KTH ROYAL INSTITUTE
OF TECHNOLOGY

Optimal Transport for Causal Discovery

Ruibo Tu, Kun Zhang, Hedvig Kjellström, Cheng Zhang

Part I: Introduction of causal discovery

- Limits of traditional methods
- Functional causal models (more recent), their advantages, and limits

Limits of traditional methods

X	Y	Z	W
1.5	2.4	9	8.1
4.5	4.3	3.2	1.9
...	...	...	...
9.6	7.5	4.7	8.8

The bivariate case: determine the causal direction

Sunrise time Wake-up time

Determine the causal direction purely based on the observational data?
PC? GES?

To further identify the causal direction:

Functional causal models (FCMs):

- Assumptions: functional forms and data distributions
- Only in the causal direction, they are satisfied

Functional Causal Models

$y = f(x, \epsilon), \quad X \perp \epsilon$

- Effect is a function of the direct cause and the noise
- Independent noise

Non-identifiability of the general form:
Always construct a model in the reverse causal direction which satisfies the FCM assumptions

$x = \tilde{f}\,(y, \tilde{\epsilon}), \quad Y \perp \tilde{\epsilon}$

Require more proper assumptions

Hyvärinen, Aapo, and Petteri Pajunen. "Nonlinear independent component analysis: Existence and uniqueness results." Neural networks 12, no. 3 (1999): 429-439.
Zhang, Kun, Zhikun Wang, Jiji Zhang, and Bernhard Schölkopf. "On estimation of functional causal models: general results and application to the post-nonlinear causal model." ACM Transactions on Intelligent Systems and Technology (TIST) 7, no. 2 (2015): 1-22.

Identifiable Functional Causal Models

Data generated by $Y = aX + E$ (i.e., $X \rightarrow Y$):

Adjusted from Kun Zhang's tutorial

Shimizu, Shohei, Patrik O. Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. "A linear non-Gaussian acyclic model for causal discovery." Journal of Machine Learning Research 7, no. 10 (2006).

Identifiable Functional Causal Models

Additive noise models:

$y = f(x) + \epsilon, \qquad X \perp \epsilon$

Post-nonlinear models:

$y = f(g(x) + \epsilon), \quad X \perp \epsilon$

FCM-based methods

- require additional technical conditions;
- **Sensitive to the model assumptions**.
- **No good understanding of the technical assumptions**
- **No direction to further relax the assumptions**

Hoyer, Patrik O., Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. "Nonlinear causal discovery with additive noise models." In NIPS, vol. 21, pp. 689-696. 2008.
Zhang, K., and A. Hyvärinen. "On the Identifiability of the Post-Nonlinear Causal Model." In 25th Conference on Uncertainty in Artificial Intelligence (UAI 2009), pp. 647-655. AUAI Press, 2009.

Part II: Optimal transport for Causal discovery

Sunrise time Wake-up time

ID	S	W
1	04:00	04:15
2	04:20	05:00
…	…	…
N	09:00	10:00

Causal relationship

- **(A)** What do the assumptions of FCMs mean in an application?
- **(B)** How to improve the FCM methods?

Part II: Optimal transport for Causal discovery

Part II: Optimal transport for Causal discovery

Why construct dynamical systems for FCMs?

- Natural and straightforward **interpretation**
- **Freedom** for modeling the causal influences
- **New direction** for causal discovery methods

(A) Re-formulation of the general FCM form

$$\mathrm{x}_T = \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} E_x \\ f(X, E_y) \end{pmatrix} = M\left(\begin{pmatrix} E_x \\ E_y \end{pmatrix} \right) = M(\mathrm{x}_0) \tag{1}$$

Structural constraints:

i. *Map constraint:* $X = E_x$ while Y is represented by a function of X and E_y;
ii. *Independence constraint:* $E_x \perp\!\!\!\perp E_y$.

