# ANOCE: Analysis of Causal Effects with Multiple Mediators via Constrained Structural Learning

Hengrui Cai   Rui Song   Wenbin Lu

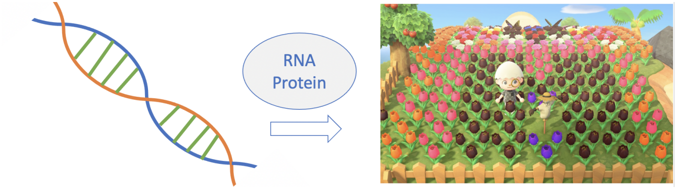
Department of Statistics, North Carolina State University, North Carolina, USA

NCSU Seminar

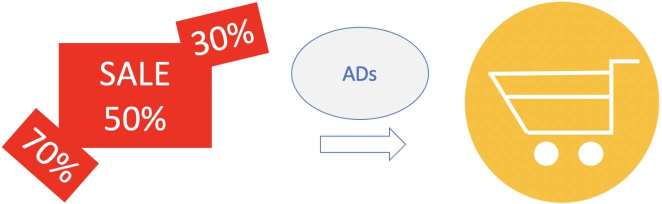
# Motivation

In the era of causal revolution, identifying the causal effect of an exposure on the outcome of interest is an important problem.

- Genetics:




- Economics:



# Example: Spread of Coronavirus Disease 2019 (COVID-19)

- Treatment $A$: Locking Wuhan down on Jan 23rd, 2020, followed by 12 other cities in Hubei, known as “2020 Hubei lockdowns”;
- Outcome $Y$: Stopping virus spreading;
- Mediators $M$: Migration scale of major Chinese cities outside Hubei, significantly reduced due to the lockdown and thus blocked the epidemic spread.


- $Q$: How to quantify the causal effects of 2020 Hubei lockdowns on reducing the COVID-19 spread via cities outside Hubei?

# Example: Spread of Coronavirus Disease 2019 (COVID-19)

- Treatment $A$: Locking Wuhan down on Jan 23rd, 2020, followed by 12 other cities in Hubei, known as "2020 Hubei lockdowns";
- Outcome $Y$: Stopping virus spreading;
- Mediators $M$: Migration scale of major Chinese cities outside Hubei, significantly reduced due to the lockdown and thus blocked the epidemic spread.


- $Q$: How to quantify the causal effects of 2020 Hubei lockdowns on reducing the COVID-19 spread via cities outside Hubei?

# Example: Spread of Coronavirus Disease 2019 (COVID-19)

- Treatment $A$: Locking Wuhan down on Jan 23rd, 2020, followed by 12 other cities in Hubei, known as "2020 Hubei lockdowns";
- Outcome $Y$: Stopping virus spreading;
- Mediators $M$: Migration scale of major Chinese cities outside Hubei, significantly reduced due to the lockdown and thus blocked the epidemic spread.


- $Q$: How to quantify the causal effects of 2020 Hubei lockdowns on reducing the COVID-19 spread via cities outside Hubei?

# Example: Spread of Coronavirus Disease 2019 (COVID-19)

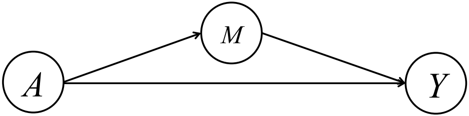
- Treatment $A$: Locking Wuhan down on Jan 23rd, 2020, followed by 12 other cities in Hubei, known as "2020 Hubei lockdowns";
- Outcome $Y$: Stopping virus spreading;
- Mediators $M$: Migration scale of major Chinese cities outside Hubei, significantly reduced due to the lockdown and thus blocked the epidemic spread.


- $Q$: How to quantify the causal effects of 2020 Hubei lockdowns on reducing the COVID-19 spread via cities outside Hubei?

# Background

Consider a **Causal Graph**:

- The exposure / treatment ($A$) may have a **Direct Effect** (DE) on the outcome ($Y$);
- and also an **Indirect Effect** (IE) that goes through a set of potential mediators ($M$);
- The total causal effect of $A$ on $Y$: TE $=$ DE $+$IE;

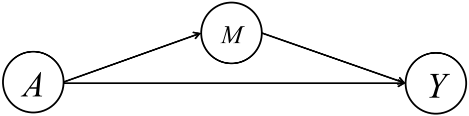



- The goal is to understand what portion of TE can be attributed to the potential $M$. **[Causal Mediation Analysis]**

# Background

Consider a **Causal Graph**:

- The exposure / treatment $(A)$ may have a **Direct Effect** (DE) on the outcome $(Y)$;
- and also an **Indirect Effect** (IE) that goes through a set of potential mediators $(M)$;
- The total causal effect of $A$ on $Y$: TE $=$ DE $+$IE;

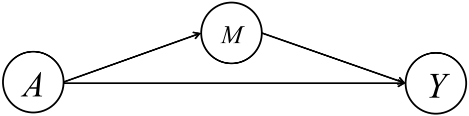



- The goal is to understand what portion of TE can be attributed to the potential $M$. **[Causal Mediation Analysis]**

