LLM-Initialized Differentiable Causal Discovery

Shiv Kampani **David Hidary** **Constantijn van der Poel**
Martin Ganahl **Brenda Miao**
SandboxAQ

Causal Discovery

Causal discovery is the NP-hard problem of learning a causal graphical model (CGM) from a set of observed data points. A CGM is a directed acyclic graph whose edges indicate causal relationships between variables or nodes. It is formulated as the following combinatorial optimization problem:

$$\max_{G \in \text{DAG}} \quad \mathcal{L}(G, \theta; X) = \frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{d} \log p(v_j = x_j^n \,;\, X, \theta, G)$$

Differentiable causal discovery (DCD) methods solve a relaxation of the problem above, which is amenable to continuous optimization techniques like gradient-descent. The relaxation is as follows ($D \gg d, h$ is a DAG-penalty):

$$\max_{\theta \in \mathbb{R}^D} \quad \mathcal{L}(A_\theta, \theta; X) = \frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{d} \log p(v_j = x_j^n \,;\, X, \theta, A_\theta) - \beta h(A_\theta)$$

Separately, LLMs have shown promise in being able to evaluate pairwise causal relationships between variables of interest, based on their parametric "background knowledge".

Key Contributions

- LLM-DCD combines LLMs with DCD, **leveraging prior knowledge** from LLMs while retaining DCD's performance and computational efficiency for causal discovery tasks.
- Achieves **state-of-the-art performance** across multiple causal discovery benchmark datasets.
- Optimizes an explicitly defined adjacency matrix, enhancing **interpretability** in causal discovery, in contrast with previous non-interpretable methods.
- LLM-DCD benefits from high-quality adjacency matrix initialization and future advancements in LLM quality, capabilities, and reasoning.

LLM-DCD Method

We modify the DCD program and introduce an ansatz function $p(v_j = x_j^n \,;\, X, A)$ (`MLE-INTERP`), which depends on elements $a_{jk} \geq 0$ of the explicitly defined adjacency matrix A as the only variational parameters.

$$\max_{A \in \mathbb{R}^{d \times d}} \quad \mathcal{L}(A; X) = \frac{1}{n} \sum_{n=1}^{N} \sum_{j=1}^{d} \log \texttt{MLE-INTERP}(x_j^n; X, A) - \alpha ||A||_1 - \beta_t |\lambda_d|$$

For 2 variables, `MLE-INTERP` is easily computed from ratios of frequency counts of observations in the training data, as follows:

$$\texttt{MLE-INTERP}(x_1; X, A) = \frac{\text{cnt}(x_1)(1 - a_{21}) + \text{cnt}(x_1, x_2) a_{21}}{N(1 - a_{21}) + \text{cnt}(x_2) a_{21}}.$$

For ≥ 2 variables, `MLE-INTERP` is the generalization of the expression above. We present the following algorithm for computing this function.

Algorithm. Computing $\texttt{MLE-INTERP}(\mathbf{x}_j^i \,;\, \mathbf{x}^i, W)$

Input: i, j, X, A
$\texttt{num}, \texttt{den} \leftarrow 0$
for $k \in [n]$ **do**
  $\texttt{numprod}, \texttt{denprod} \leftarrow 1$
  for $m \in [d]$ **do**
    $\texttt{numprod} \leftarrow \texttt{numprod} \cdot \big((1 \text{ if } \mathbf{x}_m^k = \mathbf{x}_m^i \text{ else } g(1 - a_{mj})) \text{ if } \mathbf{x}_j^k = \mathbf{x}_j^i \text{ else } 0\big)$
    $\texttt{denprod} \leftarrow \texttt{denprod} \cdot (1 \text{ if } \mathbf{x}_m^k = \mathbf{x}_m^i \text{ or } m = j \text{ else } g(1 - a_{mj}))$
  end for
  $\texttt{num} \leftarrow \texttt{num} + \texttt{numprod}, \ \texttt{den} \leftarrow \texttt{den} + \texttt{denprod}$
end for
return $(\texttt{num} \ / \ \texttt{den})$

The LLM-DCD program is solved by a gradient-ascent method (with penalty) using an out-of-the-box Adam optimizer with default hyperparameters.

Results

LLM-DCD (BFS) outperformed all baseline SBM, DCD, and LLM-based approaches on large-sized datasets, and achieved results that were comparable to the top-performing models on the small and medium-sized datasets.

We further show that initialization of the adjacency matrix in LLM-DCD affects performance, with higher quality initializations tending to result to better performance across metrics and datasets.