(A) An interesting view of FCMs

Observed variables

Independent noise

$$x_T = \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} E_x \\ f(X, E_y) \end{pmatrix} = M\left(\begin{pmatrix} E_x \\ E_y \end{pmatrix} \right) = M(x_0)$$

the transformation from the independent **noise** to the observational **data**

observational **data**

Dynamical process

← ? ← ? ← ? ← ?

independent **noise**

(A) To find suitable dynamical systems for FCMs

$$x_T = \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} E_x \\ f(X, E_y) \end{pmatrix} = M\left(\begin{pmatrix} E_x \\ E_y \end{pmatrix} \right) = M(x_0)$$

Find the dynamical systems which can be used for describing FCMs and satisfy

- FCM **constraints**: mapping M and independent noise

Underdetermined problem:

(i) Many dynamical systems satisfy the requirement Solve now in **(A)**

(ii) No access of the noise $\mathbf{x}_0$

Solve later in **(B)**

(A) Optimal transport for the underdetermined problem

Find the dynamical systems which can be used for describing FCMs and consistent with

- The FCM **constraints**: mapping M and independent noise

Suppose that we have the data of noise Ex and Ey

1. Define a proper cost function
2. Select the unique dynamical process

FCM-Constrained optimal transport problem

(A) Optimal Transport (OT)

Find couplings

1. Move samples x_0 from p_0 to samples x_T from p_T
2. The cost of moving all samples is minimum.

The solution: $x_T = M^*(x_0)$ with the smallest distance.

(A) L^p Wasserstein distance

$$c(\mathbf{x}_0, \mathbf{x}_T) = |\mathbf{x}_T - \mathbf{x}_0|^p$$

Distance^p between x_T and x_0

Prob. distribution of x_0

L^p Wasserstein distance:

$$W_p(\rho_0, \rho_T)^p = \inf_M \int |M(x_0) - x_0|^p \rho_0(x_0) dx_0$$

$$\text{s.t.} \quad \rho_0(x_0) = \rho_T(M(x_0)) \det(\nabla M(x_0))$$

Change of variable formula

(A) L² Wasserstein distance and Pressureless potential flow

Solve the L2 Wasserstein distance problem

- Find couplings
- Equivalent to find the dynamical system with the least action
- Considering time, the trajectories of samples are from a pressureless potential flow.

The time evolution and the solution: A pressureless potential flow

$$\begin{aligned} x_t &= x_0 + \frac{t}{T} v(t, x_t); \\ v(t, x_t) &= M^*(x_0) - x_0, \ t \in [0, T]. \end{aligned}$$

Benamou, Jean-David, and Yann Brenier. "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem." Numerische Mathematik 84, no. 3 (2000): 375-393.

(A) Dynamical L² Wasserstein distance

Computing the least action in fluid mechanics

Density at time t

Square of the velocity at time t at position x_t

$$W_2^2(\rho_0, \rho_T) = \inf_{\rho, \mathrm{v}} T \int \int_0^T \rho(t, \mathrm{x}_t) |\mathrm{v}(t, \mathrm{x}_t)|^2 dt d\mathrm{x}_t,$$

$$\text{s.t.} \quad \begin{cases} \rho(0, \cdot) = \rho_0, \rho(T, \cdot) = \rho_T & \text{Boundary conditions} \\ \partial_t \rho + \nabla \cdot (\rho \mathrm{v}) = 0 & \text{Continuity condition} \end{cases}$$

(A) Other dynamical systems and definitions of optimal transport

- L^2 Wasserstein distance ➔ the pressureless potential flows
- the pressureless potential flows ➔ L^2 Wasserstein distance
- the (in)compressible pressureless potential flows ?➔ other definitions of optimal transport

Benamou, Jean-David, and Yann Brenier. "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem." Numerische Mathematik 84, no. 3 (2000): 375-393.