# Background

Consider a **Causal Graph**:

- The exposure / treatment $(A)$ may have a **Direct Effect** (DE) on the outcome $(Y)$;
- and also an **Indirect Effect** (IE) that goes through a set of potential mediators $(M)$;
- The total causal effect of $A$ on $Y$: TE $=$ DE $+$IE;

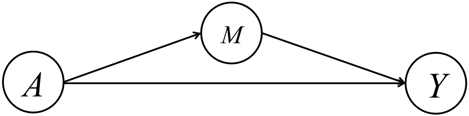



- The goal is to understand what portion of TE can be attributed to the potential $M$. **[Causal Mediation Analysis]**

# Background

Consider a **Causal Graph**:

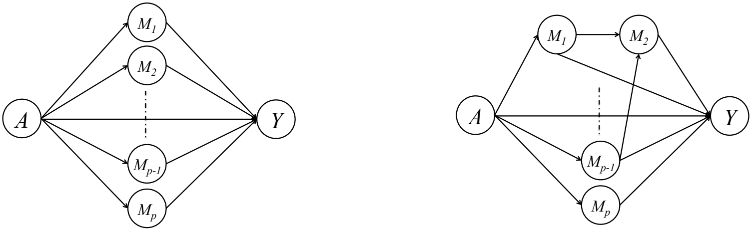
- The exposure / treatment $(A)$ may have a **Direct Effect** (DE) on the outcome $(Y)$;
- and also an **Indirect Effect** (IE) that goes through a set of potential mediators $(M)$;
- The total causal effect of $A$ on $Y$: TE $=$ DE $+$IE;




- The goal is to understand what portion of TE can be attributed to the potential $M$. **[Causal Mediation Analysis]**

# Causal Mediation Analysis

The 'interaction' means that there exists at least one mediator that is regulated by other mediator(s); otherwise, we call mediators are 'parallel'. We consider all possible causal structures with multiple mediators.




(a) A DAG with parallel mediators.

(b) A DAG with interacted mediators.

Figure 1: The illustration of DAGs with different types of mediators, where $A$ is the exposure, $\{M_1, \cdots, M_p\}$ are mediators, and $Y$ is the outcome of interest.

# Contributions

- Propose a new statistical framework to comprehensively characterize causal effects with multiple mediators: Analysis of Causal Effects (ANOCE);
  - Give an exact decomposition of the indirect effect on the level of individual mediators;
- Develop a feasible algorithm to estimate the individual mediator effect by incorporating background knowledge of the temporal causal relationship: Constrained Variational Auto-Encoder (CVAE).

# Graph Terminology

- A graph $\mathcal{G} = (X, E)$ with a node set $X$ and an edge set $E$;
- $X_i$ is a **parent of** $X_j$ if there is a directed edge from $X_i$ to $X_j$, and let the set of all parents of node $X_j$ in $\mathcal{G}$ as $PA_{X_j}(\mathcal{G})$;
- A **directed path** from $X_i$ to $X_j$ is a path between $X_i$ and $X_j$ where all edges are directed toward $X_j$, together with the directed edge $X_j$ to $X_i$ forming a **directed cycle**;
- A directed graph that does not contain directed cycles is called a **directed acyclic graph (DAG)**.
- Let $B = \{b_{i,j}\}_{1 \leq i \leq d, 1 \leq j \leq d}$ be a $d \times d$ matrix, where $b_{i,j}$ is the weight of the edge $X_i \to X_j \in E$, and $b_{i,j} = 0$ otherwise;
- $\mathcal{G} = (X, B)$ is a **weighted DAG** to characterize the causal relationship, where $X_i \to X_j$ means that $X_i$ is a **direct cause** of $X_j$.

# Assumptions

Let $Y^*(A = a, M = m)$ be the potential outcome that would be observed after receiving treatment $a$ with mediators as $m$, and $M^*(A = a)$ be the potential mediators that would be observed after receiving treatment $a$.

- (A1) the effect of $A$ on $Y$ is unconfounded, i.e., $Y^*(A = a, M = m) \perp A, \forall a, m$;
- (A2) the effect of $A$ on $M$ is unconfounded, i.e., $M^*(A = a) \perp A, \forall a$;
- (A3) the effect of $M$ on $Y$ is unconfounded given $A$, i.e., $Y^*(A = a, M = m) \perp M|A, \forall a, m$.

$X = [A, M^\top, Y]^\top$ is generated from a **linear structural equation model (LSEM)** characterized by the pair ($\mathcal{G}$, $\epsilon$), where $\epsilon$ is a random vector of jointly independent error variables, i.e.

