構造的因果モデルの類似性に基づくドメイン適応

手嶋毅志 (東京大学)

2020 年 3 月 29 日 @ Mathiine-learning

Today's content

- Exploit causal models for transfer learning:

  Few-shot domain adaptation by causal mechanism transfer.

  arXiv:2002.03497. Teshima, T., Sato, I., and Sugiyama, M.,

今日の内容とゴール

- 先日投稿した論文の，理論保証を中心にお話しします
- そこでまずは「どのようなアルゴリズムの理論保証をするか」を伝えることまでをゴールにして研究の全体像を（動機から）お話しします

Contents

1. Part I: Modeling Causality (10min)

2. Part II: Our Research (15min)

3. Theoretical analysis: details

Part I

- Our research is built on models of causality.
- Part I briefly introduce the topic of causality starting from its motivation.

Motivation: Correlation vs. causation

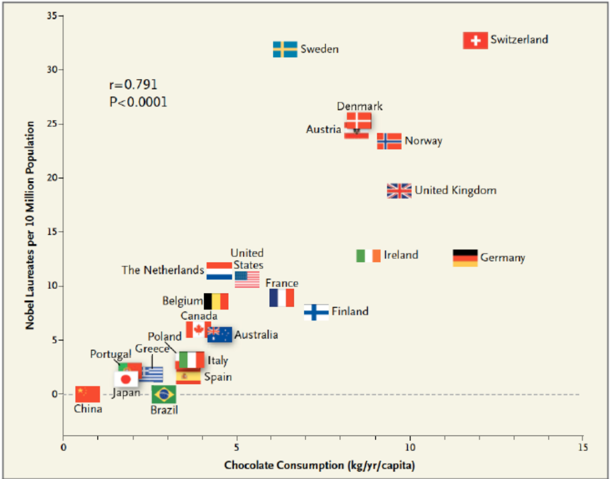


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Figure: [1]

Motivation: Correlation vs. causation

- 🤔 says: Let's eat more chocolate!
- We say: Wait! It's just correlation. Not causation!
- What's causality?

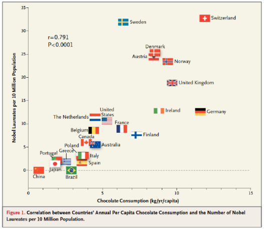
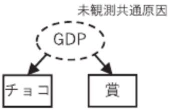


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Figure: [1]

Motivation: Correlation vs. causation

- 🤔 says: Let's eat more chocolate!
- We say: Wait! It's just correlation. Not causation!
- What's causality?



- The difference of causation vs. association (correlation) appears when we intervene in a system.

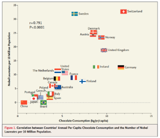


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Figure: [1]

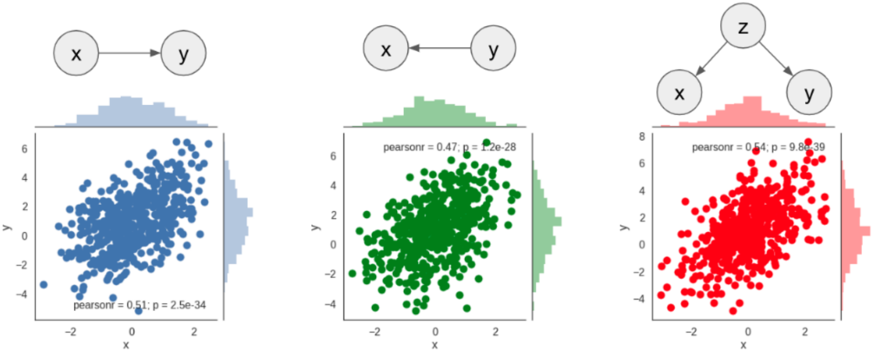
- Intervention = Manipulate a random variable (e.g., fixing its value, etc.)

Motivation: Intervention

```
x = randn()
y = x + 1 + sqrt(3)*randn()
```

```
y = 1 + 2*randn()
x = (y-1)/4 + sqrt(3)*randn()/2
```

```
z = randn()
y = z + 1 + sqrt(3)*randn()
x = z
```



- Different ways to generate the same joint distribution.

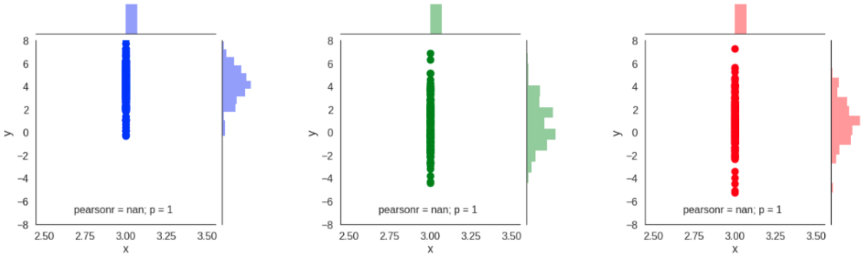
Figure: [2]

Motivation: Intervention

```
x = randn()
x = 3
y = x + 1 + sqrt(3)*randn()
x = 3
```

```
y = 1 + 2*randn()
x = 3
x = (y-1)/4 + sqrt(3)*randn()/2
x = 3
```

```
z = randn()
x = 3
x = z
x = 3
y = z + 1 + sqrt(3)*randn()
x = 3
```



- Different behavior under intervention $\mathrm{do}(X = 3)$

Figure: [2]

Modeling causality: SCMs

Structural Equation Models (SEMs) a.k.a. Structural Causal Models (SCMs) [3]

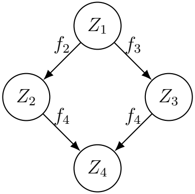
- An SEM is a tuple $(q, \mathcal{G}, \mathcal{F})$, which defines a distribution over random variables $\{Z_i\}_{i=1}^D$.

