Revisiting Differentiable Structure Learning: Inconsistency of ℓ_1 Penalty and Beyond

Kaifeng Jin[1*], **Ignavier Ng**[2], **Kun Zhang**[2,3], **Biwei Huang**[4]
[1]University of Illinois Urbana-Champaign
[2]Carnegie Mellon University
[3]Mohamed bin Zayed University of Artificial Intelligence
[4]University of California San Diego

Abstract

Recent advances in differentiable structure learning have framed the combinatorial problem of learning directed acyclic graphs as a continuous optimization problem. Various aspects, including data standardization, have been studied to identify factors that influence the empirical performance of these methods. In this work, we investigate critical limitations in differentiable structure learning methods, focusing on settings where the true structure can be identified up to Markov equivalence classes, particularly in the linear Gaussian case. While recent work highlighted potential non-convexity issues in this setting, we demonstrate and explain why the use of ℓ_1-penalized likelihood in such cases is fundamentally inconsistent, even if the global optimum of the optimization problem can be found. To resolve this limitation, we develop a hybrid differentiable structure learning method based on ℓ_0-penalized likelihood with hard acyclicity constraint, where the ℓ_0 penalty can be approximated by different techniques including Gumbel-Softmax. Specifically, we first estimate the underlying moral graph, and use it to restrict the search space of the optimization problem, which helps alleviate the non-convexity issue. Experimental results show that the proposed method enhances empirical performance both before and after data standardization, providing a more reliable path for future advancements in differentiable structure learning, especially for learning Markov equivalence classes.

Code — https://github.com/kaifeng-jin/CALM
Extended version — https://arxiv.org/abs/2410.18396

1 Introduction

Probabilistic graphical models, such as Bayesian networks, are powerful tools for capturing complex probabilistic relationships in a concise way (Pearl 1988; Koller and Friedman 2009). Their graph structures, usually encoded as Directed Acyclic Graphs (DAGs), allow efficient representation of data dependencies and have become essential in fields like health (Tennant et al. 2021) and economy (Awokuse and Bessler 2003). Traditionally, learning these structures involves discrete methodologies. Constraint-based methods, which leverage conditional independence tests, are one common approach (Spirtes and Glymour 1991; Spirtes, Glymour, and Scheines 2001). Another popular technique involves score-based methods, where the search space of potential graphs is explored based on scoring functions (Koivisto and Sood 2004; Singh and Moore 2005; Cussens 2011; Yuan and Malone 2013; Chickering 2002; Peters and Bühlmann 2014). Given the combinatorial nature of the task, greedy search strategies have been commonly used (Chickering 1996; Chickering, Heckerman, and Meek 2004).

In recent years, (Zheng et al. 2018) introduced a continuous formulation for characterizing the acyclicity constraint, effectively converting the discrete nature of the structure learning problem into one that can be approached using gradient-based optimization techniques. Although this formulation still involves nonconvex optimization, it opened the door to applying efficient gradient-based methods. This formulation has since inspired a wide range of extensions, being adapted to deal with nonlinearity (Yu et al. 2019; Lachapelle et al. 2019; Zheng et al. 2020; Ng et al. 2022b; Kalainathan et al. 2022), latent confounding (Bhattacharya et al. 2021; Bellot and van der Schaar 2021; Ng et al. 2024; Prashant et al. 2025; Sethuraman and Fekri 2025), interventional data (Brouillard et al. 2020; Faria, Martins, and Figueiredo 2022), time series data (Pamfil et al. 2020; Sun et al. 2021), and missing data (Wang et al. 2020; Gao et al. 2022). Other applications include multi-task learning (Chen et al. 2021), and working with federated learning systems (Ng and Zhang 2022; Gao et al. 2021), domain adaptation (Yang et al. 2021), and recommendation system (Wang et al. 2022).

This move toward continuous structure learning has prompted growing attention to both its theoretical underpinnings and practical performance. Researchers like (Wei, Gao, and Yu 2020) and (Ng et al. 2022a) have investigated the optimality and convergence properties of continuous, constrained optimization techniques (Zheng et al. 2018). Meanwhile, (Deng et al. 2023) provided insight into how an appropriately designed optimization scheme can reach the global minimum for least squares objectives in simple cases. Further refinements have also been proposed, with (Zhang et al. 2022) and (Bello, Aragam, and Ravikumar 2022) highlighting challenges such as gradient vanishing in existing DAG constraints (Zheng et al. 2018; Yu et al. 2019) and sug-

[*]This work was carried out primarily while Kaifeng Jin was affiliated with the University of California San Diego.

gesting potential improvements.

Recently, (Ng, Huang, and Zhang 2024) highlighted the non-convexity issues in differentiable structure learning methods, particularly in the linear Gaussian setting where the true structure can be identified up to Markov equivalence classes. While non-convexity poses major challenges in this context, we further identify another critical issue: ℓ_1-penalized likelihood is inconsistent, even if the global optimum of the optimization problem can be found. To address these limitations, we propose a method that resolves the ℓ_1 inconsistency and enhances empirical performance, both before and after data standardization, even under non-convex conditions. It is worth noting that similar issues have been investigated by (Deng et al. 2024); see Appendix I for a detailed discussion.

Contributions In this work, we tackle fundamental challenges in differentiable structure learning, particularly in the linear Gaussian case, by focusing on the limitations of penalized likelihood approaches. Our contributions include:

- We identify and demonstrate the inconsistency of using ℓ_1-penalized likelihood in differentiable structure learning methods, even if the global optimum of the optimization problem can be found, particularly when learning Markov equivalence classes in the linear Gaussian case.
- We develop a differentiable structure learning method that optimizes an ℓ_0-penalized likelihood with hard acyclicity constraints and incorporates a moral graph estimation procedure, where the ℓ_0 penalty is approximated by differentiable techniques, such as Gumbel-Softmax. We call our method CALM (**C**ontinuous and **A**cyclicity-constrained **L**0-penalized likelihood with estimated **M**oral graph). CALM not only addresses ℓ_1 inconsistency, but also results in a solution much closer to the true structure or its Markov equivalent graphs. Our method provides a more reliable path for future advancements in differentiable structure learning, especially for learning Markov equivalence classes.
- Our method performs consistently well both before and after data standardization, demonstrating its robustness.

