Overview of this work

- Goal: Learning DAG structure from observational data
- Recent approaches employ lasso-type score functions to guide this search
  - Needs parameter returning 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 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
  \[ x_i = w_{iX} + z_i, \quad \text{where } X = [x_1, \ldots, x_d] \in \mathbb{R}^{d \times 1}, \]
  - DAG adjacency matrix \( W = [w_{ij}] \in \mathbb{R}^{d \times d} \) collects the edge weights
  - \( x_i \in \mathbb{R}^d \) is a vector of mutually independent, exogenous noises
  - \( z_i = w_{iX} = w_{i1}x_1 + w_{i2}x_2 + w_{i3}x_3 + z_i \)

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

\[ \min_{W} S(W) \quad \text{subject to } \quad Q(W) \in \mathcal{D} \]  

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

Continuous optimization approach to DAG structure learning

- Acyclicity characterization using non-smooth, smooth \( h(W) : \mathbb{R}^{d \times d} \rightarrow \mathbb{R} \)
  - Relax combinatorial constraint by enforcing \( h(W) = 0 \) \( \Leftrightarrow \ Q(W) \in \mathcal{D} \)
- Pioneering NOTEARS formulation adopts \( h_{\text{NOTEARS}}(W) = \text{Tr}(W^3) - d \)
- Diagonal entries of powers of \( W \) in \( \text{encode information about cycles} \)
- Solve the smooth, continuous optimization problem
  \[ \min_{W} S(W) \quad \text{subject to } \quad h(W) = 0 \]

Continuous Linear DAG Estimation (CoLiDE)

- All exogenous variables \( z_1, \ldots, z_d \) in the linear SEM have equal variance (EV) \( \sigma^2 \)
- Inspired by the smoothing continuous lasso
  \[ \min_{W \in \mathbb{R}^{d \times d}} \frac{1}{2n} \|X - W'X\|_2^2 + \frac{d}{2} \|W\|_1 \quad \text{subject to } \quad h(W) = 0 \]

- Score \( S(W, \sigma) \) is jointly convex, \( (d/2) \) for consistency under Gaussianity
- \( \lambda \) decouples from \( \sigma \) as minimax optimality now requires \( \lambda = \sqrt{\log d/n} \)
- Solve a sequence of unconstrained problems where \( H(W) \) is viewed as a regularizer
  - Acyclicity function \( h_{\text{NOTEARS}}(W) = \text{Tr}(W^3) - d \)
  - Optimization: Given a decreasing sequence \( \mu_k \rightarrow 0 \), at step \( k \) we solve
    \[ \min_{W \in \mathbb{R}^{d \times d}} \mu_k \frac{1}{2n} \|X - W'X\|_2^2 + \frac{d}{2} \|W\|_1 + H_{\text{NOTEARS}}(W, \mu_k) \]
  - Limit \( \mu_k \rightarrow 0 \) is guaranteed to yield a DAG
  - Jointly estimates the noise level \( \sigma \) and the adjacency matrix \( W \) for each \( \mu_k \)
  - Fixing \( \sigma \) to its latest value and minimizing score function inexactly w.r.t. \( W \)
  - One iteration of gradient descent via ADAM optimizer
- Updating \( \sigma \) in closed form given the latest \( W \) via function \( \sigma = \frac{1}{(\sqrt{d} + 1)^2} \|X - W'X\|_2^2 / (d/2) \)

CoLiDE-NV

- Noise variables \( z_1, \ldots, z_d \) have non-equal variances (NV) \( \sigma_1^2, \ldots, \sigma_d^2 \)
- Mimicking the previous optimization approach, we propose CoLiDE-NV
  \[ \min_{W \in \mathbb{R}^{d \times d}} \frac{1}{2n} \text{Tr}((X - W'X)'(X - W'X)) + \frac{2}{n} \text{Tr}(\Sigma) + \frac{d}{2} \|W\|_1 + H_{\text{NOTEARS}}(W, \mu_k) \]
  - \( \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2) \) is a diagonal matrix of noise standard deviations
  - Per iteration cost is \( O(d^2) \), on par with state-of-the-art DAG learning methods

Score functions and their limitations

Regression-based

- Ordinary LS loss augmented with an \( l_2 \)-norm regularizer
  \[ S(W) = \frac{1}{2n} \|X - W'X\|_2^2 + \frac{d}{2} \|W\|_1 \]
  - 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 \( \lambda = \sigma \sqrt{2 \log d/n} \), but \( \sigma^2 \) is unknown
- Requires returning \( \lambda \), implicitly assumes equal noise variances

Likelihood-based

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

CoLiDE-EV

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