For a more formal definition, see [4].

Distribution q of independent random variables $\{S_i\}_{i=1}^D$

$$S \sim q(S_1, S_2, S_3, S_4) = \prod_{d=1}^{4} q_d(S_d)$$

Directed acyclic graph (DAG) $\mathcal{G}$ whose vertex set is $\{Z_i\}_{i=1}^D$



Functions $\mathcal{F} = \{f_d\}_{d=1}^D$

$$Z_d = f_d(Z_{\mathrm{Pa}_{\mathcal{G}}(d)}, S_d)$$

$$\begin{cases} Z_1 &= f_1(S_1), \\ Z_2 &= f_2(Z_1, S_2), \\ Z_3 &= f_3(Z_1, S_3), \\ Z_4 &= f_4(Z_2, Z_3, S_4). \end{cases}$$

Modeling causality: Intervention

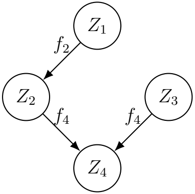
Perfect interventions [3] $\mathrm{do}(Z_I = \zeta_I)$

- Perfect intervention enforces Z_I to attain value ζ_I.
- This changes the SCM $(q, \mathcal{G}, \mathcal{F})$ into $(q, \mathcal{G}', \mathcal{F}')$ w/

$$f'_d(Z_{\mathrm{Pa}_{\mathcal{G}}(d)}, S_d) = \begin{cases} \zeta_d & \text{if } d \in I \\ f_d(Z_{\mathrm{Pa}_{\mathcal{G}}(d)}, S_d) & \text{if } d \notin I \end{cases}$$

- In the graph, all edges incoming to Z_I are removed.

$$\begin{cases} Z_1 &= f_1(S_1), \\ Z_2 &= f_2(Z_1, S_2), \\ Z_3 &= \zeta_3, \\ Z_4 &= f_4(Z_2, Z_3, S_4). \end{cases}$$



SEM: Formal definition [4]

Definition ([Wright, 1921], [Pearl, 2000], [Bongers et al., 2018])

A Structural Causal Model (SCM), also known as Structural Equation Model (SEM), is a tuple $\mathcal{M} = \langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{E}}, \boldsymbol{f}, \mathbb{P}_{\boldsymbol{\mathcal{E}}} \rangle$ with:

1. a product of standard measurable spaces $\boldsymbol{\mathcal{X}} = \prod_{i \in \mathcal{I}} \mathcal{X}_i$ (domains of the endogenous variables)
2. a product of standard measurable spaces $\boldsymbol{\mathcal{E}} = \prod_{j \in \mathcal{J}} \mathcal{E}_j$ (domains of the exogenous variables)
3. a measurable mapping $\boldsymbol{f} : \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{E}} \to \boldsymbol{\mathcal{X}}$ (the causal mechanism)
4. a product probability measure $\mathbb{P}_{\boldsymbol{\mathcal{E}}} = \prod_{j \in \mathcal{J}} \mathbb{P}_{\mathcal{E}_j}$ on $\boldsymbol{\mathcal{E}}$ (the exogenous distribution)

Definition

A pair of random variables $(\boldsymbol{X}, \boldsymbol{E})$ is a solution of SCM $\mathcal{M}$ if $\mathbb{P}^{\boldsymbol{E}} = \mathbb{P}_{\boldsymbol{\mathcal{E}}}$ and the structural equations $\boldsymbol{X} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{E})$ hold a.s..

SEM: Formal definition [4]

Definition

The components of the causal mechanism usually do not depend on *all* variables: for $i \in \mathcal{I}$,

$$X_i = f_i(\boldsymbol{x}_{\mathrm{pa}_i^{\mathcal{I}}}, \boldsymbol{e}_{\mathrm{pa}_i^{\mathcal{J}}})$$

where f_i only depends on $\mathrm{pa}_i^{\mathcal{I}} \subseteq \mathcal{I}$ (the endogenous parents of i) and $\mathrm{pa}_i^{\mathcal{J}} \subseteq \mathcal{J}$ (the exogenous parents of i).

Definition

The augmented graph $\mathcal{G}^a(\mathcal{M})$ of SCM $\mathcal{M}$ is a directed graph with nodes $\mathcal{I}\dot{\cup}\mathcal{J}$ and an edge $k \to i$ iff $k \in \mathrm{pa}_i^{\mathcal{I}}\dot{\cup}\mathrm{pa}_i^{\mathcal{J}}$ is a parent of $i \in \mathcal{I}$.