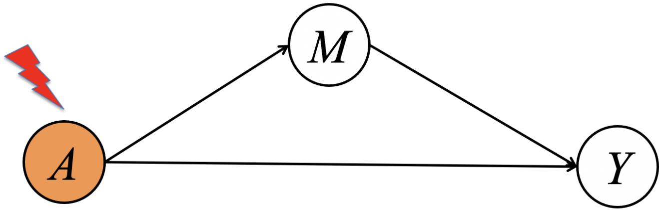
$$X = B^\top X + \epsilon.$$

# Preliminary of Causal Effects

**Total Effect ($TE$):**

$$TE = \partial E\{Y|do(A=a)\}/\partial a = E\{Y|do(A=a+1)\} - E\{Y|do(A=a)\},$$

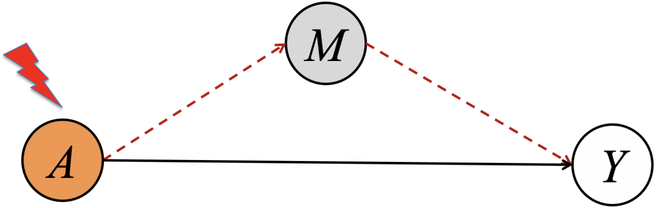
where $do(A=a)$ is a mathematical operator to simulate physical interventions that hold $A$ constant as $a$ while keeping the rest of the model unchanged.



# Preliminary of Causal Effects

Natural Direct Effect ($DE$) not mediated by mediators:

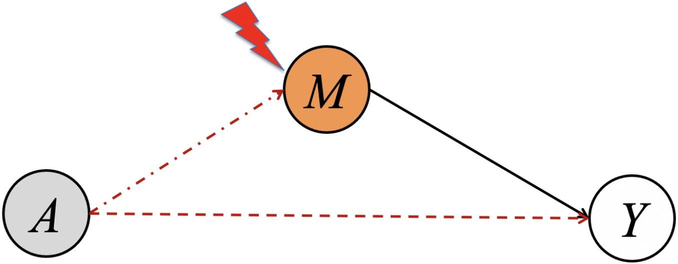
$$DE = E\{Y|do(A = a+1, M = m^{(a)})\} - E\{Y|do(A = a)\}.$$



# Preliminary of Causal Effects

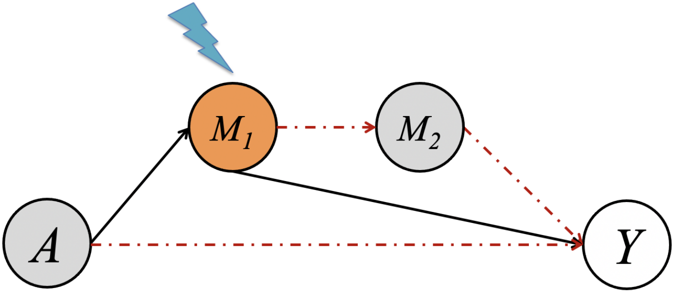
Natural Indirect Effect ($IE$) regulated by mediators:

$$IE = E\{Y|do(A = a, M = m^{(a+1)})\} - E\{Y|do(A = a)\}.$$



# Natural Direct Effect for Individual Mediators

**Interpretation:** The natural direct effect for individual mediator is the causal effect through a particular mediator from the treatment on the outcome that is not regulated by its descendent mediators.

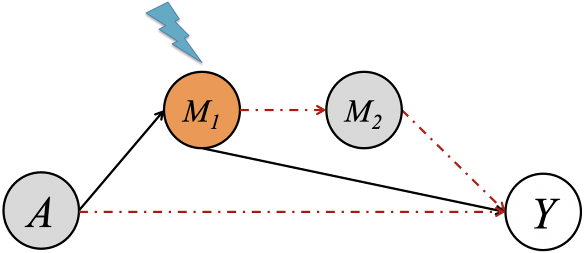


# Natural Direct Effect for Individual Mediators

Natural direct effect for $M_i$:

$$DM_i = \Big[E\{M_i|do(A=a+1)\} - E\{M_i|do(A=a)\}\Big] \times \Big[E\{Y|do(A=a, M_i=m_i^{(a)}+1, \Omega_i=o_i^{(a)})\} - E\{Y|do(A=a)\}\Big],$$

where $m_i^{(a)}$ is the value of $M_i$ if setting $do(A=a)$, $\Omega_i = M \setminus M_i$ is the set of mediators except $M_i$, and $o_i^{(a)}$ is the value of $\Omega_i$ if setting $do(A=a)$.

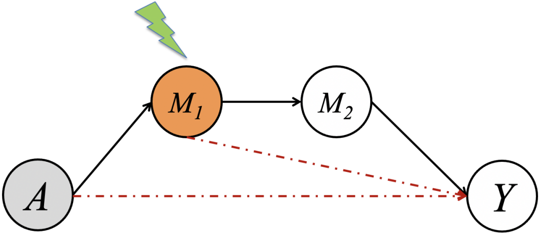


# Natural Indirect Effect for Individual Mediators

Natural indirect effect for $M_i$:

$$IM_i = \Big[E\{M_i|do(A=a+1)\} - E\{M_i|do(A=a)\}\Big] \times \Big[E\{Y|do(A=a, M_i=m_i^{(a)}+1)\} - E\{Y|do(A=a, M_i=m_i^{(a)}+1, \Omega_i=o_i^{(a)})\}\Big].$$

**Interpretation:** Capture the indirect effect of a particular mediator on the outcome regulated by its descendent mediators.



