boldsymbol{\Sigma}_{\boldsymbol{X}} \odot \boldsymbol{C}^\intercal$. As a result, $\boldsymbol{W}_{\boldsymbol{Z}}^* \odot \boldsymbol{\Sigma}_{\boldsymbol{Z}} = \boldsymbol{C} \odot \boldsymbol{W}_{\boldsymbol{X}}^* \odot \boldsymbol{\Sigma}_{\boldsymbol{X}} \odot \boldsymbol{C}^\intercal = \boldsymbol{W}_{\boldsymbol{X}}^* \odot \boldsymbol{\Sigma}_{\boldsymbol{X}} \odot (\boldsymbol{C} \odot \boldsymbol{C}^\intercal) = \boldsymbol{W}_{\boldsymbol{X}}^* \odot \boldsymbol{\Sigma}_{\boldsymbol{X}}$.

Finally, both $\boldsymbol{W}_{\boldsymbol{X}}^*$ and $\boldsymbol{C} \odot \boldsymbol{W}_{\boldsymbol{X}}^*$ are acyclic, so the third term of the objective is zero for both.

Now suppose there is a solution $\tilde{\boldsymbol{W}}$ on $\boldsymbol{Z}$ for which $\boldsymbol{\Theta}(\boldsymbol{Z}, \tilde{\boldsymbol{W}}) < \boldsymbol{\Theta}(\boldsymbol{X}, \boldsymbol{W}_{\boldsymbol{X}}^*)$. This would imply that $\boldsymbol{\Theta}(\boldsymbol{X}, \boldsymbol{C}^\intercal \odot \tilde{\boldsymbol{W}}) < \boldsymbol{\Theta}(\boldsymbol{X}, \boldsymbol{W}_{\boldsymbol{X}}^*)$. This is a contradiction, since $\boldsymbol{W}_{\boldsymbol{X}}^*$ is the unique minimizer of $\boldsymbol{\Theta}(\boldsymbol{X}, \boldsymbol{W})$ with respect to $\boldsymbol{W}$. □

D ADDITIONAL SIMULATION RESULTS

D.1 Experiments on $n = 100$ and $n = 1000$

The results are provided in Table 4 and Table 5.

Table 4: **Accuracy results.** ASHD of NOTIME versus benchmarks on lognormal, exponential and t_3 noise with $n = 100$. Each panel in the table represents one noise type.

scaletype	**original**				**equal**				**reverse**			
d	3	4	5	10	3	4	5	10	3	4	5	10
lognormal noise												
NOTEARS	**0.1**	1.2	2.4	**16.8**	1.4	2.9	3.9	21.4	1.6	3.3	5.2	27
GES	1.2	3.3	5.3	32.6	1.2	3.3	5.3	32.6	1.2	3.3	5.3	32.6
SortNRegress	1.2	3.2	6.8	47.5	1.2	3.2	6.8	47.5	1.2	3.2	6.8	47.5
NOTIME-zero	0.5	2.2	3.6	22	0.5	2.4	4	22	**0.6**	2.8	4.5	22
NOTIME-ICA	0.5	**0.9**	3.1	20.5	**0.3**	1.1	3.4	**19.7**	**0.6**	**1.4**	3.6	**19.9**
NOTIME(m)-zero	0.3	**0.9**	**2.4**	21.9	0.7	**1**	**2.5**	22.6	0.8	1.7	**3.2**	23.7
exponential noise												
NOTEARS	**0.2**	**0.7**	**1.6**	**16.4**	1.3	2.9	3.8	21.9	1.4	3	4.5	22.4
GES	1.2	2.9	5.3	33.3	1.2	2.9	5.3	33.3	1.2	2.9	5.3	33.3
SortNRegress	1.1	3	6	48.3	1.1	3	6	48.3	1.1	3	6	48.3
NOTIME-zero	0.3	1.8	2.9	22	**0.2**	2.3	2.8	22	**0.2**	2.7	3.7	22
NOTIME-ICA	0.5	1	3.7	20.3	0.4	**1.2**	2.9	**20.3**	**0.2**	**1.9**	3.6	**20.8**
NOTIME(m)-zero	0.6	1.3	2.7	22.8	0.4	**1.2**	**2.4**	23.1	0.5	2.1	**3.3**	23.5
t_3 noise												
NOTEARS	**0.1**	**0.8**	**2.3**	**17.1**	1.4	2.6	3.8	**21.6**	1.5	3.3	4.7	24.8
GES	1.2	2.4	5.1	35.1	1.2	2.4	5.1	35.1	1.2	2.4	5.1	35.1
SortNRegress	1.6	3.5	7.4	47.9	1.6	3.5	7.4	47.9	1.6	3.5	7.4	47.9
NOTIME-zero	1.3	2.6	4.5	22	1.3	3	4.5	22	1.2	3	4.4	**22**
NOTIME-ICA	1.1	1.9	3.8	21.3	**1.1**	**2.3**	**3.6**	22	**0.9**	**2.7**	**4.2**	**22**
NOTIME(m)-zero	1.7	3	3.8	21.5	1.7	3.5	4.2	22.1	1.9	3.3	4.5	22.5

