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

Chengchun Shi 1 and Yunzhe Zhou 2 and Lexin Li 2

¹London School of Economics and Political Science

²University of California at Berkeley

In this talk, we will focus on...

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



• 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.

- 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
 - sUpervised learning based on neural networks
 - Distributional generator based on Generative AdveRsarial networks (GANs, Goodfellow et al., 2014)

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

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

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

• Testing hypotheses:

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

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

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

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

Theorem

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

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

Main idea (Cont'd)

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

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

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

W is the coefficient matrix in the first layer.

- Compute \widehat{AC}_j and set $\mathcal{M} = \widehat{AC}_j \{k\}$.
- 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

- $h_b \rightarrow \mathsf{E}\{h_b(X_k, X_{\mathcal{M}-\{k\}}) | X_{\mathcal{M}-\{k\}}\}$
- 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 $\mathsf{E}\{h_b(X_k, X_{\mathcal{M}-\{k\}})|X_{\mathcal{M}-\{k\}}\}$ 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)

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=4	5, k=1	4	j=4	15, k=1	5	j=5	0, k=1	4	
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	ł	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=4	5, k=1	2	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=4	5, k=1	4	j=4	5, k=1	5	j=5	50, k=1	4
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	1	j=3	85, k=3	0	j=4	0, k=1	5
Edge Hypothesis	j=	$\frac{35, k=4}{\mathcal{H}_1}$	1	j=3	35, k=3 <i>H</i> 1	0	j=4	\mathcal{H}_1	5
Edge Hypothesis Method	j= SUGAR	35, k=4 \mathcal{H}_1 DRT	LRT	j=3 SUGAR	\mathcal{H}_1 DRT	0 LRT	j=4 SUGAR	0, k=1 \mathcal{H}_1 DRT	5 LRT
Edge Hypothesis Method $\alpha = 0.05$	j= SUGAR 0.534	35, k=4 H_1 DRT 0.082	LRT 1.000	j=3 SUGAR 0.992	${{\cal H}_1 \over {\cal D} { m RT}}$	0 LRT 0.204	j=4 SUGAR 0.550	0, k=1 \mathcal{H}_1 DRT 0.204	5 LRT 0.102
EdgeHypothesisMethod $\alpha = 0.05$ $\alpha = 0.10$	j= SUGAR 0.534 0.546	35, k=4 \mathcal{H}_1 DRT 0.082 0.126	LRT 1.000 1.000	j=3 SUGAR 0.992 0.992	85, k=3 <i>H</i> 1 DRT 0.728 0.818	0 LRT 0.204 0.290	j=4 SUGAR 0.550 0.550	0, k=1 <i>H</i> 1 DRT 0.204 0.264	5 LRT 0.102 0.180
$\begin{tabular}{c} Edge \\ \hline Hypothesis \\ \hline Method \\ \hline α = 0.05 \\ \hline α = 0.10 \\ \hline $Edge \end{tabular}$	j= SUGAR 0.534 0.546 j=4	35, k=4 \mathcal{H}_1 DRT 0.082 0.126 \mathcal{H}_5 , k=1	LRT 1.000 1.000 2	j=3 SUGAR 0.992 0.992 j=4	35, k=3 <i>H</i> ₁ DRT 0.728 0.818 5, k=1	0 LRT 0.204 0.290 3	j=4 SUGAR 0.550 0.550 j=5	0, k=1 H ₁ DRT 0.204 0.264 50, k=1	5 LRT 0.102 0.180 3
$\begin{tabular}{c} Edge \\ Hypothesis \\ \hline Method \\ \hline α = 0.05 \\ \hline α = 0.10 \\ \hline $Edge \\ Hypothesis \\ \end{tabular}$	j= SUGAR 0.534 0.546 j=4	35, $k=4$ H_1 DRT 0.082 0.126 $J_5, k=1$ H_1	LRT 1.000 1.000 2	j=3 SUGAR 0.992 0.992 j=4	\mathcal{H}_1 \mathcal{H}_1 DRT 0.728 0.818 \mathcal{H}_5 , k=1 \mathcal{H}_1	0 LRT 0.204 0.290 3	j=4 SUGAR 0.550 0.550 j=5	$ \begin{array}{c} {\rm H}_0, \ {\rm k}{=}1 \\ {\cal H}_1 \\ {\rm DRT} \\ {\rm 0.204} \\ {\rm 0.264} \\ {\rm 00, \ {\rm k}{=}1} \\ {\cal H}_1 \end{array} $	5 LRT 0.102 0.180 3
$\begin{array}{c} Edge \\ Hypothesis \\ Method \\ \alpha = 0.05 \\ \alpha = 0.10 \\ Edge \\ Hypothesis \\ Method \end{array}$	j= SUGAR 0.534 0.546 j=4 SUGAR	$\begin{array}{c} 35, \ k=4\\ \hline \mathcal{H}_1 \\ \hline DRT \\ 0.082 \\ 0.126 \\ \hline 5, \ k=1 \\ \hline \mathcal{H}_1 \\ \hline DRT \end{array}$	LRT 1.000 1.000 2 LRT	j=3 SUGAR 0.992 0.992 j=4 SUGAR	35, k=3 <i>H</i> ₁ DRT 0.728 0.818 5, k=1 <i>H</i> ₁ DRT	0 LRT 0.204 0.290 3 LRT	j=4 SUGAR 0.550 0.550 j=5 SUGAR	0, k=1 \mathcal{H}_1 DRT 0.204 0.264 0, k=1 \mathcal{H}_1 DRT	5 LRT 0.102 0.180 3 LRT
$\begin{array}{c} Edge \\ Hypothesis \\ Method \\ \alpha = 0.05 \\ \alpha = 0.10 \\ Edge \\ Hypothesis \\ Method \\ \alpha = 0.05 \end{array}$	j= SUGAR 0.534 0.546 j=4 SUGAR 0.946	$\begin{array}{c} 35, \ k=4\\ \mathcal{H}_1 \\ \ DRT \\ 0.082\\ 0.126\\ 5, \ k=1\\ \mathcal{H}_1 \\ \ DRT \\ 0.524 \end{array}$	LRT 1.000 1.000 2 LRT 0.988	j=3 SUGAR 0.992 0.992 j=4 SUGAR 0.808	5, k=3 H ₁ DRT 0.728 0.818 5, k=1 H ₁ DRT 0.248	0 LRT 0.204 0.290 3 LRT 0.832	j=4 SUGAR 0.550 0.550 j=5 SUGAR 0.670	0, k=1 H_1 0.204 0.264 0, k=1 H_1 0.188	5 LRT 0.102 0.180 3 LRT 0.730

