DAG-GNN: DAG Structure Learning with Graph Neural Networks

Yue Yu[1], Jie Chen[2,3], Tian Gao[3], Mo Yu[3]

[1]Department of Mathematics, Lehigh University, USA
[2]MIT-IBM Watson AI Lab, USA
[3]IBM Research, USA

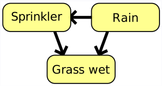
ICML 2019
June 13th, 2019

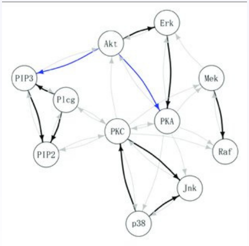
Motivation

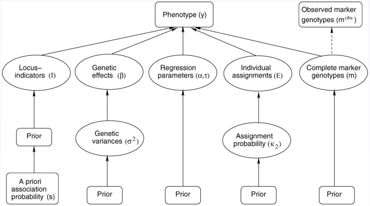
The DAG learning problem is a vital part in causal inference:

- Let $A \in \mathbb{R}^{m \times m}$ be the unknown weighted adjacency matrix of a DAG with m nodes.
- Given n identically distributed (i.i.d.) samples $X^k \in \mathbb{R}^{m \times d}$, from a distribution corresponding to A.
- Our focus is to recovery the directed acyclic graph (DAG) A from $X = \{X^1, \cdots, X^n\}$.

However, DAG learning is proven to be NP-hard.







Motivation

Conventional DAG learning methods:

- Perform score-and-search for discrete variables: with a constraint stating that the graph must be acyclic.
- Make a parametric (e.g. Gaussian) assumption for continuous variables: may result in model misspecification.

An equivalent acyclicity constraint was proposed by Zheng et al[1] **(NOTEARS)** for linear Structural Equation Model (SEM), by imposing a continuous penalty function

$$h(A) = \mathrm{tr}(\exp(A \circ A)) - m.$$

We followed the framework of [1] to **formulate the problem as a continuous optimization**, with the following major contributions:

1. We developed **a deep generative model (VAE) parameterized by a novel graph neural network architecture (DAG-GNN)**.
2. We proposed an **alternative constraint** $h(A)$.
3. The model is capable to capture **complex distributions** of data and to sample from them, and **naturally handles various data types**.

[1] Zheng, X., Aragam, B., Ravikumar, P. K., & Xing, E. P. (2018). DAGs with NO TEARS: Continuous Optimization for Structure Learning. In Advances in Neural Information Processing Systems (pp. 9472-9483).

Model Learning with Variational Autoencoder (VAE)

Our method learns the weighted adjacency matrix A of a DAG by using a deep generative model through maximizing the evidence lower bound (ELBO)

$$L_{\text{ELBO}} = \frac{1}{n}\sum_{k=1}^{n} L_{\text{ELBO}}^{k},$$

$$L_{\text{ELBO}}^{k} \equiv -D_{\text{KL}}\Big(q(Z|X^k)\,||\,p(Z)\Big) + E_{q(Z|X^k)}\Big[\log p(X^k|Z)\Big].$$

The ELBO lends itself to a VAE: given X^k, the encoder (inference model) encodes it into a latent variable Z with density $q(Z|X^k)$; and the decoder (generative model) reconstructs X^k from Z with density $p(X^k|Z)$.

Inspired by the linear SEM model

$$X = A^T X + Z, \text{ or, equivalently, } X = (I - A^T)^{-1} Z,$$

we propose a new graph neural network architecture for the decoder

$$\hat{X} = f_2((I - A^T)^{-1} f_1(Z)),$$

and the corresponding encoder

$$Z = f_4((I - A^T) f_3(X)).$$

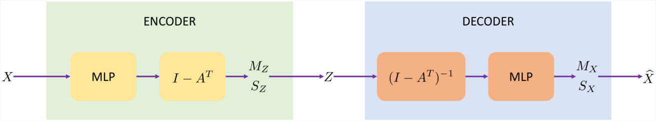
Graph Neural Network (GNN) Architecture

For the inference model (encoder) $Z = f_4((I - A^T)f_3(X))$: we let f_3 be a multilayer perceptron (MLP) and f_4 be the identity mapping. Then the variational posterior $q(Z|X)$ is a factored Gaussian with mean M_Z and standard deviation S_Z:

$$[M_Z|\log S_Z] = (I - A^T)\mathrm{MLP}(X, W^1, W^2) := (I - A^T)\mathrm{ReLU}(XW^1)W^2.$$

For the generative model (decoder) $\hat{X} = f_2((I - A^T)^{-1}f_1(Z))$: we let f_1 be the identity mapping and f_2 be an MLP. Then the likelihood $p(X|Z)$ is a factored Gaussian with mean M_X and standard deviation S_X:

$$[M_X|\log S_X] = \mathrm{MLP}((I - A^T)^{-1}Z, W^3, W^4) := \mathrm{ReLU}((I - A^T)^{-1}ZW^3)W^4.$$



A Robust Acyclicity Constraint

To further guarantee that the learnt A is a acyclic, we propose an (alternative) equality constraint when maximizing the ELBO.