Table 5: **Accuracy results.** ASHD of NOTIME versus benchmarks on lognormal, exponential, and t_3 noise with $n = 1000$. Each panel in the table represents one noise type, analogous to table 4

scaletype	**original**				**equal**				**reverse**			
d	3	4	5	10	3	4	5	10	3	4	5	10
lognormal noise												
NOTEARS	**0**	0.5	1.8	**16.2**	1.4	3	3.8	22	2.1	3.9	5.6	27.7
GES	1	2.5	3.7	35	1	2.5	3.7	35	1	2.5	3.7	35
SortNRegress	1.6	2.6	7.5	52.4	1.6	2.6	7.5	52.4	1.6	2.6	7.5	52.4
NOTIME-zero	0.1	0.8	1.5	17.5	**0.1**	**0.5**	1.2	20.9	**0.1**	0.9	**1.1**	22
NOTIME-ICA	0.1	**0.4**	**1.1**	**16.2**	**0.1**	**0.5**	**1.1**	**16.9**	**0.1**	**0.5**	1.2	**17.2**
NOTIME(m)-zero	0.2	0.6	2.3	19.6	0.2	0.6	2.1	21.1	0.3	1.2	2.2	21.5
exponential noise												
NOTEARS	**0**	0.6	**1.2**	**15.8**	1.3	3	4.1	21.3	1.2	3	4.5	22.4
GES	1.2	2.7	5.2	33.6	1.2	2.7	5.2	33.6	1.2	2.7	5.2	33.6
SortNRegress	1.4	3.7	8	52.1	1.4	3.7	8	52.1	1.4	3.7	8	52.1
NOTIME-zero	0.1	1	1.4	16.3	**0.1**	1	1.4	20.4	0.2	1.2	1.4	22
NOTIME-ICA	0.1	**0.6**	**1.2**	15.9	**0.1**	**0.6**	**1.2**	**16.1**	**0.1**	**0.7**	**1.3**	**17.7**
NOTIME(m)-zero	0.3	0.9	1.9	20.3	0.3	1	1.6	19.5	0.5	1	2.2	21.1
t_3 noise												
NOTEARS	**0**	**0.4**	**1.1**	**15.7**	0.8	2.7	4.2	21.3	1.5	3.4	5.2	24.6
GES	1	2.5	4.7	34.6	1	2.5	4.7	34.6	1	2.5	4.7	34.6
SortNRegress	1.6	3.5	7.4	47.9	1.6	3.5	7.4	47.9	1.6	3.5	7.4	47.9
NOTIME-zero	0.3	1.2	2.5	18.7	0.3	**1.2**	**2.4**	21.8	**0.3**	**1.2**	2.7	22
NOTIME-ICA	0.3	1.2	2.2	17.4	0.3	1.3	2.5	**17.2**	**0.3**	**1.2**	**2.6**	**18.6**
NOTIME(m)-zero	0.2	1.8	2.5	21.5	**0.2**	2.1	2.5	21.7	0.5	2.4	3.4	24

D.2 Standard Errors

Tables 6 to 13 below present the standard errors of the estimated mean structural Hamming distances of the empirical experiments.

