NODAGS-Flow: Nonlinear Cyclic Causal Structure Learning

Muralikrishnna G. Sethuraman[1], Romain Lopez[2, 3], Rahul Mohan[2], Faramarz Fekri[1], Tommaso Biancalani[2], and Jan-Christian Hütter[2]

[1]School of ECE, Georgia Institute of Technology

[2]Division of Research and Early Development, Genentech

[3]Department of Genetics, Stanford University

Introduction

Problem Statement

- Causal understanding of real-world systems is crucial for prediction under unseen interventions.
- With a few notable exceptions, most **causal discovery** (CD) methods rely of the structure being a **directed acyclic graph** (DAG).
- While DAG assumption allows for regularization of search space, it is not very realistic in practice.
- Recent advances in biological assays allow for large-scale **interventions** over gene networks, enabling investigations over potential feedback loops.

Contributions

- Novel framework for CD that allows for **cycles** and flexible **nonlinear causal relations**. And is consistent without explicitly enforcing a DAG constraint, leading to efficient optimization algorithms.
- **Maximum likelihood estimation** (MLE) based graph recovery utilizing **contractive residual flows**.

Problem Setup

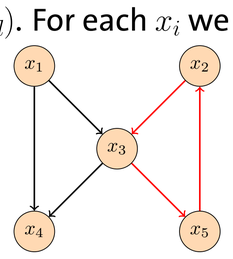
Structural Equation Model

A directed graphical (DG) model $G = (V, E)$ can be represented using a **structural equation model** (SEM), with node observations $x = (x_1, \dots, x_d)$. For each x_i we have

$$x_i = f_i(x_{\mathsf{pa}(i)}) + \varepsilon_i$$

Vectorized form

$$x = f(x) + \varepsilon.$$



Assumptions:

1. No confounders.
2. $p_E(\varepsilon)$ is known.
3. The mapping $x \mapsto \varepsilon = (\mathsf{id} - f)$ is invertible.
4. $(\mathsf{id} - f)$ and $(\mathsf{id} - f)^{-1}$ are differentiable.

Under these assumptions, the probability density of x is given by

$$p_X(x) = p_E\big((\mathsf{id} - f)(x)\big)\,|\det J_{(\mathsf{id}-f)}(x)|,$$

where $J_{(\mathsf{id}-f)}$ denotes the Jacobian matrix of $(\mathsf{id} - f)$.

Modeling Interventions

SEM:

$$x_3 = \tilde{e}$$
$$x_i = f_i(x_{\mathsf{pa}(i)}) + \varepsilon_i \quad i \neq 3;$$

Given a family of interventions $\mathcal{I}_k$ the density of x:

$$p_X(x) \propto p_E\big([(\mathsf{id} - \boldsymbol{U}_k f)(x)]_{\mathcal{U}_k}\big)\,|\det J_{(\mathsf{id}-\boldsymbol{U}_k f)}(x)|,$$

- $\mathcal{U}_k$ - set of purely observed nodes.
- $\boldsymbol{U}_k$ - diagonal matrix with 1 corresponding to $\mathcal{U}_k$ and 0 everywhere else.

Differential Causal Discovery

DCDI[2]: $G \in$ DAGs

$$\max_{\theta} \underbrace{\sum_{j=1,i=1}^{n,d} \log p_\theta\Big(x_i^j | x^j_{\mathsf{pa}(i)}\Big)}_{\text{data likelihood}} - \overbrace{\lambda\Omega(\theta)}^{\text{regularizer}} \quad \text{s.t} \quad \underbrace{\mathsf{Tr}e^{A_\theta} = d}_{\text{DAG constraint}}$$

- The DAG constraint is expensive to compute and needs augmented lagrangian method for optimization.
- The graph in consideration has to be a DAG for the likelihood to be computed correctly.

