Testing Directed Acyclic Graph via Structural, Supervised and Generative Adversarial Learning

Chengchun Shi [1] and Yunzhe Zhou [2] and Lexin Li [2]

[1]London School of Economics and Political Science

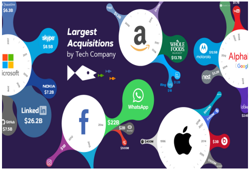
[2]University of California at Berkeley

In this talk, we will focus on...

- **Directed acyclic graph** (DAG) is an important tool to characterize pairwise directional relations.

(a) Genetics

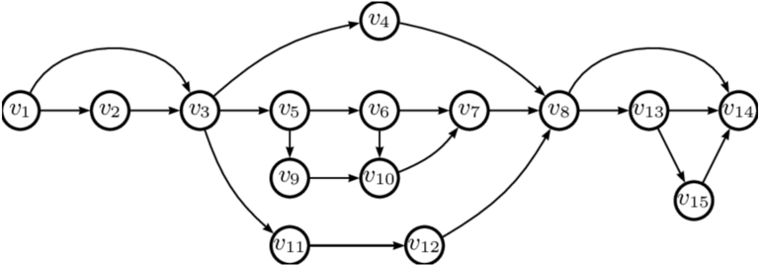
(b) Neuroscience



(c) Tech Company

- It leads to **causal interpretations** when the no unmeasured confounders assumption holds.

Example



—Taken from Brankovic et al. (2015)

- Edges are unidirectional
- No directed cycles

Existing literature on DAG estimation

- Challenge: the DAG constraint
- Statistics
  - PC algorithm (Spirtes et al., 2000)
  - ℓ_0 penalization (van de Geer & Bühlmann, 2013)
  - Surrogate constraint (Yuan et al., 2019)
- Computer science
  - Continuous optimization (Zheng et al., 2018)
  - Variational autoencoder (Yu et al., 2019)
  - Neural networks (Zheng et al., 2020)
  - Reinforcement learning (Zhu et al., 2020)

Existing literature on DAG inference

- **DAG inference** (e.g., hypothesis testing) has been less explored.
- Some existing work focused on **linear DAGs**.
  - De-biased inference (Jankovà and van de Geer, 2019)
  - Constrained likelihood ratio test (Li, et al., 2020)
- Objective: develop inference methods for **general DAGs** in **high-dimensions**.

Our proposal: SUGAR

- Challenge: nonlinearity & high-dimensionality
- Proposal: employ modern machine learning techniques (e.g., deep neural networks)
  - DAG **S**tructure learning based on neural networks or reinforcement learning
  - s**U**pervised learning based on neural networks
  - Distributional generator based on **G**enerative **A**dve**R**sarial networks (GANs, Goodfellow et al., 2014)

Problem formulation

- DAG model: **Additive noise model** (Peters et al., 2014)

$$X_j = f_j(X_{\mathrm{PA}_j}) + e_j,$$

  where X_j denotes the jth node in the DAG.
  Ensures **identifiability** under mild conditions.

- **Testing hypotheses**:

$$\mathcal{H}_0(j,k) : k \notin \mathrm{PA}_j \quad vs \quad \mathcal{H}_1(j,k) : k \in \mathrm{PA}_j.$$

- **Data**: $\{\mathbb{X}_{i,t,j}\}_{i,t,j}$
  - i indexes the subject;
  - t indexes the time point;
  - j indexes the node.

Main idea

A key quantity $I(j,k|\mathcal{M},h)$ defined as

$$E\{X_j - E(X_j|X_{\mathcal{M}-\{k\}})\}[h(X_k, X_{\mathcal{M}-\{k\}}) - E\{h(X_k, X_{\mathcal{M}-\{k\}})|X_{\mathcal{M}-\{k\}}\}].$$

Theorem

Under certain assumptions, $\mathcal{H}_0(j,k)$ holds if and only if there exists some $\mathcal{M}$ such that $j \notin \mathcal{M}$, $PA_j \in \mathcal{M}$, $\mathcal{M} \cap DS_j = \emptyset$,

$$I(j,k|\mathcal{M},h) = 0, \quad \forall h.$$

Main idea (Cont'd)

- Test statistic
  - Construct a series of measures $\{I(j, k|\mathcal{M}, h_b) : 1 \leq b \leq B\}$
  - Standardize these measures and take the maximum
- The main algorithm
  - The set $\mathcal{M}$ that satisfies the desired condition (**DAG structural learning**)
  - The conditional mean function $\mathrm{E}(X_j|X_{\mathcal{M}-\{k\}})$ (**Supervised learning**)
  - The functional maps each h_b to $\mathrm{E}\{h_b(X_k, X_{\mathcal{M}-\{k\}})|X_{\mathcal{M}-\{k\}}\}$ (**Generative adversarial networks**)
  - Couple three learners with data-splitting and cross-fitting to ensure the validity of the test (Chernozhukov et al., 2018)