2 Background

In this section, we introduce our problem setting and, while reviewing the hard and soft DAG constraints, we also revisit NOTEARS (Zheng et al. 2018) and GOLEM (Ng, Ghassami, and Zhang 2020).

2.1 Problem Setting

Setup In this work, we focus on linear Gaussian Structural Equation Models (SEMs), where the variables $X = (X_1, \ldots, X_d)$ follow linear relationships represented by a DAG. The model is expressed as $X = B^\top X + N$. Here, $B \in \mathbb{R}^{d\times d}$ is the weighted adjacency matrix encoding the relationships between variables. Specifically, an entry $B_{ij} \neq 0$ indicates a directed edge from X_j to X_i. The noise vector $N = (N_1, \ldots, N_d)$ consists of independent noise terms, each corresponding to a variable X_i. The noise terms are assumed to follow a normal distribution with diagonal covariance matrix $\Omega = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, where σ_i^2 represents the variance of N_i. Given a DAG, a moral graph is an undirected graph obtained by removing the directions of the edges in the DAG and connecting all pairs of parents of common children.

Unlike many existing methods that assume equal noise variances (Zheng et al. 2018; Yu et al. 2019; Zhang et al. 2022; Bello, Aragam, and Ravikumar 2022), in this study, we focus on the general non-equal noise variance (NV) case, where the variances $\sigma_1^2, \ldots, \sigma_d^2$ are not assumed to be equal. Our goal is to estimate the DAG G or its Markov equivalence class (MEC) from the data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, consisting of n i.i.d. samples drawn from the distribution $P(X)$.

2.2 Hard and Soft DAG Constraint

In the context of DAG learning, the DAG constraint, denoted as $h(B)$, ensures that the learned structure is a DAG when $h(B) = 0$. NOTEARS (Zheng et al. 2018) employs a hard DAG constraint, whereas GOLEM (Ng, Ghassami, and Zhang 2020) introduces a soft DAG constraint.

NOTEARS and hard DAG constraint NOTEARS solves the following constrained optimization problem

$$\begin{aligned} \min_{B\in\mathbb{R}^{d\times d}} \quad & \ell(B;\mathbf{X}) := \frac{1}{2n}\|\mathbf{X} - \mathbf{X}B\|_F^2 + \lambda\|B\|_1 \\ \text{subject to} \quad & h(B) = 0. \end{aligned}$$

Here, $\ell(B; X)$ is the least squares loss with ℓ_1 penalty, and $h(B) = 0$ enforces the hard DAG constraint. The constrained optimization problem can be solved using standard algorithms such as augmented Lagrangian method (Wright 2006), quadratic penalty method (Ng et al. 2022a), and barrier method (Bello, Aragam, and Ravikumar 2022).

GOLEM and soft DAG constraint The GOLEM framework aims to maximize the likelihood of the observed data under the assumption of a linear Gaussian model. There are two formulations in GOLEM, one assuming equal noise variance across variables (GOLEM-EV), and the other allowing for non-equal noise variance (GOLEM-NV).

Unlike NOTEARS, GOLEM adopts the soft DAG constraint, making the problem unconstrained. In other words, GOLEM incorporates $h(B)$ as an additional penalty term in the score function, controlled by the hyperparameter λ_2. Here, we only review the the non-equal noise variance formulation, GOLEM-NV, which is the focus of this paper. Assuming that $\mathbf{X}$ is centered and that the sample covariance matrix $\Sigma = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$, GOLEM-NV's unconstrained optimization problem is

$$\begin{aligned} \min_{B\in\mathbb{R}^{d\times d}} \quad & \mathcal{L}(B;\Sigma) + \lambda_1\|B\|_1 + \lambda_2 h(B), \\ \text{where} \quad \mathcal{L}(B;\Sigma) &= \frac{1}{2}\sum_{i=1}^{d}\log\left(\left((I-B)^\top\Sigma(I-B)\right)_{i,i}\right) \\ &\quad - \log\left|\det(I-B)\right|. \end{aligned}$$

Here, $\mathcal{L}(B;\Sigma)$ is the likelihood function of linear Gaussian directed graphical models. This allows us to use B and Σ to express GOLEM-NV's formulation.

Metric	B^*	B_{ℓ_1}	Proportion of $\tilde{B}$ with $\|\tilde{B}\|_1 < \|B^*\|_1$
Average ℓ_1 norm	10.04 ± 0.04	4.22 ± 0.03	77.86% ± 0.46%
Average ℓ_0 norm (Number of Edges)	8.0 ± 0.0	22.74 ± 0.15	N/A
Average SHD of CPDAG	0 ± 0.0	19.97 ± 0.17	N/A

Table 1: Comparison of $\tilde{B}$s which generate Σ^* with True DAG B^*. The results are averaged over 1,000 simulated B^*s, with standard errors (SE) reported alongside mean values. The "Proportion" column reflects the average percentage of DAGs $\tilde{B}$ with ℓ_1 norms smaller than that of B^* among $d!$ $\tilde{B}$s per B^*.

3 Inconsistency of ℓ_1 Penalty in Structure Learning

In this section, we explore the inconsistency of the ℓ_1 penalty in structure learning by comparing its behavior with ℓ_0 in linear Gaussian cases. We demonstrate this inconsistency through extensive experiments and conclude with a counterexample that highlights the issue.

3.1 ℓ_0 vs ℓ_1 Penalty in Linear Gaussian Cases

In structure learning for linear Gaussian cases, score-based methods, specifically with the BIC score (Schwarz 1978; Chickering 2002), often aim to recover the sparsest underlying DAG that best explains the observed data (Singh and Moore 2005; Cussens 2011; Yuan and Malone 2013; Chickering 2002). In the asymptotic case where the population covariance matrix, denoted by Σ^*, is available, this can, loosely speaking, be formulated as

$$\begin{aligned} \min_{B,\Omega} \quad & \|B\|_0 \\ \text{subject to} \quad & (I-B)^{-\top}\Omega(I-B)^{-1} = \Sigma^* \\ \text{and} \quad & B \text{ represents a DAG.} \end{aligned}$$

Recall that B denotes the weighted adjacency matrix representing the structure and Ω is the diagonal matrix of noise variances. That is, the goal is to minimize the ℓ_0 norm of B, i.e., the number of edges, while maximizing the likelihood by satisfying the covariance constraint and ensuring that B is a valid DAG. In other words, the objective is to recover the sparsest DAG, $\hat{B}$, along with its corresponding $\hat{\Omega}$, that can generate the observed covariance matrix Σ^* (i.e., $(I-\hat{B})^{-\top}\hat{\Omega}(I-\hat{B})^{-1} = \Sigma^*$). Under the sparsest Markov faithfulness assumption (Raskutti and Uhler 2018), the estimated $\hat{B}$ will be Markov equivalent to the true graph B^*.

