CoLiDE: Concomitant Linear DAG Estimation

Seyed Saman Saboksayr*, Gonzalo Mateos*, and Mariano Tepper†
*Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA
†Intel Labs, Hillsboro, OR, USA
{ssaboksa, gmateosb}@ur.rochester.edu, and mariano.tepper@intel.com

Overview of this work

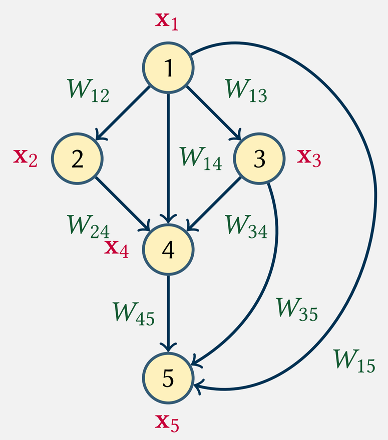
- **Goal:** Learning DAG structure from observational data
- Recent approaches employ lasso-type score functions to guide this search
  - ⇒ Needs parameter retuning if the *unknown* exogenous noise variances change
  - ⇒ Implicitly rely on limiting assumptions of equal noise variances
- **Contribution:** New convex score function for learning of linear DAGs
  - ⇒ Incorporates concomitant estimation of scale parameters
  - ⇒ Minimum (or no) recalibration effort across diverse problem instances
  - ⇒ Superior performance in tests with simulated and real-world data

What are DAGs and how to learn their connectivity structure?

- Directed graph $\mathcal{G}$ without cycles increasingly prominent in ML applications
  - ⇒ May encode causal relationships within complex systems
  - ⇒ Employ directed edges to link causes and their immediate effects
- Causal structure underlying a group of variables is often unknown
  - ⇒ Need to address the task of inferring DAGs from observational data
- A Markovian linear structural equation model (SEM) consists of

$$\mathbf{x}_i = \mathbf{w}_i^\top \mathbf{X} + \mathbf{z}_i, \quad \text{where} \quad \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_d] \in \mathbb{R}^{n \times d}$$

  - ⇒ DAG adjacency matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_d] \in \mathbb{R}^{d \times d}$ collects the edge weights
  - ⇒ $\mathbf{z}_i \in \mathbb{R}^n$ is a vector of mutually independent, exogenous noises
  - ⇒ Ex: $\mathbf{x}_4 = \mathbf{w}_4^\top \mathbf{X} + \mathbf{z}_4 = W_{14}\mathbf{x}_1 + W_{24}\mathbf{x}_2 + W_{34}\mathbf{x}_3 + \mathbf{z}_4$



Problem statement: Given data $\mathbf{X}$ adhering to a linear SEM, learn the latent DAG $\mathcal{G} \in \mathbb{D}$, i.e., estimate its adjacency matrix $\mathbf{W}$ by minimizing the score function $\mathcal{S}$, namely

$$\min_{\mathbf{W}} \ \mathcal{S}(\mathbf{W}) \ \text{ subject to } \ \mathcal{G}(\mathbf{W}) \in \mathbb{D}$$

- Learning a DAG solely from observational data $\mathbf{X}$ is in general NP-hard
  - ⇒ Combinatorial acyclicity constraint $\mathcal{G}(\mathbf{W}) \in \mathbb{D}$ is difficult to enforce
  - ⇒ Multiple DAGs can generate the same observational distribution

Continuous optimization approach to DAG structure learning

- Acyclicity characterization using nonconvex, smooth $\mathcal{H}(\mathbf{W}) : \mathbb{R}^{d \times d} \mapsto \mathbb{R}$
  - ⇒ Relax combinatorial constraint by enforcing $\mathcal{H}(\mathbf{W}) = 0 \iff \mathcal{G}(\mathbf{W}) \in \mathbb{D}$
- Pioneering NOTEARS formulation adopts $\mathcal{H}_{\text{expm}}(\mathbf{W}) = \operatorname{Tr}(e^{\mathbf{W} \circ \mathbf{W}}) - d$
  - ⇒ Diagonal entries of powers of $\mathbf{W} \circ \mathbf{W}$ encode information about cycles
- Solve the smooth, continuous optimization problem

$$\min_{\mathbf{W}} \ \mathcal{S}(\mathbf{W}) \ \text{ subject to } \ \mathcal{H}(\mathbf{W}) = 0$$

- Q: What is a proper score function to guide the search?