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=4	5, k=1	4	j=4	5, k=1	5	j=5	0, k=1	4	
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	
	j=35, k=4									
Edge	j=	35, k=4	ł	j=3	85, k=3	0	j=4	10, k=1	5	
Edge Hypothesis	j=	35, k=4 H1	ļ	j=3	85, k=3 <i>H</i> 1	0	j=4	\mathcal{H}_1	5	
Edge Hypothesis Method	j= SUGAR	$\begin{array}{c} 35, \ k=4 \\ \mathcal{H}_1 \\ \ DRT \end{array}$	LRT	j=3 SUGAR	\mathcal{H}_1 DRT	0 LRT	j=4 SUGAR	0, k=1 \mathcal{H}_1 DRT	5 LRT	
$\begin{tabular}{c} Edge \\ Hypothesis \\ \hline Method \\ \hline α = 0.05 \end{tabular}$	j= SUGAR 0.534	35, k=4 <i>H</i> 1 DRT 0.082	LRT 1.000	j=3 SUGAR 0.992	85, k=3 <i>H</i> 1 DRT 0.728	0 LRT 0.204	j=4 SUGAR 0.550	0, k=1 H ₁ DRT 0.204	5 LRT 0.102	
EdgeHypothesisMethod $\alpha = 0.05$ $\alpha = 0.10$	j= SUGAR 0.534 0.546	35, k=4 \mathcal{H}_1 DRT 0.082 0.126	LRT 1.000 1.000	j=3 SUGAR 0.992 0.992	5, k=3 <i>H</i> 1 DRT 0.728 0.818	0 LRT 0.204 0.290	j=4 SUGAR 0.550 0.550	0, k=1 \mathcal{H}_1 DRT 0.204 0.264	5 LRT 0.102 0.180	
EdgeHypothesisMethod $\alpha = 0.05$ $\alpha = 0.10$ Edge	j= SUGAR 0.534 0.546 j=4	35, k=4 H ₁ DRT 0.082 0.126 5, k=1	LRT 1.000 1.000 2	j=3 SUGAR 0.992 0.992 j=4	35, k=3 H ₁ DRT 0.728 0.818 5, k=1	0 LRT 0.204 0.290 3	j=4 SUGAR 0.550 0.550 j=5	0, k=1 H ₁ DRT 0.204 0.264 50, k=1	5 LRT 0.102 0.180 3	
$\begin{tabular}{c} Edge \\ Hypothesis \\ Method \\ \hline α = 0.05 \\ \hline α = 0.10 \\ edge \\ Hypothesis \\ \end{tabular}$	j= SUGAR 0.534 0.546 j=4	$\begin{array}{c c} 35, \ k=4\\ \hline \mathcal{H}_1 \\ \hline DRT \\ 0.082 \\ 0.126 \\ \hline 5, \ k=1\\ \hline \mathcal{H}_1 \end{array}$	LRT 1.000 1.000 2	j=3 SUGAR 0.992 0.992 j=4	$\begin{array}{c} 35, \ k=3\\ \hline \mathcal{H}_1\\ \hline DRT\\ 0.728\\ \hline 0.818\\ \hline 5, \ k=1\\ \hline \mathcal{H}_1 \end{array}$	0 LRT 0.204 0.290 3	j=4 SUGAR 0.550 0.550 j=5	$ \begin{array}{c} 0, \ k=1 \\ \hline \mathcal{H}_1 \\ \hline DRT \\ 0.204 \\ \hline 0.264 \\ \hline 0, \ k=1 \\ \hline \mathcal{H}_1 \end{array} $	5 LRT 0.102 0.180 3	