Table 6: Standard errors of NOTIME versus benchmarks on ER1 with lognormal, exponential, and t_3 noise with $n = 1000$. Each panel in the table represents one noise type, analogous to Table 2 in the main paper.

scaletype	**original**					**equal**					**reverse**				
d	10	20	30	50	100	10	20	30	50	100	10	20	30	50	100
lognormal noise															
ICA-LiNGAM	0.7	1.8	3.2	4.4	6.5	3.8	9.2	11.8	21.6	25.0	4.9	12.7	18.0	27.1	48.8
GES	1.7	4.2	7.7	17.4	36.8	1.7	4.2	7.7	17.4	36.8	1.7	4.2	7.7	17.4	36.8
DAGMA	0.9	0.7	0.0	1.9	1.8	2.7	4.9	3.5	3.6	7.9	1.8	2.4	2.3	5.8	10.1
NOTEARS	1.0	1.9	2.1	4.2	3.7	0.0	6.1	5.3	6.4	10.5	2.1	3.7	4.7	8.6	19.6
GOLEM-NV	2.6	3.5	5.2	5.5	7.9	3.5	5.0	8.0	7.0	12.4	2.6	3.4	3.9	7.9	9.2
TOPO	1.5	1.3	1.7	2.7	4.1	5.8	4.1	14.5	13.8	22.7	4.4	9.6	29.9	40.9	120.1
NOTIME-ICA	3.8	8.6	11.7	8.9	7.1	3.8	8.7	11.7	14.1	17.7	3.9	8.7	11.7	36.0	23.3
exponential noise															
ICA-LiNGAM	0.3	0.5	1.4	1.8	44.7	2.0	9.8	15.0	25.9	72.3	5.6	8.2	16.6	25.8	37.0
GES	3.4	4.8	3.2	11.3	31.3	3.4	4.8	3.2	11.3	31.3	3.4	4.8	3.2	11.3	31.3
DAGMA	0.6	0.0	1.4	1.2	0.6	2.2	5.5	4.0	5.8	9.2	0.4	0.9	1.0	3.2	4.8
NOTEARS	0.6	0.6	3.1	2.0	2.0	1.8	6.4	3.9	6.4	10.7	0.6	0.9	1.3	4.2	5.8
GOLEM-NV	2.3	3.0	7.4	3.8	7.1	3.8	4.0	8.5	5.3	15.0	3.1	4.3	2.9	6.9	9.4
TOPO	1.9	2.0	1.6	0.9	1.4	3.9	3.6	11.8	9.4	20.5	6.8	9.0	33.2	32.1	129.1
NOTIME-ICA	1.5	7.7	9.5	9.0	19.5	1.8	9.1	15.3	13.7	0.0	0.0	8.9	13.3	17.0	0.0
t_3 *noise*															
ICA-LiNGAM	0.0	1.1	1.3	2.0	3.7	2.5	12.6	10.9	11.9	44.0	5.9	15.3	24.2	27.5	55.2
GES	2.9	4.3	8.0	11.4	31.1	2.9	4.3	8.0	11.4	31.1	2.9	4.3	8.0	11.4	31.1
DAGMA	0.0	1.3	0.0	1.2	0.7	2.7	4.1	3.0	3.8	7.6	1.5	2.2	2.9	4.4	6.7
NOTEARS	1.2	1.8	1.8	2.1	5.5	2.7	6.5	3.4	4.4	9.6	2.3	2.7	4.1	5.1	13.0
GOLEM-NV	2.8	2.5	3.8	4.5	7.8	4.2	4.2	8.1	5.8	11.6	3.3	4.2	5.7	5.6	10.9
TOPO	0.3	1.8	1.7	1.1	5.4	2.4	4.1	9.8	11.2	18.7	4.7	6.9	15.9	23.9	136.0
NOTIME-ICA	0.7	11.2	8.4	6.6	14.8	0.7	11.1	8.8	7.7	28.3	3.3	11.2	9.3	10.6	0.0

