Causal Order Discovery based on Monotonic SCMs

Ali Izadi
Department of Computing Science
Simon Fraser University
ali_izadi@sfu.ca

Martin Ester
Department of Computing Science
Simon Fraser University
ester@sfu.ca

Abstract

In this paper, we consider the problem of causal order discovery within the framework of monotonic Structural Causal Models (SCMs), which have gained attention for their potential to enable causal inference and causal discovery from observational data. While existing approaches either assume prior knowledge about the causal order or use complex optimization techniques to impose sparsity in the Jacobian of Triangular Monotonic Increasing maps, our work introduces a novel sequential procedure that directly identifies the causal order by iteratively detecting the root variable. This method eliminates the need for sparsity assumptions and the associated optimization challenges, enabling the identification of a unique SCM without the need for multiple independence tests to break the Markov equivalence class. We demonstrate the effectiveness of our approach in sequentially finding the root variable, comparing it to methods that maximize Jacobian sparsity.

1 Introduction

The problem of finding the causal relationships among a set of variables, or causal discovery, has been extensively studied to address interventional and counterfactual queries (causal inference) [1]. These two type of queries play a crucial role in biology [2, 3], economics [3], philosophy [4], and machine learning [5]. Structural Causal Models (SCMs) are among the most well-known frameworks for mathematically formulating the problem of causal discovery [6]. SCMs model causal relationships by assigning each effect variable as a function of its causes and an unobserved noise. They enable the answering of interventional and counterfactual queries, as well as the learning of causal relationships from observational data.

In general, one needs interventional data to learn the true, unique SCMs. However, for special types of SCMs the task of causal discovery is possible from just observational data. The identifiability of SCMs has been investigated by proposing different functional assumptions between causes and effect. These types of SCMs, including those with linear functions with non-Gaussian noise [7], nonlinear functions with additive independent noise [8], and post-nonlinear models have been proven to be identifiable from observational data [9].

Recent research has demonstrated that Monotonic SCMs can be identified using triangular monotonic increasing (TMI) maps, provided the causal order of these maps is known [10]. Monotonic SCMs are a class of SCMs characterized by nonlinear functions between causes and effects, where each function is monotone with respect to unobserved noise. Given the true order of the maps, the Jacobian of normalizing flows transformation as TMI maps [11] has been shown to effectively learn Monotonic SCMs [12]. Although knowing the order is a restrictive assumption for causal discovery, normalizing flows have been employed for causal inference [12]. Nevertheless, TMI maps can facilitate causal discovery based on Monotonic SCMs without requiring the order [13]. It has been established that the true order of TMI maps corresponds to an order where the Jacobian of TMI maps is maximally sparse [13]. This insight leads to a two-step optimization procedure for estimating TMI maps with sparse Jacobians at each iteration of searching over orders (permutations). This introduces significant

Accepted to the NeurIPS 2024 Workshop on Causal Representation Learning.

complexity to the optimization efficiency of this procedure, with many local minima, as discussed in [13], without proposing an efficient method to overcome this issue. However, this approach identifies only the Markov Equivalent Class (MEC) of the true SCM, and multiple independence tests need to be performed to identify the unique SCM.

In this paper, we propose a procedure for learning the causal order of Monotonic SCMs based on the idea of sequentially identifying the root of the graph. For Monotonic SCMs, the Jacobian vector of the map of the root variable with respect to all other variables contains all zero values. Instead of enforcing sparsity on the Jacobian of the entire TMI maps, we determine the order by sequentially discovering the roots using the zeroness of the Jacobian of the root map. This approach enables the learning of the true order of the SCMs, making SCMs identifiable [10] and breaking the Markov Equivalent Class (MEC), eliminating the need for independence tests. Additionally, this sequential procedure, without any sparsity assumption, simplifies the complex two-step process of learning TMI maps and orders. In summary, the main contributions of this paper are as follows:

- We introduce a novel sequential procedure for discovering causal order in Monotonic SCMs by leveraging the Jacobian of TMI maps to identify the root variable, and to the best of our knowledge, this represents the first efficient approach for causal discovery based on Monotonic SCMs.
- We conduct several experiments to demonstrate the efficiency of our approach in finding the correct order by sequentially discovering the root variable. Our method is compared to approaches that aim to learn the order with the maximally sparse Jacobian.

2 Related Works

Nonlinear SCMs: The approach to causal discovery for nonlinear SCMs using the Jacobian began with [14], but it has focused on additive nonlinear noise models rather than monotonic SCMs. In contrast, the connection between nonlinear ICA and causal discovery for nonlinear SCMs has been explored in [15, 16]. However, identifying the true unique graph through ICA requires counterfactual data, which is a restrictive assumption compared to our proposed method that relies on observational data. While the identifiability of monotonic SCMs was introduced through the concept of indeterminacy in generative models [10], knowing the causal order is essential. Our paper, however, focuses on determining this causal order for monotonic SCMs. Additionally, the assumption of bijectivity for nonlinear SCMs has been demonstrated to be useful for counterfactual identifiability results in Monotonic SCMs. A recent fixed-point approach for causal discovery based on nonlinear SCMs involves amortized causal order discovery, but does not specifically address monotonic SCMs [17].

