Seminar Thesis

Differentiable constraint function (NOTEARS)

Department of Statistics
Ludwig-Maximilians-Universität München

Iker Cumplido Esteban

Munich, January 29th, 2025

Submitted in partial fulfillment of the requirements for the degree of M. Sc. Data Science
Supervised by Simon Rittel

Abstract

Causal relationships are present in many areas of our daily lives, making causal discovery an important field of research. However, identifying the causal structure of directed acyclic graphs (DAGs) remains a challenging problem. Traditional algorithms, such as the PC-algorithm, depend on discrete combinatorial constraints to ensure that the resulting graph after optimization is a valid DAG. In this thesis, we explore the approach proposed by Zheng et al. (2018), which introduces a novel algorithm called Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning (NOTEARS). Unlike previous methods, NOTEARS replaces the discrete constraint with a continuous, differentiable function, enabling more efficient and scalable learning of DAG structures. Through multiple experiments based on related studies, we assess the robustness and limitations of the NOTEARS algorithm, demonstrating its potential advantages and areas for improvement.

Contents

1 Introduction

Causality plays a fundamental role in a wide range of disciplines, such as biology or finance. Identifying how one event influences another allows researchers to predict outcomes and propose targeted interventions. While experiments can be used to uncover hidden causal relations, they are often time-consuming, expensive, or impractical. As a result, researchers have explored alternative methods, known as causal discovery, which focus on uncovering causal relationships using only observational data.

Causal structures are typically represented as directed acyclic graphs (DAGs), where a directed edge from A to B indicates that A is a direct cause of B. Ensuring acyclicity is critical for interpreting cause-and-effect relationships, since the presence of cycles would allow variables to indirectly cause themselves, complicating causal inference. Traditional algorithms for causal discovery have been therefore designed to search over the combinatorial space of DAGs. However, this search space grows superexponentially with the number of nodes, making the problem NP-hard.

The score-based NOTEARS algorithm (Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning) proposed by Zheng et al. (2018) addresses this issue by reformulating the discrete search over DAGs as a continuous optimization problem. The algorithm assumes a linear functional form and uses the least squares error for the optimization process. Its key innovation lies in transforming the acyclicity constraint into a differentiable function. Instead of iterating through possible graph structures, NOTEARS learns a weighted adjacency matrix W whose entries can be optimized with standard numerical solvers. The acyclicity constraint is imposed by a smooth function $h(W)$ that is zero if and only if W describes a DAG.

This formulation has several advantages:

- Scalability: Continuous optimization methods can handle higher dimensions better than local heuristic searches over discrete structures.
- Flexible data handling: NOTEARS does not rely on specific assumptions about the data, as it can handle both discrete and continuous variables.
- Accessibility: Standard optimization algorithms, extensively studied in the literature, can be used to solve the resulting constrained optimization problem.

Despite these benefits, NOTEARS has some limitations:

- Linear functional form: Linearity is a strong assumption about the underlying data-generating mechanism, which can lead to poor performance if real-world processes deviate significantly from it.
- Nonconvex optimization: The continuous acyclicity constraint, while efficient, does not eliminate the nonconvex nature of the problem. This remains a challenge.

- Restricted adaptability to real-world data: Related work has shown the limitations of NOTEARS with real-world data. One reason is varsortability, which quantifies how well the order of increasing marginal variance in data aligns with the causal order in a DAG. This problem will be analyzed in detail in section 4.

These limitations motivate ongoing research to refine and extend NOTEARS. In this work, we will discuss in detail how NOTEARS works, explore its core assumptions, and critically address its practical and theoretical challenges.

2 Related Work

2.1 Traditional methods

Causal discovery aims to infer the causal structure among a set of variables using observational data. Glymour et al. (2019) categorizes existing techniques into three main groups: constraint-based methods, score-based methods, and functional causal model-based methods.

- **Constraint-based methods**: These approaches discover causal structures by testing for conditional independencies between variables. For example, the PC algorithm, proposed by Spirtes and Glymour (1991), iteratively removes edges from a fully connected graph based on conditional independence tests. Constraint-based methods typically assume faithfulness, meaning that all observed independencies are due to the underlying causal structure. The statistical independence tests rely on certain distributional assumptions, and can fail if these are not met or the number of observations is reduced.

- **Score-based methods**: These methods formulate causal discovery as an optimization problem, where the goal is to find the DAG that maximizes a predefined score function, such as the Bayesian Information Criterion (BIC). One example is the Greedy Equivalence Search (GES) by Chickering (2002), which combines forward and backward greedy steps to explore the space of Markov equivalence classes of DAGs. These methods are powerful but can become computationally challenging when searching over the combinatorial space of all possible DAGs. In addition, greedy or local search strategies can get stuck in suboptimal structures.

- **Functional causal model-based methods**: These approaches assume specific functional forms for causal relationships (e.g., linear or nonlinear) and exploit asymmetries in the data to infer causal direction. For instance, the LiNGAM algorithm, proposed by Shimizu et al. (2006), identifies causal structures under the assumption of linear relationships and non-Gaussian noise. Extensions of functional causal models, such as Additive Noise Models (ANMs) introduced by Hoyer et al. (2008), can handle nonlinear relationships.

While being a score-based optimization method, NOTEARS also incorporates elements from functional causal models through its flexible parametrization of causal relationships.

It assumes that causal relationships can be represented as a linear combination of parent variables with the addition of unobserved noise. The flexibility of NOTEARS lies in its ability to adapt the form of this noise to fit various data distributions. It represents a significant advancement within the score-based paradigm by shifting the discrete search over DAGs to a continuous optimization framework.

2.2 NOTEARS Extensions

The original NOTEARS algorithm assumes a linear functional form for causal relationships. However, to accommodate more complex and nonlinear dependencies, several extensions have been proposed. One such extension is presented by the same authors of the original algorithm in Zheng et al. (2020), where they develop a nonlinear version of NOTEARS that incorporates partial derivatives to encode dependencies, maintaining the continuous acyclicity constraint. This approach allows the algorithm to learn a wider range of relationships, including additive models, and uses neural network architectures to capture nonlinear dependencies. This method is tested empirically in section 4. Additionally, other researchers have proposed neural network-based extensions, such as Yu et al. (2019), Lachapelle et al. (2019), and Huang et al. (2018), which further enhance the flexibility and applicability of NOTEARS in various settings.

Another significant extension is the GOLEM algorithm (Graph Optimization with Laplacian Eigenvalue Method) introduced by Ng et al. (2020). GOLEM improves upon the original NOTEARS by reformulating the acyclicity constraint using the Laplacian matrix of the graph. This new formulation simplifies the optimization process, making it more computationally efficient and scalable to high-dimensional datasets. Moreover, GOLEM directly optimizes the graph adjacency matrix with fewer assumptions regarding the noise distribution, enabling the discovery of causal structures in more diverse scenarios. Alongside GOLEM, the authors in Bello et al. (2022) introduced DAGMA (Directed Acyclic Graph Maximization), which employs an alternative acyclicity constraint based on the trace-exponential (log-det) characterization rather than the matrix exponential. This modification avoids the explicit computation of the matrix exponential and provides a solution which is numerically stable and computationally efficient, especially for larger graphs.