Score functions and their limitations

Regression-based

- Ordinary LS loss augmented with an ℓ_1-norm regularizer

$$\mathcal{S}(\mathbf{W}) = \frac{1}{2n}\|\mathbf{X} - \mathbf{W}^\top \mathbf{X}\|_F^2 + \lambda \|\mathbf{W}\|_1$$

  - ⇒ Computational efficiency, robustness, and even consistency
- Similar to multi-task lasso, when the response and design matrices coincide
  - ⇒ Optimal rates for lasso hinge on selecting $\lambda \asymp \sigma\sqrt{\log d / n}$, but σ^2 is unknown
- Requires retuning λ, implicitly assumes equal noise variances

Likelihood-based

- Desirable statistical properties, amenable to different exogenous noise variances
  - ⇒ Requires retuning sparsity parameter, prior knowledge on noise distribution
  - ⇒ Gaussian profile log-likelihood (GOLEM) is not decomposable

Concomitant linear DAG estimation (CoLiDE)

CoLiDE-EV

- All exogenous variables $\mathbf{z}_1, \ldots, \mathbf{z}_d$ in the linear SEM have equal variance (EV) σ^2
- Inspired by the smoothed concomitant lasso

$$\min_{\mathbf{W}, \sigma \geq \sigma_0} \underbrace{\frac{1}{2n\sigma}\|\mathbf{X} - \mathbf{W}^\top \mathbf{X}\|_F^2 + \frac{d\sigma}{2} + \lambda \|\mathbf{W}\|_1}_{:= \mathcal{S}(\mathbf{W}, \sigma)} \quad \text{subject to} \quad \mathcal{H}(\mathbf{W}) = 0$$

- Score $\mathcal{S}(\mathbf{W}, \sigma)$ is jointly convex, $(d\sigma)/2$ for consistency under Gaussianity
  - ⇒ λ decouples from σ as minimax optimality now requires $\lambda \asymp \sqrt{\log d / n}$
- Solve a sequence of unconstrained problems where $\mathcal{H}$ is viewed as a regularizer
  - ⇒ Acyclicity function $\mathcal{H}_{\text{ldet}}(\mathbf{W}, s) = d \log(s) - \log(\det(s\mathbf{I} - \mathbf{W} \circ \mathbf{W}))$
- **Optimization:** Given a decreasing sequence $\mu_k \to 0$, at step k we solve

$$\min_{\mathbf{W}, \sigma \geq \sigma_0} \mu_k \left[\frac{1}{2n\sigma}\|\mathbf{X} - \mathbf{W}^\top \mathbf{X}\|_F^2 + \frac{d\sigma}{2} + \lambda \|\mathbf{W}\|_1 \right] + \mathcal{H}_{\text{ldet}}(\mathbf{W}, s_k)$$

  - ⇒ Limit $\mu_k \to 0$ is guaranteed to yield a DAG
  - ⇒ Jointly estimates the noise level σ and the adjacency matrix $\mathbf{W}$ for each μ_k
- Fixing σ to its latest value and minimizing score function inexactly w.r.t. $\mathbf{W}$
  - ⇒ One iteration of gradient descent via ADAM optimizer
- Updating σ in closed form given the latest $\mathbf{W}$ via $\hat{\sigma} = \max\left(\frac{1}{\sqrt{nd}}\|\mathbf{X} - \mathbf{W}^\top \mathbf{X}\|_F, \sigma_0\right)$

CoLiDE-NV

- Noise variables $\mathbf{z}_1, \ldots, \mathbf{z}_d$ have non-equal variances (NV) $\sigma_1^2, \ldots, \sigma_d^2$
- Mimicking the previous optimization approach, we propose CoLiDE-NV