Causal order discovery: Many approaches in causal discovery focus on searching the space of causal orders rather than the space of DAGs. These methods first identify the topological order of the causal graph and then apply an edge pruning algorithm to obtain the final causal graph [18, 19]. RL-based [20] and Permutation based [21, 22] determine the causal order using a greedy approach, which is not applicable for high-dimensional data due to its combinatorial nature.

Another set of order-based methods attempts to find the topological order based on discovering permutation matrix [23, 24, 25]. While these methods make the problem of causal discovery fully differentiable by proposing a procedure for learning the combinatorial permutation matrix, their approaches suffer from searching in the space of permutation matrices, which is challenging to optimize due to its combinatorial nature. Another line of research focuses on using the score or gradient of probability distribution of data for causal order discovery [26, 27, 28, 29, 30]. In contrast to score-bases methods that leverage the variance of the Hessian of the probability distribution for selecting leaves of graph iteratively, our method uses Jacobian of TMI maps for finding the roots iteratively. However, unlike the proposed method, none of the methods assume Monotonic SCMs.

TMI maps and Normalizing flows for causal discovery: Flow-based models have been employed for causal discovery under the assumptions of location-scale noise models [31, 32]. Our method differs from these two, as we assume the monotonic SCMs. Another group of methods develops the theory of normalizing flows for causal inference and proposes an algorithm for causal discovery using normalizing flows [12]. However, this algorithm assumes that the topological order of the graph is given as input, whereas the focus of our paper is on discovering this causal order. TMI maps have also been used to learn conditional independence through sparsity, but this application is limited to

undirected settings in graphical models, not for causal discovery [33, 34, 35]. While TMI maps have been applied to causal discovery, their use has been confined to identifying the MEC equivalence class based on conditional independencies derived from TMI maps [36, 13]. In contrast, our work focuses on discovering the causal order of the graph, without relying on conditional independencies or outputting the Markov equivalence class.

3 Background

3.1 Monotonic Structural Causal Models

Given a d-dimensional observational dataset $X \in \mathcal{R}^{n \times d}$, where X contains n i.i.d. samples. The causal relationships among the set of variables are captured by a Directed Acyclic Graph (DAG) $\mathcal{G}$, which is represented by the adjacency matrix $A \in \{0,1\}^{d \times d}$. The adjacency matrix A models the causal relationship of each random variable x_i and its parents $Pa(x_i)$. We can write the generative process of x_i in the form of Equation 1

$$x_i = f_i(Pa_{\mathcal{G}}(x_i), u_i), \quad i = 1, ..., d \tag{1}$$

where f_i is an arbitrary function, $Pa_{\mathcal{G}}(x_i)$ denotes the parents of variable x_i in DAG $\mathcal{G}$, and the exogenous noise u_i is independent of $Pa_{\mathcal{G}}(x_i)$. Also, the different noise variables u_i are mutually independent, which can be arbitrary independent noise distributions. For Monotonic SCMs, we assume that each f_i is monotone in u_i [10, 12, 13].

3.2 Normalizing flows and TMI maps

Normalizing flows are a class of generative models that transform the probability distribution of a random variable $X \in \mathcal{R}^d$ into the probability distribution of a random variable $U \in \mathcal{R}^d$ through a series of invertible and differentiable transformations $T(x)$ [37, 38]. A key aspect of this transformation process is the use of Triangular Monotonic Increasing (TMI) maps. A TMI map is a function $\mathcal{R}^d \to \mathcal{R}^d$ that is both triangular—where each output component T_i depends on the $x_{1:i}$ components of the input, as expressed by

$$T(x_1, x_2, \ldots, x_d) = (T_1(x_1), T_2(x_1, x_2), \ldots, T_d(x_1, x_2, \ldots, x_d)), \tag{2}$$

and monotonic increasing w.r.t x_i [39, 40], ensuring that the transformation is invertible. The triangular structure of TMI maps allows for efficient computation of the inverse map and the Jacobian determinant, which are essential for calculating the likelihood in normalizing flows.

4 Proposed Method

4.1 Problem definition

The objective of this paper is to determine the causal order of variables (Refer to Appendix A for the formal definition of causal order). Although a DAG $\mathcal{G}$ may have multiple causal orderings, for any given causal ordering π, we can construct a fully connected DAG $\mathcal{G}^{\pi}_{\text{full}}$ where, for every pair of nodes $\pi(i)$ and $\pi(j)$, there is a directed edge $\pi(i) \to \pi(j)$ if and only if $i < j$ [18]. We define the set of true causal orderings Π as:

$$\Pi = \{\pi | \mathcal{G}^{\pi}_{\text{full}} \text{ is a super-graph of } \mathcal{G}\} \tag{3}$$

Under the assumption of a monotonic SCM (Section 3.1), the set Π is identifiable. Therefore, the goal of causal order discovery is to find a $\pi \in \Pi$ [41].

Assumptions: i) Each function f_i in Equation 1 is continuous, invertible, and therefore monotonic; ii) The graph is acyclic; iii) There are no hidden confounders, so the noise variables u_i are mutually independent (causal sufficiency).

4.2 Overview of the method

Our proposed method finds the causal order of the nodes by iteratively identifying the roots of the graph, removing each root from the graph, and repeating this process until only one variable remains. The topological order of the graph is then determined as the order in which roots were found.