Many previous work, including GOLEM, replace the ℓ_0 penalty with the more tractable ℓ_1 penalty. The corresponding optimization problem becomes

$$\begin{aligned} \min_{B,\Omega} \quad & \|B\|_1 \\ \text{subject to} \quad & (I-B)^{-\top}\Omega(I-B)^{-1} = \Sigma^* \\ \text{and} \quad & B \text{ represents a DAG.} \end{aligned} \tag{1}$$

The ℓ_1 penalty encourages smaller edge weights but introduces a key inconsistency: it does not guarantee true sparsity in the resulting structure. Minimizing the ℓ_1 norm may lead to a denser structure with more edges than the solution from minimizing the ℓ_0 norm. This occurs because ℓ_1 favors edges with small absolute values, even if they represent spurious edges. In some cases, the sum of the absolute values of more edges can be smaller than that of fewer, larger-magnitude edges, which leads ℓ_1-based methods to include unnecessary edges. As a result, the learned structure may deviate from the true DAG or its Markov equivalence class. Taking GOLEM as an example, even if the covariance constraint $(I-B)^{-\top}\Omega(I-B)^{-1} = \Sigma^*$ is satisfied in both ℓ_0- and ℓ_1-based formulations, the structural properties of the solution can differ significantly.

In Section 3.2, we demonstrate this with a large number of experiments, showing that a large proportion of DAGs $\tilde{B}$ satisfying the covariance constraint $(I-\tilde{B})^{-\top}\Omega(I-\tilde{B})^{-1} = \Sigma^*$ have smaller ℓ_1 norms than the true graph B^*. Moreover, in these DAGs satisfying the covariance constraint and having smaller ℓ_1 norms than the true graph B^*, we can always find ones with much larger ℓ_0 values than B^*, meaning they do not correspond to the true graph B^* or its Markov equivalence class, with the structural Hamming distance (SHD) computed over the true and estimated CPDAG (Completed Partially Directed Acyclic Graph), which we refer to as SHD of CPDAG in this paper, far from zero.

3.2 Experiment: Assessing the Inconsistency of ℓ_1 Penalty

In this section, we demonstrate and evaluate the inconsistency of the ℓ_1 penalty in likelihood-based GOLEM through experiments. We generated 1,000 true DAGs B^*, compute their corresponding covariance matrices Σ^* under infinite sample conditions. For each B^*, we use Cholesky decomposition to generate large number of DAGs $\tilde{B}$ that can generate the same Σ^*. Our goal is to identify the $\tilde{B}$ with the minimum ℓ_1 norm among these $\tilde{B}$s, denoted as B_{ℓ_1}, and compare it with B^* in terms of ℓ_1 norm, ℓ_0 norm (i.e., edge count), and record its SHD of CPDAG. Additionally, we also record the proportion of $\tilde{B}$s that satisfy the covariance constraint (i.e., generate the same Σ^*) but have a smaller ℓ_1 norm than B^*, to give an intuitive sense of the extent of ℓ_1 inconsistency. It is worth noting that, instead of using B^* itself as the reference, one could also use the DAG in the Markov equivalence class of B^* that has the smallest ℓ_1 norm.

True DAG generation and covariance matrix computation We generate 1000 8-node ($d = 8$) ER1 graphs B^*s . The data is generated with a fixed noise ratio of 16, where the variances of two randomly selected noise vari-

ables are set to 1 and 16, respectively. The variances of the remaining noise variables are sampled uniformly from the range $[1, 16]$. The edge weights are uniformly sampled from $[-2, -0.5] \cup [0.5, 2]$. For each B^*, we compute the corresponding population covariance matrix Σ^* under infinite samples using the equation $\Sigma^* = (I-B^*)^{-\top}\Omega^*(I-B^*)^{-1}$.

Generating DAGs which meet covariance constraint Following the idea of the sparsest permutation approach developed by (Raskutti and Uhler 2018), for each true covariance matrix Σ^*, we generate all possible $d!$ permutations of its rows and columns. For each permuted covariance matrix, we apply Cholesky decomposition to find a DAG $\tilde{B}$ that generates the permuted covariance matrix. After that, we restore $\tilde{B}$ to the original variable order. This ensures that all $\tilde{B}$ satisfies the covariance constraint $(I - \tilde{B})^{-\top}\tilde{\Omega}(I - \tilde{B})^{-1} = \Sigma^*$, while remaining a valid DAG. After the above steps, for each of the 1,000 true B^*, we identified $d!$ different DAGs $\tilde{B}$ that all generate Σ^*. (As implied by (Raskutti and Uhler 2018), this procedure of iterating over all $d!$ permutations exhaustively covers all possible DAGs satisfying the covariance constraint.)

Metrics and analysis For each true DAG B^*, we analyze the following metrics across all $d!$ DAGs $\tilde{B}$ that satisfy the covariance constraint: (1) ℓ_1 norm comparison: we calculate the ℓ_1 norm of each $\tilde{B}$ and record the proportion of $\tilde{B}$s whose ℓ_1 norm is smaller than that of B^*; (2) selecting $\tilde{B}$ with the minimum ℓ_1 norm: among the $d!$ $\tilde{B}$s, we select the one with the smallest ℓ_1 norm, denoted as B_{ℓ_1}. We then compare B_{ℓ_1} with B^* in terms of ℓ_1, and record its edge count and SHD of CPDAG (to test its distance to B^* or its Markov equivalence class).

