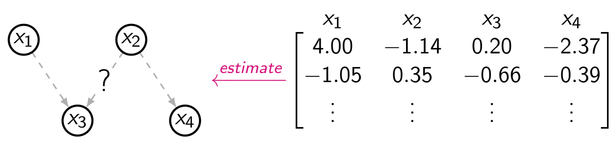
Markov Equivalence and Consistency in Differentiable Structure Learning

Chang Deng†, Kevin Bello†‡ , Pradeep Ravikumar‡ , Bryon Aragam†
† The University of Chicago ‡ Carnegie Mellon University

Background

- **Main question**: Given data $\mathbf{X}$, how to learn a DAG G that best fits the data?



This is referred to as "Causal Discovery".

- Differentiable structure learning formulates this as an optimization problem,

$$\min_{B\in\mathbb{R}^{p\times p}} s(B;\mathbf{X}) \quad \text{subject to} \quad h(B)=0. \qquad (1)$$

Constraint: $h(B)=0 \Leftrightarrow B$ is a DAG.

Issue #1: Limitations of Current Score

⭐ Score function $s(B;\mathbf{X})$ plays a crucial role in (1).

- Least squares (LS) loss (aka "reconstruction loss") has known theoretical limitations; is not compatible with traditional causal notions (faithfulness, Markov, etc.)
- ℓ_1-regularized log-likelihood as score function leads to *biased estimation*.
- ℓ_0-regularized log-likelihood is *non-differentiable*
- There is a lack of a unified score function that can:

(I) Guarantee a meaningful learned structure.
(II) Enable unbiased parameter estimation.
(III) Maintain the differentiability of optimization (1).

Can this be accomplished?

Issue #2 Scale-invariance

- LS loss is also *not scale-invariant.* i.e. re-scaling the data $\mathbf{X}$ can dramatically change the output.
- This has been used to argue that differentiable DAG learning with the LS loss is not scale-invariant.

Our Contributions

- We identify the correct score function to solve **Issue #1**.
- We show our score is **scale-invariant** to solve **Issue #2**.
- Solving (1) gives the **sparsest** DAG structure that generates the data, and all solutions belong to the same Markov Equivalence class.
- Experiments on linear, nonlinear, non-Gaussian data
- Code+implementation: github.com/Duntrain/dagrad

General Linear Gaussian Model

- Data generated by linear SEM with Gaussian Noise

$$X = B^\top X + N \qquad (2)$$

$$N\sim\mathcal{N}(0,\Omega),\quad \Omega=\operatorname{diag}(\omega_1^2,\ldots,\omega_p^2)$$

$$X\sim\mathcal{N}(0,\Sigma)\quad \Sigma=\Sigma_f(B,\Omega):=(I-B)^\top\Omega(I-B)^{-1}$$

- Model is **unidentifiable.** $\mathcal{E}(\Sigma):=\{(B,\Omega):\Sigma_f(B,\Omega)=\Sigma\}$
- The simplest structure is preferred

$$\mathcal{E}_{\min}(\Sigma):\{(B,\Omega): s_B\le s_{\tilde{B}},\forall(\tilde{B},\tilde{\Omega})\in\mathcal{E}(\Sigma),(B,\Omega)\in\mathcal{E}(\Sigma)\}$$

- How can we penalize the number of edges using differentiable penalty?

$$\text{quasi-MCP: } p_{\lambda,\delta}(t)=\lambda[(|t|-\frac{t^2}{2\delta})\mathbf{1}(|t|<\delta)+\frac{\delta}{2}\mathbf{1}(|t|>\delta)]$$

- Score function consists of **NLL+quasi-MCP**

$$s(B,\Omega;\lambda,\delta,\mathbf{X})=\ell_n(B,\Omega)+p_{\lambda,\delta}(B)$$

- Define global optimizers of optimization above

$$\mathcal{O}_{n,\lambda,\delta}=\{(B^*,\Omega^*):(B^*,\Omega^*)\text{ is a minimizer of (1)}\}$$

Theoretical Guarantees

Theorem 1 (MEC): X follows model (2). For sufficient small $\lambda,\delta>0$ (independent of n), it holds that $P(\mathcal{O}_{n,\delta,\lambda}=\mathcal{E}_{\min}(\Sigma^0))\to 1$, as $n\to\infty$. If $P(X)$ is faithful to G^0. Then, $P(\mathcal{O}_{n,\delta,\lambda}=\mathcal{M}(G^0))\to 1$, as $n\to\infty$.