NODAGS-Flow: $G \in$ DGs (Maximum Likelihood Est.)

$$\max_{\theta,\Lambda} \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{M}'\sim\boldsymbol{M}_\phi} \mathcal{L}(\theta, \Lambda^{-1} \circ f_\theta \circ \Lambda, \boldsymbol{M}')}_{\text{data likelihood}} - \lambda \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{M}'\sim\boldsymbol{M}_\phi} [\|\boldsymbol{M}'\|_1]}_{\text{regularizer}}.$$

- Lack of DAG constraint simplifies the optimization. (No augmented lagrangians).
- True likelihood is computed at all the times.

NODAGS-Flow

Modeling Causal Mechanism

- The causal mechanism f is modeled as a **contractive neural networks** (NN), with input masks $\boldsymbol{M} \in \{0, 1\}^{d\times d}$ encoding dependencies. This guarantees invertibility.

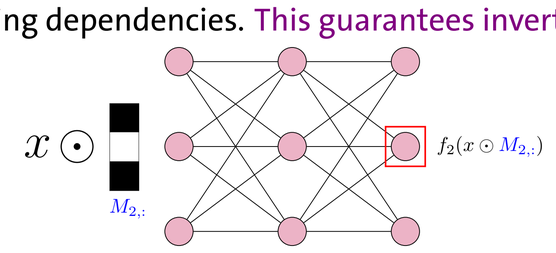


Figure 1: Neural Net

Contractive function: $f : \mathbb{R}^d \to \mathbb{R}^d$ is < 1-Lipschitz

Contractivity is enforced by rescaling NN weights.

- $x \mapsto \varepsilon = (\mathsf{id} - f)$ forms a **residual flow** network.

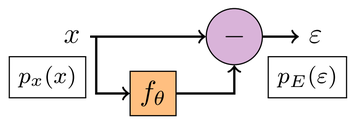


Figure 2: Residual Flow

$$p_X(x) = p_E\big((\mathsf{id} - f_\theta)(x)\big) \times |\det J_{(\mathsf{id}-f_\theta)}(x)|$$

Computing Log-det-Jacobian

- Power series expansion

$$\log|\det J_{(\mathsf{id}-f)}(x)| = -\sum_{k=1}^{\infty} \frac{1}{k} \mathsf{Tr}\big\{J_f^k(x)\big\}$$

The power series converges when f is contractive.

- Hutchinson estimator[1]: $\mathsf{Tr}\big\{J_f^k(x)\big\} = \mathbb{E}_w[w^\top J_f^k(x) w]$, where $\mathbb{E}w = 0$, and $\mathbb{E}w^2 = 1$.
- Removing bias[3]: sample $n \sim P(N)$ and re-weight.

$$\log|\det J_{(\mathsf{id}-f)}(x)| = -\mathop{\mathbb{E}}_{n,w}\left[\sum_{k=1}^{n} \frac{w^\top J_f^k(x) w}{k \cdot P(N \geq k)}\right].$$

- The data likelihood for M interventions $\{\mathcal{I}_k\}_{k=1}^M$

$$\mathcal{L}\Big(\theta, f_\theta, \boldsymbol{M}\Big) = \sum_{k=1}^{M}\sum_{i=1}^{N_k}\Bigg[\log p_{E,\theta}\Big([(\mathsf{id}-\boldsymbol{U}_k f_\theta)(x^{(k,i)})]_{\mathcal{U}_k}\Big) - \mathop{\mathbb{E}}_{n,w}\left\{\sum_{r=1}^{n} \frac{w^\top\big[J^r_{\boldsymbol{U}_k f_\theta}\big(x^{(i,k)}\big)\big]w}{r \cdot P(N \geq r)}\right\}\Bigg],$$

- Adding a preconditioning term Λ allows NODAGS-Flow to learn non-contractive DAGs.
- When SEM is linear, given $\{\mathcal{I}_k\}_{k=1}^M$, NODAGS-Flow is consistent, assymptotically, upto an interventional-equivalence class.