Definition

The graph $\mathcal{G}(\mathcal{M})$ of SCM $\mathcal{M}$ is a DMG with nodes $\mathcal{I}$, directed edges $k \to i$ iff $k \in \mathrm{pa}_i^{\mathcal{I}}$, and bidirected edges $k \leftrightarrow i$ iff $\mathrm{pa}_i^{\mathcal{J}} \cap \mathrm{pa}_k^{\mathcal{J}} \neq \emptyset$.

SEM: Formal definition [4]

Definition

An SCM $\mathcal{M}$ is said to be uniquely solvable w.r.t. $\mathcal{O} \subseteq \mathcal{I}$ if there exists a measurable mapping $\boldsymbol{g}_{\mathcal{O}} : \boldsymbol{\mathcal{X}}_{(\mathrm{pa}_{\mathcal{H}}(\mathcal{O})\setminus\mathcal{O})\cap\mathcal{I}} \times \boldsymbol{\mathcal{E}}_{\mathrm{pa}_{\mathcal{H}}(\mathcal{O})\cap\mathcal{J}} \to \boldsymbol{\mathcal{X}}_{\mathcal{O}}$ such that for $\mathbb{P}_{\boldsymbol{\mathcal{E}}}$-almost every $\boldsymbol{e}$ for all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$:

$$\boldsymbol{x}_{\mathcal{O}} = \boldsymbol{g}_{\mathcal{O}}(\boldsymbol{x}_{(\mathrm{pa}_{\mathcal{H}}(\mathcal{O})\setminus\mathcal{O})\cap\mathcal{I}}, \boldsymbol{e}_{\mathrm{pa}_{\mathcal{H}}(\mathcal{O})\cap\mathcal{J}}) \quad \Longleftrightarrow \quad \boldsymbol{x}_{\mathcal{O}} = \boldsymbol{f}_{\mathcal{O}}(\boldsymbol{x}, \boldsymbol{e}).$$

(Loosely speaking: if the structural equations for $\mathcal{O}$ provide a unique solution for $\boldsymbol{x}_{\mathcal{O}}$ in terms of the other variables).

SEM: Formal definition [4]

Definition

We call an SCM $\mathcal{M}$ simple if it is uniquely solvable with respect to any subset $\mathcal{O} \subseteq \mathcal{I}$.

Lemma

If $\mathcal{G}(\mathcal{M})$ is acyclic, $\mathcal{M}$ is simple.

- The class of simple SCMs extends the class of acyclic SCMs by allowing for (weak) cyclic causal relations, while preserving most of the simplicity and convenience of acyclic SCMs.
- The theory for non-simple SCMs is considerably more involved [Bongers et al., 2018].
- Simple SCMs induce modular SCMs (mSCMs) [Forré and Mooij, 2017].

Summary of Part I

- Causal inference requires more information (additional assumptions) than joint distributions of data.
- One of the causal models: SEMs (structural equation models).
- Causal discovery (estimation of SEMs/GCMs) has seen continuous progress in the past decades.

Contents

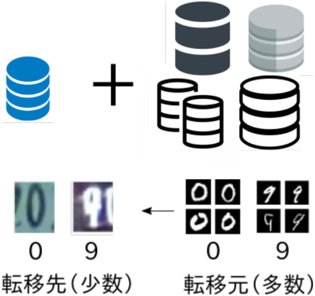
1. Part I: Modeling Causality (10min)

2. Part II: Our Research (15min)

3. Theoretical analysis: details

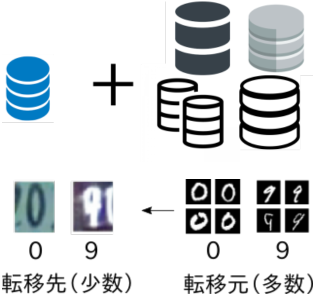
Motivation: Transfer learning

- Data is scarce resource. We want to exploit as much info as possible.
- Use data from related but different prob. distributions = Domain adaptation (DA)



Motivation: Transfer assumption (TA)

- Of course, we need some form of an assumption (transfer assumption; TA) to relate $p_{\mathrm{src}(k)}$ and p_{tar}. What commonality to exploit?
- (Without an assumption, DA cannot be justified)



Motivation: Idea of our TA

- Our TA: Causal mechanism is identical b/w domains.
- Humans care about causality (partially) because, once discovered, it applies to different systems.

Motivating example: Regional disease prediction

- Predict disease risk from medical records. [5]

National reporting of Census Data (1995-2004)
LEGEND
Census conducted
No census conducted
Source: United Nations Demographic Yearbook Database, 2005

- Common pathological mechanism across regions.
- Data distributions may vary for different lifestyles.

Figure: https://slideplayer.com/slide/15283414/

Problem setup: Formulation 1/3

In this work, we focus on regression under...

1. Homogeneous (i.e., all domains in the same space)

   $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^{D-1} \times \mathbb{R}$

2. Multi-source (i.e., we have multiple source domains)

   $\mathcal{D}_k = \{(x_{k,i}, y_{k,i})\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{src}(k)} \ (k = 1, \ldots, K)$ (n_k is large)

3. Few-shot (i.e., labeled but few target domain data)

   $\{(x_{\mathrm{tar},i}, y_{\mathrm{tar},i})\}_{i=1}^{n_{\mathrm{tar}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{tar}}$ (n_{tar} is small)

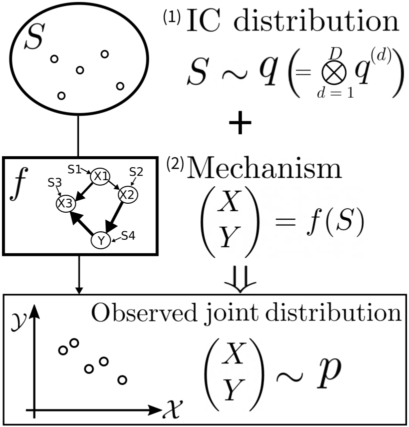
setting.