In the following sections, we detail our methodology for identifying the roots of the graph. This process involves training one-dimensional conditional normalizing flows to model the functions f_i of a monotonic SCM, and using the Jacobian of each map ($\nabla T_i(x_i)$) as a criterion for detecting the graph's root. Next, we introduce an algorithm for sequentially finding the roots to determine the causal order. Once the causal order is discovered, one can either train a d-dimensional normalizing flow to identify the adjacency matrix [12], or apply the CAM pruning method [19] to prune the edges. This approach uniquely discovers the causal graph without relying on independence tests, unlike other methods that depend on the faithfulness assumption and the sparsity of the Jacobian to identify the Markov equivalence class and break it using independence tests [13].

4.3 Finding the root

We assume d independent multi-cause models, i.e. monotonic SCM, according to equation 4 as follows:

$$x_i = f_i(X\backslash\{x_i\}, u_i), \quad i = 1, ..., d \tag{4}$$

where $X\backslash\{x_i\}$ are all variables except variable x_i, and f_i are monotonic in u_i. Thus, $f_i^{-1}(X\backslash\{x_i\}, x_i)$ is also monotonic increasing w.r.t x_i, and it is a monotonic increasing map. As a result, for each f_i^{-1} we train one-dimensional conditional normalizing flow $T_i(x_i, X\backslash\{x_i\})$, where each T_i transforms x_i to u_i conditioned on $X\backslash\{x_i\}$ and each T_i is a monotone increasing map for each x_i.

Theorem 1 *Suppose $X \in \mathcal{R}^{n\times d}$ is generated based on a monotonic SCM according to Equation 1. Then, for the multi-cause model 4, $\nabla_{X\backslash\{x_i\}} T_i(x_i, X\backslash\{x_i\}) = \tilde{\mathbf{0}}$, if and only if x_i is a root in $\mathcal{G}$.*

Sketch of the Proof: The proof leverages Theorem 1 from [12], which establishes that, under the assumption of a known causal order, the Jacobian of the transformation T is $I - A$, where A is a lower triangular matrix. Therefore, for the root variable x_i, the Jacobian $\nabla_{X\backslash\{x_i\}} T_i(x_i, X\backslash\{x_i\}) = \tilde{\mathbf{0}}$ is zero due to the structure of A. Refer to Appendix B for a complete proof of Theorem 1.

To practically implement the criterion in Theorem 1, we detect the root by finding the maximum absolute value in each row of the Jacobian and selecting the variable corresponding to the row with the smallest of these maximum values.

$$\text{Root variable} = \arg\min_{i=1,\dots,d} \max \left|\nabla_{X\backslash\{x_i\}} T_i(x_i, X\backslash\{x_i\})\right| \tag{5}$$

While other aggregation schemes, such as summation or Hotelling's T-squared test [42], can be applied here, we found the min-max criterion to be more stable in our experiments for detecting the root.

4.4 Causal order discovery

Once the first root of the graph is detected, we remove the root variable from the current set of variables and add it as the first element in the causal order. We then continue finding subsequent roots until the entire topological order is determined. Refer to Appendix C for a detailed description of the causal order discovery algorithm, as summarized in Algorithm 1.

5 Empirical evaluation

Synthetic data generation Following the previous works [13, 15, 16], the synthetic data was generated by first constructing a DAG with d nodes, where edges were included with a probability of $\frac{2}{d-1}$. Edges were then randomly assigned to the nodes. Synthetic samples were drawn from a standard Gaussian distribution and passed through an MLP with 8 layers. The MLP's weights were masked at each layer based on the generated lower-triangular adjacency matrix, ensuring that connections aligned with the DAG structure. The weights were also randomly sampled and constrained to be positive to enforce monotonicity, with nonlinearity introduced through the leaky ReLU activation function. Additionally, [43] mentioned that the true topological order may be discovered by sorting the marginal variance of each variable. To prevent this phenomenon, the dataset was standardized.

Evaluation metric To assess the performance of causal order discovery, we employed the Count Backward (CB) metric, introduced in [26]. CB counts the number of backward edges in the ground truth graph with respect to the discovered topological order. This metric quantifies the number of errors in edges of the form $i \to j$ in the true ground truth graph $\mathcal{G}$, where $\pi_i > \pi_j$ in the discovered topological order π. CB equals zero when the predicted topological order is correct, and it equals the number of edges in the true graph $\mathcal{G}$ when the predicted topological order is the reverse of the true order.

See Appendix D for a detailed explanation of the architectural design and hyperparameters of Algorithm 1 for causal order discovery.

5.1 Sequential or Permutation based Causal order discovery?

Experimental design As a baseline, we implemented a procedure for causal order discovery in monotonic SCMs based on learning the permutation matrix for variable ordering and enforcing sparsity of the Jacobian through l_1 regularization on the Jacobian of the TMI transformation, following the approach introduced in [13]. For learning the permutation matrix, we used the Gumbel-Sinkhorn network with a temperature parameter t [44]. See Appendix E for a detailed explanation of the permutation based algorithm. As discussed in [13], satisfying both sparsity and learning the permutation is a complex optimization problem. This challenge is evident in setting the temperature parameter (t) and the penalty coefficient of the regularization term (λ), as it leads to many local minima. In our results, we compare the proposed method with the baseline using different values of t and λ to demonstrate the superiority of our approach in terms of optimization efficiency, avoiding many local minima.