Experiments

Gaussian Structural Causal Models

- **Training**: 1-node interventions.
- **GT**: generated using Erdős-Rényi random graph model (20 nodes).
- **Testing**: 2-node unseen interventions.

Baseline	SEM	Graph
NOTEARS[7]	Linear	DAG
DCDI[2]	Nonlinear	DAG
GOLEM[6]	Linear	DAG
LLC[5]	Linear	Cyclic

Table 1: Baseline Characteristics

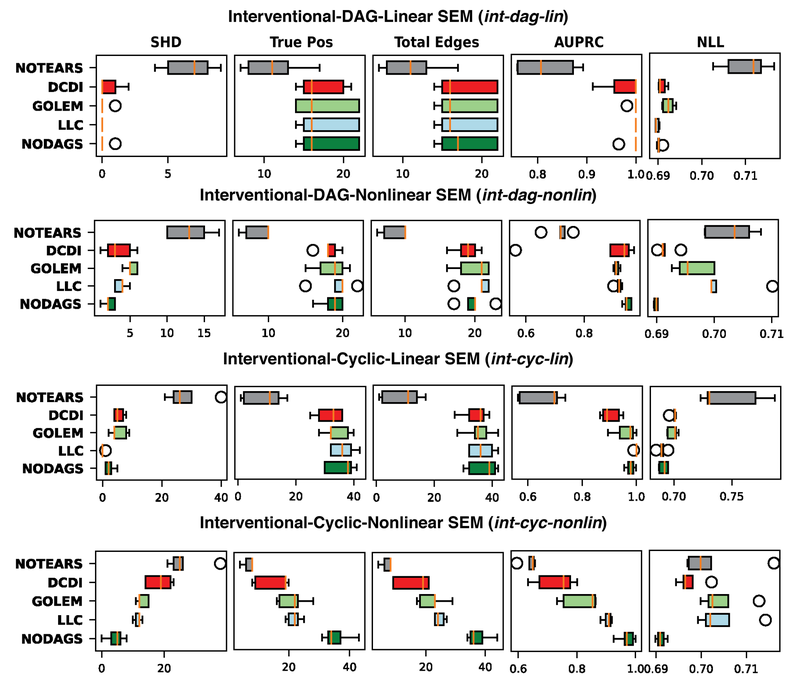


Figure 3: Results on synthetic data

NODAGS-Flow beats all the baselines on nonlinear cyclic SEMs and is competitive with the baselines on other settings.

Real World Data (Perturb-CITE-seq)

- 218,331 melanoma cells split over 3 sets.
- Subset to 61 genes.
- **Evaluation metrics**: interventional negative log-likelihood (I-NLL) and mean absolute error (I-MAE) on holdout interventions.

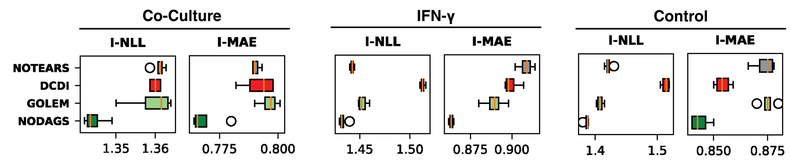


Figure 4: Results on Perturb-CITE-seq dataset [4]

NODAGS-Flow beats all the baselines on all the data sets.

References

[1] Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International Conference on Machine Learning*, pages 573–582. PMLR, 2019.

[2] Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

[3] Ricky TQ Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Chris J Frangieh, Johannes C Melms, Pratiksha I Thakore, Kathryn R Geiger-Schuller, Patricia Ho, Adrienne M Luoma, Brian Cleary, Livnat Jerby-Arnon, Shruti Malu, Michael S Cuoco, et al. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nature genetics*, 53(3):332–341, 2021.

[5] Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Learning linear cyclic causal models with latent variables. *The Journal of Machine Learning Research*, 13(1):3387–3439, 2012.

[6] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

[7] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.