Another extension is the DYNOTEARS method proposed by Pamfil et al. (2020), which specifically targets time-series data where causal relationships evolve over time. DYNOTEARS adapts NOTEARS to model dynamic Bayesian networks (DBNs), which are widely used to represent discrete-time temporal dynamics in directed graphical models. In addition to these extensions, Zhu et al. (2019) combined NOTEARS with Reinforcement Learning (RL) to improve the search for optimal DAGs. Their approach employs an encoder-decoder neural network model to generate candidate adjacency matrices based on the observed data. By using RL as a search strategy, this extension enhances NOTEARS' ability to explore the space of possible DAGs effectively.

3 Method

In this section, the NOTEARS algorithm is introduced. It reformulates the combinatorial DAG constraint into a smooth, differentiable function $h(W)$, enabling the use of continuous optimization techniques. By combining this with a suitable score function $F(W)$ and the augmented Lagrangian method, the algorithm provides an efficient and scalable approach for causal discovery. After optimization, thresholding the resulting weight matrix produces a DAG structure suitable for causal inference tasks. The following subsections will discuss its assumptions and detail the steps of the algorithm.

3.1 Assumptions

The NOTEARS algorithm relies on multiple assumptions to ensure a reliable and consistent performance discovering DAGs. These assumptions, along with their implications and limitations, are outlined below:

1. Causal Sufficiency (No Latent Confounders): Reichenbach (1956) introduced the common cause principle, which states: *If two random variables X and Y are statistically dependent ($X \not\perp\!\!\!\perp Y$), then there exists a third variable Z that causally influences both. (As a special case, Z may coincide with either X or Y.) Furthermore, this variable Z screens X and Y from each other in the sense that given Z, they become independent, $X \perp\!\!\!\perp Y \mid Z$.*

Causal sufficiency assumes that there are no unobserved (latent) variables that simultaneously affect two or more observed variables. This ensures that any detected relationship between variables is caused by direct effects within the observed data, and not by hidden confounders. Without causal sufficiency, the algorithm would be unable to reliably identify causal dependencies, as unobserved latent variables could introduce false associations or hide true causal relationships.

2. Acyclicity: The underlying true causal graph is assumed to be a DAG. This ensures there are no feedback loops or cycles, simplifying the interpretation of causal relationships.

3. Linearity of Relationships: NOTEARS models causal relationships using linear Structural Equation Models (SEMs). Each variable X_j is expressed as a linear function of its parent variables $\mathrm{Pa}(X_j)$ (see section 3.3 for more details). While this linear assumption simplifies computation and interpretation, extensions of NOTEARS exist for nonlinear relationships, which are better suited for more complex dependencies (see section 2).

4. Identifiability and Additive Noise: Each variable X_j is influenced by its parent variables $\mathrm{Pa}(X_j)$ and an independent additive noise term z_j, where z_j is uncorrelated with $\mathrm{Pa}(X_j)$. This assumption allows the model to distinguish causal effects from randomness. The authors claim that the linear model with additive noise is identifiable and consistent for both Gaussian and non-Gaussian noise. However, the specific conditions for this to be true depend on the noise distribution:

- Gaussian noise: Identifiability holds only if the variances of the noise variables are known and equal (see Peters and Bühlmann (2014)), which in practice is not realistic unless we work with synthetic data.
- Non-Gaussian noise: The linear SEM is fully identifiable as the non-Gaussian noise introduces asymmetries that enable the unique determination of causal directions, as demonstrated in the LiNGAM framework by Shimizu et al. (2006).

For linear Gaussian SEMs without information about the variance, the method can only identify the causal graph up to its Markov equivalence class. This means that certain causal directions remain ambiguous (e.g., distinguishing between $A \to B$ and $B \to A$), as all graphs within the same equivalence class imply the same set of conditional independencies.

5. Large Sample Size: The algorithm assumes that the dataset is sufficiently large to accurately estimate the causal structure. This ensures that the statistical properties of the observed data are reliable.

6. Regularization: NOTEARS presumes that the true causal graph is sparse, meaning most variables have limited direct causes. This sparsity is enforced through ℓ_1-regularization, which penalizes the magnitude of W (see section 3.5). This assumption aligns with the observation that causal systems in many real-world applications are typically sparse.

No Assumption of Faithfulness: Unlike many constraint-based methods (e.g., the PC algorithm), NOTEARS, as a score-based method, does not rely on the faithfulness assumption, which states that all conditional independencies are directly caused by the structure of the causal graph, with no coincidences or cancellations. By using linear SEMs and the least-squares loss, NOTEARS has been shown in Zheng et al. (2018) to recover the true DAG with high probability, even for finite samples and high-dimensional data, without needing faithfulness. This is a significant advantage, as faithfulness is often violated in practice due to unexpected independencies or cancellations.

3.2 General optimization Problem

Let $X \in \mathbb{R}^{n\times d}$ represent a data matrix containing n i.i.d. observations of the random vector $X = (X_1, \ldots, X_d)$, where d is the number of variables under analysis. Each observation corresponds to a vector with values assigned to the d variables. Additionally, let D denote the discrete space of all DAGs $G = (V, E)$ with d nodes. The objective is to use the observation matrix X to identify the optimal DAG $G \in D$ that best represents the joint distribution $P(X)$ in the form of a Bayesian network.

We model X using a Structural Equation Model (SEM) defined by a weighted adjacency matrix $W \in \mathbb{R}^{d\times d}$. Each nonzero entry W_{kj} represents a potential causal edge from variable X_k to variable X_j. A general score-based learning approach can be written as:

$$\min_{W\in\mathbb{R}^{d\times d}} F(W) \quad \text{subject to} \quad G(W) \in \text{DAGs}, \tag{1}$$

where $F(W)$ is a score (or loss) function reflecting the fit of the graph $G(W)$ to the observed data, and the constraint $G(W) \in$ DAGs enforces acyclicity. Since directly searching over the discrete space of DAGs is combinatorial and grows super-exponentially in d, NOTEARS reformulates (1) into a continuous problem:

$$\min_{W \in \mathbb{R}^{d \times d}} F(W) \quad \text{subject to} \quad h(W) = 0, \tag{2}$$

where $h(W) : \mathbb{R}^{d \times d} \to \mathbb{R}$ is a differentiable function that equals zero if and only if W corresponds to a DAG. So, instead of operating on the discrete space D, we reformulate the problem to operate on $\mathbb{R}^{d \times d}$, the continuous space of $d \times d$ real matrices, enabling efficient optimization. This new definition of the acyclicity constraint as a smooth function allows NOTEARS for continuous optimization over W using standard techniques, making the implementation efficient. This approach also benefits non-expert users, as it eliminates the need for prior expertise in graphical model theory. However, it is important to note that (2) remains a nonconvex optimization problem, which implies that the algorithm may converge to local minima instead of the global one during the optimization process.

3.3 Structural Equation Model (SEM) and Loss Function

A SEM provides a general framework for describing how each variable X_j is generated from its parents. Typically, we assume:

$$X_j = f_j\big(\mathrm{Pa}(X_j); W\big) + \epsilon_j, \tag{3}$$

where $\mathrm{Pa}(X_j)$ denotes the set of parents of X_j, $f_j(\cdot)$ is a function parameterized by W, and ϵ_j is a noise term independent of $\mathrm{Pa}(X_j)$. In the linear NOTEARS setting proposed by Zheng et al. (2018), each X_j is modeled as:

$$X_j = \sum_{k=1}^{d} W_{kj} X_k + \epsilon_j, \tag{4}$$

where $W_{kj} = 0$ if X_k is not a parent of X_j, and $W_{kj} \neq 0$ indicates the edge $X_k \to X_j$. The independent noise ϵ_j can be Gaussian or non-Gaussian. To optimize this SEM model and recover the underlying DAG, a loss function is required to score the fit of the adjacency matrix W. The authors propose using the least-squares (LS) loss function as the score $F(W)$, which is defined as

$$F(W) = \frac{1}{2n} \|X - XW\|_F^2,$$

where $\| \cdot \|_F$ represents the Frobenius norm. This loss measures the total squared error between the observed data and the data reconstructed using the SEM with the adjacency matrix W. The LS loss is chosen for its consistency in both Gaussian and non-Gaussian SEMs under finite samples and high-dimensional settings ($d \gg n$), according to the authors. A few articles like Van de Geer and Bühlmann (2013) and Loh and Bühlmann (2014) are cited as a proof. However, this consistency is limited on the availability of prior knowledge or constraints on the noise variances. In cases where noise variances remain unknown, the recovered DAG under the Gaussian setting may only be identifiable up to its Markov equivalence class, limiting the ability to infer causal directions. Furthermore, while theoretical guarantees exist for finite-sample settings, smaller datasets may still lead to suboptimal estimates.

Additionally, the algorithm incorporates an ℓ_1-regularization term to enforce sparsity in the adjacency matrix W. The regularized loss function is defined as:

$$F_W = \frac{1}{2n}\|X - XW\|_F^2 + \lambda\|W\|_1, \tag{5}$$

where $\|W\|_1 = \sum_{j=1}^{d}\sum_{k=1}^{d}|W_{kj}|$ is the ℓ_1-norm of W, and $\lambda > 0$ is a regularization parameter controlling the trade-off between data fit and sparsity. By minimizing F_W, the algorithm aims to find a sparse DAG that balances accuracy and simplicity. The term $\frac{1}{n}$ normalizes the loss by the number of observations, ensuring that it reflects the average error per data point, while the factor $\frac{1}{2}$ simplifies gradient computations during optimization.

3.4 Characterization of Acyclicity

The most important breakthrough of NOTEARS is the derivation of a smooth function $h(W)$ that is zero if and only if W corresponds to a DAG. The following characterization is described in Theorem 1 of Zheng et al. (2018):

$$h(W) = \text{trace}\big(e^{W \odot W}\big) - d, \tag{6}$$

where $\odot$ is the elementwise (Hadamard) product and $e^{(\cdot)}$ is the matrix exponential. This definition allows to quantify the "DAG-ness" of the graph, as larger values represent a stronger violation of the DAG constraint. The authors of NOTEARS show that $h(W) = 0$ if and only if W encodes a DAG (i.e., no directed cycles). This can be proven as follows (the hadamard product is left out for simplicity):

The matrix exponential is given by the infinite series

$$e^W = \sum_{k=0}^{\infty}\frac{W^k}{k!}, \quad \text{and hence} \quad \text{tr}\big(e^W\big) = \sum_{k=0}^{\infty}\frac{1}{k!}\,\text{tr}\big(W^k\big).$$

(1) Acyclic $\Rightarrow$ $h(W) = 0$: If W represents a DAG, it has no directed cycles. This implies $\text{tr}(W^k) = 0$ for all $k \geq 1$, because there are no closed walks of length k. The only term contributing to $\text{tr}(e^W)$ is from $k = 0$, for which $W^0 = I$ with $\text{tr}(I) = d$. Therefore,

$$\text{tr}\big(e^W\big) = d, \quad \text{so} \quad h(W) = \text{tr}\big(e^W\big) - d = 0.$$

(2) Cyclic $\Rightarrow$ $h(W) > 0$: If there is at least one directed cycle in W, then some power W^k with $k \geq 1$ must have a nonzero diagonal entry representing a length-k closed walk. Consequently, $\text{tr}(W^k) > 0$ for one or more k. Summing these positive contributions causes

$$\text{tr}(e^W) > d \quad \Longrightarrow \quad h(W) = \text{tr}\big(e^W\big) - d > 0.$$

As demonstrated, W is acyclic if and only if $h(W) = 0$, and cyclic if and only if $h(W) > 0$. To illustrate this, Appendix A provides two examples of 2×2 matrices: one representing an acyclic graph and the other representing a cyclic graph.

3.5 Optimization

Chapter 3.4 introduces the smooth acyclicity constraint proposed by Zheng et al. (2018). This novel definition transforms the problem from a discrete constrained optimization framework into a continuous one, as stated in (2). Despite this transformation, the problem remains nonconvex due to the constraint $h(W) = 0$. Therefore, the optimization process only guarantees convergence to a local solution. To solve the constrained optimization problem, the authors propose using the augmented Lagrangian method. This approach reformulates the equality-constrained program (ECP):

$$\min_{W \in \mathbb{R}^{d \times d}} F(W) \quad \text{subject to} \quad h(W) = 0,$$

by introducing a quadratic penalty term, resulting in:

$$\min_{W \in \mathbb{R}^{d \times d}} F(W) + \frac{\rho}{2} \big| h(W) \big|^2 \quad \text{subject to} \quad h(W) = 0,$$

where $\rho > 0$ is the penalty parameter. The augmented Lagrangian is then defined as:

$$\mathcal{L}(W, \alpha, \rho) = F(W) + \frac{\rho}{2} \big| h(W) \big|^2 + \alpha\, h(W), \tag{7}$$

where:

- α is the Lagrange multiplier enforcing the equality constraint.
- ρ penalizes violations of $h(W) \approx 0$.

The augmented Lagrangian method improves upon the standard Lagrangian approach by adding a quadratic penalty term. This additional term enhances numerical stability during optimization, especially when dealing with equality constraints, as it strongly penalizes constraint violations and ensures better convergence (Nemirovsky (1999)). To update α, the dual function $D(\alpha)$ is used:

$$D(\alpha) = \min_{W \in \mathbb{R}^{d \times d}} \mathcal{L}(W, \alpha, \rho).$$

The goal is to maximize $D(\alpha)$, i.e., solve:

$$\max_{\alpha \in \mathbb{R}} D(\alpha).$$

The dual ascent method updates α iteratively based on the gradient:

$$\alpha_{t+1} = \alpha_t + \rho\, h(W_{t+1}),$$

where W_{t+1} is the solution to the primal subproblem at the t-th iteration.
The optimization process involves the following iterative steps:

1. **Primal update**: Minimize $\mathcal{L}(W, \alpha, \rho)$ with respect to W:

   $$W_{t+1} \leftarrow \arg\min_{W} \mathcal{L}^{\rho}(W, \alpha_t),$$

   using gradient-based methods such as L-BFGS or Adam.

2. **Dual ascent**: Update the Lagrange multiplier α to enforce the equality constraint:

$$\alpha_{t+1} \leftarrow \alpha_t + \rho \cdot h(W_{t+1}).$$

These steps are repeated until the algorithm converges to a solution W^*, which minimizes $F(W)$ while satisfying the acyclicity constraint $h(W^*) = 0$. The complete original algorithm is the following:

Algorithm 1 NOTEARS Algorithm Zheng et al. (2018)

```
Require: Initial guess $(W_0, \alpha_0)$, progress rate $c \in (0,1)$, tolerance $\epsilon > 0$, threshold $\omega > 0$.
for $t = 0, 1, 2, \ldots$ do
    (a) Solve primal $W_{t+1} \leftarrow \arg\min_W \mathcal{L}^\rho(W, \alpha_t)$ with $\rho$ such that $h(W_{t+1}) < ch(W_t)$.
    (b) Dual ascent $\alpha_{t+1} \leftarrow \alpha_t + \rho h(W_{t+1})$.
    (c) If $h(W_{t+1}) < \epsilon$, set $\widetilde{W}_{\mathrm{ECP}} = W_{t+1}$ and break.
end for
Return the thresholded matrix:
    $$\widehat{W} := \widetilde{W}_{\mathrm{ECP}} \circ 1\big(|\widetilde{W}_{\mathrm{ECP}}| > \omega\big).$$
```

To solve the unconstrained subproblem $\min_W \mathcal{L}(W, \alpha, \rho)$, gradient-based methods such as L-BFGS or Adam are employed. These methods are particularly effective in minimizing the augmented Lagrangian and guiding W toward a solution that both minimizes the objective function $F(W)$ and satisfies the acyclicity constraint. When sparsity regularization is applied, the proximal quasi-Newton (PQN) method can be used to handle the non-smooth ℓ_1-penalty term (see Zheng et al. (2018) for more detail). The PQN method is computationally efficient due to the use of low-rank approximations of the Hessian, allowing the algorithm to scale well even for high-dimensional problems.

3.6 Thresholding

After the continuous optimization converges, we obtain a real-valued matrix W^* which encodes the causal direction and strength of dependencies. In order to identify the final edges, we convert W^* into a binary adjacency matrix:

$$\widehat{A}_{kj} = \begin{cases} 1, & \text{if } |W^*_{kj}| \; > \; \delta, \\ 0, & \text{otherwise.} \end{cases}$$

The threshold hyperparameter δ must be chosen in advance. Thresholding is crucial for interpretability as large weights indicate stronger dependencies, while small weights often represent noise. The choice of δ can affect the precision of recovered edges, so it should be selected carefully. This is empirically tested in section 4.

3.7 Nonlinear Extension

Beyond the linear model, the authors of NOTEARS propose a generalized function model in Zheng et al. (2020), which extends the framework to capture nonlinear causal relationships. The general SEM equation stated in 3 highlights that any flexible function can be used to learn the causal relationships between variables. In this context, multilayer perceptrons (MLPs) are employed as f_j to model complex and nonlinear dependencies.

However, unlike the linear case, the adjacency matrix W cannot be directly interpreted from the MLP weights due to the layered structure and nonlinear transformations within the network. To address this, the algorithm computes the adjacency matrix $W(f)$ using partial derivatives of the MLP outputs with respect to their inputs. This technique quantifies the influence of one variable on another and ensures that the learned adjacency matrix $W(f)$ reflects meaningful causal relationships, even in the presence of nonlinear dependencies. This extension preserves the core principles of NOTEARS while extending its applicability to real-world datasets with complex structures. The method is empirically proven in 4.2.

4 Experiments

In this section, two sets of experiments are presented to (i) assess specific limitations of the NOTEARS algorithm (particularly regarding the *varsortability* critique by Reisach et al. (2021)) and (ii) compare its linear and nonlinear variants under different assumptions. Another focus lies on how hyperparameters affect the performance of NOTEARS.

4.1 Experiment 1: Varsortability

Motivation. Reisach et al. (2021) highlight that causal structures in simulated data can often be trivially discovered due to *varsortability*, a property that measures how well the causal order in a DAG aligns with the order of increasing marginal variances of the variables. In causal systems, the variance of a variable X_j is influenced by the variances of its parent variables X_i, the weights of the causal edges $w_{i \to j}$, and the noise term z_j. This can lead to an accumulation effect, where variables further downstream in the causal order inherit variance contributions from all their ancestors, compounded by their own noise. In synthetic data, this variance accumulation frequently results in a high varsortability situation which allows methods like NOTEARS to appear to perform well by exploiting the variance structure, rather than correctly identifying causal relationships. The authors of Reisach et al. (2021) mathematically prove in their Appendix A that this behavior can cause standard mean-squared-error-based directional inference methods to fail when the noise variance of the cause is not sufficiently smaller than that of the effect. The goal of Experiment 1 is to empirically verify this limitation by testing whether causal direction inference breaks down under specific noise variance conditions. Additionally, we aim to explore how this accumulating variance effect influences the performance of NOTEARS on synthetic datasets.

Setup. As in Reisach et al. (2021), consider a simple 2-variable model $A \to B$, generated by:

$$A = N_A, \quad B = w\,A + N_B,$$

where N_A and N_B are independent noise terms with zero mean and variances V_A and V_B, respectively. The inference method compares two linear regression models:

$$(1) \quad B \approx w_A\, A \quad \text{vs.} \quad (2) \quad A \approx w_B\, B,$$

and chooses the direction with the smaller overall MSE. The theoretical result states that, for consistent direction inference, one needs:

$$(1 - w^2)\, V_A < V_B.$$

If this inequality is not satisfied, the method incorrectly reverses the causal direction. In the experiments, we vary:

- Causal Weight (w): Values in $\{0.3, 0.5, 0.7, 0.9\}$.
- Noise Ratio (V_A/V_B): Swept across a grid of values from 0.1 to 10.0.
- Noise Distributions: Gaussian, exponential, and Gumbel.
- Sample Size and Repetitions: A sample size of 2000, with 1000 repetitions to reduce random effects.
- Standardization: Whether standardizing (A, B) after simulation affects the inference accuracy.

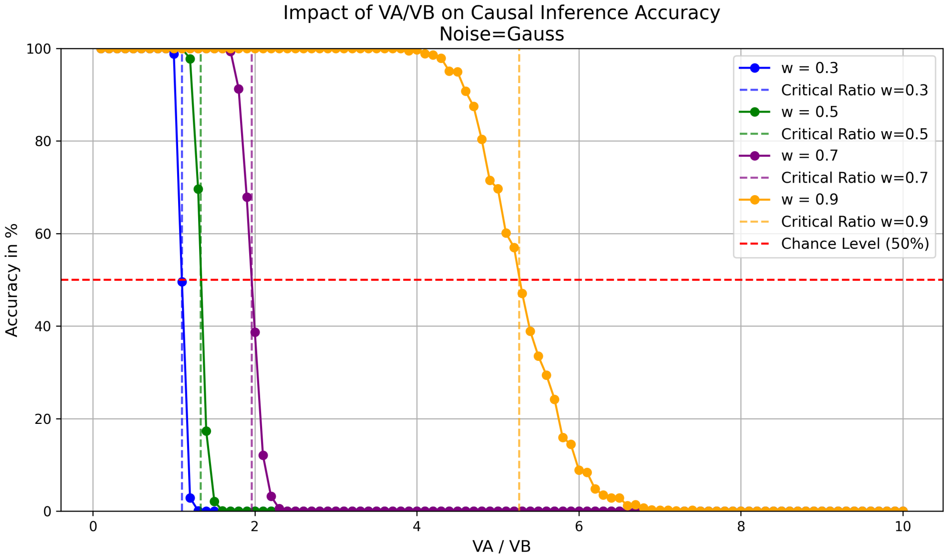

Figure 1: Impact of the ratio V_A/V_B on causal inference accuracy under Gaussian noise for different causal weights w. The solid lines represent the accuracy trends, while the dashed vertical lines mark the corresponding critical variance ratio. The red dashed line indicates the chance level (50%).

4.1.1 Experiment 1.1: Impact of V_A/V_B Ratio on Accuracy

In this experiment, we examine the impact of the variance ratio V_A/V_B on causal inference accuracy under Gaussian noise (Figure 1). The causal weight w is varied in $\{0.3, 0.5, 0.7, 0.9\}$, and the inference method evaluates multiple times whether $A \rightarrow B$ or $B \rightarrow A$ is the correct direction.

Observations. For lower noise variance ratios ($V_A/V_B \ll 1$), the accuracy is very high for all the causal weights. However, accuracy drops sharply once V_A/V_B exceeds a certain threshold. Stronger causal weights (e.g., $w = 0.9$) shift this threshold to higher variance ratios. For weaker causal relationships (e.g., $w = 0.3$), accuracy quickly deteriorates when the ratio nears its theoretical boundary.

Interpretation. These observations confirm the theoretical statement by Reisach et al. (2021): if $(1-w^2)\,V_A \geq V_B$, inference often fails. Larger w values allow a larger threshold, meaning the method remains accurate for a wider range of V_A/V_B.

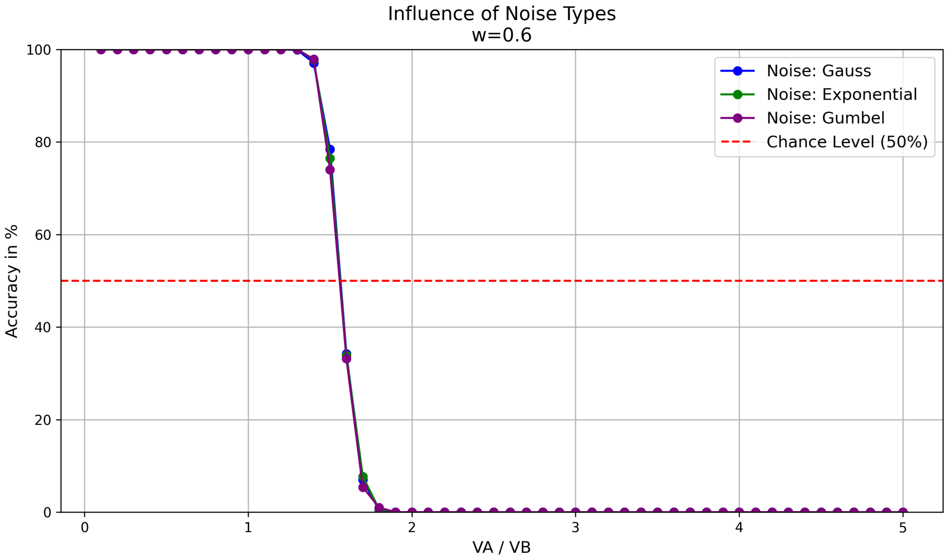

Figure 2: Influence of noise distributions on causal inference accuracy for a fixed causal weight $w = 0.6$. The accuracy trends for Gaussian, exponential, and Gumbel noise are shown. The red dashed line represents the chance level (50%).

4.1.2 Experiment 1.2: Consistency Across Noise Distributions

To check robustness, we replace Gaussian noise with exponential or Gumbel noise. The causal weight is fixed at $w = 0.6$, and V_A/V_B is varied as before.

Observations. Figure 2 shows that all tested distributions yield similar accuracy curves. The trend remains consistent as inference is strong when $V_A \ll V_B$ and deteriorates otherwise.

Interpretation. Even under non-Gaussian noise, the key ratio $(1 - w^2)\, V_A < V_B$ continues to work. While the proof in Reisach et al. (2021) is derived under Gaussian assumptions, this experiment suggests that the phenomenon is not restricted to that setup.

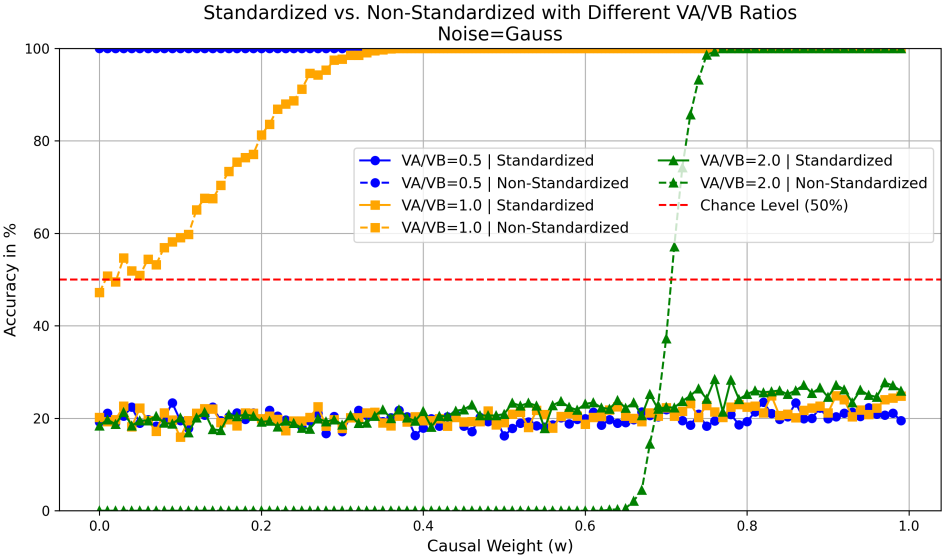

Figure 3: Comparison of standardized and non-standardized data on causal inference accuracy for different V_A/V_B ratios and causal weights w. The red dashed line indicates the chance level (50%).