# Expressions of Causal Effects under LSEM

**The LSEM under Assumptions (A1-A3)**

$$\begin{bmatrix} A \\ M \\ Y \end{bmatrix} = B^\top \begin{bmatrix} A \\ M \\ Y \end{bmatrix} + \epsilon = \begin{bmatrix} 0 & \mathbf{0}_{p\times 1} & 0 \\ \alpha & B_M^\top & 0 \\ \gamma & \beta^\top & 0 \end{bmatrix} \begin{bmatrix} A \\ M \\ Y \end{bmatrix} + \begin{bmatrix} \epsilon_A \\ \epsilon_{M_p} \\ \epsilon_Y \end{bmatrix},$$

where $\gamma$ is a scalar, $\alpha$, $\beta$, and $\mathbf{0}_{p\times 1}$ are $p \times 1$ vectors, $B_M$ is a $p \times p$ matrix, and $\epsilon \equiv [\epsilon_A, \epsilon_M^\top, \epsilon_Y]^\top$.

Here, by assumptions (A1-A3), we have the exposure $A$ has **no parents** and the outcome $Y$ has **no descendants**, so equivalently, the **first row** and the **last column** of $B^\top$ are all zeros (i.e. the first column and the last row of B are all zeros).

# Expressions of Causal Effects under LSEM

**Theorem 1**

*Under assumptions (A1-A3) and the LSEM, we have:*
*1). the natural direct effect is* $DE = \gamma$*;*
*2). the natural indirect effect is* $IE = \beta^\top (I_p - B_M^\top)^{-1}\alpha$*, where* $I_p$ *is a* $p \times p$ *identity matrix;*
*3). the total effect of* $A$ *on* $Y$ *is* $TE = \gamma + \beta^\top (I_p - B_M^\top)^{-1}\alpha$*;*
*4). the natural direct effect of* $M_i$ *on* $Y$ *is* $DM_i = \beta_i \{(I_p - B_M^\top)^{-1}\alpha\}_i$*, where* $\beta_i$ *is the* $i$*-th element of the vector* $\beta$ *and corresponds to the weight of the edge* $M_i \to Y$*, and* $\{(I_p - B_M^\top)^{-1}\alpha\}_i$ *is the* $i$*-th element of the vector* $(I_p - B_M^\top)^{-1}\alpha$ *and corresponds to the total effect of* $A$ *on* $M_i$*, i.e.* $E\{M_i|do(A = a+1)\} - E\{M_i|do(A = a)\}$.

# Exactly Decomposition of Indirect Effect via Mediators

**Theorem 2**

*Under assumptions (A1-A3), the $IE$ can be decomposed through $DM$s as:*

$$IE = \sum_{i=1}^{p} DM_i.$$

The proposed natural direct effect of individual mediators **(DM) exactly decomposes the indirect effect (IE)** of the exposure on the outcome.

# Analysis of Causal Effects Table

Table 1: Table of Analysis of Causal Effects (ANOCE Table).

| Source | Degree of freedom | Causal effects |
|---|---|---|
| Direct effect from $A$ | $1$ | $DE$ |
| Indirect effect via $M$ | $p$ | $IE$ |
| $\begin{cases} M_1 \\ M_2 \\ \vdots \\ M_p \end{cases}$ | $\begin{cases} 1 \\ 1 \\ \vdots \\ 1 \end{cases}$ | $\begin{cases} DM_1 \\ DM_2 \\ \vdots \\ DM_p \end{cases}$ |
| Total | $1+p$ | $TE$ |

- $TE = DE + IE$ in Pearl et al. (2009);
- Theorem 2: $IE = \sum_{i=1}^{p} DM_i$.

# How to Estimate the Causal Effects: Learn DAG

- The LSEM $X = B^\top X + \epsilon$ can be rewritten as $(I_{p+2} - B^\top)X = \epsilon$, where $I_{p+2}$ is an identity matrix. Inversely, we have

$$X = (I_{p+2} - B^\top)^{-1}\epsilon.$$

- Following the Variational Auto-Encoder (VAE) in Yu et al. (2019), treat the random error $\epsilon$ as the **independent latent variables** to generate $X$, by two multilayer perceptrons as encoder and decoder, with weights as $\theta$.
- Adopt the **acyclicity constraint** on $B$ as,

$$h_1(B) \equiv \mathrm{tr}\left[(I_{p+2} + tB \bullet B)^{p+2}\right] - (p+2) = 0,$$

where $\mathrm{tr}(\cdot)$ is the trace of a matrix, $t$ is a hyperparameter that depends on an estimation of the largest eigenvalue of $B$, and $\bullet$ denotes for the element-wise square.