(A) FCM-constrained L² Wasserstein distance

$$x_T = \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} E_x \\ f(X, E_y) \end{pmatrix} = M\left(\begin{pmatrix} E_x \\ E_y \end{pmatrix} \right) = M(x_0)$$

Find the dynamical systems which can be used for describing FCMs and consistent with

- The FCM constraints: mapping M and independent noise
- **Cost function:** Suppose that the least action principle is satisfied
- Suppose that the noise data are available

  Solve later in **(B)**

Solving such **FCM-constraine**d L² Wasserstein distance

- Finding the unique dynamical system for a FCM

(A) FCM-constrained L^2 Wasserstein distance

$$x_T = \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} E_x \\ f(X, E_y) \end{pmatrix} = M\left(\begin{pmatrix} E_x \\ E_y \end{pmatrix} \right) = M(x_0) \tag{1}$$

Structural constraints:

1. *Map constraint:* $X = E_x$ while Y is represented by a function of X and E_y;
2. *Independence constraint:* $E_x \perp\!\!\!\perp E_y$.

FCM constraints

Proposition 1. *Under constraints* (I) *and* (II), *the square of the* L^2 *Wasserstein distance between* p_0 *and* p_T *is*

$$W_2^2(p_0, p_T) = \mathbb{E}_{E_x}\left[W_2^2\left(p(E_y), p(Y|E_x)\right)\right]. \tag{7}$$

- Bivariate causal discovery is a 2D problem.
- The 2D optimal transport problem can be solved by solving the 1D problem.
- Efficiency (sorting operation empirically): Closed-form solution for the 1D optimal transport problem

(A) Couplings and velocity fields given by the constrained L2 Wasserstein distance

The time evolution and the solution: A pressureless potential flow

$$
\begin{aligned}
x_t &= x_0 + \frac{t}{T} v(t, x_t); \\
v(t, x_t) &= M^*(x_0) - x_0, \; t \in [0, T].
\end{aligned}
$$

(A) Dynamical interpretation of FCMs

A pressureless flow

- Only move up and down

Functional causal models:

$Y = f(X, E_y)$ $\quad X \perp E_y$

(A) Dynamical interpretation of ANMs

A pressureless flow
- Only move up and down
- **Volume-preserving**
- Parallelly move

Additive noise models:

$Y = f(X) + E_y \qquad X \perp E_y$

(A) ANMs: Divergence of the velocity fields

Corollary: the corresponding dynamical systems of ANMs

Under constraints (I)—(III), the probability density $p(t, \mathrm{x}_t)$ is time-invariant and $\mathrm{div}\ \mathrm{v}(t, \mathrm{x}_t) = 0$, $\forall t \in [0, T]$, $\mathrm{x}_t \in \mathbb{R}^2$, which leads to a volume-preserving pressureless potential flow.

- $\mathrm{v}(t, \mathrm{x}_t) = [v_x(t, x_t), v_y(t, y_t)]' = \mathrm{v}(0, \mathrm{x}_0) = M^*(\mathrm{x}_0) - \mathrm{x}_0$ and $t \in [0, T]$.
- $\mathrm{div}\ \mathrm{v}(t, \mathrm{x}_t) = \frac{\partial v_x(t, x_t)}{\partial x_t} + \frac{\partial v_y(t, y_t)}{\partial y_t} = 0, \forall t \in [0, T]$, $x_t, y_t \in \mathbb{R}$.
- Relationship between $p(t, \mathrm{x}_t)$ and $\mathrm{div}\ \mathrm{v}(t, \mathrm{x}_t)$ (Eqn. (4) in Appendix).

(B) Causal discovery criterion for ANMs

Proposition: Criterion for determining causal direction in ANMs

Under the conditions of identifiable ANMs and contraints (I)—(III), $D(\mathrm{v})$ or $D_{\mathtt{var}}(\mathrm{v})$ is equal to 0 iif X is the direct cause of Y.

- $D(\mathrm{v}) = \mathbb{E}_{\mathrm{x}_0}[\,|\mathrm{div}\ \mathrm{v}(0, \mathrm{x}_0)|^2]$ or $D_{\mathtt{var}}(\mathrm{v}) = \mathbb{E}_X[\mathbb{V}[V_{y|X}|X]]$ (variance based).
- $\mathrm{v}(t, \mathrm{x}_t) = [v_x(t, x_t), v_y(t, y_t)]' = \mathrm{v}(0, \mathrm{x}_0) = M^*(\mathrm{x}_0) - \mathrm{x}_0$ and $t \in [0, T]$.