Table 7: Standard errors of NOTIME versus differentiable methods initialized by ICA-LiNGAM on ER1 with lognormal, exponential, and t_3 noise with $n = 1000$. Each panel in the table represents one noise type, analogous to Table 3 in the main paper.

scaletype	**original**					**equal**					**reverse**				
d	10	20	30	50	100	10	20	30	50	100	10	20	30	50	100
lognormal noise															
NOTEARS-ICA	1.0	1.0	2.0	2.3	4.1	2.7	5.2	1.9	5.6	8.5	2.3	3.1	3.4	6.2	9.6
DAGMA-ICA	0.9	0.7	0.0	31.0	1.8	2.6	5.1	2.1	8.3	12.0	2.4	3.4	6.3	8.2	30.9
GOLEM-NV-ICA	0.0	0.0	6.0	0.3	1.2	1.3	6.4	8.7	14.0	12.8	3.3	8.6	11.3	11.4	10.1
NOTIME-ICA	3.8	8.6	11.7	8.9	7.1	3.8	8.7	11.7	14.1	17.7	3.9	8.7	11.7	36.0	23.3
exponential noise															
NOTEARS-ICA	0.6	0.6	1.5	1.4	1.3	2.2	4.6	4.2	5.4	8.3	1.7	2.2	2.5	3.2	5.0
DAGMA-ICA	0.6	0.0	1.5	14.9	0.6	1.9	5.3	3.9	12.8	9.0	1.1	2.2	12.3	16.4	68.3
GOLEM-NV-ICA	0.0	0.0	0.9	0.0	18.5	1.8	8.9	11.8	13.1	19.2	4.1	5.1	10.1	11.7	13.0
NOTIME-ICA	1.5	7.7	9.5	9.0	19.5	1.8	9.1	15.3	13.7	0.0	0.0	8.9	13.3	17.0	0.0
t_3 noise															
NOTEARS-ICA	1.5	1.3	1.7	1.9	4.0	2.6	3.7	2.7	3.6	7.9	2.2	1.9	3.0	3.9	7.2
DAGMA-ICA	0.0	1.3	0.0	3.9	0.7	2.6	4.2	3.8	10.2	12.1	2.4	2.2	2.9	12.5	41.8
GOLEM-NV-ICA	0.0	0.0	0.0	0.3	0.7	2.3	8.8	7.1	7.8	24.1	4.1	9.5	8.8	8.7	14.5
NOTIME-ICA	0.7	11.2	8.4	6.6	14.8	0.7	11.1	8.8	7.7	28.3	3.3	11.2	9.3	10.6	0.0

Table 8: Standard errors of ASHD of NOTIME and competing methods on lnorm noise with n = 100

scaletype	original				equal				reverse			
d	3	4	5	10	3	4	5	10	3	4	5	10
NOTEARS	0	0.41	0.34	1.51	0.22	0.4	0.39	0.7	0.34	0.63	0.55	1.44
NOTIME-zero	0.15	0.58	0.57	0.68	0.15	0.64	0.61	1.16	0.13	0.69	0.53	1.21
NOTIME-ICA	0.13	0.5	0.33	1.22	0	0.55	0.47	1.31	0.1	0.54	0.56	1.24
NOTIME(m)-zero	0.15	0.29	0.48	1.09	0.26	0.33	0.51	1.13	0.25	0.37	0.52	0.97

Table 9: Standard errors of ASHD of NOTIME and competing methods on exp noise with n = 100

scaletype	original				equal				reverse			
d	3	4	5	10	3	4	5	10	3	4	5	10
NOTEARS	0.1	0.22	0.21	1.16	0.26	0.38	0.59	1.07	0.21	0.49	0.45	1.35
NOTIME-zero	0.1	0.27	0.47	0.99	0.1	0.41	0.48	1.16	0	0.45	0.59	1.16
NOTIME-ICA	0	0.16	0.38	1.16	0	0.17	0.45	1.29	0	0.38	0.52	1.58
NOTIME(m)-zero	0.22	0.28	0.4	1.35	0.22	0.26	0.39	1.35	0.16	0.31	0.52	1.42