Theorem 2(Scale-invariant): Let $\mathbf{Z}$ be the standardized version of $\mathbf{X}$. For all small $\lambda,\delta\ge 0$ (independent of n) $\mathcal{G}(\mathcal{O}_{n,\delta,\lambda}(\mathbf{X}))=\mathcal{G}(\mathcal{O}_{n,\delta,\lambda}(\mathbf{Z}))$ **for all n.** For all small $\lambda,\delta>0$,

$$P[\mathcal{G}(\mathcal{O}_{n,\delta,\lambda}(\mathbf{X}))=\mathcal{G}(\mathcal{O}_{n,\delta,\lambda}(\mathbf{Z}))=\mathcal{G}(\mathcal{E}_{\min}(\Sigma_f(B^0,\Omega^0)))]\to 1,\text{ as } n\to\infty$$

General Model

- Let $X\sim P(X;\psi^0,\xi^0)$, define the equivalence class and minimal equivalence class

$$\mathcal{E}(\psi^0,\xi^0)=\{(\psi,\xi):P(x;\psi,\xi)=P(x;\psi^0,\xi^0),\forall x\in\mathbb{R}^p\}$$

$$\mathcal{E}_{\min}(\psi^0,\xi^0)=\{(\psi,\xi):s_{B(\psi)}\le s_{B(\tilde{\psi})},\forall(\tilde{\psi},\tilde{\xi})\in\mathcal{E}(\psi^0,\xi^0),(\psi,\xi)\in\mathcal{E}(\psi^0,\xi^0)\}$$

- Score functions: **NLL with quasi-MCP**

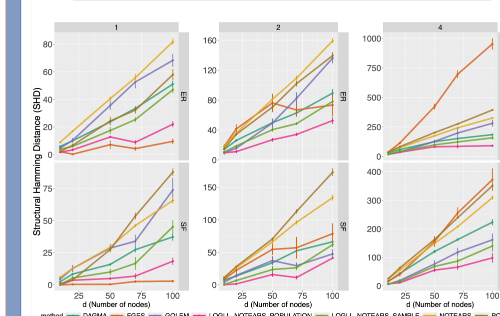
$$\min_{\psi,\xi}\ell_n(\psi,\xi)+p_{\lambda,\delta}(B(\psi))\text{ subject to } h(B(\psi))=0 \qquad (3)$$

$$\mathcal{O}_{n,\lambda,\delta}=\{(\psi^*,\xi^*):(\psi^*,\xi^*)\text{ is minimizer of (3)}\}$$

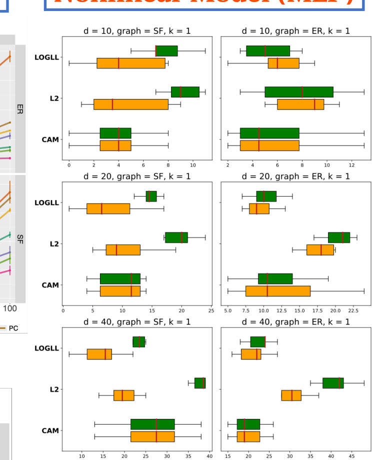
Theorem 3 (MEC): Let $X\sim P(X;\psi^0,\xi^0)$, under certain assumptions. For sufficient small $\lambda,\delta>0$ (independent of n), then $P(\mathcal{O}_{n,\delta,\lambda}=\mathcal{E}_{\min}(\psi^0,\xi^0))\to 1$, as $n\to\infty$. If additionally $P(X)$ is faithful to G^0, then $P(\mathcal{O}_{n,\delta,\lambda}=\mathcal{M}(G^0))\to 1$, as $n\to\infty$.

Experiments

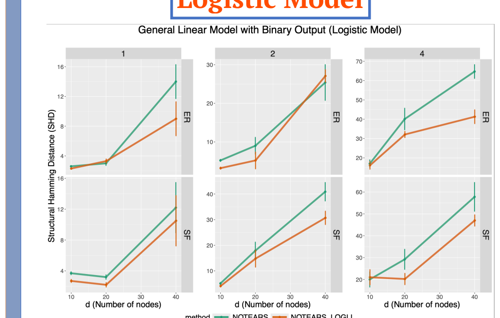
General Linear Gaussian Model



Nonlinear Model (MLP)



Logistic Model



Code: More in Paper! Paper: