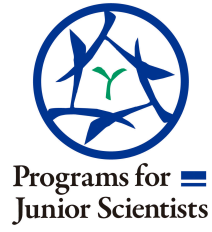
Few-shot Domain Adaptation by Causal Mechanism Transfer

手嶋 毅志 [12]

[1] 東京大学 [2] 理研 AIP



Joint w/ Issei Sato[12], and Masashi Sugiyama[21]

(本研究は理研の大学院生リサーチ・アソシエイト制度の下での成果です。)

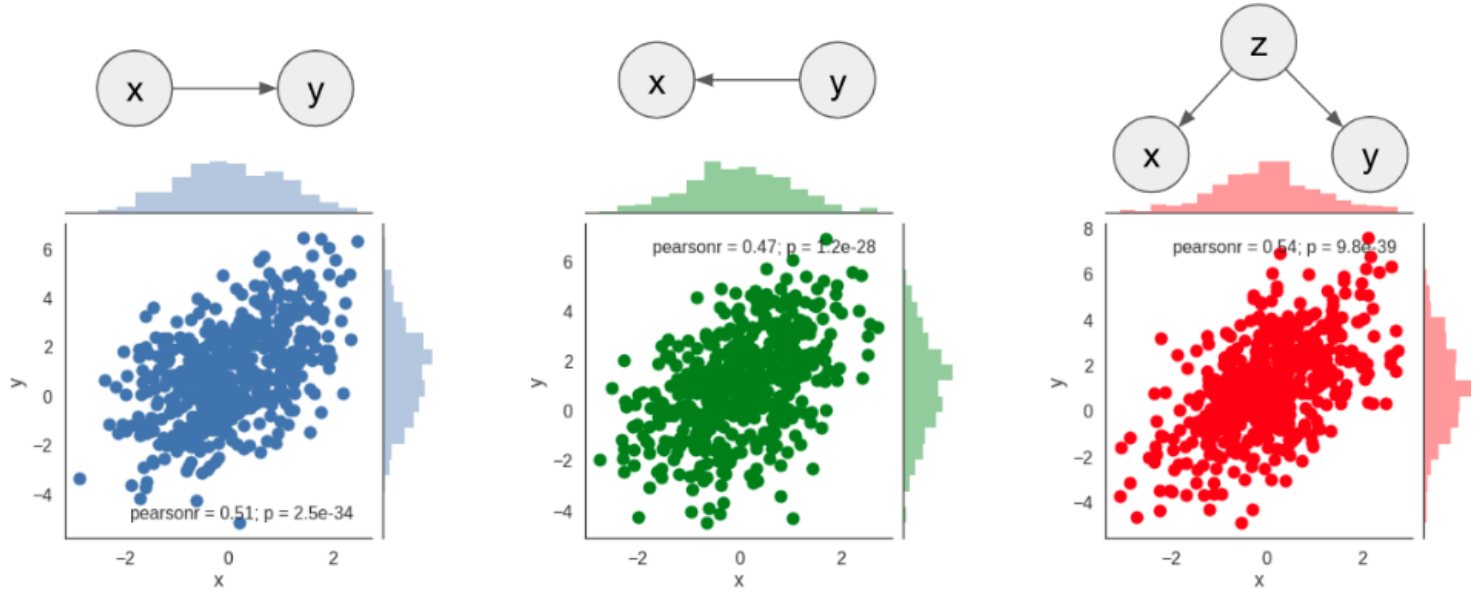
https://bit.ly/2C4EEye

本トークの構成

- Part 1. 因果モデルとは何かをご紹介
  - “Pearl の” 因果モデル (今回の論文でも重要)
- Part 2. 因果的機械学習を概観 (関連分野)
  - 反実仮想/介入効果推定「しない」因果的機械学習
- Part 3. 今回の論文の内容を紹介 (英語スライド)

Part 1.
Pearlの因果モデル

- 機械学習モデル：「データの確率分布」を考える
- 因果モデル：更に背後の「生成過程」まで考える
- 因果モデルには主に Pearl 流と Rubin 流の 2 つがある



なぜ「因果モデル」を考えるか

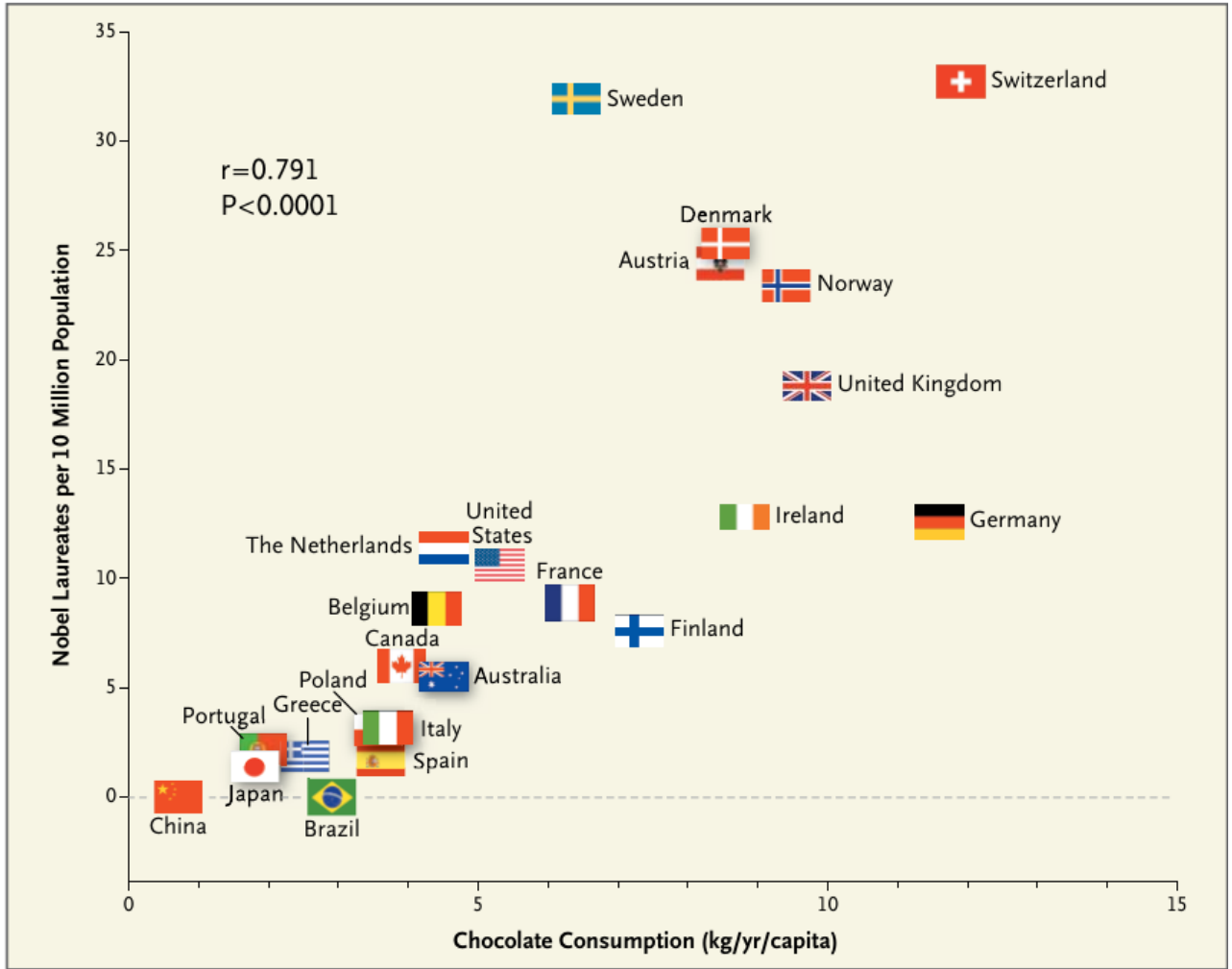


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Figure: [1]