Table 10: Standard errors of ASHD of NOTIME and competing methods on t3 noise with n = 100

scaletype	original				equal				reverse			
d	3	4	5	10	3	4	5	10	3	4	5	10
NOTEARS	0	0.42	0.45	1.1	0.2	0.5	0.4	1.2	0.21	0.63	0.54	1.45
NOTIME-zero	0.34	0.47	0.52	1.37	0.34	0.41	0.54	1.19	0.34	0.5	0.67	1.16
NOTIME-ICA	0.26	0.52	0.41	0.85	0.27	0.45	0.48	0.81	0.27	0.52	0.64	0.8
NOTIME(m)-zero	0.37	0.37	0.54	0.94	0.37	0.38	0.63	0.82	0.2	0.43	0.67	0.99

Table 11: Standard errors of ASHD of NOTIME and competing methods on lnorm noise with n = 1000

scaletype	original				equal				reverse			
d	3	4	5	10	3	4	5	10	3	4	5	10
NOTEARS	0.1	0.22	0.43	1.37	0.26	0.51	0.49	1.12	0.3	0.71	0.65	1.81
NOTIME-zero	0.1	0.4	0.49	1.1	0.1	0.22	0.43	1.29	0.1	0.51	0.38	1.22
NOTIME-ICA	0.1	0.22	0.37	1.16	0.1	0.22	0.37	1.2	0.1	0.27	0.38	1.39
NOTIME(m)-zero	0.13	0.34	0.92	1.53	0.13	0.34	0.84	1.49	0.21	0.44	0.72	1.77

Table 12: Standard errors of ASHD of NOTIME and competing methods on exp noise with n = 1000

scaletype	original				equal				reverse			
d	3	4	5	10	3	4	5	10	3	4	5	10
NOTEARS	0	0.22	0.25	1.03	0.26	0.42	0.48	1.08	0.43	0.49	0.45	1.25
NOTIME-zero	0.1	0.47	0.43	0.93	0.1	0.47	0.49	1.29	0.1	0.67	0.49	1.45
NOTIME-ICA	0.1	0.34	0.39	1.07	0.1	0.34	0.39	1.2	0.1	0.42	0.37	1.35
NOTIME(m)-zero	0.21	0.6	0.79	2.06	0.21	0.6	0.6	2.18	0.22	0.45	0.57	2.09

Table 13: Standard errors of ASHD of NOTIME and competing methods on t3 noise with n = 1000

scaletype	original				equal				reverse			
d	3	4	5	10	3	4	5	10	3	4	5	10
NOTEARS	0.1	0.22	0.31	1.43	0.3	0.5	0.37	1.25	0.21	0.73	0.47	1.45
NOTIME-zero	0.21	0.63	0.48	1.38	0.21	0.6	0.52	1.28	0.1	0.67	0.58	1.16
NOTIME-ICA	0.1	0.41	0.54	1.27	0.1	0.47	0.7	1.28	0.1	0.47	0.73	1.51
NOTIME(m)-zero	0.13	0.7	0.88	2.15	0.13	0.76	0.83	2.54	0.16	0.86	0.83	2.2

D.3 Role of Parameters

In this Section, we discuss the choice of parameters and how they affect the performance of NOTIME.

D.3.1 γ in Equation (2)

To prevent power issues of $\widehat{\text{dHSIC}}$ as discussed in Section 3, we add a parameter γ to rescale the statistics in high dimensions. Specifically, we choose from the following in the experiments in Section 4:

Table 14: Choices of γ in the experiments.