Results The results for datasets with dimensions 4 and 10 are presented in Figure 1. Each column in each plot corresponds to a box plot of the given metric over runs of the algorithm for 10 random datasets. The results show that the proposed algorithm consistently outperforms the permutation-based method in terms of Count Backward (CB). For both dimensions, different combinations of temperature t and regularization λ yield varying results, highlighting the presence of multiple local minima in this approach. In contrast, the proposed method demonstrates superior performance without the need to fine-tune hyperparameters related to ordering and sparsity.

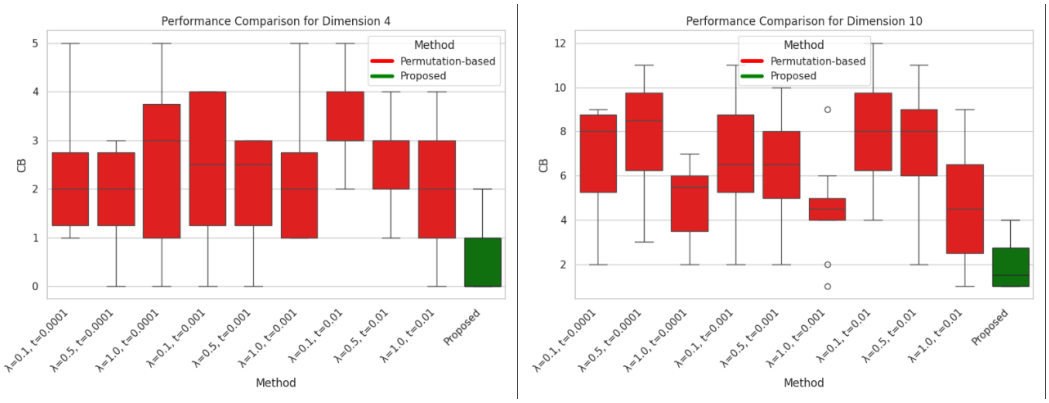

Figure 1: Performance comparison of sequential order discovery vs permutation-based order discovery

5.2 Ablation Analysis

As discussed in [12], we conduct an ablation analysis on the number of flows for both our sequential method and the permutation-based method to balance model complexity and causal consistency. This analysis helps us understand how increasing the number of normalizing flow layers affects causal inconsistency, the likelihood of encountering shortcuts, and the risk of local minima.

Both methods were executed on 10 random synthetic datasets, and the mean and standard deviation of the results from these 10 runs for the *CB* metric are shown in Figure 2. In the permutation-based approach, the values of t and λ are set to 0.0001 and 0.5, respectively. The results indicate that for dimension 10, increasing the number of flow layers results in a higher *CB* for the proposed

method, but a lower *CB* for the permutation-based method. The results for the proposed method are intuitive, as adding more layers of normalizing flow introduces causal inconsistency. However, for the permutation-based method, the performance improves, likely due to the increased complexity of the flow helping to overcome the challenges of the optimization process, as discussed in the previous section.

A similar phenomenon occurred for dimension 4, where increasing the number of flow layers from 1 to 2 improved performance, but the reverse effect was observed with 3 flow layers. This suggests a trade-off between model complexity and causal consistency.

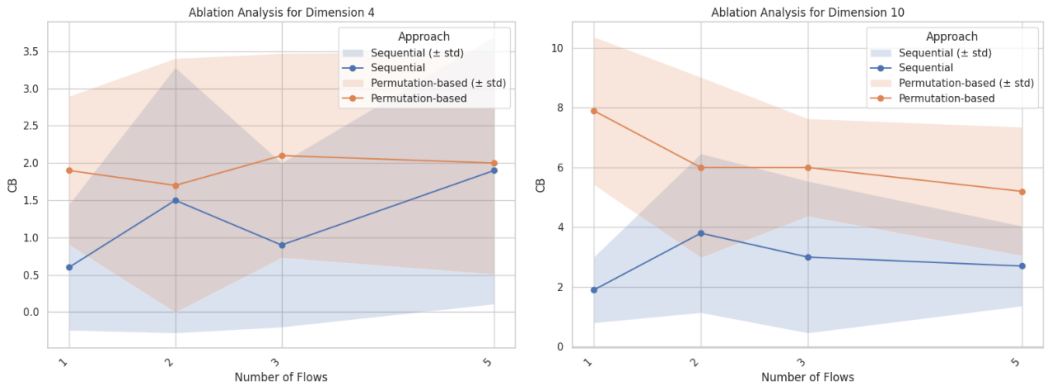

Figure 2: Ablation analysis on the number of flow layers and hidden layers of MLP.

In addition to the above results, detailed evaluations of the proposed method against other causal order discovery approaches, such as SCORE [26] and Direct-Lingam [45], on synthetic and real world datasets are available in Appendix F.