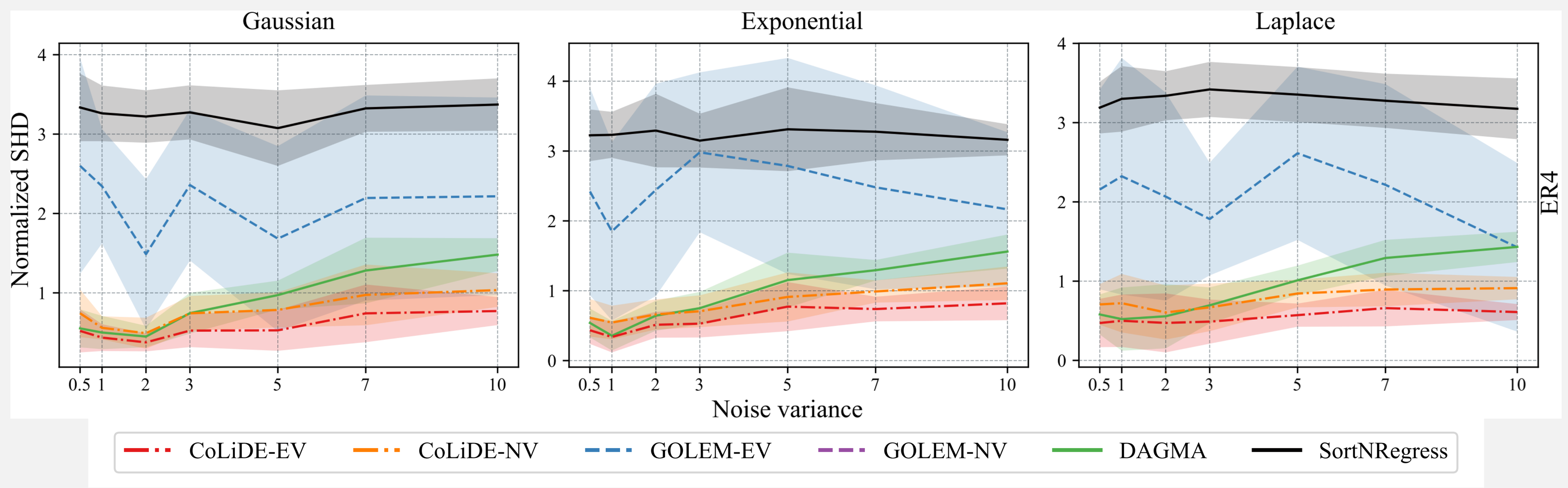
$$\min_{\mathbf{W}, \Sigma \geq \Sigma_0} \mu_k \left[\frac{1}{2n} \operatorname{Tr}\left((\mathbf{X} - \mathbf{W}^\top \mathbf{X})^\top \Sigma^{-1} (\mathbf{X} - \mathbf{W}^\top \mathbf{X}) \right) + \frac{1}{2}\operatorname{Tr}(\Sigma) + \lambda \|\mathbf{W}\|_1 \right] + \mathcal{H}_{\text{ldet}}(\mathbf{W}, s_k)$$

  - ⇒ $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_d)$ is a diagonal matrix of noise standard deviations
- Per iteration cost is $\mathcal{O}(d^3)$, on par with state-of-the-art DAG learning methods