4.1.3 Experiment 1.3: Impact of Standardization on the Causal Weight

In a final test for Experiment 1, we examine how standardizing the data before inference affects the influence of the causal weight on the accuracy. In general, as seen in Figure 1, a stronger causal weight leads to better performance even under bad noise variance ratios. We compare inference on (A, B) in their original scales versus standardizing each variable to zero mean and unit variance.

Observations. Figure 3 shows that in the non-standardized case, lower variance ratios produce higher accuracy, and increasing w further extends the range of good performance, which is expected as seen in previous experiments. However, once the data are standardized, the overall accuracy remains relatively constant and does not depend on either V_A/V_B or w.

Interpretation. By standardizing, we remove scale information that is crucial for the MSE-based direction test. The inference procedure effectively loses the variance structure that helps identify the true direction. Consequently, standardization reduces the impact of the causal weight on the inference outcome, potentially reducing the accuracy of the NOTEARS linear method.

4.1.4 Experiment Outcome

Overall, Figures 1, 2, and 3 confirm the statement by Reisach et al. Reisach et al. (2021). If the simulation setup aligns variable order with their variances (high varsortability), certain causal discovery methods can appear to perform extremely well. However, performance can drop with variance ratios that violate the theoretical threshold, revealing the limitation of purely MSE-based directional inference. It is important to emphasize that in real-world data, the noise variances V_A and V_B are not directly observed. Unlike in synthetic experiments, where we can explicitly control these variances, in practical settings they remain unknown and must be inferred indirectly. Thus, the reliance of MSE-based methods on variance ratios makes them particularly vulnerable to issues arising from unobserved noise variance in real-world applications.

4.2 Experiment 2: Performance Comparison and Hyperparameter Assessment

Motivation. We next investigate two additional limitations of NOTEARS. First, we compare the *linear* and *nonlinear* variants under both standardized and non-standardized data, examining the impact of different data-generating mechanisms (strictly linear vs. nonlinear). Second, we assess how two key hyperparameters—the regularization weight λ and the binarization threshold—affect structural recovery.

Setup. We simulate data with $d = 10$ variables from an Erdős-Rényi random graph (ER) with $s_0 = 20$ edges and $n = 1000$ samples. Independent Gaussian noise with scale 1.0 is added. Each experiment is evaluated using standard structural metrics: False Discovery Rate (FDR), False Positive Rate (FPR), True Positive Rate (TPR), and Structural Hamming Distance (SHD).

4.2.1 Experiment 2.1: Linear vs. Nonlinear under Standardization

We first compare the standard linear NOTEARS with its nonlinear variant by Zheng et al. (2020) on non-standardized vs. standardized data generated from a linear SEM. This experiment is motivated by the statement of the authors in Kaiser and Sipos (2021), where they argue the lack of scale-invariance of the NOTEARS model.

Observations. Figure 4 shows that the linear method recovers the causal structure almost perfectly on non-standardized data, reaching high TPR and low FDR. The nonlinear variant performs slightly worse in this purely linear setting but still yields a good TPR. However, once the data are standardized, the performance of both models drops significantly, with a marked increase in FDR and lower TPR.

Interpretation. The results confirm that standardization can remove crucial scale information, which aligns with the *varsortability* concerns stated in Reisach et al. (2021). While the nonlinear extension can capture more general relationships, it does not alleviate the standardization issue of the linear NOTEARS.

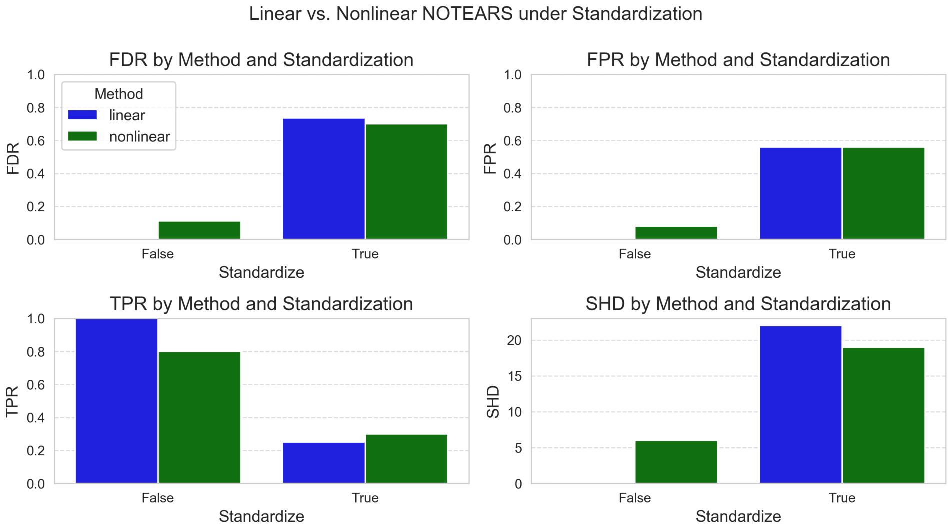

Figure 4: Comparison of False Discovery Rate (FDR), False Positive Rate (FPR), True Positive Rate (TPR), and Structural Hamming Distance (SHD) for linear and nonlinear NOTEARS under standardized and non-standardized conditions.

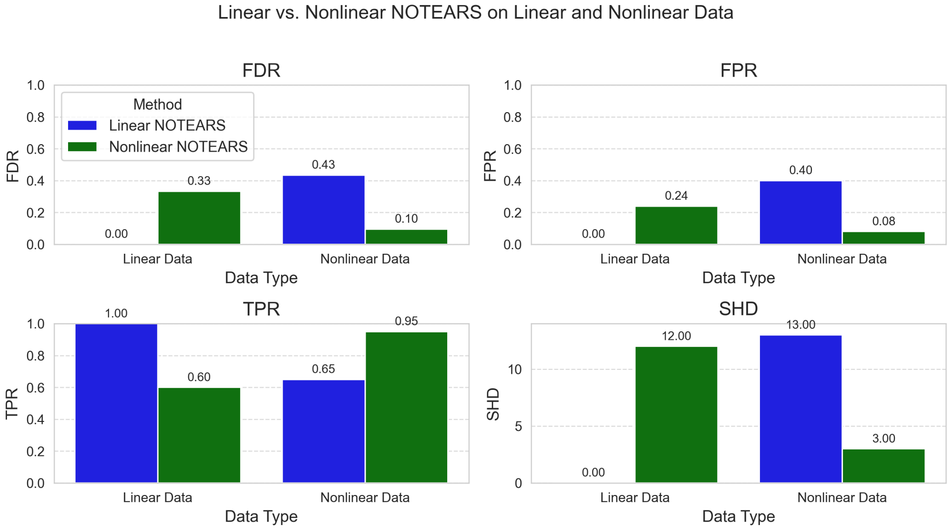

Figure 5: Performance of linear and nonlinear NOTEARS on datasets with linear and nonlinear relationships. Metrics include False Discovery Rate (FDR), False Positive Rate (FPR), True Positive Rate (TPR), and Structural Hamming Distance (SHD).