# Identification Constraint: Temporal Causal Relationship

To incorporate **prior knowledge** of the temporal causal relationship among variables, we propose an identification constraint that indicates the topological order of the exposure and the outcome.

## Identification Constraint

Under unconfounded assumptions, the exposure $A$ has no parents ($PA_A(\mathcal{G}) = \varnothing$), and the outcome $Y$ has no descendants ($Y \notin PA_X(\mathcal{G})$):

$$h_2(B) \equiv \sum_{i=1}^{p+2} |b_{i,1}| + \sum_{j=2}^{p+2} |b_{p+2,j}| = 0,$$

where $b_{i,j}$ is the element of the matrix $B$ in $i$-th row and $j$-th column.

The above constraint forces the topological order of $A$ as 1 while of $Y$ as $p+2$, under which the DAG is searched within a restricted regime.

# Constrained VAE for ANOCE

Algorithm

## Objective Function: Evidence Lower Bound with Two Constraints

$$\begin{cases} \min\limits_{B,\theta} \quad f(B,\theta) = \frac{1}{p+2}\sum_{i=1}^{p+2} D_{KL}\{q(\epsilon|X_i)||p(\epsilon)\} - E_{q(\epsilon|X_i)}\{\log p(X_i|\epsilon)\}, \\ s.t. \quad h_1(B) = 0 \quad \text{and} \quad h_2(B) = 0, \end{cases}$$

where $D_{KL}(\cdot||\cdot)$ is the Kullback-Leibler divergence, $p(\epsilon)$ is the prior distribution of $\epsilon$, $q(\epsilon|X_i)$ is the reconstructed empirical posterior distribution of $\epsilon$, and $p(X_i|\epsilon)$ is the likelihood function.

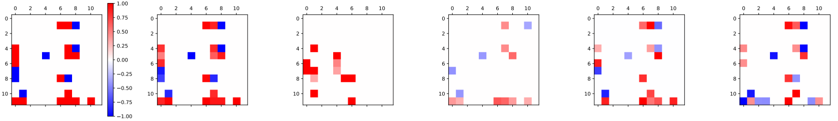
## Loss Function by Augmented Lagrangian

$$L_{c,d}(B,\theta,\lambda_1,\lambda_2) = f(B,\theta) + \lambda_1 h_1(B) + \lambda_2 h_2(B) + c|h_1(B)|^2 + d|h_2(B)|^2,$$

where $\lambda_1$ and $\lambda_2$ are Lagrange multipliers, and $c$ and $d$ are penalty terms.

Compare to popular causal discovery methods:

- the **PC** (Spirtes et al. 2000);
- the **ICA-LiNGAM** (Shimizu et al. 2006);
- the **NOTEA**RS (Zheng et al. 2018);
- the DAG-**GNN** (Yu et al. 2019).




(a) True. (b) ANOCE (c) PC (d) ICA (e) NOTEA (f) GNN

# Real Data Analysis: COVID-19 Outbreak

The Hubei lockdowns not only **directly** blocked infected people leaving from Hubei but also stimulated a decreased migration outside Hubei and thus **indirectly** prevented the spreading of the virus in other parts of China.


- Exposure $A$ is if Hubei is on lockdown: 0 for unlocked (before and on Jan 23rd) and 1 for locked (on and after Jan 24th);
- Potential mediators $M$ are candidate cities:
  - Contain most potential infected people;
  - Use the **daily migration scale index** (MSI, the migration magnitude of large groups of people from one geographical area to another, data from Baidu Qianxi) as the value of each mediator.

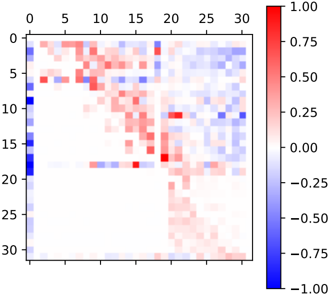
# Dataset Description

Data from the National Health Commission of China:

- Outcome $Y$ characterizes the severity of the virus spreading:
  - Due to the diagnose and incubation period of COVID-19 (Lauer et al. 2020);
  - $Y$ is **the increasing rate of confirmed cases out of Hubei** with a one-week delay:
  - $Y_t = \frac{\text{Confirmed cases out of Hubei}_{t+8} - \text{Confirmed cases out of Hubei}_{t+7}}{\text{Confirmed cases out of Hubei}_{t+7}}$.
- Selected period: $t$ **is chosen from Jan 12th to Feb 20th, 2020**.
  - Jan 19th, 2020: earliest date with an available confirmed cases (to compute $Y_{t=1}$ on Jan 12th);
  - After Feb 20th, 2020: the pandemic was under control outside Hubei with the evidence of the work resumption in China.
- The final dataset yields **a total of 38 records**.