$\begin{tabular}{lllllllllllllllllllllllllllllllllll$	j= SUGAR 0.534 0.546 j=4 SUGAR	$\begin{array}{c} 35, \ k=4\\ \mathcal{H}_1 \\ \ DRT \\ 0.082 \\ 0.126 \\ I5, \ k=1 \\ \mathcal{H}_1 \\ \ DRT \end{array}$	LRT 1.000 1.000 2 LRT	j=3 SUGAR 0.992 0.992 j= ² SUGAR	$\begin{array}{c} \text{35, k=3} \\ \hline \mathcal{H}_1 \\ \hline \text{DRT} \\ 0.728 \\ \hline 0.818 \\ \hline \text{45, k=1} \\ \hline \mathcal{H}_1 \\ \hline \text{DRT} \end{array}$	0 LRT 0.204 0.290 3 LRT	j=4 SUGAR 0.550 0.550 j=5 SUGAR	$\begin{array}{c} 0, \ k=1 \\ \hline \mathcal{H}_1 \\ \hline DRT \\ 0.204 \\ 0.264 \\ \hline 0, \ k=1 \\ \hline \mathcal{H}_1 \\ \hline DRT \end{array}$	5 LRT 0.102 0.180 3 LRT	
$\begin{array}{c} {\sf Edge} \\ {\sf Hypothesis} \\ {\sf Method} \\ \alpha = 0.05 \\ \alpha = 0.10 \\ {\sf Edge} \\ {\sf Hypothesis} \\ {\sf Method} \\ \alpha = 0.05 \end{array}$	j= SUGAR 0.534 0.546 j=4 SUGAR 0.946	$\begin{array}{c} 35, \ k=4\\ \mathcal{H}_1 \\ \ DRT \\ 0.082 \\ 0.126 \\ I5, \ k=1\\ \mathcal{H}_1 \\ \ DRT \\ 0.524 \end{array}$	LRT 1.000 1.000 2 LRT 0.988	j=3 SUGAR 0.992 0.992 j=4 SUGAR 0.808	$\begin{array}{c} \text{35, k=3} \\ \hline \mathcal{H}_1 \\ \hline \text{DRT} \\ 0.728 \\ \hline 0.818 \\ \hline \text{5, k=1} \\ \hline \mathcal{H}_1 \\ \hline \text{DRT} \\ \hline 0.248 \end{array}$	0 LRT 0.204 0.290 3 LRT 0.832	j=4 SUGAR 0.550 0.550 j=5 SUGAR 0.670	$\begin{array}{c} 0, \ k=1 \\ \hline \mathcal{H}_1 \\ \hline DRT \\ 0.204 \\ 0.264 \\ \hline 0, \ k=1 \\ \hline \mathcal{H}_1 \\ \hline DRT \\ 0.188 \end{array}$	5 LRT 0.102 0.180 3 LRT 0.730	

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		Default mode		Visual		Fronto-parietal	
	(13)		(58)		(31)		(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		Default mode		Visual		Fronto-parietal	
	(13)		(58)		(31)		(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		Defau	lt mode	Visual		Fronto-parietal	
	(13)		(58)		(31)		(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 N, large T

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! ③