Theorem: Let $A \in \mathbb{R}^{m \times m}$ be the (possibly negatively) weighted adjacency matrix of a directed graph. For any $\alpha > 0$, the graph is acyclic if and only if

$$h(A) = \mathrm{tr}[(I + \alpha A \circ A)^m] - m = 0.$$

Here α may be treated as a hyperparameter.

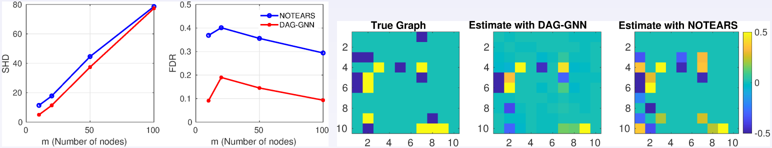
When the eigenvalues of $A \circ A$ have a large magnitude, by taking sufficiently small constant α, $(I + \alpha A \circ A)^m$ is more stable than $\exp(A \circ A)$:

Theorem: Let $\alpha = c/m > 0$ for some c. Then for any complex λ, we have $(1 + \alpha|\lambda|)^m \leq e^{c|\lambda|}$.

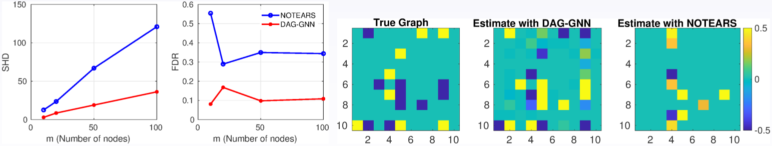
In practice, α **depends on** m **and an estimation of the largest eigenvalue of** $A \circ A$ in magnitude.

Nonlinear and vector value datasets

- **Nonlinear synthetic data**: generated by $X = A^T \cos(X + \mathbf{1}) + Z$:



- **Vector value data** $X^k \in \mathbb{R}^{m\times d}$, $d > 1$: generated by $\tilde{x} = A^T\tilde{x} + \tilde{z}$, $x^k = u^k\tilde{x} + v^k + z^k$ and $X = [x^1|x^2|\cdots|x^d]$:



Discrete value datasets

The proposed model naturally handles **discrete variables**. Assuming that each variable has a finite support of cardinality d, let $p(X|Z)$ be a factored categorical distribution with probability matrix P_X, one embedding layer is added to the encoder and the decoder is modified as:

$$P_X = \mathrm{softmax}(\mathrm{MLP}((I - A^T)^{-1}Z, W^3, W^4)).$$

The solver is compared with the state-of-the-art exact DAG solver GOPNILP[2] on 3 benchmark datasets:

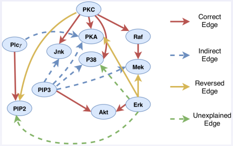
Dataset	m	Groundtruth	GOPNILP	DAG-GNN
Child	20	-1.27e+4	-1.27e+4	-1.38e+4
Alarm	37	-1.07e+4	-1.12e+4	-1.28e+4
Pigs	441	-3.48e+5	-3.50e+5	-3.69e+5

Table : BIC scores on benchmark datasets of discrete variables.

[2] Cussens, J., Haws, D., & Studeny, M. (2017). Polyhedral aspects of score equivalence in Bayesian network structure learning. Mathematical Programming, 164(1-2), 285-324.

Applied to a **bioinformatics dataset**[3] for the discovery of a protein signaling network:

Method	SHD	# Predicted edges
FGS	22	17
NOTEARS	22	16
DAG-GNN	19	18



Applied to a **knowledge base (KB) schema dataset**[4]. The nodes of which are relations and the edges indicate whether one relation suggests another.

film/ProducedBy	⇒	film/Country
film/ProductionCompanies	⇒	film/Country
person/Nationality	⇒	person/Languages
person/PlaceOfBirth	⇒	person/Languages
person/PlaceOfBirth	⇒	person/Nationality
person/PlaceLivedLocation	⇒	person/Nationality

[3] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. Science, 308(5721), 523-529.

[4] Toutanova, K., Chen, D., Pantel, P., Poon, H., Choudhury, P., & Gamon, M. (2015). Representing text for joint embedding of text and knowledge bases. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 1499-1509).

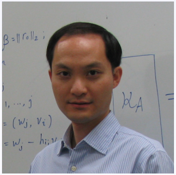
Thank you for your attention.

The code is available at `https://github.com/fishmoon1234/DAG-GNN`.

For further details and questions, please come to our poster session:
This evening 06:30 – 09:00 PM, Pacific Ballroom #215.

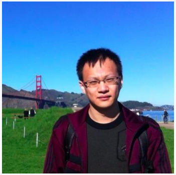
Acknowledgement

- **Collaborators:**



Jie Chen

Tian Gao



Mo Yu

- **Funding support:**
NSF CAREER award DMS1753031, Lehigh FRG program.