Results

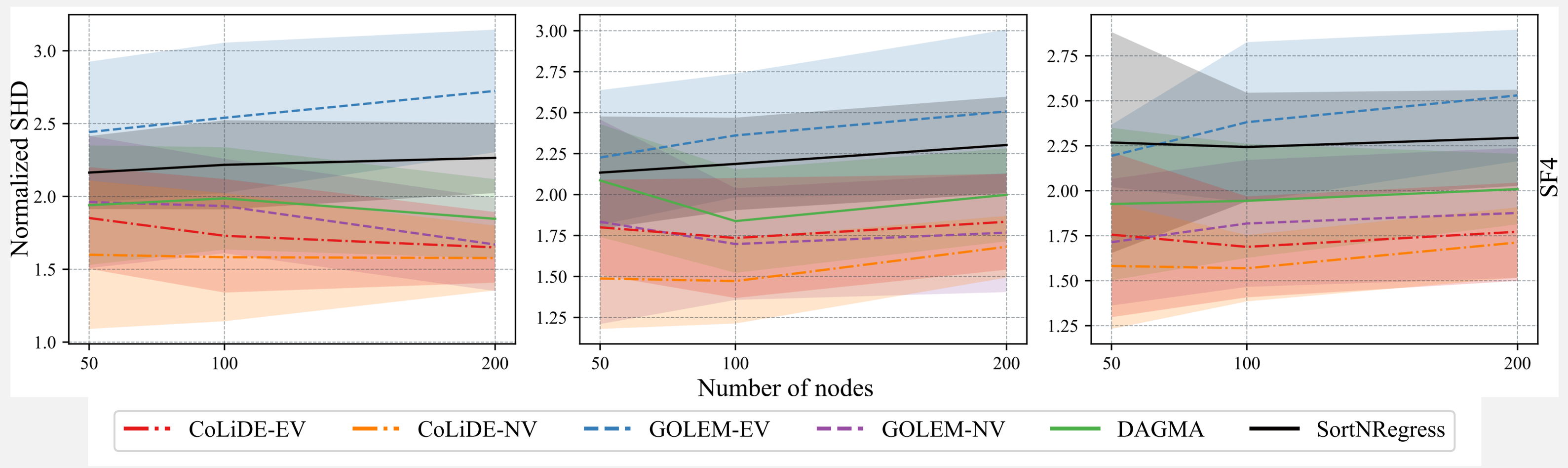
Equal variance experiments

- Impact of noise levels varying from 0.5 to 10 on DAG recovery performance
- 200-node ER graphs with weighted edges $\mathcal{E} \in [-2, -0.5] \cup [0.5, 2]$
- Simulate $n = 1000$ samples considering diverse noise distributions via linear SEM
- SHD counts number of edge corrections required to reach true graph



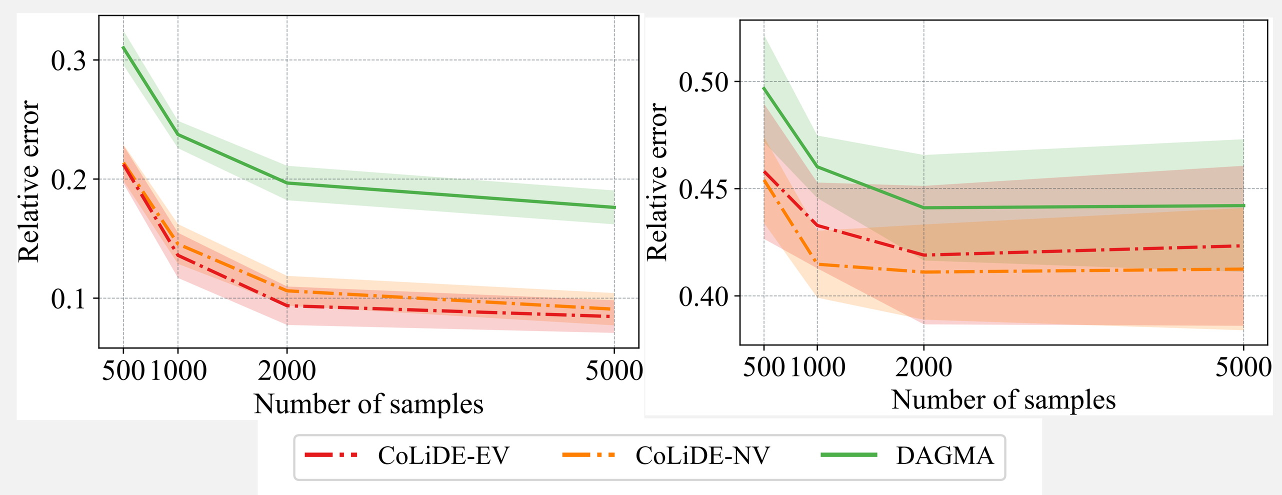
Non-equal variance experiments

- Noise variance of each node is uniformly drawn from $[0.5, 10]$
- SF graphs with weighted edges $\mathcal{E} \in [-1, -0.25] \cup [0.25, 1]$



Noise estimation experiments

- Methods that do not explicitly estimate the noise, we use $\hat{\sigma}_i^2 = \frac{1}{n}\|x_i - \hat{\mathbf{w}}_i^\top \mathbf{x}\|_2^2$
- 200-node ER; simulate Linear SEM with Gaussian noise; EV (left) and NV (right)



Sachs dataset

	GOLEM-EV	GOLEM-NV	DAGMA	SortNRegress	DAGuerreotype	GES	CoLiDE-EV	CoLiDE-NV
SHD	22	15	16	13	14	13	13	12
SID	49	58	52	47	50	56	47	46
SHD-C	19	11	15	13	12	11	13	14
FDR	0.83	0.66	0.5	0.61	0.57	0.5	0.54	0.53
TPR	0.11	0.11	0.05	0.29	0.17	0.23	0.29	0.35

References and GitHub page

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with no tears: Continuous optimization for structure learning. In *Proc. Adv. Neural. Inf. Process. Syst.*, 2018.
Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization. In *Proc. Adv. Neural. Inf. Process. Syst.*, 2022.
Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. In *Proc. Adv. Neural. Inf. Process. Syst.*, 2020.