Goal: accurate predictor for the target distribution

Find $g : \mathbb{R}^{D-1} \to \mathbb{R}$ s.t. $R(g) := \mathbb{E}_{\mathrm{tar}}\ell(g, X, Y)$ is minimal.

ℓ: loss function

Problem setup: Formulation 2/3



[1] IC distribution

$$S \sim q\left(= \bigotimes_{d=1}^{D} q^{(d)}\right)$$

+

[2] Mechanism

$$\begin{pmatrix} X \\ Y \end{pmatrix} = f(S)$$

$$\Downarrow$$

Observed joint distribution

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim p$$

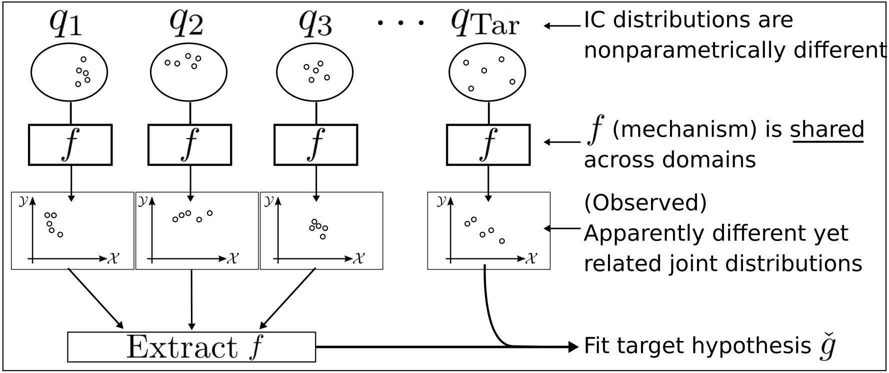
- Relation to structural causal models: Solve SEMs by imputation.

- Nonlinear independent component analysis (NLICA) model.

1. ICs $S = (S^{(d)})_{d \in 1:D}$ are sampled from q.
2. Invertible f transforms S into $(X, Y) = f(S)$.

- Each pair (f, q) defines a joint distribution p.

Problem setup: Formulation 3/3



IC distributions are nonparametrically different

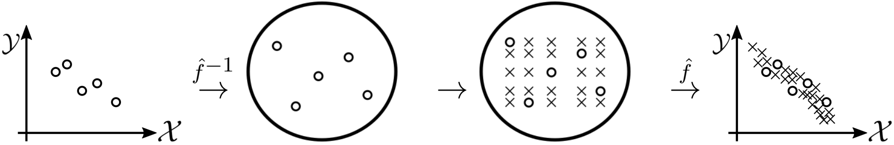
f (mechanism) is shared across domains

(Observed) Apparently different yet related joint distributions

Fit target hypothesis $\check{g}$

- Assumption of common generative mechanism.
- Capture common generative mechanism ⇒ Enable DA among seemingly very different distributions without parametric assumptions.

Proposed method: Algorithm



1. $\hat{f} \leftarrow \mathrm{ICA}(\mathcal{D}_1, \ldots, \mathcal{D}_K)$ NLICA on source domains[1]
2. $\hat{s}_{\mathrm{tar},i} \leftarrow \hat{f}^{-1}(x_{\mathrm{tar},i}, y_{\mathrm{tar},i})$ Extract IC in target domain
3. $\{\bar{s}_j\}_{j=1}^{n_{\mathrm{tar}}^D} \leftarrow \mathrm{Shuffle}(\{\hat{s}_{\mathrm{tar},i}\}_i)$ Shuffle IC of target domain
4. $\{\bar{x}_j, \bar{y}_j\}_{j=1}^{n_{\mathrm{tar}}^D} \leftarrow \hat{f}(\{\bar{s}_j\}_j)$ Get augmented target data[2]
5. $\hat{R}_{\mathrm{aug}}(g) := \frac{1}{n_{\mathrm{tar}}^D} \sum_{j=1}^{n_{\mathrm{tar}}^D} \ell(g, \bar{x}_j, \bar{y}_j)$ Augmented emp. risk
6. $\boxed{\hat{g}_{\mathrm{aug}} \in \mathrm{argmin}_{g \in \mathcal{G}} \hat{R}_{\mathrm{aug}}(g)}$ Train on the augmented data

[1] Multi-source required for nonlinear ICA with generalized contrastive learning.

[2] Inverse is possible if we model $\hat{f}$ by invertible neural networks.