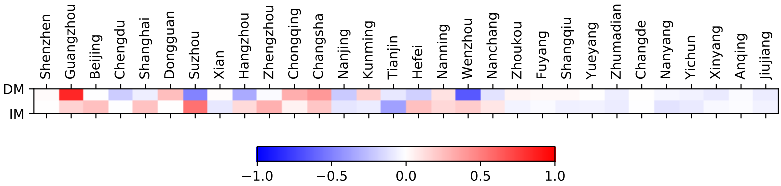
# Overall Summary

- By locking Hubei down, China successfully reduced 49.7% of the daily new cases outside Hubei;
- 84% of which is the **indirect effect** contributed via the reduced migration of major cities out of Hubei;
- the rest 16% owes to the **direct effect** of Hubei lockdowns since infected people were constrained in Hubei after the lockdown.
- Therefore, the lockdown is effective in reducing the COVID-19 spread.



# On the Level of Individual City

- The total indirect effect of the lockdown ($IE$) can be further broken down by cities' direct effects ($DM$s);
- $DM$: the intensity of transmission within a particular city;
- $IM$: the secondary migration from a particular city to other places;
- A positive effect means spreading virus while negative means control.
- Cities are ordered by their **cumulative MSI** during the data period.



# Spreading Network


Figure 3: The spreading network among selected cities (Beijing, Shanghai, Guangzhou, Shenzhen, Chengdu, Chongqing, Zhengzhou, Changsha, and Xinyang).

# Thank You!

Figure 4: QR code for our Github of ANOCE-CVAE (https://github.com/anoce-cvae/ANOCE-CVAE).

# Algorithm: Analysis of Causal Effects via Constrained VAE (ANOCE-CVAE)

**Global**: Dataset $X = \{A, M, Y\}$, sample size $n$, dimension of mediators $p$, max iteration $K$, number of epoch $H$, original learning rate $r_0$, tolerance of constrain to zero $\delta$, parameter update bound $U$, tuning parameters $\rho$ and $\omega$, and penalty terms $c$ and $d$;

**Local**: mean and standard variance of $\epsilon$ $\mu_\epsilon$ and $\sigma_\epsilon$, mean and standard variance of $X$ $\mu_X$ and $\sigma_X$, weights in multilayer perceptrons of encoder and decoder $\theta = \{W^{(1)}, W^{(2)}, W^{(3)}, W^{(4)}\}$, Lagrange multipliers $\lambda_1$ and $\lambda_2$, penalty terms $c$ and $d$, matrix $B_{(p+2)\times(p+2)}$, Loss $L$, old and new value for first constraint $h_1^{\text{old}}$ and $h_1^{\text{new}}$, for second constraint $h_2^{\text{old}}$ and $h_2^{\text{new}}$, and learning rate $r$;

**Output**: estimated matrix $\widehat{B}$, total effect $TE$, natural direct and indirect effect $DE$ and $IE$, natural direct and indirect effect for mediator $DM$ and $IM$.

# Part One: Generate $\widehat{B}$ via Constrained VAE

I. Initialization: $\lambda_1 \leftarrow 0$; $\lambda_2 \leftarrow 0$; $c \leftarrow 1$; $d \leftarrow 1$; $r \leftarrow r_0$; $B = \mathbf{0}_{(p+2)\times(p+2)}$;
$h_1^{\text{old}} \leftarrow \infty$; $h_2^{\text{old}} \leftarrow \infty$;

II. For step $k$, $k = 1, \cdots, K$:

A. While $c \times d < U$ :

a). For epoch $i$, $i = 1, \cdots, H$:

1. Build Encoder $(\mu_\epsilon, \sigma_\epsilon) \leftarrow (I_{p+2} - B^\top) MLP\{X, W^{(1)}, W^{(2)}\}$;
2. Build Decoder $(\mu_X, \sigma_X) \leftarrow MLP\{(I_{p+2} - B^\top)^{-1}\epsilon, W^{(3)}, W^{(4)}\}$;
3. Calculate values of constrain functions $h_1^{\text{new}} \leftarrow h_1(B)$ and $h_2^{\text{new}} \leftarrow h_2(B)$, and the loss function $L \leftarrow L_{c,d}(B, W^{(1)}, W^{(2)}, W^{(3)}, W^{(4)}, \lambda_1, \lambda_2)$;
4. Use backward to update parameters $\{B, W^{(1)}, W^{(2)}, W^{(3)}, W^{(4)}\}$;
5. Update learning rate $r$;

b). If $h_1^{\text{new}} > \rho h_1^{\text{old}}$ and $h_2^{\text{new}} > \rho h_2^{\text{old}}$: $c \leftarrow c \times \omega$; $d \leftarrow d \times \omega$;
Elseif $h_1^{\text{new}} > \rho h_1^{\text{old}}$ and $h_2^{\text{new}} < \rho h_2^{\text{old}}$: $c \leftarrow c \times \omega$;
Elseif $h_1^{\text{new}} < \rho h_1^{\text{old}}$ and $h_2^{\text{new}} > \rho h_2^{\text{old}}$: $d \leftarrow d \times \omega$;
Else: Break;

B. $h_1^{\text{old}} \leftarrow h_1^{\text{new}}$; $h_2^{\text{old}} \leftarrow h_2^{\text{new}}$; $\lambda_1 \leftarrow \lambda_1 \times h_1^{\text{new}}$; $\lambda_2 \leftarrow \lambda_2 \times h_2^{\text{new}}$;

C. If $h_1^{\text{new}} < \delta$ and $h_2^{\text{new}} < \delta$: Break;

III. Output $\widehat{B} \leftarrow B$;

# Part Two: Estimate Causal Effects based on $\widehat{B}$

I. According to Equation (1):

A. Get $\widehat{\gamma}$ as the direct effect $DE$;

B. Get $\widehat{\alpha}$ as the effect of $A$ on $M$, $\widehat{\beta}$, and the inside matrix $\widehat{B}_M$;

II. Get $\widehat{\zeta} \equiv (I_p - B_M^\top)^{-1}\widehat{\alpha}$ that represents the causal effect of $A$ on $M$;

III. Get $\widehat{\beta}^\top\widehat{\zeta}$ that represents the total natural indirect effect $IE$;

For each mediator $M_i$, $i = 1, \cdots, p$:

Define the natural direct effect for $M_i$ as $DM[i] = \widehat{\alpha}[i]\widehat{\zeta}[i]$;

IV. Get the natural indirect effect for mediator:

For each mediator $M_i$, $i = 1, \cdots, p$:

A. Delete $M_i$ from the matrix $\widehat{B}$ and get $\widehat{B_i'}$;

B. Repeat step II. with reduced matrix $\widehat{B_i'}$ and get $\widehat{\beta'}$ and $\widehat{\zeta'}$;

C. Calculate the effect difference as the total mediation effect $\widehat{\beta}^\top\widehat{\zeta} - \widehat{\beta'}^\top\widehat{\zeta'}$

D. Define the natural indirect effect for $M_i$ as $IM[i] = \{\widehat{\beta}^\top\widehat{\zeta} - \widehat{\beta'}^\top\widehat{\zeta'}\} - DM[i]$;

V. Define the total effect $TE = \widehat{\gamma} + \widehat{\beta}^\top\widehat{\zeta}$.

Back

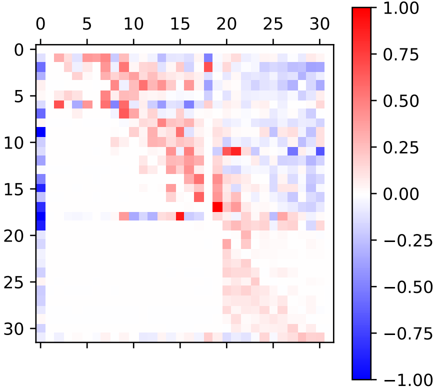
# Results under Different Causal Discovery Methods

Evaluation metrics: the false discovery rate (**FDR**), the true positive rate (**TPR**), and the structural Hamming distance (**SHD**).

| Methods | Case | ER1 | ER2 | ER4 | SF1 | SF2 | SF4 |
|---|---|---|---|---|---|---|---|
| ANOCE-CVAE | FDR | 0.00 (0.14) | 0.00 (0.09) | 0.14 (0.03) | 0.21 (0.06) | 0.17 (0.05) | 0.13 (0.06) |
| | TPR | 0.50 (0.00) | 1.00 (0.08) | 0.93 (0.05) | 1.00 (0.03) | 0.96 (0.05) | 0.81 (0.06) |
| | SHD | 5.00 (1.01) | 0.00 (3.57) | 8.00 (2.44) | 3.00 (1.49) | 6.00 (2.15) | 10.00 (2.65) |
| PC | FDR | 0.00 (0.10) | 0.50 (0.04) | 0.23 (0.05) | 0.00 (0.01) | 0.29 (0.04) | 0.27 (0.05) |
| | TPR | 0.40 (0.01) | 0.26 (0.01) | 0.41 (0.04) | 1.00 (0.00) | 0.46 (0.02) | 0.34 (0.02) |
| | SHD | 6.00 (0.55) | 19.00 (0.51) | 29.00 (2.01) | 0.00 (0.17) | 19.00 (1.19) | 25.00 (1.18) |
| ICA-LiNGAM | FDR | 0.00 (0.18) | 0.08 (0.16) | 0.15 (0.12) | 0.00 (0.14) | 0.00 (0.18) | 0.00 (0.14) |
| | TPR | 0.40 (0.12) | 0.52 (0.14) | 0.41 (0.10) | 0.64 (0.17) | 0.50 (0.17) | 0.47 (0.13) |
| | SHD | 6.00 (1.50) | 12.00 (4.08) | 26.00 (4.67) | 4.00 (2.98) | 13.00 (6.29) | 17.00 (5.36) |
| NOTEARS | FDR | 0.00 (0.02) | 0.00 (0.06) | 0.04 (0.06) | 0.00 (0.00) | 0.00 (0.07) | 0.04 (0.03) |
| | TPR | 0.50 (0.00) | 0.78 (0.09) | 0.63 (0.08) | 1.00 (0.00) | 0.58 (0.08) | 0.72 (0.07) |
| | SHD | 5.00 (0.10) | 5.00 (2.29) | 15.00 (3.90) | 0.00 (0.00) | 11.00 (2.82) | 9.00 (2.59) |
| DAG-GNN | FDR | 0.29 (0.07) | 0.15 (0.06) | 0.13 (0.04) | 0.29 (0.07) | 0.13 (0.06) | 0.11 (0.05) |
| | TPR | 0.50 (0.00) | 0.74 (0.06) | 0.80 (0.06) | 0.91 (0.06) | 0.77 (0.07) | 0.75 (0.05) |
| | SHD | 7.00 (0.49) | 9.00 (2.22) | 12.00 (3.08) | 5.00 (1.66) | 9.00 (2.23) | 10.00 (1.71) |