4.2.2 Experiment 2.2: Linear vs. Nonlinear on Linear and Nonlinear Data

Next we simulate data from strictly linear and explicitly nonlinear SEMs. We apply both the linear and nonlinear NOTEARS versions to each dataset.

Observations. Figure 5 illustrates that in the linear data scenario, the linear NOTEARS model achieves superior accuracy (lower FDR, higher TPR) compared to the nonlinear version. On the contrary, when the data-generating process is nonlinear, the nonlinear NOTEARS outperforms the linear algorithm, producing fewer false discoveries and achieving a lower SHD.

Interpretation. When the underlying mechanisms are linear, the nonlinear extension may be slightly less effective, possibly due to overfitting or unnecessary complexity. For nonlinear processes, however, the linear model systematically misses edges or introduces false positives, while the nonlinear approach captures the underlying relationships more accurately.

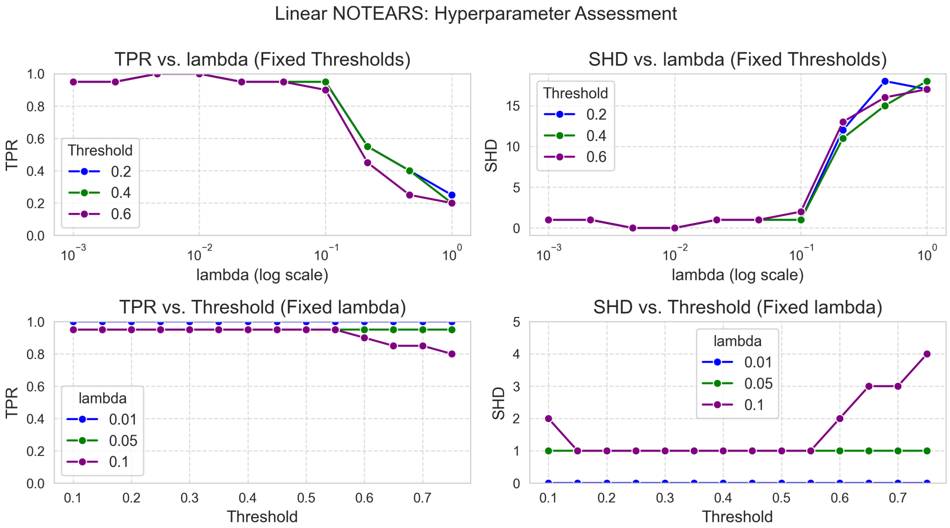

Figure 6: Evaluation of hyperparameter effects on the performance of linear NOTEARS. The top-left and top-right plots show the impact of the regularization parameter λ on True Positive Rate (TPR) and Structural Hamming Distance (SHD), respectively, for fixed thresholds.

4.2.3 Experiment 2.3: Hyperparameter Assessment

We finally examine how the regularization parameter λ (controlling sparsity) and the threshold for binarizing the learned weighted adjacency matrix affect performance in the linear NOTEARS.

Observations. As shown in Figure 6, increasing λ significantly reduces TPR while increasing the SHD. Excessively large regularization forces the model to drop too many edges. By contrast, varying the binarization threshold has a modest effect on TPR and SHD in these experiments.

Interpretation. A careful balance must be struck for λ: too little regularization can result in false edges, whereas too much can eliminate true causal relationships. In contrast, in our tests, threshold selection appears less critical, although in other settings or data scales it might need closer tuning.

4.2.4 Experiment Outcome

Across all sub-experiments, we find that (i) standardization degrades performance for both the linear and nonlinear NOTEARS variants, reflecting the varsortability challenge; (ii) linear NOTEARS performs best on linear data, while the nonlinear extension is a better option when the true generative mechanisms are nonlinear; and (iii) the choice of regularization weight λ is crucial for balancing sparsity and accuracy, whereas thresholding shows limited sensitivity for these synthetic datasets. These results emphasize the importance of aligning model assumptions with the data-generating process and carefully tuning hyperparameters to achieve reliable causal structure recovery.

5 Conclusions

This report presented the NOTEARS algorithm for causal discovery, focusing on both its theoretical foundation and practical considerations. We discussed how NOTEARS uses a smooth acyclicity constraint to transform a discrete DAG search problem into a continuous optimization task, making it scalable for high-dimensional data. Our experiments reveal several important takeaways:

- **Varsortability Issue**: In line with prior findings, our two-node experiments confirmed that NOTEARS and related MSE-based methods can use variance-driven patterns that artificially boost performance on synthetic data. When the noise variance ratio violates certain thresholds, inference accuracy can drop significantly. This issue becomes even more problematic with standardization, highlighting the importance of careful data preprocessing.

- **Linear vs. Nonlinear Extensions**: The nonlinear variant of NOTEARS is better suited to capturing complex, nonlinear data-generating processes. However, in cases where the true relationships are purely linear, the standard linear version performs better. This emphasizes the importance of understanding the nature of the data to select the appropriate model variant.

- **Hyperparameter Sensitivity**: Our results show that the regularization weight λ plays a crucial role in controlling model sparsity and recovering edges. While thresholding had a relatively minor effect on performance in these tests, its impact could vary depending on the dataset or noise levels.

- **Practical Implications**: Despite its strengths, NOTEARS relies on MSE-based objectives and makes certain strong assumptions, such as the absence of latent confounders and knowledge of noise variances. In real-world scenarios, where noise variances are often unknown or unobservable, this can lead to ambiguities or inaccuracies in the inferred structures. Researchers should carefully consider these limitations when interpreting the results.

Overall, NOTEARS demonstrates significant advantages, particularly in terms of computational efficiency and adaptability. However, our analysis indicates that its performance in real-world scenarios may not match its success with synthetic data. Continued research into more robust algorithms, including approaches that address variance-related challenges, is essential to improve causal discovery in practical applications.

A Appendix: Example of the Acyclicity Constraint

In this annex, we provide a concrete example demonstrating how the acyclicity constraint

$$h(W) \;=\; \mathrm{tr}\big(e^{W \circ W}\big) - d$$

works for a small 2×2 weighted adjacency matrix W. The difference between an acyclic matrix and a cyclic one is shown.

A.1 Acyclic Example

Matrix Definition. Consider the following matrix $W^{(\mathrm{A})}$ with two variables ($d = 2$):

$$W^{(\mathrm{A})} \;=\; \begin{bmatrix} 0 & 0.5 \\ 0 & 0 \end{bmatrix}.$$

Here, $X_1 \to X_2$ is the only directed edge, so we expect $W^{(\mathrm{A})}$ to be acyclic.