なぜ「因果モデル」を考えるか

- ？？？「チョコレートを食べよう！」

- 我々🤔「落ち着こう…因果関係があるかは怪しい」
  ＝チョコを増やせばノーベル賞が増えるとは思えない

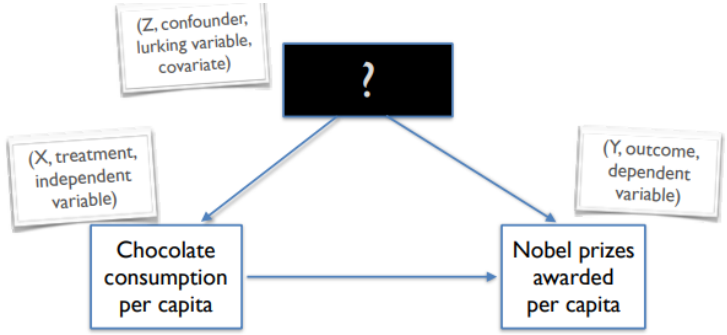


Figure: [2]

- 最初のプロットを見ているだけだと議論できない
- これを数理的にモデル化/データで説明したい

前提：2 つの因果フレームワーク

Rubin の潜在反応フレームワーク [3]

- 主眼：介入効果/反実仮想の定式化
- 方針：データ生成過程のモデルを小さく保ちたい

Pearl の構造的因果モデルフレームワーク [4]

- 主眼：関与する変数全体の挙動の定式化
- 方針：All-in-one なモデルを作る

- ここでは Pearl の枠組みにフォーカス.
- 構造方程式・因果グラフが主な道具

前提：因果モデル（構造方程式）

構造方程式モデル [4–6]

「因果関係は，何らかの関数関係を通して決定論的に記述できる」という考えに基づいて構築されたデータ生成過程のモデル †(次スライド)

† 日本語の説明：https://yhiss.hatenablog.com/entry/2020/04/12/170309

前提：因果モデル（構造方程式）

構造方程式モデル (SEM)/構造的因果モデル (SCM) [4–6]

- データ $\boldsymbol{Z} = \{Z_d\}_{d=1}^D$ の生成過程を $(\mathcal{F}, q)$ でモデル化.

関数たち $\mathcal{F} = \{f_d\}_{d=1}^D$

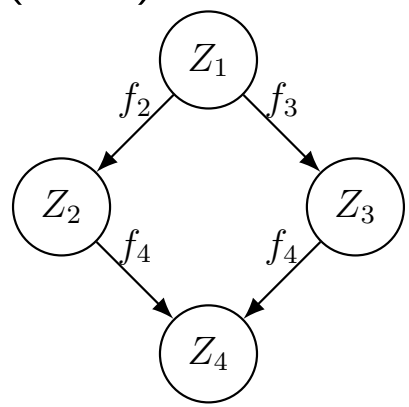
$$\begin{cases} Z_1 &= f_1(\mathrm{pa}_1, S_1) \\ &\vdots \\ Z_d &= f_d(\mathrm{pa}_d, S_d) \\ &\vdots \\ Z_D &= f_D(\mathrm{pa}_D, S_D) \end{cases}$$

- $\mathrm{pa}_d \subset \boldsymbol{Z}$ は Z_d の直接的原因と解釈される変数集合

独立確率変数たち $\boldsymbol{S} = \{S_i\}_{i=1}^D$ の分布 q

$$q(S) = \prod_{d=1}^{D} q_d(S_d)$$

(Directed acyclic graph (DAG) $\mathcal{G}$ encoding $\mathcal{F}$)



- $\mathcal{F}$ を定性的に表現したグラフ

- 初めは $\boldsymbol{S}$ のみ確率的. $\mathcal{F}$ を通して $\boldsymbol{Z}$ が確率的になる.

Simplified: acyclic, Markovian, same-dimensional latent. See [6, 7].

前提：構造的因果モデルの使用法 (☆)

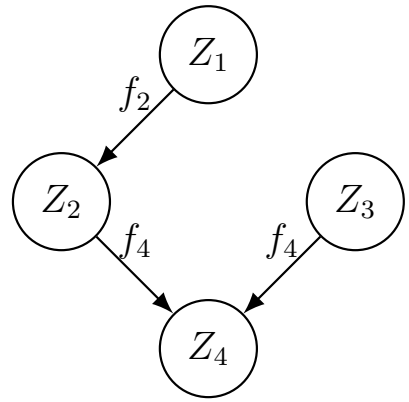
完全介入 [4] $\mathrm{do}(Z_I = \zeta_I)$

- 「介入」の定式化が出来る
- Z_I の値を ζ_I にする介入 $(q, \mathcal{F}) \rightsquigarrow (q, \mathcal{F}')$

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

- グラフでは Z_I に向かう矢印が削除.

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



- 反実仮想分布は $p(Z|\mathrm{do}(Z_I = \zeta_I)) \leftarrow (q, \mathcal{F}')$.

因果グラフだけでも出来ること (☆) [4]

$p(\boldsymbol{Z})$ の条件付き独立性の情報はグラフ $\mathcal{G}$ だけで分かる

1. 局所マルコフ条件 (親を条件付けると非子孫と独立)

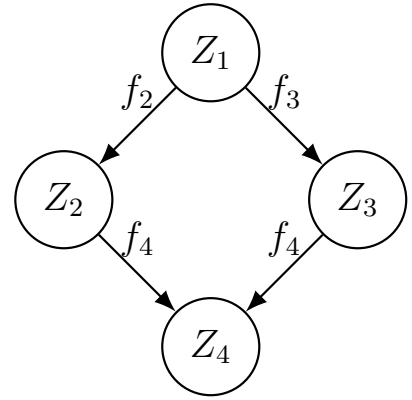
$$Z_d \perp\!\!\!\perp \mathrm{ND}_d | \mathrm{pa}_d \quad (\mathrm{ND}: \text{ non-descendants})$$

2. 大域マルコフ条件 (“d-separation” →条件付独立)

$$\text{d-sep}(\boldsymbol{Z}^{(1)} || \boldsymbol{Z}^{(2)}; \boldsymbol{Z}^{(3)}) \Rightarrow \boldsymbol{Z}^{(1)} \perp\!\!\!\perp \boldsymbol{Z}^{(2)} | \boldsymbol{Z}^{(3)}$$

3. 「因果的」条件付き分布分解

$$p(Z_1, \ldots, Z_D) = \prod_{d=1}^{D} p(Z_d | \mathrm{pa}_d)$$



($p(\boldsymbol{Z}) = \prod_{d=1}^{D} p(Z_d | Z_1, \ldots, Z_{d-1})$ までならいつでも出来る)

Part 2.