Experimental results After running experiments for 1000 B^*s, we summarized the result in Table 1. Table 1 shows a comparison between the true DAG B^* and the B_{ℓ_1} that generate the same covariance matrix Σ^*. On average, for each true DAG B^*, 77.86% of the $d!$ DAGs $\tilde{B}$ satisfying the covariance constraint have smaller ℓ_1 norms than B^*. In the 1,000 runnings, the average ℓ_1 norm of B_{ℓ_1} is 4.22, which is smaller than the average ℓ_1 norm of B^*, which is 10.04. The average ℓ_0 norm (number of edges) of B_{ℓ_1} is 22.74, which is larger than the ℓ_0 norm (number of edges) of B^*, which is 8. The average SHD of CPDAG between B_{ℓ_1} and B^* is 19.97. In addition, in each running, B_{ℓ_1} consistently has a smaller ℓ_1 norm than B^*, a larger ℓ_0 norm (number of edges), and a SHD of CPDAG greater than zero, indicating that B_{ℓ_1} is structurally different from B^* and its Markov equivalence class. These results demonstrate two key points: (1) a significant proportion of DAGs $\tilde{B}$ that satisfy the covariance constraint have smaller ℓ_1 norms than B^*, and (2) in each running, we can find a counterexample (i.e., the B_{ℓ_1}) where the ℓ_1 norm of B_{ℓ_1} is smaller than that of B^*, while the ℓ_0 norm of B_{ℓ_1} is larger than that of B^*, and the resulting DAG is not equivalent to B^* or its Markov equivalence class. This supports the conclusion that ℓ_1-based solutions are inconsistent in recovering the true structure.

3.3 Case Study: A Specific Counterexample

We present a 3-node counterexample to demonstrate the inconsistency of ℓ_1 penalty in structure learning. Specifically, we compare a true weighted adjacency matrix B^* with an estimated one $\tilde{B}$, and show that although both matrices can generate the same covariance matrix, their ℓ_0 norm (edge count), ℓ_1 norm, and structural differences, measured by SHD of CPDAG, reveal the inconsistency of the ℓ_1 penalty.

The true adjacency matrix B^* and its corresponding noise covariance matrices Ω^* are given as:

$$B^* = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}, \quad \Omega^* = \begin{bmatrix} 16 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

As B^* represents a v-structure, there is only one element in its Markov equivalence class. The estimated adjacency matrix $\tilde{B}$ and its corresponding noise covariance matrix $\tilde{\Omega}$ are:

$$\tilde{B} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{10} \\ 0 & 0 & -\frac{1}{5} \\ 0 & 0 & 0 \end{bmatrix}, \quad \tilde{\Omega} = \begin{bmatrix} 16 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & \frac{4}{5} \end{bmatrix}.$$

Both matrices B^* and $\tilde{B}$, along with their respective noise covariance matrices, generate the same covariance matrix:

$$\Sigma^* = \begin{bmatrix} 16 & 8 & 0 \\ 8 & 9 & -1 \\ 0 & -1 & 1 \end{bmatrix}.$$

We have $\|B^*\|_0 = 2$ and and $\|\tilde{B}\|_0 = 3$, indicating that B^* is sparser than $\tilde{B}$. However, when considering the ℓ_1 norm, we observe that: $\|B^*\|_1 = \frac{3}{2}$ and $\|\tilde{B}\|_1 = \frac{4}{5}$. That is, although $\tilde{B}$ has a higher ℓ_0 norm, it achieves a lower ℓ_1 norm, highlighting the inconsistency between the two norms. Therefore, the optimization problem in Eq. (1) may return $\tilde{B}$, which is clearly not Markov equivalent to B^*.

This counterexample demonstrates the inconsistency of the ℓ_1 penalty: it may lead to solutions with smaller total edge weights (resulting in a lower ℓ_1 norm), but these solutions may still have more edges (a higher ℓ_0 norm) and deviate from the true DAG structure and its Markov equivalence class, even if these solutions can generate the same covariance matrix as the ground truth DAG.

4 Continuous and Acyclicity-Constrained ℓ_0-penalized Likelihood with Estimated Moral Graph

In Section 2.2, we reviewed the GOLEM-NV formulation proposed by (Ng, Ghassami, and Zhang 2020), which aims to maximize the data likelihood utilizing an ℓ_1 penalty and soft DAG constraint. We refer to this original model as GOLEM-NV-ℓ_1 throughout this paper. The problem formulation can be expressed as

$$\min_{B \in \mathbb{R}^{d \times d}} \quad \mathcal{L}(B; \Sigma) + \lambda_1 \|B\|_1 + \lambda_2 h(B).$$

However, as pointed out by (Ng, Huang, and Zhang 2024), GOLEM-NV-ℓ_1 suffers from significant non-convexity, often leading to suboptimal local minima with poor performance, both before and after data standardization. Moreover, as we demonstrated in Section 3, the ℓ_1 penalty leads to

	SHD of CPDAG	Precision of Skeleton	Recall of Skeleton
Soft Constraints Without Moral	33.8 ± 2.7	0.98 ± 0.01	0.43 ± 0.05
Soft Constraints With Moral	7.6 ± 2.5	0.98 ± 0.01	0.95 ± 0.03
Hard Constraints Without Moral	16.7 ± 3.4	0.88 ± 0.03	0.97 ± 0.01
Hard Constraints With Moral (CALM)	5.5 ± 1.9	0.98 ± 0.01	0.99 ± 0.00

Table 2: Impact of moral graph and soft/hard DAG constraints for 50-node ER1 graphs under data standardization

	SHD of CPDAG	Precision of Skeleton	Recall of Skeleton
CALM-Non-Standardized	9.9 ± 3.4	0.95 ± 0.02	0.99 ± 0.01
CALM-Standardized	5.5 ± 1.9	0.98 ± 0.01	0.99 ± 0.00

Table 3: Impact of Data Standardization on CALM for 50-node ER1 graphs

inconsistent solutions. To address these limitations, we propose CALM (**C**ontinuous and **A**cyclicity-constrained **L**0-penalized likelihood with estimated **M**oral graph), a differentiable structure learning method that optimizes an ℓ_0-penalized likelihood with hard DAG constraints and incorporates moral graphs. Our experiments demonstrate that CALM significantly improves performance compared to the original GOLEM-NV-ℓ_1, yielding results much closer to the true DAG or its Markov equivalence class. In the following subsections, we will provide a comprehensive introduction and analysis of CALM.

4.1 The CALM Approach

ℓ_0 penalty and its approximation with Gumbel Softmax CALM begins with applying an ℓ_0 penalty to regularize the likelihood, enforcing sparsity in the learned adjacency matrix. Inspired by (Ng et al. 2022b; Kalainathan et al. 2022), we use Gumbel Softmax (Jang, Gu, and Poole 2017) as a technique to achieve the approximation of ℓ_0 penalty in our approach, as it proved to be the most effective and robust ℓ_0 approximation among those we experimented with in Appendix B. When using Gumbel Softmax to approximate ℓ_0 penalty, CALM starts by representing the learned adjacency matrix B as an element-wise product of a learned mask, $g_\tau(U) \in \mathbb{R}^{d\times d}$, which determines the presence of edges, and a learned parameter matrix, $P \in \mathbb{R}^{d\times d}$, which learns the weights of the edges. The mask $g_\tau(U)$ is generated using the Gumbel-Softmax approach. Here, $U_{i,j}$ represents the logits, and a logistic noise $G_{i,j} \sim \text{Logistic}(0, 1)$ is added to $U_{i,j}$, producing $g_\tau(U_{i,j}) = \sigma((U_{i,j} + G_{i,j})/\tau)$, where τ is the temperature that controls the smoothness of the Softmax, and $\sigma(\cdot)$ is the logistic sigmoid function. As the optimization process proceeds, the values of $g_\tau(U_{i,j})$ approach either 0 or 1, approximating an ℓ_0 penalty.

Incorporating the moral graph and hard DAG constraints Furthermore, CALM incorporates a learned moral graph $M \in \{0, 1\}^{d\times d}$ to restrict the optimization to edges within the moral graph, thus reducing the search space. Note that similar idea has been used in various existing works such as (Loh and Bühlmann 2014; Nazaret et al. 2024). This moral graph acts as a filter over the Gumbel-Softmax mask, allowing only edges present in the moral graph. The final learned B can be represented as $B = M \circ g_\tau(U) \circ P$, incorporating both the sparsity from the Gumbel-Softmax mask and the structural constraints from the moral graph. Here, B contains the edge weights for the structure, determined by the mask $M \circ g_\tau(U)$.

CALM's objective function, incorporating the Gumbel-Softmax mask, moral graph, and hard DAG constraints into the GOLEM-NV-ℓ_1 formulation, is given by

$$\begin{aligned} \min_{U\in\mathbb{R}^{d\times d}, P\in\mathbb{R}^{d\times d}} \quad & \mathcal{L}(M \circ g_\tau(U) \circ P; \Sigma) + \lambda_1 \|M \circ g_\tau(U)\|_1 \\ \text{subject to} \quad & h(M \circ g_\tau(U)) = 0. \end{aligned}$$

where $\mathcal{L}(M \circ g_\tau(U) \circ P; \Sigma)$ is the likelihood term, and the $\lambda_1 \|M \circ g_\tau(U)\|_1$ term approximates the ℓ_0 penalty for sparsity. Here, both the ℓ_0 penalty for sparsity and the DAG constraints are applied to the final learned mask $M \circ g_\tau(U)$, which determines the presence of edges.

It is worth noting that there exist structure learning approaches that adopted Gumbel Softmax, but they focus on nonlinear data (Ng et al. 2022b) or interventional data (Brouillard et al. 2020). Our work targets the linear Gaussian case and incorporates a moral graph to reduce the search space and mitigate non-convexity.

4.2 Experimental Setup

Across all experiments in section 4, we simulate Erdös–Rényi graphs (Erdös and Rényi 1959) with kd edges, denoted as ERk graphs, with edge weights uniformly sampled from $[-2, -0.5] \cup [0.5, 2]$. For all experiments, the data is generated with a fixed noise ratio of 16. Specifically, the variances of two randomly selected noise variables are set to 1 and 16, respectively, while the variances of the remaining noise variables are sampled uniformly from the range $[1, 16]$. This setting ensures a realistic variation in noise across the variables, aligning with the assumptions of non-equal noise variances (NV) (the only exception is the additional experiments in Appendix F, where we compare CALM against other baselines under the equal noise variance data cases).

Further implementation details of our experiments in section 4 are in Appendix A. Algorithm 1 shows CALM's pseudocode. Appendix A.2 details the implementation and parameter selection of the Gumbel-Softmax ℓ_0 penalty and the

Algorithm 1: CALM

Input: Sample covariance matrix Σ (computed from centered data matrix $\mathbf{X}$)
Obtain moral graph M by IAMB algorithm
Initialize parameters: U, P, penalty weight ρ, step size, maximum iterations $k_{\max}$
for $k = 1$ **to** $k_{\max}$ **do**
    Solve CALM's objective function (quadratic penalty method) via Adam optimizer, updating U and P
    Increase penalty weight for DAG constraint: $\rho \leftarrow 3\rho$
    if DAG constraint value $< 10^{-8}$ **then**
        break
    end if
end for
return $B = M \circ g_\tau(U) \circ P$

hard DAG constraint in CALM. It also discusses our optimization scheme, highlighting the benefits and computational complexity of the quadratic penalty method (QPM), and evaluates the practical quality of the Gumbel-Softmax ℓ_0 approximation. Moral graph estimation method, its parameter choice and quality evaluation are detailed in Appendix A.1, while Appendix A.3 summarises CALM's small-scale hyper-parameter tuning and the computing infrastructure. Appendix B compares three ℓ_0 approximations with the original GOLEM-NV-ℓ_1; all three ℓ_0 variants outperform GOLEM-NV-ℓ_1, and the Gumbel-Softmax approximation is the most robust and consistently effective, so we adopt it as our representative implementation. From this point forward, unless otherwise specified, we use "CALM" to refers to the specific version of the method where the ℓ_0-penalty is approximated using Gumbel Softmax.

For each scenario in the following sections, we conducted 10 experiments and calculated the mean SHD of CPDAG, precision of skeleton, recall of skeleton, and their standard errors.