Given by solving the constrained L2 Wasserstein distance

(B) Compute the divergence measure

$$D(\mathbf{v}) = \int_{\mathbb{R}^2} |div\ \mathbf{v}|^2 p_0(\mathbf{x}_0) d\mathbf{x}_0 = \mathbb{E}_{\mathbf{x}_0}[\,|div\ \mathbf{v}|^2],$$

$$D_{\mathtt{var}}(\mathbf{v}) = \mathbb{E}_X[\mathbb{V}[V_{y|X}|X]],$$

$$D_{\mathtt{var}}(\mathbf{v}) \approx \frac{1}{N} \sum_{x \in \{x_i\}_N} \frac{\left\|\mathrm{sort}(\overrightarrow{y_x}) - \mathrm{sort}(\overrightarrow{e_y}) - \mathrm{ave}(\overrightarrow{y_x} - \overrightarrow{e_y})\right\|_2^2}{N_x - 1},$$

E_y?

1. Parameterize a general distribution as the unknown noise distribution
2. Estimate the parameters by minimizing the divergence measure

   Only the ground-true distribution gives the zero divergence measure

$$e_y^i = f_\theta^{\mathtt{noise}}(e_y^{\mathtt{source}}) = \theta \times e_y^{\mathtt{source}} \text{ and } e_y^{\mathtt{source}} \sim \mathcal{N}(0,1)/\mathcal{U}(0,1),$$

Optimize over the parameters of the noise distribution

Algorithm 1: DIVOT: divergence measure with optimal transport for causal direction determination.

Input: data $\{(x_i, y_i)\}$ and noise distribution $p(E; \theta_0)$
Output: $X \to Y$ or $Y \to X$.

Def `DIVOT`$(\{(x_i, y_i)\}, p(E; \theta_0))$ **:**
  data, $\theta = \{(x_i, y_i)\}, \theta_0$
  $\mathcal{L}_{X \to Y}$ = Div(data, $p(E; \theta)$)
  data, $\theta = \{(y_i, x_i)\}, \theta_0$
  $\mathcal{L}_{Y \to X}$ = Div(data, $p(E; \theta)$)
if $\mathcal{L}_{Y \to X} < \mathcal{L}_{X \to Y}$ **then**
  return $Y \to X$
else
  return $X \to Y$

Def `Div`(*data*, $p(E; \theta)$) **:**
  noise = Sampling($p(E; \theta)$)
  data^s, noise^s = OT(data, noise)
  loss = $\min_\theta$ VarDiv(data^s, noise^s)
  return loss/Var($p(E; \theta^*)$)

(B) The divergence measure for causal discovery

Experimental results

Figure: Performance on synthetic data generated from different ANMs with different sample size.

Experimental results: Robustness

Figure: Robustness to prior misspecification: The performance of DIVOT with different hypothesized distributions of $p(E_y; \theta)$ on the datasets with uniform distribution noise $E_y \sim \mathcal{U}(0, 1)$.

Experimental results: Tuebingen cause-effect pair datasets

Table 1: Accuracy (%) of recovering the true causal direction on the Tübingen cause-effect pair dataset (Mooij et al., 2016) compared with the reported results of (Stegle et al., 2010) in the upper table and (Zhang et al., 2015a) in the lower table.

Ours (PNL)	Ours (ANM)	LiNGAM	ANM-Gauss	ANM-MML	ANM-HSIC	PNL	GPI-HSIC	GPI-MML	IGCI
78.9±3.9	71±4	62±3	45±3	68±1	68±3	67±4	62±4	72±2	76±1

Ours (PNL)	Ours (ANM)	ANM	WGP-Gauss	WGP-MoG	PNL-MLP	GPI	IGCI
76.6±4.5	67±3	63	67	73	70	72	73

🎉 Thank you for listening 🍾