反実仮想「しない」因果的機械学習

- 最近の発展・方向性を概観
- 機械学習による因果推論 (causality by ML)
- 機械学習のための因果推論 (causality for ML)

機械学習による因果推論 (by ML)

- 因果探索
- 因果変数構築

By-ML: 因果探索 [8]

大まかな流れ

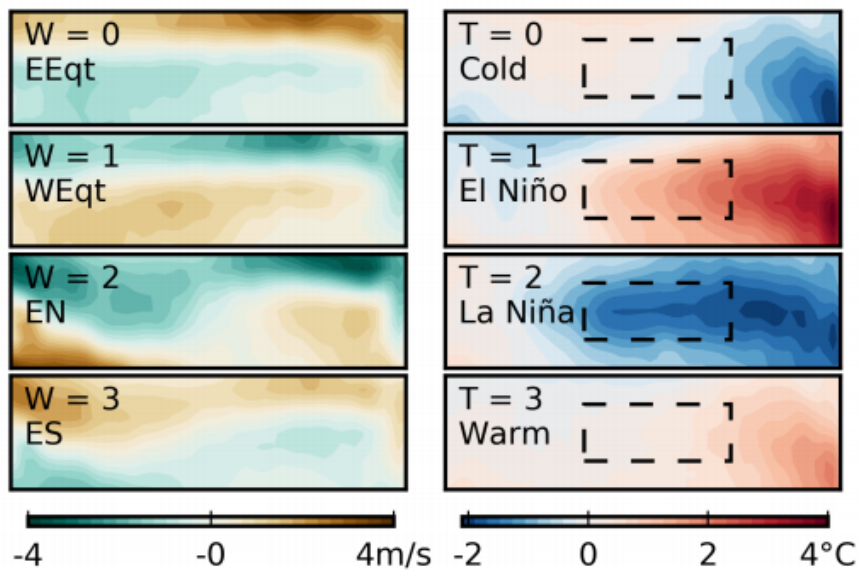
- 第一次：条件付き独立性を使う
  - Constraint based [8], Score based [8, 9]
- 第二次：SEM を制約する
  - Linear + non-Gaussian (LiNGAM) [10, 11]
- 第三次：原因・結果の非対称性を使う
  - ICA，アルゴリズム複雑性，情報幾何，... [8, 12–15]

- データ→条件付き独立性の一覧→当てはまるグラフ ("忠実性" の仮定のもと)
  - 😟同値類までしか分からない．特に $(Z_1 \leftarrow Z_2) \sim (Z_1 \rightarrow Z_2)$
- そのため近年の手法は特に 2 変数のケースを重視

By-ML: 因果変数構築

因果変数構築 (causal feature learning) [18]

- 粒度の細かい情報（例：画像のピクセル）から「因果変数」として何を使えばよいか学習
- 「因果変数」を学習する基準を公理的に定義 [16]
- エルニーニョ現象を気象データから再発見 [17]



機械学習のための因果推論 (for ML) [19, 20]

- 指導原理として：「因果メカニズムの独立性」
  1. 「情報の有無」の判断基準/正当化
  2. 「分布変化はスパースに起きる」
- 「個別的な問題設定を分析する」道具として
  (具体的に因果グラフを描ける場面を考察する)
- 「不変性」「安定性」の論拠として

For-ML: 指導原理 (因果メカニズムの独立性) [20]

「因果的」条件付き分布分解

$$p(Z_1, \ldots, Z_D) = \prod_{d=1}^{D} p(Z_d | \mathrm{pa}_d)$$

($p(\boldsymbol{Z}) = \prod_{d=1}^{D} p(Z_d | Z_1, \ldots, Z_{d-1})$ ならいつでも出来る)

1. 「情報の有無」の判断基準/正当化
   $p(Z_d|\mathrm{pa}_d)$ から他の $p(Z_{d'}|\mathrm{pa}_{d'})$ の知識は得られない
2. 「分布変化はスパースに起きる」
   $p(Z_d|\mathrm{pa}_d)$ が変化しても他の $p(Z_{d'}|\mathrm{pa}_{d'})$ は不変

For-ML: 「情報の有無」の指導原理

記念碑的論文：Causal and Anti-causal learning [21]

- $X \to Y$ か？ $X \leftarrow Y$ か？
- 半教師付き学習は $X \leftarrow Y$ の場合に有効と予想
  - $X \to Y$ なら $p(X)$ は $p(Y|X)$ の情報を含まない
  - →過去の実験結果を集計したところ予想どおり

教師なし転移学習手法の「直観的正当化」 [22, 23]

- 「$X \leftarrow Y$ なら $p(X)$ にも $p(Y|X)$ の情報がある」
- $p(X)$ だけでパラメータ推定することを正当化
  教師無し転移学習 (転移元データ：(X, Y)，転移先データ：X) で便利

For-ML: 「分布変化はスパースに起きる」指導原理17/37

ドメイン汎化 [24]

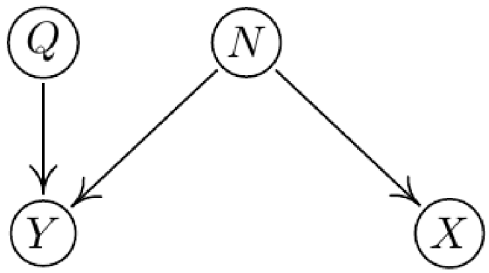
- 仮定「$p(Y|\boldsymbol{X}_S)$ が異なる転移元分布で不変なら，転移先でも不変」
- それぞれの条件付き分布$=$「要素」は独立に変化

パラメトリックな分布変化モデルの正当化 [22]

- $X \leftarrow Y$ なら $p(X|Y)$ の変化と $p(Y)$ の変化をそれぞれモデル化
- クラス Y ごとの X の分布が線型変換で移り合う

Half-sibling regression [25]

- ゴール：系外惑星探索 (Exoplanet search)



  - 恒星が一瞬暗くなる→惑星候補の発見
  - 測定器 (Kelper) 由来でも星が暗くなる．この現象は遠い星で同時に起きる．

「介入分布への転移学習」[26]

- 状況設定：SCM は不変．介入により分布変化
- 介入前のデータを用いて介入後の分布に対し予測
- 因果グラフを推定→賢く変数選択

For-ML: 「不変性」「安定性」の論拠として

不変リスク最小化 (IRM) [27]

- 一つの予測器がどの分布でも「最適」になるよう特徴学習

$$\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H}\\ w:\mathcal{H}\to\mathcal{Y}}} \sum_{e\in\mathcal{E}_{\mathrm{tr}}} R^e(w\circ\Phi)$$

$$\text{subject to} \quad w \in \operatorname*{arg\,min}_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w}\circ\Phi),\ \text{for all } e\in\mathcal{E}_{\mathrm{tr}}.$$

- 線型な SEM(を少し一般化したもの) が不変という仮定のもとで手法が正当化される十分条件を解析

- 本研究はこの文脈が比較的近い

その他のリソース

- Machine Learning Summer School (MLSS) の Tutorial が充実 (特に Tubingen で開催されるとき). 動画も結構残っている.