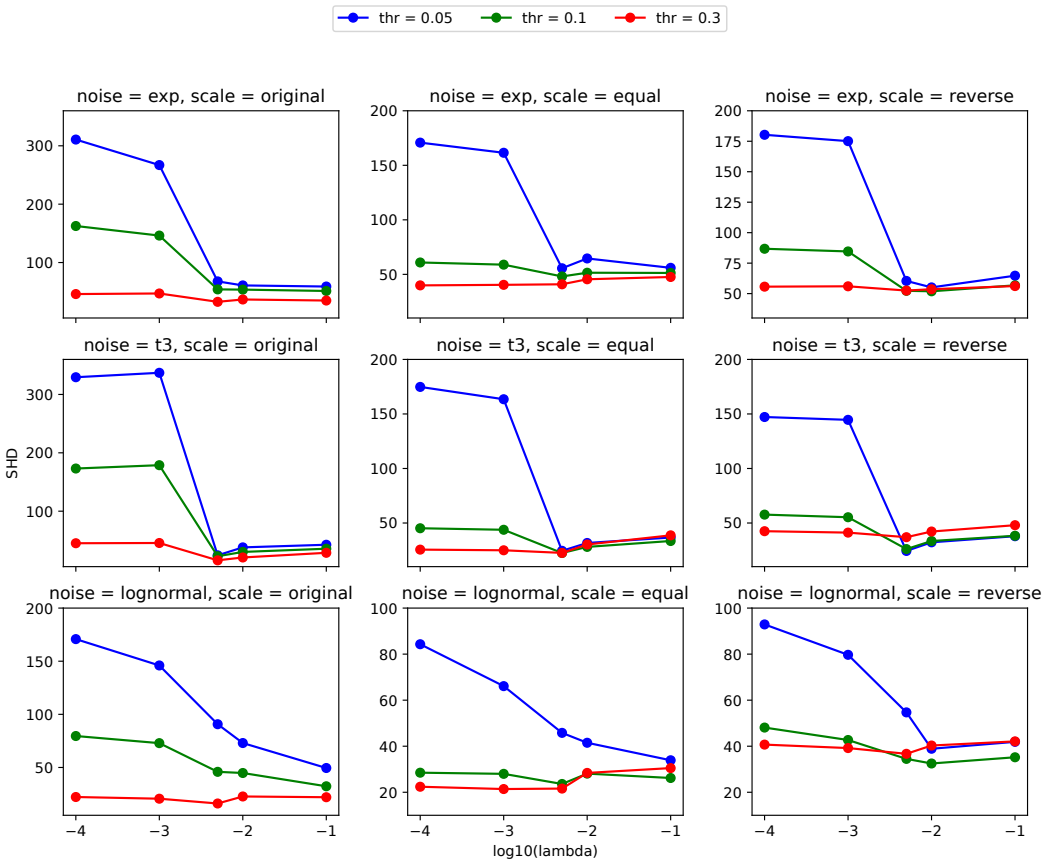
d	10	20	30	50	100
γ	$\{1, 1e^{-1}\}$	$\{1, 1e^{1}\}$	$\{1, 1e^{4}\}$	$\{1e^{11}, 1e^{13}\}$	$\{1e^{13}, 1e^{20}\}$

D.3.2 Role of the Thresholding and Sparsity Parameter

In this section, we discuss the role of thresholding and sparsity parameters for NOTIME-ICA (default dHSIC loss with ICA-LiNGAM initialization) on datasets withsample size $n = 1000$ and dimension $d = 50$, using parameters $\lambda \in \{0.0001, 0.001, 0.005, 0.01, 0.1\}$ and threshold $\in \{0.05, 0.1, 0.3\}$. Figure 2 shows the results.

In general, using $\lambda = 0.005$ provides good performance. A reduction in λ is expected compared to the experiments on low-dimensional data, as the number of parameters to regularize increases. However, the choice of the thresholding parameter is more data-driven. A large threshold generally performs well with raw or equally scaled data, while small thresholds are required for reversely scaled data. This can be induced by the data-generating process, where rescaling changes the magnitude of the coefficients.