Theoretical analysis: goal

- Understand the statistical properties of the proposed risk estimator

$$\hat{R}_{\mathrm{aug}}(g) := \frac{1}{n_{\mathrm{tar}}^{D}} \sum_{j=1}^{n_{\mathrm{tar}}^{D}} \ell(g, \bar{x}_j, \bar{y}_j)$$

and that of its minimizer $\hat{g}_{\mathrm{aug}}$:

$$R(\hat{g}_{\mathrm{aug}}) - R(g^*).$$

Theoretical analysis: Research questions

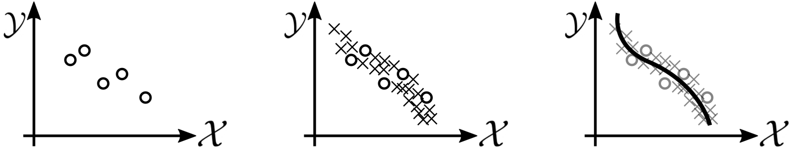
Theoretical Q&A:

Q1. What does it mean to exploit independence?

A1. When $\hat{f} = f$ (ideal case), $\hat{R}_{\mathrm{aug}}(g)$ is the uniformly minimum variance unbiased estimator of the target risk. Essentially, it should help in terms of variance.

Q2. Estimating f induces error. What are the trade-offs?

A2. $\hat{f} \neq f \Rightarrow$ 😄 Mitigate overfitting. 😟 Introduce bias.



Theoretical analysis: Strategy

- Interpret $\hat{R}_{\mathrm{aug}}(g)$ as the von-Mises statistic (process). (When $\hat{f} = f$, it is also the generalized U-statistic.)
- Define $\tilde{\ell}(s_1, \ldots, s_D) = \ell(g, \hat{f}(s_1^{(1)}, \ldots, s_D^{(D)}))$. Then,

$$\hat{R}_{\mathrm{aug}}(g) = \frac{1}{n^D} \sum_{i_1=1}^{n} \cdots \sum_{i_D=1}^{n} \tilde{\ell}(S_{i_1}, \ldots, S_{i_D}).$$

- This is the V-statistic [6] of

$$\check{Q}^D \tilde{\ell} := \int \tilde{\ell}(s_1, \cdots, s_D) \check{q}(s_1) \cdots \check{q}(s_D) \mathrm{d}s_1 \cdots \mathrm{d}s_D.$$

$$\check{Q} := (\hat{f}^{-1} \circ f)_\sharp Q_{\mathrm{Tar}}$$

Theoretical analysis: If $\hat{f} = f$

Q1. What does it mean to exploit independence?

$\mathcal{Q}$: set of independent continuous distributions over $\mathbb{R}^D$.

Theorem: minimum variance property

- Assume $\hat{f} = f$. Then $\hat{R}_{\mathrm{aug}}(g)$ is the (unique) UMVUE (uniformly minimum variance unbiased estimator) of $R(g)$ on $\mathcal{Q}$. That is,
- $\forall \hat{R}(g)$:unbiased, $\forall q \in \mathcal{Q}$, $\mathrm{Var}(\hat{R}_{\mathrm{aug}}(g)) \leq \mathrm{Var}(\hat{R}(g))$
- Special case: $\mathrm{Var}(\hat{R}_{\mathrm{aug}}(g)) \leq \mathrm{Var}(\hat{R}_{\mathrm{ERM}}(g))$

Why? (Details are skipped)

Reinterpret $\hat{R}_{\mathrm{aug}}(g)$ as generalized U-statistic [6] of $R(g)$.

Generalized U-Statistic

Lemma (Generalized U-statistic is UMVUE [7])

Consider a regular statistical functional with kernel $\psi : \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_L} \to \mathbb{R}$*:*

$$\theta(q) := \int \psi(\begin{pmatrix} x_1^{(1)} \\ \vdots \\ x_{k_1}^{(1)} \end{pmatrix}, \ldots, \begin{pmatrix} x_1^{(L)} \\ \vdots \\ x_{k_L}^{(L)} \end{pmatrix}) \prod_{j=1}^{k_1} q_1(x_j^{(1)}) \mathrm{d}x_j^{(1)} \cdots \prod_{j=1}^{k_L} q_L(x_j^{(L)}) \mathrm{d}x_j^{(L)}.$$

Its generalized U-statistic given samples $\{x_i^{(l)}\}_{i=1}^{n_l} \overset{\textit{i.i.d.}}{\sim} q_l$ *is*

$$\mathrm{GU}_{(n_1,\ldots,n_L)}^{(k_1,\ldots,k_L)} \psi := \frac{1}{\prod_l \binom{n_l}{k_l}} \sum_{\mathrm{All}} \psi\left(\left(x_{i_1^{(1)}}^{(1)}, \ldots, x_{i_{k_1}^{(1)}}^{(1)}\right), \ldots, \left(x_{i_1^{(L)}}^{(L)}, \ldots, x_{i_{k_L}^{(L)}}^{(L)}\right)\right).$$

Then, $\mathrm{GU}_{(n_1,\ldots,n_L)}^{(k_1,\ldots,k_L)} \psi$ *is the uniformly minimum variance unbiased estimator of* θ *on* $\mathcal{Q}$*.*

Theoretical analysis: If $\hat{f} \neq f$

Q2. What happens when $\hat{f} \neq f$?

Theorem: generalization error bound [3]

Under appropriate assumptions, with probability at least $1 - (\delta + \delta')$,

$\|\cdot\|_{W^{1,1}}$: $(1,1)$-Sobolev norm

$$R(\hat{g}_{\mathrm{aug}}) - R(g^*)$$
$$\leq \underbrace{C \sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|_{W^{1,1}}}_{\text{Approximation error}} + \underbrace{4D\mathfrak{R}(\mathcal{G}) + 2DB_\ell \sqrt{\frac{\log 2/\delta}{2n}}}_{\text{Estimation error}} + \text{Higher order terms.}$$

[3] This also provides a bound on the negative transfer.

Theoretical analysis: If $\hat{f} \neq f$

Theorem: generalization error bound

$$R(\hat{g}_{\mathrm{aug}}) - R(g^*)$$

$$\leq \underbrace{C \sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|_{W^{1,1}}}_{\text{Approximation error}} + \underbrace{4D\mathfrak{R}(\mathcal{G}) + 2DB_\ell \sqrt{\frac{\log 2/\delta}{2n}}}_{\text{Estimation error}} + \text{Higher order terms.}$$

- Effective Rademacher complexity:

$$\mathfrak{R}(\mathcal{G}) := \frac{1}{n}\mathbb{E}_{\hat{S}}\mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \left| \sum_{i=1}^{n} \sigma_i \mathbb{E}_{S'_2,\ldots,S'_D}[\tilde{\ell}(\hat{s}_i, S'_2, \ldots, S'_D)] \right| \right],$$

  - $\tilde{\ell}(s_1, \ldots, s_D) := \frac{1}{D!}\sum_{\pi \in \mathfrak{S}_D} \ell(g, \hat{f}(s_{\pi(1)}^{(1)}, \ldots, s_{\pi(D)}^{(D)}))$,
  - $\{\sigma_i\}_{i=1}^{n}$: Independent sign variables, $\mathbb{E}_{\hat{S}}$: Expectation w.r.t. $\{\hat{s}_i\}_{i=1}^{n_{\mathrm{Tar}}}$, $\mathfrak{S}_D$: degree-D symmetric group.

Theoretical analysis

- 提案手法のリスク推定量は「データ点を増やすことでモデルの複雑度を落とし，より複雑な仮説を安心してフィットできるようにしている」と考えられる．
- 但しいくら安心してフィットできても，ずれた点を生成していたら誤った場所にフィットしてしまう．

Theoretical analysis

- 提案法のリスク推定量最小化に基づく ERM の汎化誤差解析をする (estimation error bound).
- 「複雑度を下げる効果」は V-統計量/U-統計量の理論を経由して出てくる Rademacher complexity を通じて見る.
- 「バイアスの度合い」は $\hat{f}$ によって誘導される誤差の伝搬を，$\|\check{q} - q\|_{L^1}$ を経由して見る.

ここからは論文を投影します

Appendix

Motivation: Causal modeling

- Causality is about the data generating process.
- Statistical machine learning: knowing the joint distribution solves most problems.
- Causal inference (e.g., make predictions under intervention): Joint dist. + assumptions on the data generation process are required.
- Data may have intrinsic causal structure (cf. the chocolate-Nobel example). The structure can be useful for ML. . . ?

Modeling causality: Two frameworks

There are mainly two frameworks. [8]

"Rubin's model": Potential outcome framework

- Model random variables before and after intervention, e.g., $\mathbb{E}[\mathrm{Nobel}_{\mathrm{AdChocolate}} - \mathrm{Nobel}_{\mathrm{NoAd}}]$

"Pearl's model": Structural causal models / Causal Bayesian networks (today)

- Model data generation process by functional relations.

- These are related [3, 4, 9]. Both models have some form of causal assumptions.

Motivation: Existing TAs

Transfer Assumption (TA)	AD	NP	Suited app. example
(1) Parametric dist. family [10] or shift [11–14].	✓	-	Remote sensing [11].
(2) Invariant dist [15] $p(Y\|\mathcal{T}(X))$ Covariate shift $\mathcal{T} = \mathrm{Id}$[17] Transfer component $\mathcal{T}$ [18] Feature selection $\mathcal{T}$ [19, 20] TarS [11, 21] $p(X\|Y)$ R-vine copulas [22].	-	✓	BCI [16]
(3) Discrepancy [23–28] / IPM [29] $+$ *ideal joint hypothesis* [25]	-	✓	Computer vision [29]
(4) Param-transfer [30]	✓	✓	Computer vision [30, 31]
(Ours) Mechanism	✓	✓	Medical records [5]

- AD: adaptation among Apparently Different distributions is accommodated.
- NP: Non-Parametrically flexible.

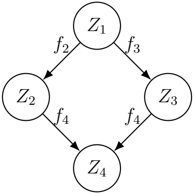
Motivation: Existing TAs

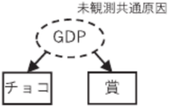
Transfer Assumption (TA)	AD	NP	Suited app. example
(1) Parametric dist. family [10] or shift [11–14].	✓	-	Remote sensing [11].
(2) Invariant dist [15] $p(Y\|\mathcal{T}(X))$ Covariate shift $\mathcal{T} = \mathrm{Id}$[17] Transfer component $\mathcal{T}$ [18]	-	✓	BCI [16]

- Different TAs have different (targeted) application fields.
- Compared to previously proposed TAs (approach-wise)...
  - Adaptation among apparently different distributions is accommodated.
  - Does not rely on parametric assumptions.

Modeling causality: GCMs

- Graph $\mathcal{G}$ has rich info. to enable causal inference (e.g., $p(Z_4|\mathrm{do}(Z_3 = 3))$).
- Knowing the whole $(q, \mathcal{G}, \mathcal{F})$ is not always necessary.
- Bayesian network of the induced distribution of $\{Z_d\}_{d=1}^{D}$.
  - We can read out conditional independence relations among variables.





Modeling causality: Estimation [32]

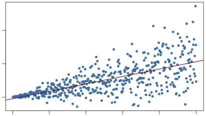
Approach	Example	Ref.
(1) Constraint-/Score-based	PC, FCI, GES	[32]
(2) "Functional constraint"-based	ANM, PNL	[32]
(3) "ICA"-based	LiNGAM	[32, 33]
Others	JCI	[34–36]

1. Estimate equivalence class of $\mathcal{G}$. Generic but <u>cannot</u> distinguish $Z_1 \leftarrow Z_2$ vs. $Z_1 \rightarrow Z_2$.
2. Estimate $\mathcal{G}$ by restricting function class of $\mathcal{F}$.
3. Non-Gaussianity/auxiliary information.

* This is only an incomplete list.

Understanding the assumption

- Simple example of such a data generation process:
  - Regression with (heteroskedastic) noise ($x \neq 0$ a.s.)



$$\begin{cases} X = S_1 \\ Y = h(X) + XS_2 \end{cases} \qquad \left(\Leftrightarrow \begin{cases} S_1 = X \\ S_2 = (Y - h(X))/ \end{cases} \right.$$

- Even if f is shared, $p_{\mathrm{src}(k)}(y|x)$ and $p_{\mathrm{tar}}(y|x)$ can be very different when $q_{\mathrm{src}(k)}$ and q_{tar} are different.

$$p(y|x) = \int p(y|s)p(s|x)\mathrm{d}s = \int \underbrace{p(y|s)p(x|s)}_{\text{Invariant}} \underbrace{\frac{q(s)}{p(z)}}_{\text{Variant}} \mathrm{d}s$$

Nonlinear ICA (1/2)

Problem: Independent Component Analysis

- Assume observed r.v. $X \in \mathbb{R}^D$ is an unknown transformation f (smooth and invertible) of the (dim-wise indep.) latent r.v. $S \in \mathbb{R}^D$ as $X = f(S)$.
- Goal: retrieve the inverse f^{-1} and the independent components $\{S^{(d)}\}_{d=1}^{D}$ based on observed X.

- Linear $f \Rightarrow$ well-established.
- Nonlinear $f \Rightarrow$ impossible in one-sample i.i.d. setting [37].

Nonlinear ICA (2/2)

- Nonlinear ICA has been realized [33, 38–40].
- Exploit auxiliary info (e.g. temporal dependence)[4].

Generalized contrastive learning [33] for NLICA

- Data has auxiliary variable (u): $\{(X_i, u_i)\}_{i=1}^n$
- Latent prior is conditioned on u: $p(s|u) = \prod_d q^{(d)}(s^{(d)}|u)$
- Train binary classifier $r(x, u) = \sigma(\sum_{d=1}^D \psi_d(h(x)_d, u))$ to distinguish $(x_i, u_i) : +1$ vs. $(x_i, \tilde{u}) : -1$.

σ: sigmoid

- Then, given sufficient theoretical conditions, $h : \mathcal{X} \to \mathbb{R}^D$ consistently estimates f $(n \to \infty)$.

[4] In our case, we use the source domain ID (k) as the auxiliary information.

References

[1] S. Shimizu, 統計的因果探索. Tokyo: 講談社, 2017.

[2] F. Huszár, *Causal Inference 2: Illustrating Interventions via a Toy Example*, Jan. 2019.

[3] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd. Cambridge, U.K. ; New York: Cambridge University Press, 2009.

[4] J. Mooij, *MLSS 2019: Causality*, 2019.

[5] P. Yadav, M. Steinbach, V. Kumar, and G. Simon, 'Mining electronic health records (EHRs): A survey', *ACM Computing Surveys*, vol. 50, no. 6, pp. 1–40, 2018.

[6] A. J. Lee, *U-Statistics: Theory and Practice*. New York: M. Dekker, 1990.

[7] S. Clémençon, I. Colin, and A. Bellet, 'Scaling-up empirical risk minimization: Optimization of incomplete U-statistics', *Journal of Machine Learning Research*, vol. 17, no. 76, pp. 1–36, 2016.

[8] S. Yasui, 効果検証入門〜正しい比較のための因果推論／計量経済学の基礎. 2020, 株式会社ホクソエム 監修.

References (cont.)

[9] M. Kuroki and F. Kobayashi, ‘構造的因果モデルについて’, 計量生物学, vol. 32, no. 2, pp. 119–144, 2012.

[10] A. J. Storkey and M. Sugiyama, ‘Mixture regression for covariate shift’, in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds., MIT Press, 2007, pp. 1337–1344.

[11] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, ‘Domain adaptation under target and conditional shift’, in *Proceedings of the 30th International Conference on Machine Learning*, 2013, pp. 819–827.