2013	Dominik Janzing & Bernhard Schölkopf	[Slides], [Videos]
2015	Bernhard Schölkopf and Jonas Peters	[Slides], [Script]
2016	Jonas Peters	[Slides], [Videos]
2017	Dominik Janzing	[Slides]
2017	Bernhard Schölkopf	[Slides]
2017	David Lopez-paz	[Video]
2019	Joris Mooij	[Slides], [Exercises]
2020	Bernhard Schölkopf	[Slides], [Videos]
2020	Stefan Bauer	[Slides], [Videos]

- MLSS 2020 Tutorial slides 3 ページ目の読書案内
- Causal Reinforcement Learning Tutorial (ICML 2020) (Project page)

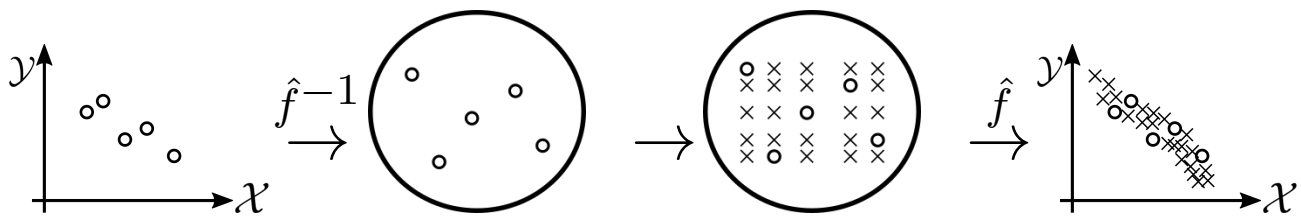
Part 3.

今回の論文

Teshima, T., Sato, I., and Sugiyama, M., (2020)

Few-shot domain adaptation

by causal mechanism transfer.

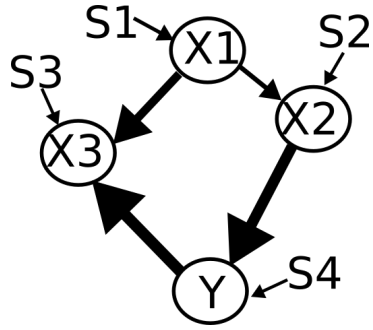


Overview of this work

Domain adaptation

Q. When is it possible?
Transfer assumption (TA)?

Causal mechanism



A. Common causal mechanism as the relation.

Summary

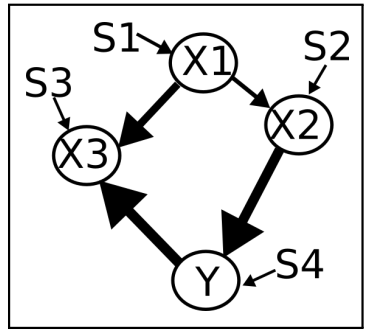
Common data generating (causal) mechanism can be a foundation for domain adaptation.

Preliminary: Causal model (SEMs)

Structural Equation Models (SEMs)[1, 2] [4]

- Generative model for the joint distribution of data.
- Consists of $(\mathcal{F}, q)$ where $\mathcal{F}$: deterministic functions

$$\mathcal{F} = \begin{cases} X_1 &= f_1'(\mathrm{pa}_1, S_1) \\ X_2 &= f_2'(\mathrm{pa}_2, S_2) \\ X_3 &= f_3'(\mathrm{pa}_3, S_3) \\ Y &= f_4'(\mathrm{pa}_4, S_4) \end{cases}$$



and q: independent distribution of $(S_1, \ldots, S_D)$.

[1] More precisely, NPSEM-IE (Nonparametric SEM with Independent Errors).

[2] Acyclicity is assumed.

Nonlinear-ICA and reduced-form SEM

Reduced form: Structural equations solved for $(\mathbf{X}, Y)$.

$$\begin{cases} X_1 &= f_1'(\mathrm{pa}_1, S_1) \\ X_2 &= f_2'(\mathrm{pa}_2, S_2) \\ X_3 &= f_3'(\mathrm{pa}_3, S_3) \\ Y &= f_4'(\mathrm{pa}_4, S_4) \end{cases}$$

Structural equations

$$\Longrightarrow$$

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ Y \end{pmatrix} = f \begin{pmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{pmatrix}$$

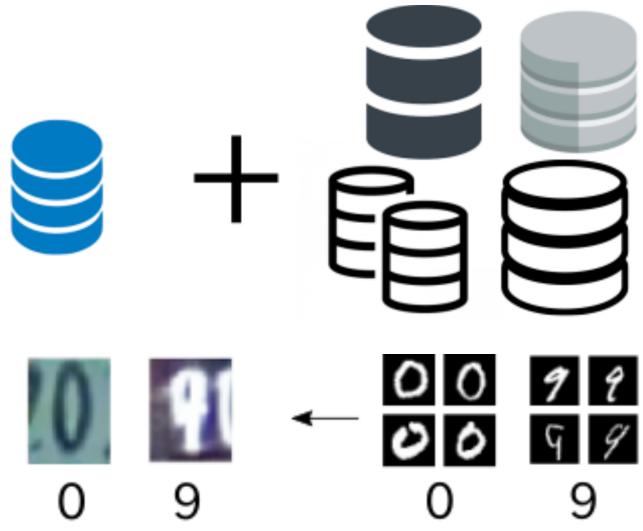
Reduced form

- Under certain *identification conditions*, nonlinear-ICA[3] methods can estimate f (we use it in our method).

[3] ICA = Independent component analysis.

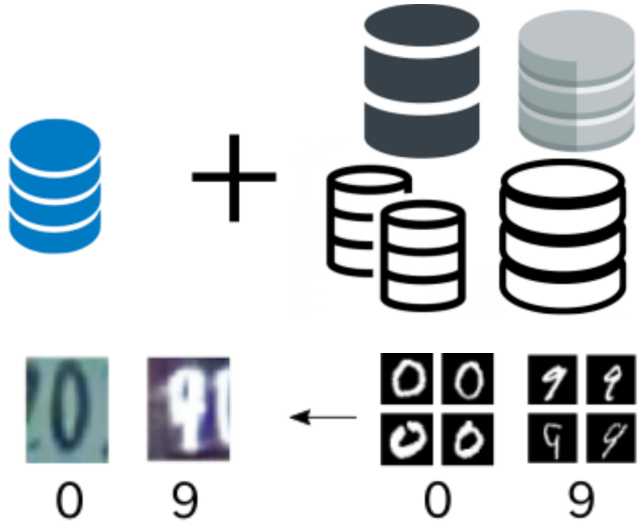
Motivation: Transfer learning

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



Motivation: Transfer assumption (TA)

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



- Common data generating (causal) mechanism can be a foundation for domain adaptation.

Intuition of our TA

Intuition

Humans care about finding causal knowledge because, once discovered, it applies to different systems.