6 Discussion

Limitations While our method addresses several challenges in causal order discovery for Monotonic SCMs, it assumes that the underlying causal relationships adhere strictly to the monotonicity condition. This assumption may not hold in more complex or real-world scenarios where the relationships between variables are non-monotonic or involve interactions that violate the monotonicity condition. Additionally, our sequential root-identification process, although efficient, may be susceptible to error propagation, especially in high-dimensional settings or when the data is noisy, which could lead to incorrect causal orderings. Additionally, the min-max criterion may not be suitable for high-dimensional settings, as it can overlook variables whose Jacobian values are close to zero. This issue likely arises from the difficulty of accurately estimating the Jacobian using a single-layer normalizing flow. Higher-dimensional cases require more complex functions, but these often struggle with numerous local minima when estimating the Jacobian, leading to large Jacobian values even for the root variable.

Future Works As the number of variables increases, the computational complexity of iteratively identifying the root variable may become a limiting factor. Future research could focus on developing scalable algorithms to improve efficiency in high-dimensional settings. Another key direction for future work involves increasing the number of normalizing flow layers without compromising causal consistency or getting trapped in local minima. Using more complex normalizing flows can lead to better Jacobian estimation for accurately detecting the root of the graph. Additionally, alternative methods for estimating triangular maps could help overcome the challenges of computing the Jacobian for root variable identification.

Moreover, exploring the benefits of the proposed causal order discovery approach for the overall problem of causal graph discovery would be valuable. One could incorporate different pruning algorithms, such as CAM [19], or leverage identifiability results of monotonic SCMs to extract the adjacency matrix of causal relationships, given the causal graph order, by utilizing a single d-dimensional normalizing flow [12].

Several studies address the challenges associated with optimization-based algorithms for causal discovery, which can be applied to causal discovery based on monotonic SCMs. For instance, the analysis of sparsity penalties and thresholding discussed in [46] provides insights into overcoming the limitations of permutation-based methods. Additionally, ideas such as topological swaps introduced in [47] can help tackle the difficulties faced in order-based causal discovery. Furthermore, the notion of global optimality proposed in [48] offers strategies for avoiding local minima in causal discovery methods based on monotonic SCMs.

7 Conclusion

In this paper, we introduced a novel sequential procedure for causal order discovery within Monotonic Structural Causal Models (SCMs). Unlike existing methods that rely on complex optimization techniques and sparsity assumptions, our approach iteratively identifies the root variable by leveraging the zeroness of the Jacobian of Triangular Monotonic Increasing (TMI) maps. This method not only simplifies the causal discovery process but also ensures the identification of a unique SCM without requiring multiple independence tests to break the Markov equivalence class. We conducted several experiments comparing our method to those that maximize Jacobian sparsity, demonstrating its efficiency in correctly identifying the causal order.

References

[1] Bernhard Scholkopf, Jonas Peters, and Dominik Janzing. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

[2] Alessio Zanga, Elif Ozkirimli, and Fabio Stella. A survey on causal discovery: theory and practice. *International Journal of Approximate Reasoning*, 151:101–129, 2022.

[3] Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

[4] Daniel Malinsky and David Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470, 2018.

[5] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

[6] Judea Pearl. *Causality: Models, reasoning, and inference*. Cambridge University Press, 2009.

[7] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

[8] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

[9] Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, volume 9, page 647–655. AUAI Press, 2009.

[10] Quanhan Xi and Benjamin Bloem-Reddy. Indeterminacy in generative models: Characterization and strong identifiability. In *International Conference on Artificial Intelligence and Statistics*, pages 6912–6939. PMLR, 2023.

[11] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.

[12] Adrián Javaloy, Pablo Sánchez-Martín, and Isabel Valera. Causal normalizing flows: from theory to practice. *Advances in Neural Information Processing Systems*, 36, 2024.

[13] Quanhan Xi, Sebastian Gonzalez, and Benjamin Bloem-Reddy. Triangular monotonic generative models can perform causal discovery. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023.

[14] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. Pmlr, 2020.

[15] Patrik Reizinger, Yash Sharma, Matthias Bethge, Bernhard Schölkopf, Ferenc Huszár, and Wieland Brendel. Jacobian-based causal discovery with nonlinear ica. *Transactions on Machine Learning Research*, 2023.

[16] Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with general non-linear relationships using non-linear ica. In *Uncertainty in artificial intelligence*, pages 186–195. PMLR, 2020.

[17] Meyer Scetbon, Joel Jennings, Agrin Hilmkil, Cheng Zhang, and Chao Ma. A fixed-point approach for causal generative modeling. In *Forty-first International Conference on Machine Learning*, 2024.

[18] Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research JMLR*, 15(58):2009–2053, 2014.

[19] Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526 – 2556, 2014.

[20] Xiaoqiang Wang, Yali Du, Shengyu Zhu, Liangjun Ke, Zhitang Chen, Jianye Hao, and Jun Wang. Ordering-based causal discovery with reinforcement learning. *arXiv preprint arXiv:2105.06631*, 2021.

[21] Garvesh Raskutti and Caroline Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1):e183, 2018.

[22] Chandler Squires, Joshua Amaniampong, and Caroline Uhler. Efficient permutation discovery in causal dags. *arXiv preprint arXiv:2011.03610*, 2020.

[23] Bertrand Charpentier, Simon Kibler, and Stephan Günnemann. Differentiable dag sampling. *arXiv preprint arXiv:2203.08509*, 2022.