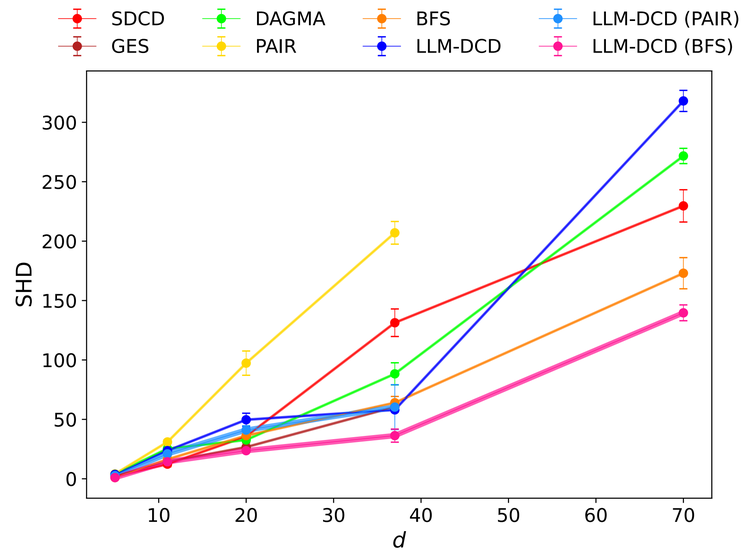

Figure 1. We report performance of LLM-DCD and other models using structural Hamming distance (SHD) between the predicted and true CGMs. Lower SHD indicates better performance.

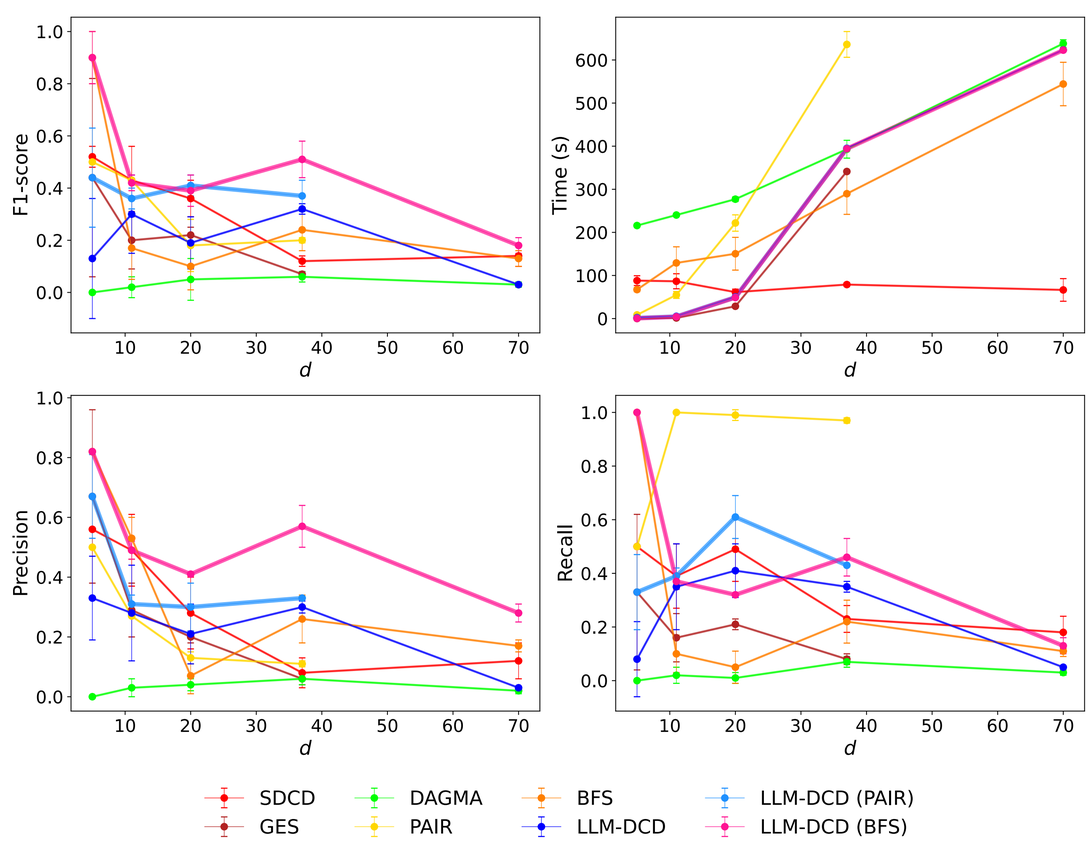

Figure 2. F1-score, precision, recall, and runtime on observational datasets of different sizes (d). Results are not reported for methods with intractable runtimes on large-sized datasets.

Limitations

- **Time complexity**: Less efficient than the recently developed SDCD but future implementations may improve on this using techniques inspired from SDCD.
- **Limited real-world scope**: All methods were tested on benchmarking datasets; our work lacks evaluation on real-world datasets.

Future Directions

- Explore how LLM size and reasoning capabilities influence adjacency matrix initialization and LLM-DCD performance.
- Apply LLM-DCD to real-world datasets and fields including drug discovery, epidemiology, genetics, and economics.
- Improve the time-complexity of LLM-DCD to that of SDCD.

SANDBOXAQ

PDF