Motivating example: Regional disease prediction

- Predict disease risk from medical records. [28]
- Data distributions may vary for different lifestyles.
- Common pathological mechanism across regions.

Formulation 1/3: Data and goal

Basic setup: regression domain adaptation

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., multiple source domains)

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

3. Few-shot supervised (i.e., target data with labels)

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

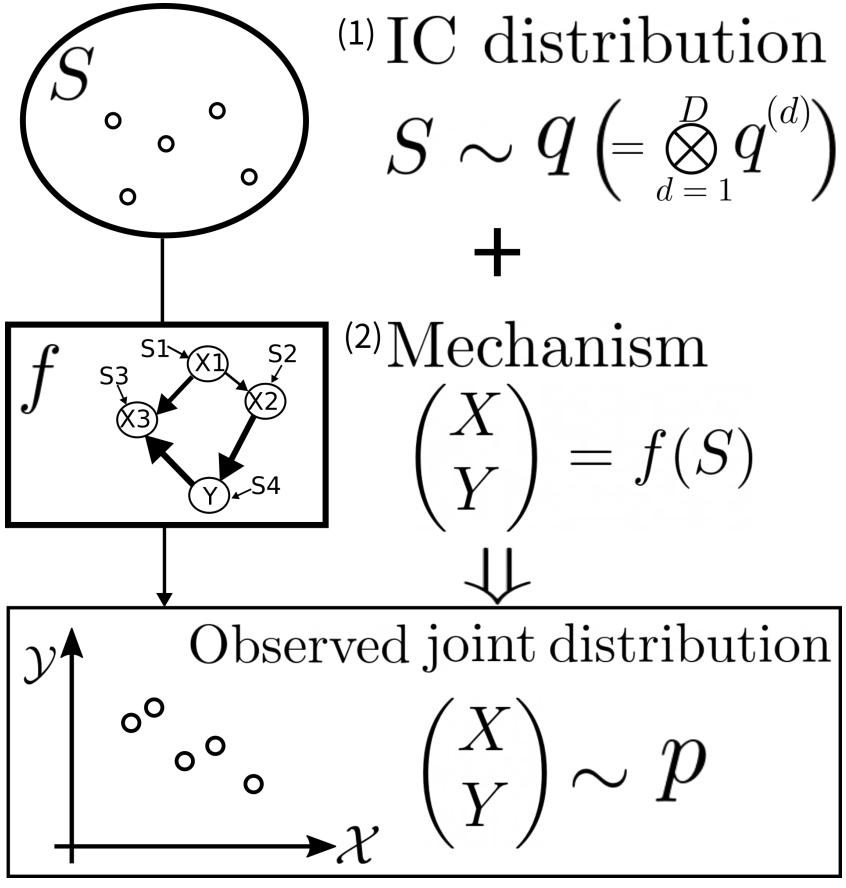
Goal: accurate predictor for the target distribution

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)

Formulation 2/3: Each domain

- Each domain follows a nonlinear-ICA model.



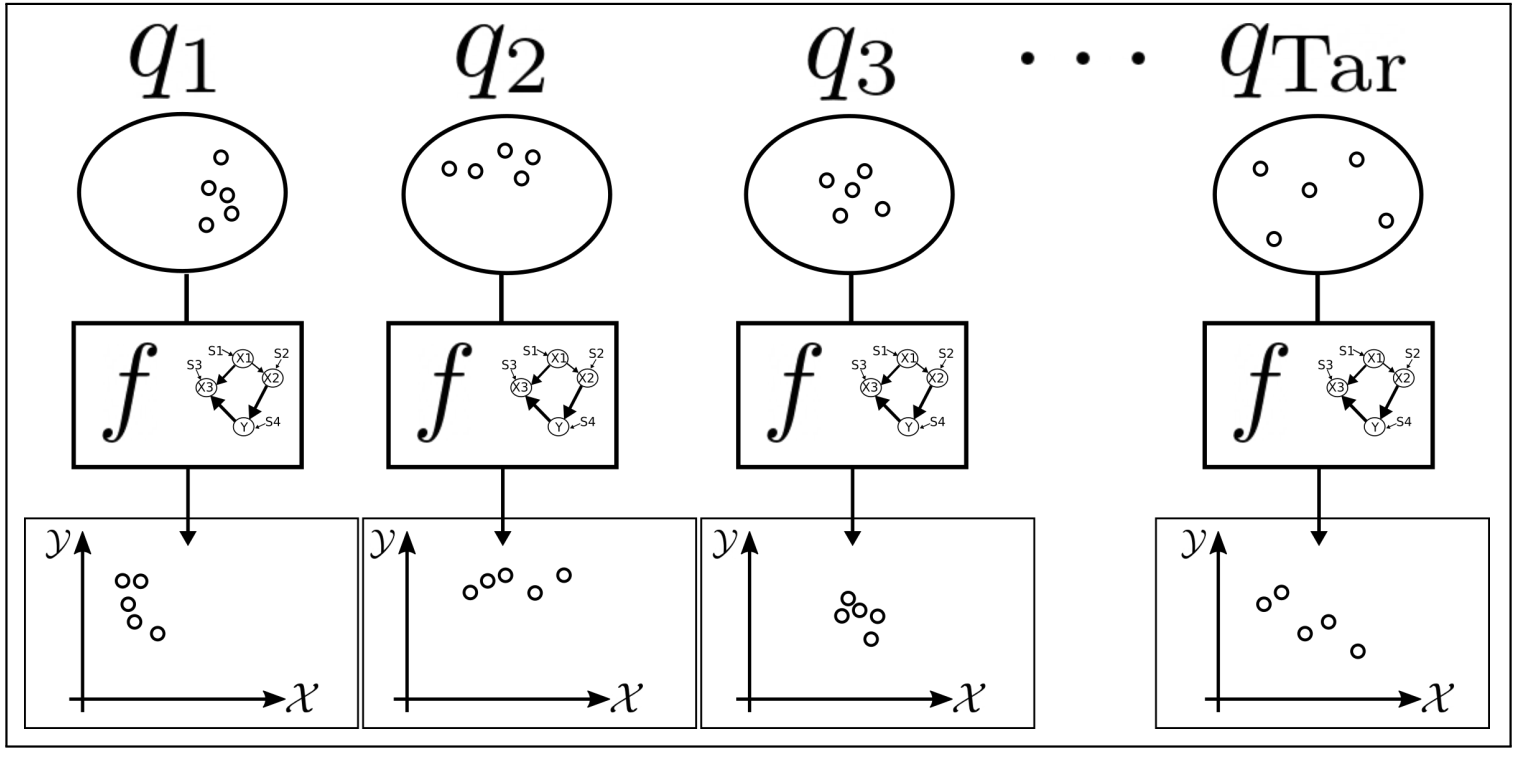
Dist. p consists of (f, q)

1. D-dimensional ICs S are sampled from q.
2. Invertible f transforms S into $(X, Y) = f(S)$.

- f can be estimated by ICA under assumptions.
- f corresponds to the reduced form of an SEM.