Step 1: neural structural learning (Zheng et al., 2020)

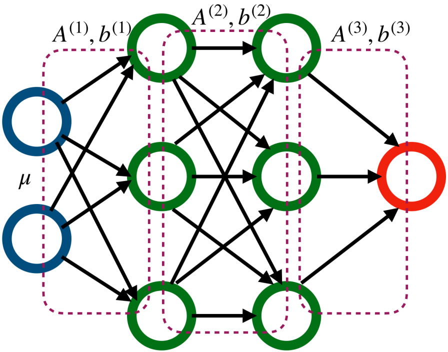
- Use a multilayer perceptron (MLP) to model nonlinearity
- Use a novel characterization of the DAG constraint

$$\mathrm{trace}\{\exp(W \circ W)\} = \text{dimension of the DAG},$$

  W is the coefficient matrix in the first layer.
- Compute $\widehat{\mathrm{AC}}_j$ and set $\mathcal{M} = \widehat{\mathrm{AC}}_j - \{k\}$.
- Requires **order consistency**, weaker than **DAG selection consistency**.

Step 2: deep learning

- Use the Scikit-learn MLP regressor to learn the conditional mean function

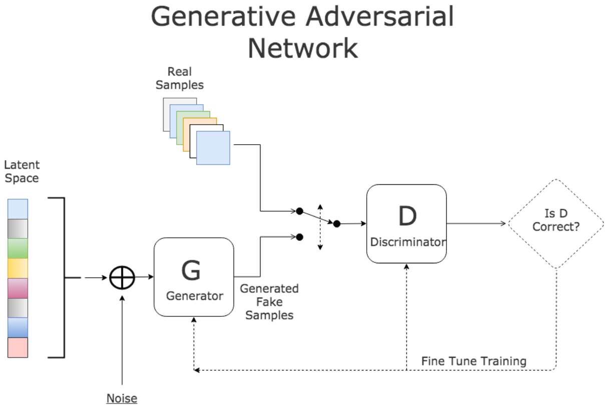


Step 3: generative adversarial networks

- $h_b \to \mathrm{E}\{h_b(X_k, X_{\mathcal{M}-\{k\}})|X_{\mathcal{M}-\{k\}}\}$
- Naive solution: separately apply supervised learning B times. **Computationally intensive** for large B.
- Learn a **distributional generator**
  - Input: $X_{\mathcal{M}-\{k\}}$
  - Output: $\{X_k^{(m)}\}_{m=1}^M$
  - Minimize the discrepancy between $X_k|X_{\mathcal{M}-k}$ and $X_k^{(m)}|X_{\mathcal{M}-k}$
- Approximate $\mathrm{E}\{h_b(X_k, X_{\mathcal{M}-\{k\}})|X_{\mathcal{M}-\{k\}}\}$ by

$$\frac{1}{M}\sum_{m=1}^{M} h_b(X_k^{(m)}, X_{\mathcal{M}-\{k\}})$$

Step 3: generative adversarial networks (Cont'd)



- We use the Sinkhorn GANs (Cuturi, 2013; Genevay et al., 2016)

Simulations

Competing tests

- **Likelihood ratio test** (LRT, Li et al., 2020)
- **Doubly robust test** (DRT), a hybrid test that combines our proposal with double regression conditional independence test (Shah and Peters, 2018)

Settings

- Nonlinear associations
- (Dimension, Sparsity) $= (50, 0.1)$, $(100, 0.04)$, $(150, 0.02)$

Simulations (Cont'd)