Figure 2: SHD for NOTIME-ICA trained on different parameters. The x-axis represents $log_{10}\lambda$.



E DISCUSSION ON COMPUTATION TIME AND COMPLEXITY

Our method can be sped up by using faster measures of independence. There are many such options, including distance covariance [Jin and Matteson, 2018] and multivariance [Böttcher et al., 2019], both with complexity $\mathcal{O}(n^2 d)$, though none are quite as fast as the MSE. However, we found dHSIC to be the best performer, see Section D, and given $n = 1000$, the advantage of $\mathcal{O}(n^2 d)$ over the complexity of dHSIC $\mathcal{O}(nd^2)$ is not that significant. Therefore, we put it forward as our primary proposal, given our resultes in Section D.1.

We are convinced that there will be a price to pay for identifiability guarantees, and most likely part of it will be due to the increased complexity of measuring independence rather than sums of squares. The running time of the proposed NOTIME algorithm is about one order of magnitude slower than NOTEARS in most of our simulation setups, which we believe to be very manageable for practical use cases, and well worth the added benefit of identifiable DAGs.

Below we provide runtime comparisons for NOTEARS and NOTIME-ICA. We choose $\gamma = 1e^{-1}, 1, 1e^{4}, 1e^{11}, 1e^{13}$ respectively for $d = 10, 20, 30, 50, 100$.

Table 15: Comparison for the runtime (seconds) of NOTEARS and NOTIME-ICA for different dimensions d in median $\pm$ IQR/2.

d	NOTEARS	NOTIME-ICA
10	0.95 ± 0.58	18.70 ± 34.47
20	3.74 ± 3.50	12.44 ± 60.13
30	10.79 ± 9.61	508.35 ± 2088.42
50	53.35 ± 3.60	1957.00 ± 4096.50
100	235.56 ± 252.69	4215.92 ± 1481.41

F ANALYSIS OF THE AIRFOIL DATA

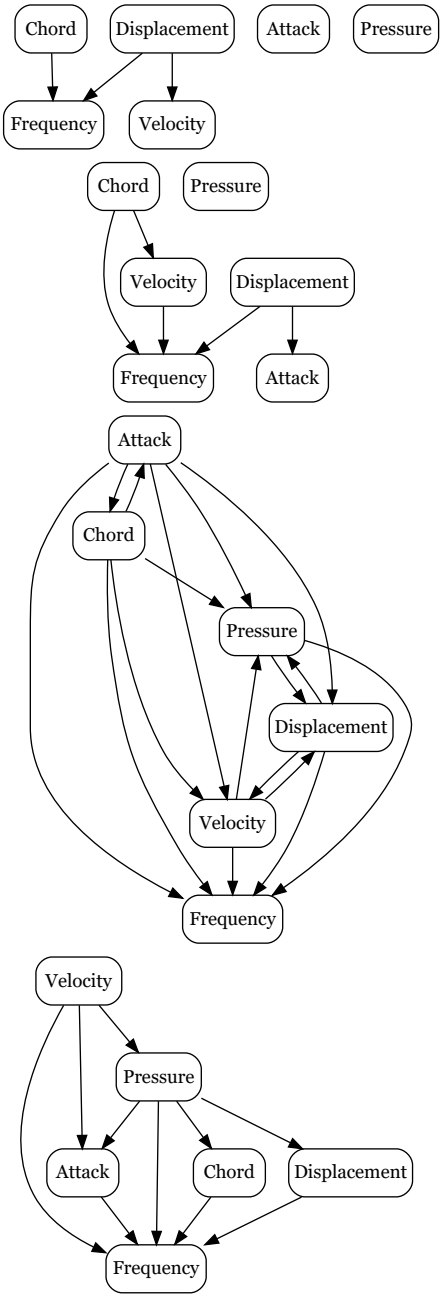


Figure 3: Prediction on the Airfoil dataset. From top to bottom: NOTIME-zero, NOTIME-ICA, NOTEARS with default parameters, NOTEARS with $\lambda = 5$.