4.3 Impact of Moral Graph and Soft/Hard DAG Constraints

Recall that NOTEARS adopts a hard DAG constraint while GOLEM uses a soft DAG constraint. Here, we evaluate the impact of incorporating the moral graph and using either soft or hard DAG constraints on the results of the ℓ_0-penalized likelihood optimization. We consider linear Gaussian models with 50 variables and ER1 graphs. Here, we focus on the nonconvexity aspect of the optimization problem, and thus set the sample size to infinity to eliminate finite sample errors (the way we achieve infinite samples is in Appendix A.5). The experiment results for finite samples are included in Section 4.5. Following (Reisach, Seiler, and Weichwald 2021; Kaiser and Sipos 2022), we standardize the data. All experiments were conducted with the Gumbel-Softmax approach to approximate ℓ_0 penalty. The implementation details of Gumbel-Softmax-based ℓ_0 penalty and the hard DAG constraints is in Appendix A.2. The implementation details of soft DAG constraints is in Appendix A.4.

Comparison of soft and hard DAG constraints From the results in Table 2, we observe that using hard DAG constraints leads to a lower SHD of CPDAG compared to soft DAG constraints, regardless of whether the moral graph is incorporated. This suggests that, even when both formulation with soft and hard DAG constraints converge to local optima, the hard DAG constraint results are closer to the true adjacency matrix B or its Markov equivalence class.

One explanation for the improved performance of hard DAG constraints is the use of a quadratic penalty method (QPM) (Ng et al. 2022a). In this framework, the hyperparameter ρ, which controls the weight of the DAG constraint, is progressively increased during optimization, with each ρ value triggering a full subproblem optimization. This leads to a more refined optimization process. In contrast, the soft DAG constraint uses a fixed ρ, resulting in only one subiteration and possibly worse convergence. Additionally, hard constraints ensure that the final graph is always a DAG, eliminating the need for post-processing, whereas soft constraints often require post-processing to enforce acyclicity (Ng, Ghassami, and Zhang 2020), which may introduce errors and increase SHD of CPDAG.

Impact of including the moral graph Table 2 also shows that incorporating the moral graph improves performance in both soft and hard DAG constraint settings, with a notably lower SHD of CPDAG. The moral graph reduces the search space by focusing on edges within it, which is especially beneficial in higher-dimensional settings like our 50-node experiments, where the reduction in search space is more substantial. This significantly simplifies the optimization process and leads to better convergence towards the true adjacency matrix or its Markov equivalence class.

In summary, CALM, combining hard DAG constraints and the moral graph, delivers the best results.

4.4 Impact of Data Standardization

(Ng, Huang, and Zhang 2024) previously pointed out that the original GOLEM-NV-ℓ_1 formulation performed poorly both before and after data standardization. To further evaluate the robustness of CALM, we conduct experiments to compare its performance with (CALM-Standardized) and without data standardization (CALM-Non-Standardized). We use infinite samples to eliminate finite sample error and consider a 50-node linear Gaussian model with ER1 graphs.

In Table 3, CALM shows consistently low SHD of CPDAG before and after data standardization, demonstrating its stability and robustness across both standardized and non-standardized data. Interestingly, this is in constrast with the observation by (Reisach, Seiler, and Weichwald 2021; Kaiser and Sipos 2022) that differentiable structure learning methods do not perform well after data standardization, which further validate the robustness of our method.

4.5 Comparison with Baseline Methods

We finally compare the performance of CALM against several baseline methods, including three differentiable methods, the original GOLEM-NV-ℓ_1, NOTEARS, and DAGMA (Bello, Aragam, and Ravikumar 2022), and two discrete

		50-node ER1 graphs		50-node ER4 graphs		100-node ER1 graphs	
	Method	SHD of CPDAG	Recall of Skeleton	SHD of CPDAG	Recall of Skeleton	SHD of CPDAG	Recall of Skeleton
Differentiable	CALM	12.1 ± 2.7	0.98 ± 0.00	168.8 ± 8.3	0.67 ± 0.02	26.7 ± 3.6	0.99 ± 0.00
	GOLEM-NV-ℓ_1	60.0 ± 3.9	0.65 ± 0.07	211.6 ± 4.2	0.22 ± 0.02	120.5 ± 6.7	0.75 ± 0.06
	NOTEARS	46.3 ± 1.9	0.81 ± 0.02	209.5 ± 1.2	0.15 ± 0.01	87.4 ± 2.9	0.74 ± 0.03
	DAGMA	73.3 ± 4.0	0.95 ± 0.01	253.0 ± 7.5	0.33 ± 0.02	152.6 ± 3.9	0.95 ± 0.01
Discrete	PC	11.0 ± 1.4	0.92 ± 0.01	200.5 ± 2.1	0.22 ± 0.01	24.7 ± 1.6	0.89 ± 0.01
	FGES	8.4 ± 2.4	0.98 ± 0.00	425.6 ± 23.2	0.80 ± 0.01	12.1 ± 1.7	0.98 ± 0.00

Table 4: Comparison with baseline methods for 50-node ER1, 50-node ER4, and 100-node ER1 graphs under data standardization, using 1000 samples. Results for 10^6 samples under data standardization (including SHD of CPDAG, Precision of Skeleton, and Recall of Skeleton) are in Appendix C.

methods, PC (Spirtes and Glymour 1991) and FGES (Ramsey et al. 2017) (see Appendix A.6 for baseline methods' implementation details). Here, we report results with data standardization for a sample size of $n = 1000$ in the main text; due to space limit, we only report SHD of CPDAG and Recall of Skeleton here. Additional results with data standardization for $n = 10^6$, including SHD of CPDAG, Precision of Skeleton, and Recall of Skeleton, are provided in Appendix C. We considered a 50-node linear Gaussian model with ER1 and ER4 graphs, as well as a 100-node linear Gaussian model with ER1 graphs. Here, the moral graph in CALM is estimated from finite samples. Specifically, for the 1000-sample experiments here, the moral graph is estimated from 1000 samples, while for the 10^6-sample experiments in Appendix C, it is estimated from 10^6 samples. Results on data without standardization (for both $n = 1000$ and $n = 10^6$) are reported separately in Appendix D.