Step 1: Elementwise Square $(W \circ W)$.

$$W^{(\mathrm{A})} \circ W^{(\mathrm{A})} \;=\; \begin{bmatrix} 0 & 0.25 \\ 0 & 0 \end{bmatrix}.$$

Step 2: Matrix Exponential (Taylor Series Approximation). Let $M^{(\mathrm{A})} = W^{(\mathrm{A})} \circ W^{(\mathrm{A})}$. We approximate:

$$e^{M^{(\mathrm{A})}} \;\approx\; I + M^{(\mathrm{A})} + \frac{(M^{(\mathrm{A})})^2}{2!}.$$

- $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.
- $(M^{(\mathrm{A})})^2 = \begin{bmatrix} 0 & 0.25 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0.25 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, since the second row is all zeros.

If we were to continue the Taylor-Series expansion, each following term would result in matrixes with zero entries, so nothing would be added to the diagonal. Thus, we can leave it like that and get:

$$e^{M^{(\mathrm{A})}} \approx \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0.25 \\ 0 & 0 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0.25 \\ 0 & 1 \end{bmatrix}.$$

Step 3: Compute the Trace and $h(W^{(\mathrm{A})})$.

$$\mathrm{tr}\big(e^{M^{(\mathrm{A})}}\big) \;=\; 1 + 1 \;=\; 2, \quad h\big(W^{(\mathrm{A})}\big) \;=\; 2 \;-\; d \;=\; 2 - 2 \;=\; 0.$$

Since $h\big(W^{(\mathrm{A})}\big) = 0$, it confirms that $W^{(\mathrm{A})}$ is acyclic, as expected.

A.2 Cyclic Example

Matrix Definition. Now, consider a matrix $W^{(\mathrm{C})}$ that introduces a cycle:

$$W^{(\mathrm{C})} \;=\; \begin{bmatrix} 0 & 0.5 \\ 0.3 & 0 \end{bmatrix}.$$

Here, $X_1 \to X_2$ and $X_2 \to X_1$, forming a cycle between the two variables.

Step 1: Elementwise Square.

$$W^{(\mathrm{C})} \circ W^{(\mathrm{C})} \;=\; \begin{bmatrix} 0 & 0.25 \\ 0.09 & 0 \end{bmatrix}.$$

Denote this as $M^{(\mathrm{C})}$.

Step 2: Matrix Exponential. Again, approximate:

$$e^{M^{(\mathrm{C})}} \;\approx\; I + M^{(\mathrm{C})} + \frac{(M^{(\mathrm{C})})^2}{2!} + \frac{(M^{(\mathrm{C})})^3}{3!}.$$

(a) Compute $(M^{(\mathrm{C})})^2$.

$$(M^{(\mathrm{C})})^2 = \begin{bmatrix} 0 & 0.25 \\ 0.09 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0.25 \\ 0.09 & 0 \end{bmatrix} = \begin{bmatrix} 0.0225 & 0 \\ 0 & 0.0225 \end{bmatrix}.$$

(b) Compute $(M^{(\mathrm{C})})^3$.

$$(M^{(\mathrm{C})})^3 = (M^{(\mathrm{C})})^2 \cdot M^{(\mathrm{C})} = \begin{bmatrix} 0.0225 & 0 \\ 0 & 0.0225 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0.25 \\ 0.09 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0.005625 \\ 0.002025 & 0 \end{bmatrix}.$$

(c) Summation.

$$e^{M^{(\mathrm{C})}} \approx \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{I} + \underbrace{\begin{bmatrix} 0 & 0.25 \\ 0.09 & 0 \end{bmatrix}}_{M^{(\mathrm{C})}} + \frac{1}{2} \underbrace{\begin{bmatrix} 0.0225 & 0 \\ 0 & 0.0225 \end{bmatrix}}_{(M^{(\mathrm{C})})^2} + \frac{1}{6} \underbrace{\begin{bmatrix} 0 & 0.005625 \\ 0.002025 & 0 \end{bmatrix}}_{(M^{(\mathrm{C})})^3}.$$

Adding these up term by term yields a matrix whose diagonal elements will exceed 1.

Step 3: Trace and $h\big(W^{(\mathrm{C})}\big)$. The summation shows that the diagonal entries become slightly larger than 1 after including the $(M^{(\mathrm{C})})^2$ and $(M^{(\mathrm{C})})^3$ terms. Therefore,

$$\mathrm{tr}\big(e^{M^{(\mathrm{C})}}\big) \;>\; 2. \quad\Longrightarrow\quad h\big(W^{(\mathrm{C})}\big) \;=\; \mathrm{tr}\big(e^{M^{(\mathrm{C})}}\big) \;-\; d \;>\; 2 - 2 \;=\; 0.$$

Because $h\big(W^{(\mathrm{C})}\big) \neq 0$, the matrix $W^{(\mathrm{C})}$ encodes a cycle, confirming that this graph is *not* acyclic.

B Electronic appendix

The code for all experiments conducted in this thesis is available at the following GitHub repository: `https://github.com/ikumpli/notears_experiments`.

References

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

Chickering, D. M. (2002). Optimal structure identification with greedy search, *Journal of machine learning research* **3**(Nov): 507–554.

Glymour, C., Zhang, K. and Spirtes, P. (2019). Review of causal discovery methods based on graphical models, *Frontiers in genetics* **10**: 524.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J. and Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models, *Advances in neural information processing systems* **21**.

Huang, B., Zhang, K., Lin, Y., Schölkopf, B. and Glymour, C. (2018). Generalized score functions for causal discovery, *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1551–1560.

Kaiser, M. and Sipos, M. (2021). Unsuitability of notears for causal graph discovery, *arXiv preprint arXiv:2104.05441* .

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

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

Nemirovsky, A. (1999). Optimization ii. numerical methods for nonlinear continuous optimization.

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

Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P. and Aragam, B. (2020). Dynotears: Structure learning from time-series data, *International Conference on Artificial Intelligence and Statistics*, Pmlr, pp. 1595–1605.

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

Reichenbach, H. (1956). *The Direction of Time*, Dover Publications, Mineola, N.Y.

Reisach, A., Seiler, C. and Weichwald, S. (2021). Beware of the simulated dag! causal discovery benchmarks may be easy to game, *Advances in Neural Information Processing Systems* **34**: 27772–27784.

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

Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs, *Social science computer review* **9**(1): 62–72.

Van de Geer, S. and Bühlmann, P. (2013). 0-penalized maximum likelihood for sparse directed acyclic graphs.

Yu, Y., Chen, J., Gao, T. and Yu, M. (2019). Dag-gnn: Dag structure learning with graph neural networks, *International conference on machine learning*, PMLR, pp. 7154–7163.

Zheng, X., Aragam, B., Ravikumar, P. K. and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning, *Advances in neural information processing systems* **31**.

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

Zhu, S., Ng, I. and Chen, Z. (2019). Causal discovery with reinforcement learning, *arXiv preprint arXiv:1906.04477* .

Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Munich, 29.01.2025

Iker.

Iker Cumplido Esteban