# Check the Reasonability by Estimated Adjacency Matrix

The first node (indexed 0) represents the Hubei lockdowns, the last node (indexed 31) is the increasing rate out of Hubei, and the middle 30 nodes (indexed 1-30) correspond to 30 selected cities.



# Interactive Graph of cities' $DM$s on the Chinese map

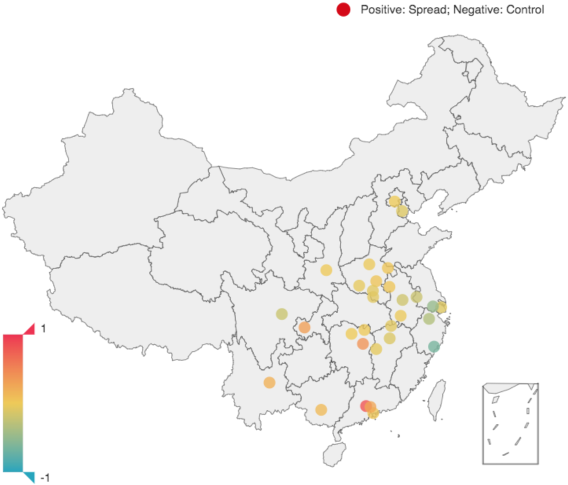



Figure 5: The estimated direct effect of cities ($DM$s).

Chakrabortty, A., Nandy, P. & Li, H. (2018), 'Inference for individual mediation effects and interventional effects in sparse high-dimensional causal graphical models', *arXiv preprint arXiv:1809.10652* .

Kalisch, M. & Bühlmann, P. (2007), 'Estimating high-dimensional directed acyclic graphs with the pc-algorithm', *Journal of Machine Learning Research* **8**(Mar), 613–636.

Kingma, D. P. & Ba, J. (2014), 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980* .

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G. & Lessler, J. (2020), 'The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application', *Annals of internal medicine* .

Maathuis, M. H., Kalisch, M., Bühlmann, P. et al. (2009), 'Estimating high-dimensional intervention effects from observational data', *The Annals of Statistics* **37**(6A), 3133–3164.

Nandy, P., Maathuis, M. H., Richardson, T. S. et al. (2017), 'Estimating the effect of joint interventions from observational data in sparse high-dimensional settings', *The Annals of Statistics* **45**(2), 647–674.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. (2017), 'Automatic differentiation in pytorch'.

Pearl, J. et al. (2009), 'Causal inference in statistics: An overview', *Statistics surveys* **3**, 96–146.

Ramsey, J., Glymour, M., Sanchez-Romero, R. & Glymour, C. (2017), 'A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images', *International journal of data science and analytics* **3**(2), 121–129.

Shah, R. D. & Peters, J. (2018), 'The hardness of conditional independence testing and the generalised covariance measure', *arXiv preprint arXiv:1804.07203* .

Shimizu, S., Hoyer, P. O., Hyvärinen, A. & Kerminen, A. (2006), 'A linear non-gaussian acyclic model for causal discovery', *Journal of Machine Learning Research* **7**(Oct), 2003–2030.

Spirtes, P., Glymour, C., Scheines, R., Kauffman, S., Aimale, V. & Wimberly, F. (2000), 'Constructing bayesian network models of gene expression networks from microarray data'.

VanderWeele, T. & Vansteelandt, S. (2014), 'Mediation analysis with multiple mediators', *Epidemiologic methods* **2**(1), 95–115.

Vansteelandt, S. & Daniel, R. M. (2017), 'Interventional effects for mediation analysis with multiple mediators', *Epidemiology (Cambridge, Mass.)* **28**(2), 258.

Yu, Y., Chen, J., Gao, T. & Yu, M. (2019), 'Dag-gnn: Dag structure learning with graph neural networks', *arXiv preprint arXiv:1904.10098* .

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.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X. et al. (2020), 'Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study', *The Lancet* .