Table 4 summarizes the performance comparison between CALM and the baseline methods. The results clearly demonstrate that CALM consistently outperforms the three differentiable methods, NOTEARS, the original GOLEM-NV-ℓ_1 and DAGMA, across all graph structures. We also performed Wilcoxon signed-rank tests on the SHD of CPDAG, confirming that CALM statistically outperforms all three differentiable baselines ($p < 0.005$ for all comparisons). This highlights the effectiveness and robustness of incorporating the Gumbel-Softmax approximation to ℓ_0, moral graph, and hard DAG constraints.

For sparse graphs such as ER1, CALM does not outperform the discrete methods PC and FGES. This is expected, as continuous optimization in linear likelihood-based formulations struggles with such high nonconvexity compared to discrete approaches. However, it is worth noting that for ER1 graphs with $n = 1000$ samples (the practical sample size considered here), the results of CALM are nearly identical to those of PC. This indicates that in practical scenarios with relatively small sample sizes, even in sparse graphs, CALM can effectively match the performance of discrete methods like PC, representing a significant breakthrough.

Furthermore, in more dense graphs, the 50-node ER4 graphs, CALM demonstrates superior performance compared to the PC and FGES methods. This result suggests that in higher-density graphs, CALM enables differentiable structure learning methods to outperform discrete methods. The comparison between sparse and dense graphs highlights an important aspect of CALM's performance. While CALM is less competitive than non-differentiable baselines like PC and FGES on sparse graphs, it demonstrates stronger performance on dense graphs. This contrast showcases CALM's ability to handle the increased complexity of dense graph structures. We also compare CALM with baselines on a broader range of graph sizes and structures, including 20-node ER4, 50-node SF4, 70-node ER4, and 200-node ER4 graphs, in Appendix E. The trade-off between runtime and performance in CALM is discussed in Appendix G. A supplementary but important discussion of improvements in CALM over existing methods is provided in Appendix H.

4.6 Real-World Data

To evaluate CALM on real-world data, we conducted experiments on the Sachs dataset (Sachs et al. 2005), which is commonly utilized in probabilistic graphical model research to analyze the expression levels of proteins and phospholipids within human cells. The dataset has $d = 11$ variables and n=853 samples, with a ground truth of 17 edges. Our method, CALM, achieved an SHD of CPDAG of 12, better than GOLEM-NV-ℓ_1 (SHD of CPDAG: 13) and NOTEARS (SHD of CPDAG: 22). These results demonstrate the strong performance of CALM on real-world data, highlighting its superiority over other differentiable methods.

5 Conclusion

Our work begins by identifying the inconsistency of ℓ_1-penalized likelihood in differentiable structure learning for the linear Gaussian case. To address this and improve performance, we propose CALM, which optimizes an ℓ_0-penalized likelihood with hard acyclicity constraints and incorporates moral graphs. Our results show that CALM, particularly with Gumbel-Softmax ℓ_0 approximation, significantly outperforms GOLEM-NV-ℓ_1, NOTEARS, and DAGMA across various graph types and sample sizes. In sparse graphs like ER1, CALM's performance rivals PC with 1000 samples, while in dense graphs like ER4, it achieves the best results among all baseline methods. CALM also maintains robust performance both before and after data standardization. Future work includes extending CALM to nonlinear models and integrating advanced optimization techniques for further improvements in linear models.

Acknowledgements

The authors would like to thank the reviewers for their constructive comments. We would like to acknowledge the support from NSF Award No. 2229881, AI Institute for Societal Decision Making (AI-SDM), the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Quris AI, Florin Court Capital, and MBZUAI-WIS Joint Program. IN acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) Postgraduate Scholarships – Doctoral program.

References

Awokuse, T. O.; and Bessler, D. A. 2003. Vector autoregressions, policy analysis, and directed acyclic graphs: an application to the US economy. *Journal of Applied Economics*, 6(1): 1–24.

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.

Bellot, A.; and van der Schaar, M. 2021. Deconfounded score method: Scoring DAGs with dense unobserved confounding. *arXiv preprint arXiv:2103.15106*.

Bhattacharya, R.; Nagarajan, T.; Malinsky, D.; and Shpitser, I. 2021. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, 2314–2322. PMLR.

Brouillard, P.; Lachapelle, S.; Lacoste, A.; Lacoste-Julien, S.; and Drouin, A. 2020. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33: 21865–21877.

Chen, X.; Sun, H.; Ellington, C.; Xing, E.; and Song, L. 2021. Multi-task learning of order-consistent causal graphs. *Advances in Neural Information Processing Systems*, 34: 11083–11095.

Chickering, D. M. 1996. Learning Bayesian networks is NP-complete. *Learning from data: Artificial intelligence and statistics V*, 121–130.

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

Chickering, M.; Heckerman, D.; and Meek, C. 2004. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5: 1287–1330.

Cussens, J. 2011. Bayesian network learning with cutting planes. In *27th Conference on Uncertainty in Artificial Intelligence*, 153–160. AUAI Press.

Deng, C.; Bello, K.; Ravikumar, P.; and Aragam, B. 2023. Global optimality in bivariate gradient-based DAG learning. *Advances in Neural Information Processing Systems*, 36: 17929–17968.

Deng, C.; Bello, K.; Ravikumar, P.; and Aragam, B. 2024. Markov equivalence and consistency in differentiable structure learning. *Advances in Neural Information Processing Systems*, 37: 91756–91797.

Erdös, P.; and Rényi, A. 1959. On Random Graphs I. *Publicationes Mathematicae*, 6: 290–297.

Faria, G. R. A.; Martins, A.; and Figueiredo, M. A. 2022. Differentiable causal discovery under latent interventions. In *Conference on Causal Learning and Reasoning*, 253–274. PMLR.

Gao, E.; Chen, J.; Shen, L.; Liu, T.; Gong, M.; and Bondell, H. 2021. FedDAG: Federated DAG structure learning. *arXiv preprint arXiv:2112.03555*.

Gao, E.; Ng, I.; Gong, M.; Shen, L.; Huang, W.; Liu, T.; Zhang, K.; and Bondell, H. 2022. Missdag: Causal discovery in the presence of missing data with continuous additive noise models. *Advances in Neural Information Processing Systems*, 35: 5024–5038.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparametrization with Gumble-Softmax. In *International Conference on Learning Representations (ICLR 2017)*. OpenReview. net.