Edge	j=35, k=5			j=35, k=31			j=40, k=16		
Hypothesis	$\mathcal{H}_0$			$\mathcal{H}_0$			$\mathcal{H}_0$		
Method	SUGAR	DRT	LRT	SUGAR	DRT	LRT	SUGAR	DRT	LRT
$\alpha = 0.05$	0.050	0.108	1.000	0.012	0.068	0.316	0.016	0.016	1.000
$\alpha = 0.10$	0.078	0.154	1.000	0.032	0.098	0.412	0.032	0.030	1.000
Edge	j=45, k=14			j=45, k=15			j=50, k=14		
Hypothesis	$\mathcal{H}_0$			$\mathcal{H}_0$			$\mathcal{H}_0$		
Method	SUGAR	DRT	LRT	SUGAR	DRT	LRT	SUGAR	DRT	LRT
$\alpha = 0.05$	0.014	0.026	1.000	0.032	0.054	0.954	0.030	0.096	1.000
$\alpha = 0.10$	0.030	0.050	1.000	0.058	0.092	0.964	0.046	0.126	1.000
Edge	j=35, k=4			j=35, k=30			j=40, k=15		
Hypothesis	$\mathcal{H}_1$			$\mathcal{H}_1$			$\mathcal{H}_1$		
Method	SUGAR	DRT	LRT	SUGAR	DRT	LRT	SUGAR	DRT	LRT
$\alpha = 0.05$	0.534	0.082	1.000	0.992	0.728	0.204	0.550	0.204	0.102
$\alpha = 0.10$	0.546	0.126	1.000	0.992	0.818	0.290	0.550	0.264	0.180
Edge	j=45, k=12			j=45, k=13			j=50, k=13		
Hypothesis	$\mathcal{H}_1$			$\mathcal{H}_1$			$\mathcal{H}_1$		
Method	SUGAR	DRT	LRT	SUGAR	DRT	LRT	SUGAR	DRT	LRT
$\alpha = 0.05$	0.946	0.524	0.988	0.808	0.248	0.832	0.670	0.188	0.730
$\alpha = 0.10$	0.948	0.616	0.996	0.816	0.318	0.870	0.672	0.252	0.824

Simulations (Cont'd)

Edge	j=35, k=5			j=35, k=31			j=40, k=16		
Hypothesis	$\mathcal{H}_0$			$\mathcal{H}_0$			$\mathcal{H}_0$		
Method	SUGAR	DRT	LRT	SUGAR	DRT	LRT	SUGAR	DRT	LRT
$\alpha = 0.05$	0.050	0.108	1.000	0.012	0.068	0.316	0.016	0.016	1.000
$\alpha = 0.10$	0.078	0.154	1.000	0.032	0.098	0.412	0.032	0.030	1.000
Edge	j=45, k=14			j=45, k=15			j=50, k=14		
Hypothesis	$\mathcal{H}_0$			$\mathcal{H}_0$			$\mathcal{H}_0$		
Method	SUGAR	DRT	LRT	SUGAR	DRT	LRT	SUGAR	DRT	LRT
$\alpha = 0.05$	0.014	0.026	1.000	0.032	0.054	0.954	0.030	0.096	1.000
$\alpha = 0.10$	0.030	0.050	1.000	0.058	0.092	0.964	0.046	0.126	1.000
Edge	j=35, k=4			j=35, k=30			j=40, k=15		
Hypothesis	$\mathcal{H}_1$			$\mathcal{H}_1$			$\mathcal{H}_1$		
Method	SUGAR	DRT	LRT	SUGAR	DRT	LRT	SUGAR	DRT	LRT
$\alpha = 0.05$	0.534	0.082	1.000	0.992	0.728	0.204	0.550	0.204	0.102
$\alpha = 0.10$	0.546	0.126	1.000	0.992	0.818	0.290	0.550	0.264	0.180
Edge	j=45, k=12			j=45, k=13			j=50, k=13		
Hypothesis	$\mathcal{H}_1$			$\mathcal{H}_1$			$\mathcal{H}_1$		
Method	SUGAR	DRT	LRT	SUGAR	DRT	LRT	SUGAR	DRT	LRT
$\alpha = 0.05$	0.946	0.524	0.988	0.808	0.248	0.832	0.670	0.188	0.730
$\alpha = 0.10$	0.948	0.616	0.996	0.816	0.318	0.870	0.672	0.252	0.824

Simulations (Cont'd)

Edge	j=35, k=5			j=35, k=31			j=40, k=16		
Hypothesis	$\mathcal{H}_0$			$\mathcal{H}_0$			$\mathcal{H}_0$		
Method	SUGAR	DRT	LRT	SUGAR	DRT	LRT	SUGAR	DRT	LRT
$\alpha = 0.05$	0.050	0.108	1.000	0.012	0.068	0.316	0.016	0.016	1.000
$\alpha = 0.10$	0.078	0.154	1.000	0.032	0.098	0.412	0.032	0.030	1.000
Edge	j=45, k=14			j=45, k=15			j=50, k=14		
Hypothesis	$\mathcal{H}_0$			$\mathcal{H}_0$			$\mathcal{H}_0$		
Method	SUGAR	DRT	LRT	SUGAR	DRT	LRT	SUGAR	DRT	LRT
$\alpha = 0.05$	0.014	0.026	1.000	0.032	0.054	0.954	0.030	0.096	1.000
$\alpha = 0.10$	0.030	0.050	1.000	0.058	0.092	0.964	0.046	0.126	1.000
Edge	j=35, k=4			j=35, k=30			j=40, k=15		
Hypothesis	$\mathcal{H}_1$			$\mathcal{H}_1$			$\mathcal{H}_1$		
Method	SUGAR	DRT	LRT	SUGAR	DRT	LRT	SUGAR	DRT	LRT
$\alpha = 0.05$	0.534	0.082	1.000	0.992	0.728	0.204	0.550	0.204	0.102
$\alpha = 0.10$	0.546	0.126	1.000	0.992	0.818	0.290	0.550	0.264	0.180
Edge	j=45, k=12			j=45, k=13			j=50, k=13		
Hypothesis	$\mathcal{H}_1$			$\mathcal{H}_1$			$\mathcal{H}_1$		
Method	SUGAR	DRT	LRT	SUGAR	DRT	LRT	SUGAR	DRT	LRT
$\alpha = 0.05$	0.946	0.524	0.988	0.808	0.248	0.832	0.670	0.188	0.730
$\alpha = 0.10$	0.948	0.616	0.996	0.816	0.318	0.870	0.672	0.252	0.824