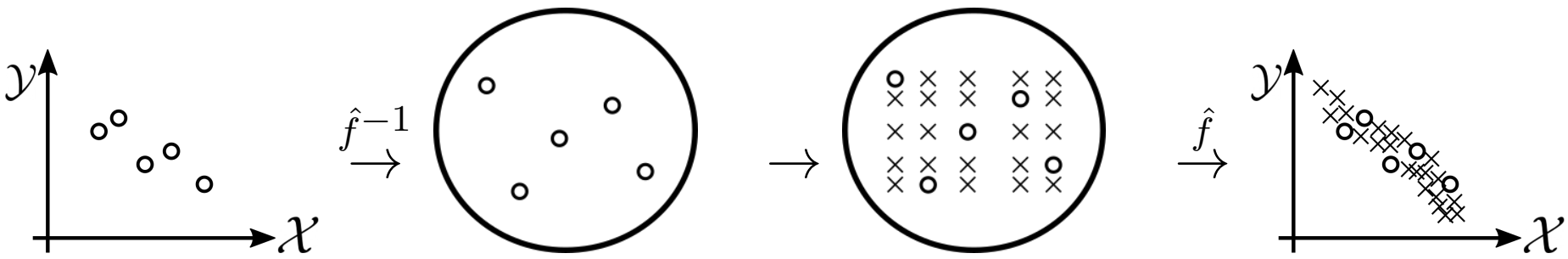
Formulation 3/3: Key assumption

- Key Assumption: generative mechanism f is common.



- Allow flexible shift in $q \rightsquigarrow$ Enables DA among seemingly very different distributions.

Proposed method: Strategy



Idea: How to exploit the assumption

1. Estimate f using source domain data (NLICA).
2. Estimate ICs of the target data using $\hat{f}^{-1}$.
3. Get "candidate ICs" by exchanging values.
   $=$ resample from emp. margins $=$ take grid points.
4. Generate target data from reshuffled ICs using $\hat{f}$.
5. Train the predictor g on the generated data.

What do you mean by exchanging?

- Select one sample for each dimension (with replacement).

$$
\begin{array}{c}
 & \hat{S}_1 & \hat{S}_2 & \cdots & \hat{S}_{n-1} & \hat{S}_n \\
1 & \hat{s}_{11} & \hat{s}_{12} & \cdots & \hat{s}_{1,n-1} & \hat{s}_{1n} \\
2 & \hat{s}_{21} & \hat{s}_{22} & \cdots & \hat{s}_{2,n-1} & \hat{s}_{2n} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
D-1 & \hat{s}_{D-1,1} & \hat{s}_{D-1,2} & \cdots & \hat{s}_{D-1,n-1} & \hat{s}_{D-1,n} \\
D & \hat{s}_{D1} & \hat{s}_{D2} & \cdots & \hat{s}_{D,n-1} & \hat{s}_{Dn}
\end{array}
\rightarrow
\begin{pmatrix}
\hat{s}_{1,n-1} \\
\hat{s}_{22} \\
\vdots \\
\hat{s}_{D-1,1} \\
\hat{s}_{D2}
\end{pmatrix}
$$

- Essentially, this means resampling from the empirical marginals (independently).

Theoretical analyses

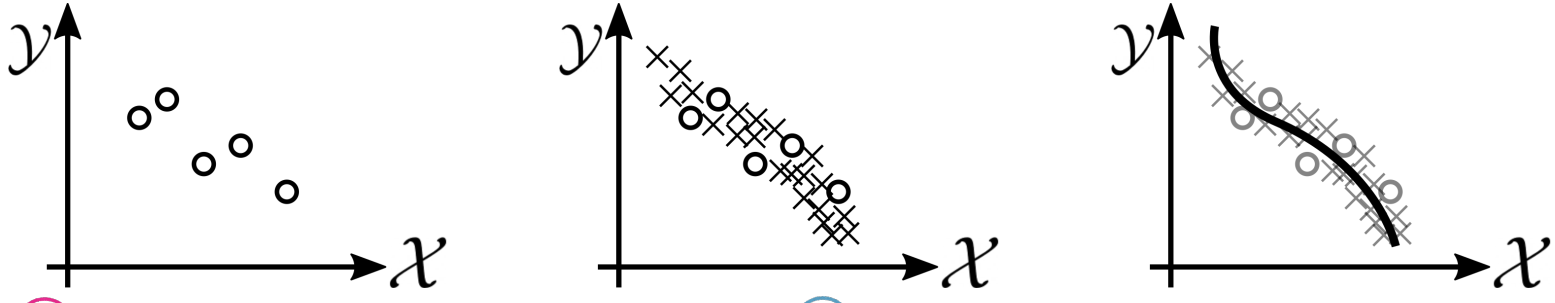
Q1. How does the method statistically help?

Theorem: If $\hat{f} = f$, the proposed risk estimator is the uniformly minimum variance unbiased risk estimator.

The method should help in terms of variance.

Q2. What happens when $\hat{f} \neq f$? What's the catch?

Theorem: generalization error bound for $\hat{f} \neq f$.



Mitigate overfitting. Introduce bias.

Experiment: Setup

- Dataset: Gasoline consumption dataset [29].
  - Panel data from econometrics (SEMs have been applied).
  - 18 countries (=domains), 19 years, $D = 4$.
- Baselines for regression domain adaptation.

Name	Compared method (predictor: KRR)
TarOnly	Train on target.
SrcOnly	Train on source.
S&TV	Train on source, CV on target.
TrAdaBoost	Boosting for few-shot regression transfer [30].
IW	Joint importance weight using RuLSIF [31].
GDM	Generalized discrepancy minimization [32].
Copula	Non-parametric R-vine copula method [33].
LOO (reference)	LOOCV error estimate.