[24] Valentina Zantedeschi, Luca Franceschi, Jean Kaddour, Matt J Kusner, and Vlad Niculae. Dag learning on the permutahedron. *arXiv preprint arXiv:2301.11898*, 2023.

[25] Chris Cundy, Aditya Grover, and Stefano Ermon. Bcd nets: Scalable variational approaches for bayesian causal discovery. *Advances in Neural Information Processing Systems*, 34:7095–7110, 2021.

[26] Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning ICML*, pages 18741–18753. PMLR, 2022.

[27] Pedro Sanchez, Xiao Liu, Alison Q O'Neil, and Sotirios A Tsaftaris. Diffusion models for causal discovery via topological ordering. *arXiv preprint arXiv:2210.06201*, 2022.

[28] Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Causal discovery with score matching on additive models with arbitrary noise. *arXiv preprint arXiv:2304.03265*, 2023.

[29] Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable causal discovery with score matching. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213, pages 752–771. PMLR, 2023.

[30] Zhenyu Zhu, Francesco Locatello, and Volkan Cevher. Sample complexity bounds for score-matching: causal discovery and generative modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

[31] Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvarinen. Causal autoregressive flows. In *International conference on artificial intelligence and statistics*, pages 3520–3528. PMLR, 2021.

[32] Hamidreza Kamkari, Vahid Zehtab, Vahid Balazadeh, and Rahul G Krishnan. Ocdaf: Ordered causal discovery with autoregressive flows. *arXiv preprint arXiv:2308.07480*, 2023.

[33] Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. Inference via low-dimensional couplings. *Journal of Machine Learning Research*, 19(66):1–71, 2018.

[34] Rebecca Morrison, Ricardo Baptista, and Youssef Marzouk. Beyond normality: Learning sparse probabilistic graphical models in the non-gaussian setting. *Advances in neural information processing systems*, 30, 2017.

[35] Ricardo Baptista, Rebecca Morrison, Olivier Zahm, and Youssef Marzouk. Learning non-gaussian graphical models via hessian scores and triangular transport. *Journal of Machine Learning Research*, 25(85):1–46, 2024.

[36] Sina Akbari, Luca Ganassali, and Negar Kiyavash. Causal discovery via monotone triangular transport maps. In *NeurIPS 2023 Workshop Optimal Transport and Machine Learning*, 2023.

[37] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

[38] Nicholas J Irons, Meyer Scetbon, Soumik Pal, and Zaid Harchaoui. Triangular flows for generative modeling: Statistical consistency, smoothness classes, and fast rates. In *International Conference on Artificial Intelligence and Statistics*, pages 10161–10195. PMLR, 2022.

[39] Vladimir Igorevich Bogachev, Aleksandr Viktorovich Kolesnikov, and Kirill Vladimirovich Medvedev. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309, 2005.

[40] Guillaume Carlier, Alfred Galichon, and Filippo Santambrogio. From knothe's transport to brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576, 2010.

[41] Ignavier Ng, Shengyu Zhu, Zhuangyan Fang, Haoyang Li, Zhitang Chen, and Jun Wang. Masked gradient-based causal structure learning. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 424–432. SIAM, 2022.

[42] Harold Hotelling et al. The generalization of student's ratio. 1931.

[43] Alexander G Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! varsortability in additive noise models. *Advances in neural information processing systems*, 13, 2021.

[44] Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. *arXiv preprint arXiv:1802.08665*, 2018.

[45] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research-JMLR*, 12(Apr):1225–1248, 2011.

[46] Ignavier Ng, Biwei Huang, and Kun Zhang. Structure learning with continuous optimization: A sober look and beyond. In *Causal Learning and Reasoning*, pages 71–105. PMLR, 2024.

[47] Chang Deng, Kevin Bello, Bryon Aragam, and Pradeep Kumar Ravikumar. Optimizing notears objectives via topological swaps. In *International Conference on Machine Learning*, pages 7563–7595. PMLR, 2023.

[48] Chang Deng, Kevin Bello, Pradeep Ravikumar, and Bryon Aragam. Global optimality in bivariate gradient-based dag learning. *Advances in Neural Information Processing Systems*, 36:17929–17968, 2023.

[49] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

[50] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[51] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Appendix

A Causal order

For a DAG $\mathcal{G}$, a permutation $\pi : \{1, ..., d\} \to \{1, ..., d\}$ is a causal topological order if:

$$\pi(i) < \pi(j) \quad \text{if} \quad j \in DE_i^G \tag{6}$$

where DE_i^G denotes the set of descendents of variable x_i in graph $\mathcal{G}$ [1]. Our proposed algorithm discovers the causal relationships between variables by finding the causal order of variables. We highlight that the causal ordering of a given graph is not necessarily unique [18].