[12] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf, ‘Domain adaptation with conditional transferable components’, in *Proceedings of the 33rd International Conference on Machine Learning*, M. F. Balcan and K. Q. Weinberger, Eds., New York, USA: PMLR, 2016, pp. 2839–2848.

[13] K. Zhang, M. Gong, and B. Schölkopf, ‘Multi-source domain adaptation: A causal view’, in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI Press, 2015, pp. 3150–3157.

References (cont.)

[14] P. Stojanov, M. Gong, J. Carbonell, and K. Zhang, ‘Data-driven approach to multiple-source domain adaptation’, in *Proceedings of Machine Learning Research*, K. Chaudhuri and M. Sugiyama, Eds., vol. 89, PMLR, 2019, pp. 3487–3496.

[15] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Eds., *Dataset Shift in Machine Learning*, ser. Neural Information Processing Series. Cambridge, Mass: MIT Press, 2009.

[16] M. Sugiyama, M. Krauledat, and K.-R. Müller, ‘Covariate shift adaptation by importance weighted cross validation’, *Journal of Machine Learning Research*, vol. 8, no. May, pp. 985–1005, 2007.

[17] H. Shimodaira, ‘Improving predictive inference under covariate shift by weighting the log-likelihood function’, *Journal of Statistical Planning and Inference*, vol. 90, no. 2, pp. 227–244, 2000.

[18] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, ‘Domain adaptation via transfer component analysis’, *IEEE Transactions on Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

References (cont.)

[19] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters, ‘Invariant models for causal transfer learning’, *Journal of Machine Learning Research*, vol. 19, no. 36, pp. 1–34, 2018.

[20] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij, ‘Domain adaptation by using causal inference to predict invariant conditional distributions’, in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Curran Associates, Inc., 2018, pp. 10 846–10 856.

[21] T. D. Nguyen, M. Christoffel, and M. Sugiyama, ‘Continuous Target Shift Adaptation in Supervised Learning’, in *Asian Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 45, PMLR, 2016, pp. 285–300.

[22] D. Lopez-paz, J. M. Hernández-lobato, and B. Schölkopf, ‘Semi-supervised domain adaptation with non-parametric copulas’, in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2012, pp. 665–673.

References (cont.)

[23] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, ‘Analysis of representations for domain adaptation’, in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds., MIT Press, 2007, pp. 137–144.

[24] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, ‘Learning bounds for domain adaptation’, in *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds., Curran Associates, Inc., 2008, pp. 129–136.

[25] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, ‘A theory of learning from different domains’, *Machine Learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[26] S. Kuroki, N. Charoenphakdee, H. Bao, J. Honda, I. Sato, and M. Sugiyama, ‘Unsupervised domain adaptation based on source-guided discrepancy’, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4122–4129.

References (cont.)

[27] Y. Zhang, T. Liu, M. Long, and M. Jordan, ‘Bridging theory and algorithm for domain adaptation’, in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., Long Beach, California, USA: PMLR, 2019, pp. 7404–7413.

[28] C. Cortes, M. Mohri, and A. M. Medina, ‘Adaptation based on generalized discrepancy’, *Journal of Machine Learning Research*, vol. 20, no. 1, pp. 1–30, 2019.

[29] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy, ‘Joint distribution optimal transportation for domain adaptation’, in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 3730–3739.

[30] W. Kumagai, ‘Learning bound for parameter transfer learning’, in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., Curran Associates, Inc., 2016, pp. 2721–2729.

References (cont.)

[31] H. Lee, R. Raina, A. Teichman, and A. Y. Ng, 'Exponential family sparse coding with applications to self-taught learning', in *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2009, pp. 1113–1119.

[32] C. Glymour, K. Zhang, and P. Spirtes, 'Review of Causal Discovery Methods Based on Graphical Models', *Frontiers in Genetics*, vol. 10, Jun. 2019.

[33] A. Hyvärinen, H. Sasaki, and R. Turner, 'Nonlinear ICA using auxiliary variables and generalized contrastive learning', in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 859–868.

[34] J. M. Mooij, S. Magliacane, and T. Claassen, 'Joint Causal Inference from Multiple Contexts', *arXiv:1611.10351 [cs, stat]*, Apr. 2019. arXiv: `1611.10351 [cs, stat]`.

[35] D. Janzing and B. Schölkopf, 'Causal inference using the algorithmic Markov condition', *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 5168–5194, Oct. 2010.

References (cont.)

[36] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf, ‘Information-geometric approach to inferring causal directions’, *Artificial Intelligence*, vol. 182, pp. 1–31, May 2012.

[37] A. Hyvärinen and P. Pajunen, ‘Nonlinear independent component analysis: Existence and uniqueness results.’, *Neural networks*, vol. 12, no. 3, pp. 429–439, 1999.

[38] A. Hyvärinen and H. Morioka, ‘Unsupervised feature extraction by time-contrastive learning and nonlinear ICA’, in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., Curran Associates, Inc., 2016, pp. 3765–3773.

[39] ——,‘Nonlinear ICA of temporally dependent stationary sources’, in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017, pp. 460–469.

[40] I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen, ‘Variational autoencoders and nonlinear ICA: A unifying framework’, *arXiv:1907.04809 [cs, stat]*, Jul. 2019. arXiv: `1907.04809 [cs, stat]`.