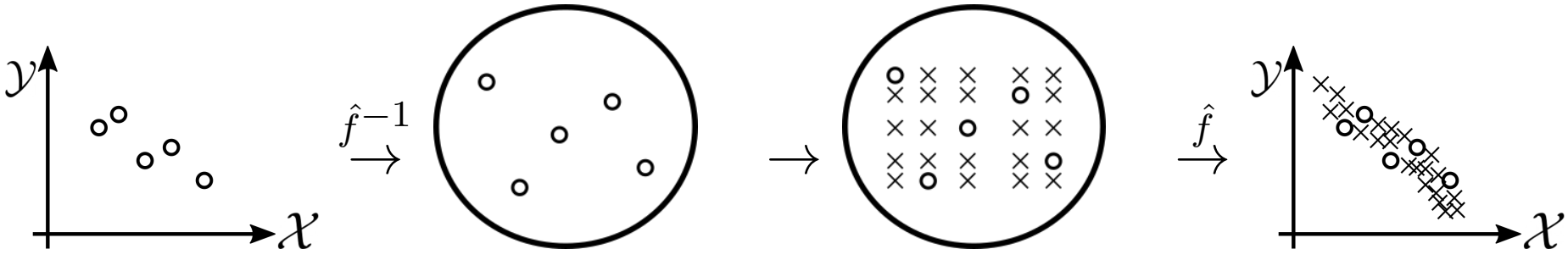
Experiment: Result

Target	(LOO)	TrgOnly	Prop	SrcOnly	S&TV	TrAda	GDM	Copula	IW(.0)	IW(.5)	IW(.95)
AUT	1	5.88 (1.60)	5.39 (1.86)	9.67 (0.57)	9.84 (0.62)	5.78 (2.15)	31.56 (1.39)	27.33 (0.77)	39.72 (0.74)	39.45 (0.72)	39.18 (0.76)
BEL	1	10.70 (7.50)	7.94 (2.19)	8.19 (0.68)	9.48 (0.91)	8.10 (1.88)	89.10 (4.12)	119.86 (2.64)	105.15 (2.96)	105.28 (2.95)	104.30 (2.95)
CAN	1	5.16 (1.36)	3.84 (0.98)	157.74 (8.83)	156.65 (10.69)	51.94 (30.06)	516.90 (4.45)	406.91 (1.59)	592.21 (1.87)	591.21 (1.84)	589.87 (1.91)
DNK	1	3.26 (0.61)	3.23 (0.63)	30.79 (0.93)	28.12 (1.67)	25.60 (13.11)	16.84 (0.85)	14.46 (0.79)	22.15 (1.10)	22.11 (1.10)	21.72 (1.07)
FRA	1	2.79 (1.10)	1.92 (0.66)	4.67 (0.41)	3.05 (0.11)	52.65 (25.83)	91.69 (1.34)	156.29 (1.96)	116.32 (1.27)	116.54 (1.25)	115.29 (1.28)
DEU	1	16.99 (8.04)	6.71 (1.23)	229.65 (9.13)	210.59 (14.99)	341.03 (157.80)	739.29 (11.81)	929.03 (4.85)	817.50 (4.60)	818.13 (4.55)	812.60 (4.57)
GRC	1	3.80 (2.21)	3.55 (1.79)	5.30 (0.90)	5.75 (0.68)	11.78 (2.36)	26.90 (1.89)	23.05 (0.53)	47.07 (1.92)	45.50 (1.82)	45.72 (2.00)
IRL	1	3.05 (0.34)	4.35 (1.25)	135.57 (5.64)	12.34 (0.58)	23.40 (17.50)	3.84 (0.22)	26.60 (0.59)	6.38 (0.13)	6.31 (0.14)	6.16 (0.13)
ITA	1	13.00 (4.15)	14.05 (4.81)	35.29 (1.83)	39.27 (2.52)	87.34 (24.05)	226.95 (11.14)	343.10 (10.04)	244.25 (8.50)	244.84 (8.58)	242.60 (8.46)
JPN	1	10.55 (4.67)	12.32 (4.95)	8.10 (1.05)	8.38 (1.07)	18.81 (4.59)	95.58 (7.89)	71.02 (5.08)	135.24 (13.57)	134.89 (13.50)	134.16 (13.43)
NLD	1	3.75 (0.80)	3.87 (0.79)	0.99 (0.06)	0.99 (0.05)	9.45 (1.43)	28.35 (1.62)	29.53 (1.58)	33.28 (1.78)	33.23 (1.77)	33.14 (1.77)
NOR	1	2.70	2.82	1.86	1.63	24.25	23.36	31.37	27.86	27.86	27.52
GBR	1	5.95 (1.86)	2.66 (0.57)	15.92 (1.02)	10.05 (1.47)	7.57 (5.10)	50.04 (1.75)	68.70 (1.25)	70.98 (1.01)	70.87 (0.99)	69.72 (1.01)
USA	1	4.98 (1.96)	1.60 (0.42)	21.53 (3.30)	12.28 (2.52)	2.06 (0.47)	308.69 (5.20)	244.90 (1.82)	462.51 (2.14)	464.75 (2.08)	465.88 (2.16)
#Best	-	2	10	2	4	0	0	0	0	0	0

Proposed $>$ *TrgOnly* when the other methods using source domain data suffer from negative transfer.

Take-home message

1. Transfer assumption of shared generative mechanism. Developed a few-shot regression DA method.
2. Proposed method extracts and uses the causal model to reduce overfitting via data augmentation.
3. Experiment with real-world data demonstrate the validity.



個人的まとめ

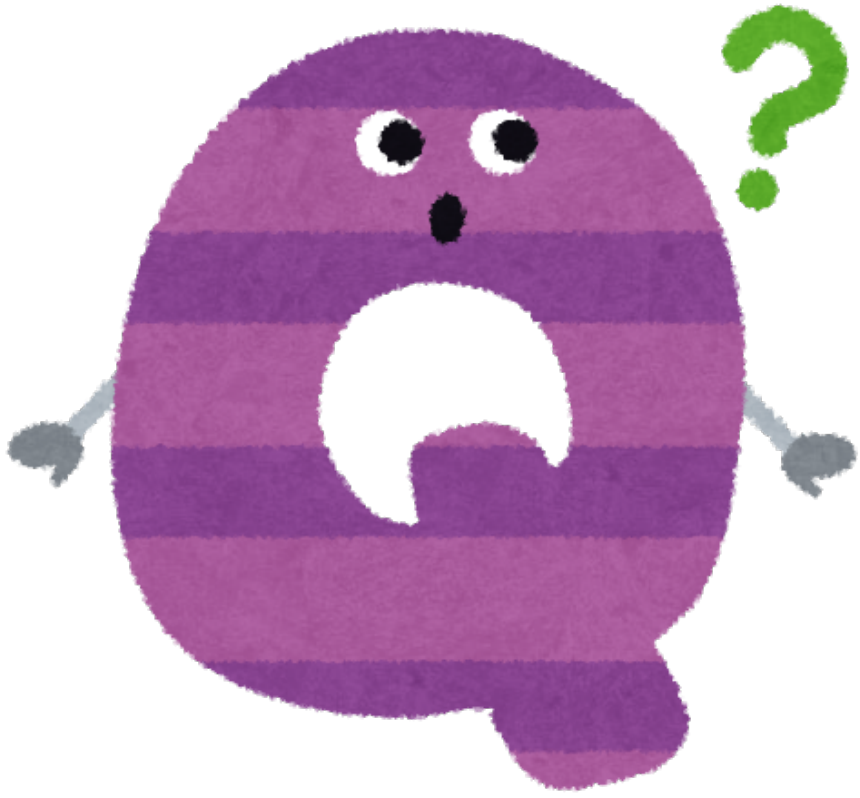
「ずっと使える知識」を獲得し活用する機械学習へ

- 手元のデータの法則性を発見するだけでなく，そこから「長く使える知識」を学習・蓄積する

- 今回のように「データ分布の背後に踏み込んで」「その背後にある知識を獲得する」というアプローチが重要だと考えています [4]

[4] 動機は Continual Learning 等に近いかもしれない

Frequently Asked Questions



How "small" is this small-data regime?

- Our experiments involved only 4 training data points.
  - $+$ 2 validation data. That's minimum requirement for our method $+$ evaluation to run.
- Even then, our method inflates the training data to $4^4 = 256$ points.
- This is the regime in which this approach is especially hopeful.

Computation time?

- The training dataset can easily explode as $\mathcal{O}(n^D)$ (though manageable for really small data).
- One can randomly take subsets of combinations (mini-batch of combinations). Theoretical justifications similar to incomplete U-statistics should be possible (future work) [34–36].
- Another approach: invariant models (neural network, marginal kernels, etc.)