Kaiser, M.; and Sipos, M. 2022. Unsuitability of NOTEARS for Causal Graph Discovery when Dealing with Dimensional Quantities. *Neural Processing Letters*, 54: 1–9.

Kalainathan, D.; Goudet, O.; Guyon, I.; Lopez-Paz, D.; and Sebag, M. 2022. Structural agnostic modeling: Adversarial learning of causal graphs. *Journal of Machine Learning Research*, 23(219): 1–62.

Koivisto, M.; and Sood, K. 2004. Exact Bayesian structure discovery in Bayesian networks. *The Journal of Machine Learning Research*, 5: 549–573.

Koller, D.; and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

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. *Journal of Machine Learning Research*, 15(88): 3065–3105.

Nazaret, A.; Hong, J.; Azizi, E.; and Blei, D. 2024. Stable Differentiable Causal Discovery. In *Proceedings of the 41st International Conference on Machine Learning*.

Ng, I.; Dong, X.; Dai, H.; Huang, B.; Spirtes, P.; and Zhang, K. 2024. Score-Based Causal Discovery of Latent Variable Causal Models. In *Forty-first International Conference on Machine Learning*.

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.

Ng, I.; Huang, B.; and Zhang, K. 2024. Structure learning with continuous optimization: A sober look and beyond. In *Causal Learning and Reasoning*, 71–105. PMLR.

Ng, I.; Lachapelle, S.; Ke, N. R.; Lacoste-Julien, S.; and Zhang, K. 2022a. On the convergence of continuous constrained optimization for structure learning. In *International Conference on Artificial Intelligence and Statistics*, 8176–8198. Pmlr.

Ng, I.; and Zhang, K. 2022. Towards federated bayesian network structure learning with continuous optimization. In *International Conference on Artificial Intelligence and Statistics*, 8095–8111. PMLR.

Ng, I.; Zhu, S.; Fang, Z.; Li, H.; Chen, Z.; and Wang, J. 2022b. Masked gradient-based causal structure learning. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, 424–432. SIAM.

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

Pearl, J. 1988. Probabilistic Reasoning in Intelligent Systems; Network of Plausible Inference. *Morgan Kaufmann, 1988*.

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

Prashant, P. P.; Ng, I.; Zhang, K.; and Huang, B. 2025. Differentiable Causal Discovery for Latent Hierarchical Causal Models. In *The Thirteenth International Conference on Learning Representations*.

Ramsey, J.; Glymour, M.; Sanchez-Romero, R.; and Glymour, C. 2017. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics*, 3: 121–129.

Raskutti, G.; and Uhler, C. 2018. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1): e183.

Reisach, A.; Seiler, C.; and Weichwald, S. 2021. Beware of the Simulated DAG! Causal Discovery Benchmarks May Be Easy to Game. In *Advances in Neural Information Processing Systems*.

Saboksayr, S. S.; Mateos, G.; and Tepper, M. 2023. CoLiDE: Concomitant Linear DAG Estimation. *arXiv preprint arXiv:2310.02895*.

Sachs, K.; Perez, O.; Pe'er, D.; Lauffenburger, D. A.; and Nolan, G. P. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529.

Scheines, R.; Spirtes, P.; Glymour, C.; Meek, C.; and Richardson, T. 1998. The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1): 65–117.

Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics*, 461–464.

Sethuraman, M. G.; and Fekri, F. 2025. Differentiable Cyclic Causal Discovery Under Unmeasured Confounders. *arXiv preprint arXiv:2508.08450*.

Singh, A. P.; and Moore, A. W. 2005. *Finding optimal Bayesian networks by dynamic programming*. Carnegie Mellon University. Center for Automated Learning and Discovery.

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

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

Sun, X.; Liu, G.; Poupart, P.; and Schulte, O. 2021. Nts-notears: Learning nonparametric temporal dags with time-series data and prior knowledge. *arXiv preprint arXiv:2109.04286*.

Tennant, P. W.; Murray, E. J.; Arnold, K. F.; Berrie, L.; Fox, M. P.; Gadd, S. C.; Harrison, W. J.; Keeble, C.; Ranker, L. R.; Textor, J.; et al. 2021. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *International journal of epidemiology*, 50(2): 620–632.

Tsamardinos, I.; Aliferis, C. F.; Statnikov, A. R.; and Statnikov, E. 2003. Algorithms for large scale Markov blanket discovery. In *FLAIRS*, volume 2, 376–81.

Wang, Y.; Menkovski, V.; Wang, H.; Du, X.; and Pechenizkiy, M. 2020. Causal discovery from incomplete data: a deep learning approach. *arXiv preprint arXiv:2001.05343*.

Wang, Z.; Chen, X.; Dong, Z.; Dai, Q.; and Wen, J.-R. 2022. Sequential Recommendation with Causal Behavior Discovery. *arXiv preprint arXiv:2204.00216*.

Wei, D.; Gao, T.; and Yu, Y. 2020. DAGs with No Fears: A closer look at continuous optimization for learning Bayesian networks. *Advances in Neural Information Processing Systems*, 33: 3895–3906.

Wright, S. J. 2006. Numerical optimization.

Yamada, Y.; Lindenbaum, O.; Negahban, S.; and Kluger, Y. 2020. Feature selection using stochastic gates. In *International conference on machine learning*, 10648–10659. PMLR.

Yang, S.; Yu, K.; Cao, F.; Liu, L.; Wang, H.; and Li, J. 2021. Learning Causal Representations for Robust Domain Adaptation. *IEEE Transactions on Knowledge and Data Engineering*.

Yu, Y.; Chen, J.; Gao, T.; and Yu, M. 2019. DAG-GNN: DAG structure learning with graph neural networks. In *International conference on machine learning*, 7154–7163. PMLR.

Yuan, C.; and Malone, B. 2013. Learning optimal Bayesian networks: A shortest path perspective. *Journal of Artificial Intelligence Research*, 48: 23–65.

Zhang, Z.; Ng, I.; Gong, D.; Liu, Y.; Abbasnejad, E.; Gong, M.; Zhang, K.; and Shi, J. Q. 2022. Truncated matrix power iteration for differentiable dag learning. *Advances in Neural Information Processing Systems*, 35: 18390–18402.

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. In *International Conference on Artificial Intelligence and Statistics*, 3414–3425. Pmlr.