B Proof of Theorem 1

Theorem 2 *Suppose $X \in \mathcal{R}^{n \times d}$ is generated based on a monotonic SCM according to Equation 1. Then, for the multi-cause model 4, $\nabla_{X \setminus \{x_i\}} T_i(x_i, X \setminus \{x_i\}) = \tilde{\mathbf{0}}$, if and only if x_i is a root in $\mathcal{G}$.*

Proof; Based on Theorem 1 in [12], if the causal order is known, the exogenous variables of the flow differ from the true ones by an invertible component-wise transformation of the variables u. Thus, $\nabla_x T(x) = I - A$, where A is a lower triangular matrix. Consequently, the first row corresponds to the root variable, and all values $\nabla T(X)_{1,2:d}$ are zero. However, since the order is unknown, we train each T_i map using a one-dimensional conditional normalizing flow. Instead of conditioning on the preceding variables $x_{1:i}$, we condition on all $X \setminus \{x_i\}$. Because A is lower triangular, $\nabla_{X \setminus \{x_i\}} T_i(x_i, X \setminus \{x_i\}) = \mathbf{0}$ for the root variable.

Motivating Example Suppose we have a monotonic SCM based on the graph $x_1 \to x_2 \to x_3$, where $x_1 = f_1(u_1)$, $x_2 = f_2(x_1, u_2)$, and $x_3 = f_3(x_2, u_3)$. Then, we have $u_1 = f_1^{-1}(x_1)$, $u_2 = f_2^{-1}(x_1, x_2)$, and $u_3 = f_3^{-1}(x_2, x_3)$. Since, $\nabla_x f^{-1}(x) = \nabla_x T(x) = I - A$, it follows that:

$$\nabla_x T(x) = \begin{bmatrix} 1 & 0 & 0 \\ * & 1 & 0 \\ 0 & * & 1 \end{bmatrix} \tag{7}$$

Thus, the first row corresponds to the root variable, and $\nabla_{x2,x3} f_1^{-1}(x_1, x_2, x_3) = \begin{bmatrix} 0 & 0 \end{bmatrix}$ while, $\nabla_{x1,x3} f_2^{-1}(x_2, x_1, x_3) \neq \tilde{\mathbf{0}}$ and $\nabla_{x1,x2} f_3^{-1}(x_3, x_1, x_2) \neq \tilde{\mathbf{0}}$.

C Causal order discovery algorithm

The algorithm for causal discovery based on Monotonic SCMs works by iteratively identifying the root variable in a dataset. Initially, the set of nodes is all variables, and the order π is empty. For each iteration, the algorithm estimates a multi-cause model for each variable, treating it as the effect and the others as its potential causes, using a one-dimensional conditional normalizing flow. The maximum Jacobian value of each variable is computed, and the variable with the smallest maximum Jacobian value is identified as the root. This variable is then added to the causal order π, and removed from the set of nodes. This process repeats until only one node remains, which is appended to complete the causal order. Algorithm 1 shows the overall procedure of the proposed method.

D Proposed Algorithm Detail

Architectural Design For each 1-dimensional normalizing flow transformation $T_i(x_i)$, we employed the neural spline flow [49], using a stack of 1 flows with 1 hidden block and 128 hidden channels. We also conditioned each $T_i(x_i)$ on $X \setminus \{x_i\}$ using an MLP comprising 1 hidden layer, each with a size of 128.

The configuration details of the hyperparameters used in our proposed method are summarized in Table 1.

E Permutation based Algorithm detail

We implement a baseline procedure for causal order discovery in Monotonic SCMs, leveraging a permutation matrix to reorder the variables while regularizing the Jacobian matrix to encourage sparsity. The approach draws upon prior work by [13]. The algorithm follows a two-step process: (1) learning a permutation matrix via the Gumbel-Sinkhorn network, and (2) enforcing sparsity on the Jacobian of the Triangular Monotonic Increasing (TMI) transformation via l_1-regularization.

Algorithm 1 Causal discovery based on Monotonic SCM.

input. Data Matrix $X \in \mathcal{R}^{n \times d}$.
output. Causal order π.
initialize. $\pi = \emptyset$, nodes = $\{1, ..., d\}$
repeat
  for $i \in nodes$ **do**
    estimate multi-cause model 4 with causes $X \backslash x_i$, and effect x_i using one-dimensional conditional normalizing flow.
    maxJac$(i) = \max \left| \nabla_{X \backslash \{x_i\}} T_i(x_i, X \backslash \{x_i\}) \right|$
  end for
  $l = \arg\min_i$ maxJac(i)
  append l to π
  $nodes \leftarrow nodes - \{l\}$
until $|\text{nodes}| = 1$
append nodes(1) to π.

Table 1: Hyperparameter Configuration for the Proposed Method

Hyperparameter	Value
Learning Rate	0.001
Batch Size	64
Number of Epochs	10
MLP Dropout Rate	0.1
MLP Hidden Layers	1
MLP Units per Layer	128
Activation Function	Leaky-ReLU
Number of flow layers	1
Number of blocks in flow	1
Number of hidden channel in flow	128
Flow base distribution	Standard Gaussian
Optimization Algorithm	Adam (weight decay=0, betas=(0.9. 0.99))

Given an input data matrix $X \in \mathcal{R}^{n \times d}$, our objective is to discover the causal order of variables by learning a permutation matrix P, which reorders X. The Gumbel-Sinkhorn network is used to generate a soft permutation matrix, which is a continuous relaxation of a discrete permutation matrix. This relaxation enables gradient-based optimization during training. The normalizing flow $T(x.P^T)$ is then trained over the permuted variables, aiming to learn a transformation that captures the causal structure of the data.