Brain effective connectivity analysis

- **Data**: Human Connectome Project (Van Essen et al., 2013)
- **Objective**: understand brain connectivity patterns of adults
- **Subjects**: individuals that undertook a story-math task
- $N = 28$, $T = 316$, $d = 127$ regions from 4 functional modules
  - auditory
  - visual
  - frontoparietal task control
  - default mode
- These modules are believed to be involved in language processing and problem solving task (Barch et al., 2013)

Brain effective connectivity analysis (Cont'd)

	Auditory (13)		Default mode (58)		Visual (31)		Fronto-parietal (25)	
	low	high	low	high	low	high	low	high
Auditory (13)	20	17	0	0	0	1	2	0
Default mode (58)	0	0	68	46	3	2	11	23
Visual (31)	0	0	3	2	56	46	0	1
Fronto-parietal (25)	2	1	11	23	0	1	22	27

- More within-module connections than between-module connections
- More within-module connections for the frontoparietal task control module for the high-performance subjects than the low-performance subjects

Brain effective connectivity analysis (Cont'd)

	Auditory (13)		Default mode (58)		Visual (31)		Fronto-parietal (25)	
	low	high	low	high	low	high	low	high
Auditory (13)	20	17	0	0	0	1	2	0
Default mode (58)	0	0	68	46	3	2	11	23
Visual (31)	0	0	3	2	56	46	0	1
Fronto-parietal (25)	2	1	11	23	0	1	22	27

- More within-module connections than between-module connections
- More within-module connections for the frontoparietal task control module for the high-performance subjects than the low-performance subjects

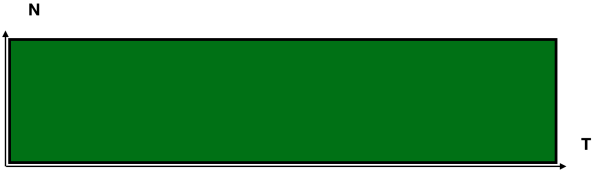
Brain effective connectivity analysis (Cont'd)

	Auditory (13)		Default mode (58)		Visual (31)		Fronto-parietal (25)	
	low	high	low	high	low	high	low	high
Auditory (13)	20	17	0	0	0	1	2	0
Default mode (58)	0	0	68	46	3	2	11	23
Visual (31)	0	0	3	2	56	46	0	1
Fronto-parietal (25)	2	1	11	23	0	1	22	27

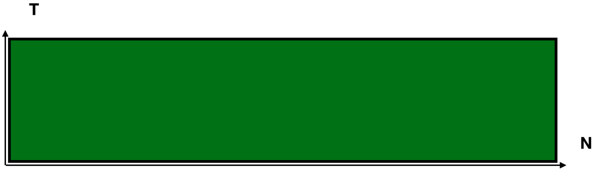
- More within-module connections than between-module connections
- More within-module connections for the frontoparietal task control module for the high-performance subjects than the low-performance subjects

Bidirectional theories

- N the number of subjects;
- T the number of time points;
- bidirectional asymptotics: a framework where either N or T grows to ∞;
- large T, small N (e.g., neuroimaging)



- large N, small T (e.g., genetics)



- large N, large T

Bidirectional theories (Cont'd)

Theorem (Size)

Under certain mild conditions, our test controls the type-I error asymptotically as either N or T diverges to infinity.

Theorem (Power)

Under certain mild conditions, the power of our test diverges to 1 as either N or T diverges to infinity.

Theorem (Order consistency)

Under certain mild conditions, the neural structural learning algorithm can consistently identify the order of the DAG, as either N or T diverges to infinity.

- Preprint: https://arxiv.org/pdf/2106.01474.pdf

Thank you! ☺