The loss function consists of two terms: (1) the log-likelihood of the normalizing flow, and (2) an l_1-norm regularization on the Jacobian of the TMI transformation. This regularization enforces sparsity, encouraging the Jacobian to reflect the triangular structure characteristic of SCMs. The loss function is formulated as follows:

$$\mathcal{L}(\theta, \beta; t, \lambda) := \log p(T_\theta^{-1}(u.P_\beta(t)^T)) + \lambda ||\nabla T_\theta(x.P_\beta(t)^T)||_1 \tag{8}$$

where θ represents the parameters of the normalizing flow, β the parameter of Gumbel-Sinkhorn network, $||\nabla T||_1$ is the l_1-norm applied to the Jacobian matrix to enforce sparsity, weighted by the coefficient λ, and $P_\beta(t)$ refers to the permutation matrix learned via the Gumbel-Sinkhorn network with temperature hyperparameter t.

Upon completion of training, the soft permutation matrix can still contain small, non-zero elements that prevent it from being a valid permutation. To address this, we apply the Hungarian algorithm [50] to convert the soft matrix into a hard permutation matrix, ensuring a valid reordering of variables. This final permutation matrix provides the estimated causal order. Algorithm 2 shows the overall procedure of the permutation-based method. To ensure a fair comparison, all shared hyperparameters with the proposed method were set identical to those listed in Table 1.

F Comparison of Causal Order Discovery

In this section, we also benchmarked our method against other causal order discovery approaches based on different types of SCMs, such as SCORE [26] and RESIT [18], which assume nonlinear additive noise, and

Algorithm 2 Causal Order Discovery in Monotonic SCMs via Permutation Matrix Learning and Jacobian Sparsity

Input: Data matrix $X \in \mathbb{R}^{n \times d}$, temperature parameter t, sparsity coefficient λ
Output: Permutation matrix P (causal order)
Initialize soft permutation matrix $P_\beta(t)$ using the Gumbel-Sinkhorn network with temperature t
repeat
    Train the normalizing flow $T_\theta(x.P_\beta(t)^\top)$
    Compute the Jacobian $\nabla T_\theta(x)$
    Compute the loss function:

$$\mathcal{L}(\theta, \beta; t, \lambda) := \log p(T_\theta^{-1}(u.P_\beta(t)^\top)) + \lambda ||\nabla T_\theta(x.P_\beta(t)^\top)||_1$$

    Update the normalizing flow parameters θ and the permutation matrix parameters β using backpropagation
until convergence of $P_\beta(t)$ and T_θ
Apply the Hungarian algorithm to convert the soft permutation matrix $P_\beta(t)$ into a hard permutation matrix
return Hard permutation matrix P (causal order)

Direct-Lingam [45], which assumes linear SCM. The comparisons were conducted using synthetic datasets containing 1000 samples across varying dimensions, including 2, 3, 4, 5, and 10. For each dimensional setting, we generated ten random datasets to ensure robustness in the evaluation. The experimental results for synthetic datasets are presented in Table 2.

For the assessment of the proposed method on real-world data, we employed dataset of protein-protein interaction networks in the human body (SACHS) [51]. This dataset includes ground truth graphs for assessment. The SACHS dataset comprises measurements of proteins in the human immune system, consisting of $d = 11$ nodes and composed of $n = 853$ data samples. This dataset's graph contains 17 edges. Additionally, the order of variables has been randomly shuffled to prevent the methods from depending on the variable order. The experimental results for SACHS dataset are presented in Table 3.

Our method achieves comparable performance to other causal order discovery approaches across various synthetic datasets, as shown in Table 2, where none of the methods demonstrate a statistically significant advantage. However, for the real-world SACHS dataset, our approach outperforms the alternatives, as seen in Table 3. One potential explanation for the challenges observed in synthetic datasets is related to difficulties in accurately estimating the Jacobian. Specifically, in cases where the root variable's Jacobian value is expected to be zero, errors in this estimation may arise, leading to high Jacobian values, even for the root variable. This issue could stem from limitations in using a single-layer normalizing flow, which may struggle to capture the underlying complexity of the data. Moving forward, increasing the complexity of the normalizing flow, or exploring alternative approaches to estimating triangular maps, could improve Jacobian estimation and lead to more accurate root detection.

Table 2: Results of different methods on various Synthetic datasets (CB metric)

Method	d=2	d=3	d=4	d=5	d=10
Direct-Lingam [45]	0.5 ± 0.53	1.0 ± 0.82	1.2 ± 0.92	1.2 ± 0.79	3.0 ± 2.0
RESIT [18]	0.2 ± 0.42	0.5 ± 0.7	0.8 ± 0.79	1.1 ± 1.1	2.7 ± 2.26
SCORE [26]	0.3 ± 0.48	0.8 ± 0.92	1.3 ± 1.49	1.6 ± 1.17	3.8 ± 1.55
Proposed	0.2 ± 0.42	1.7 ± 1.06	1.1 ± 0.74	2.0 ± 1.63	3.8 ± 1.75

Table 3: Results on Real-world dataset SACHS (CB metric)

Method	SACHS [51]
Direct-Lingam [45]	8.0
RESIT [18]	8.0
SCORE [26]	13.0
Proposed	7.0