Why need multiple source domains?

- The requirement depends on the employed NLICA method.
  - The paper employs NLICA based on generalized contrastive learning (GCL) [12] requiring multiple source domains.
- NLICA from one i.i.d. sample is known to be impossible.
  - Either: (1) strongly restrict the function class of f or (2) use auxiliary information.
- Any NLICA (or ICA) methods can be combined with our method without additional efforts.

Why V-process? Why not U-process?

- U-statistic is the minimum variance unbiased risk estimator.
- The answer is related to an edge case of statistics of generalized U-statistic (GU):
  1. If $f = \hat{f}$, then $\hat{R}_{\text{aug}}$ is the GU.
  2. If $f \neq \hat{f}$, then the theory of GU does not apply anymore, and the seemingly only alternative theory is V-statistics.
- In our setting, the risk estimator "smoothly quits" out of GU.
- I'm not aware of a theory for "V-statistics which is very close to GU."

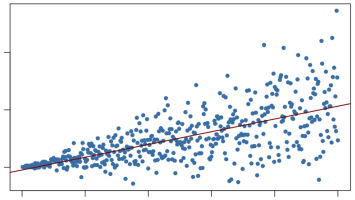
How much flexibility for fixed f? ①

Simple example: Covariate shift + target shift

$$\begin{cases} X = S_1 \\ Y = h(X) + S_2 \end{cases}$$

- Change in $S_1 \rightsquigarrow$ covariate shift
- Change in $S_2 \rightsquigarrow$ target dist. shift

A little more complex example: heteroskedastic noise



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

$$(X \neq 0)$$

How much flexibility for fixed f? ②

General case

- 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.

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

How to check the validity of assumptions?

Q. When does the assumption may hold and how to check the validity?

A. As is often the case in domain adaptation (DA), the scarcity of data disables data-driven testing of the transfer assumptions (TAs), and we need domain knowledge to judge the validity. For our TA, the intuitive interpretation as invariance of causal models (cf. lines 127-147) can be used.

What if only one source domain?

- The answer depends on the ICA method.
- GCL requires multiple source domains to operate and cannot be used with only a single source domain.
- If one can accept other identification conditions, one-sample ICA methods can be also used in the proposed approach.
  - e.g., linear ICA.
  - The theoretical analyses hold regardless of the method chosen.

Relation to previous work

Invariant risk minimization [27]

Invariant risk minimization (IRM) [27]

- Goal: *out-of-distribution (OOD) generalization*.
- Approach: Learn feature extractor that makes the optimal predictor invariant across domains.

Comparison (See [37] for more details)

- Problem setup/Assumption: no target domain data. Feature extractor that *elicits an invariant predictor* exists.
- Theory: only under a certain linearity assumption (essentially a relaxation of linear SEMs).
- Methodology: representation learning. Estimate a single best predictor after feature transformation.

Domain adaptation b/w different interventional states [26]

Domain adaptation under interventions [26]

- Goal: DA among different interventional states.
- Approach: variable selection via GCMs.

Comparison (See [37] for more details)

No strict overlap in problem settings (complementary).

- Problem setup/Assumption: Existence and identifiability of a separating set with small "incomplete information bias".
- Application: more suitable for fields with interventional experiments such as genomics.
- Methodology: variable selection. find a subset so that the conditional distribution is invariant.

Causal Generative Domain Adaptation Network (CG-DAN) [38]

Causal Generative Domain Adaptation Network [38]

- Goal: unsupervised domain adaptation.
- Approach: causal graph for generator architecture.

Comparison (See [37] for more details)

- Problem setup/Assumption: presumes *anticausal* scenario (i.e., Y is the cause of X). X given Y follows an SEM.
- Theory: no identifiability guarantee, i.e., no guarantee the learned generator is applicable across different domains.
- Methodology: estimate the GCM of X given Y. Uses it to design a generator neural network.

Related work: Other TAs

Transfer Assumption (TA)	AD	NP	Suited app. example
(1) Parametric dist. family [39] or shift [22, 23, 40, 41].	✓	-	Remote sensing [22].
(2) Invariant dist [42] $p(Y\|\mathcal{T}(X))$ Covariate shift $\mathcal{T} = \mathrm{Id}$[44] Transfer component $\mathcal{T}$ [45] Feature selection $\mathcal{T}$ [24, 26] TarS [22, 46] $p(X\|Y)$ R-vine copulas [33].	-	✓	BCI [43]
(3) Discrepancy [32, 47–51] / IPM [52] $+$ *ideal joint hypothesis* [49]	-	✓	Computer vision [52]
(4) Param-transfer [53]	✓	✓	Computer vision [53, 54]
(Ours) Mechanism	✓	✓	Medical records [28]

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

Related work: Other TAs

Transfer Assumption (TA)	AD	NP	Suited app. example
(1) Parametric dist. family [39]	✓	-	Remote sensing [22].
or shift [22, 23, 40, 41].			
(2) Invariant dist [42] $p(Y\|\mathcal{T}(X))$	-	✓	BCI [43]
Covariate shift $\mathcal{T} = \mathrm{Id}$[44]			
(Ours) Mechanism	✓	✓	Medical records [28]

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

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

Technical details

Preliminary: Causal model (SEMs)

Structural Equation Models (SEMs) [4–6]

(a.k.a. Structural Causal Models; SCMs)

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

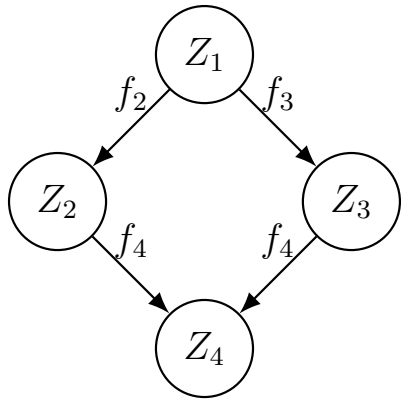
Functions $\mathcal{F} = \{f_d\}_{d=1}^D$

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

Distribution q of independent r.v.'s $\{S_i\}_{i=1}^D$

$$q(S) = \prod_{d=1}^{D} q_d(S_d)$$

(Directed acyclic graph (DAG) $\mathcal{G}$ encoding $\mathcal{F}$)



Simplified: acyclic, Markovian, same-dimensional latent. See [6, 7].

SEM: Formal definition [7]

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 [7]

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 [7]

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 [7]

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].

Invertible neural networks (INNs)

- Neural networks that are invertible by design.
- We used Glow architecture [55]

Affine coupling layer

- Coupling layer: keep some dimensions unchanged.

$$\begin{pmatrix} x_{1:d} \\ x_{d+1:D} \end{pmatrix} \mapsto \begin{pmatrix} x_{1:d} \\ s(x_{1:d}) \odot x_{d+1:D} + t(x_{1:d}) \end{pmatrix}$$

- Exact inversion using an analytic formula (recompute s and t from $x_{1:d}$)

Generalized contrastive learning

- Nonlinear ICA has been realized [12, 56–58].
- Exploit auxiliary info (e.g. temporal dependence)[5].

Generalized contrastive learning [12] 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 $(n \to \infty)$.

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

Identification condition of GCL

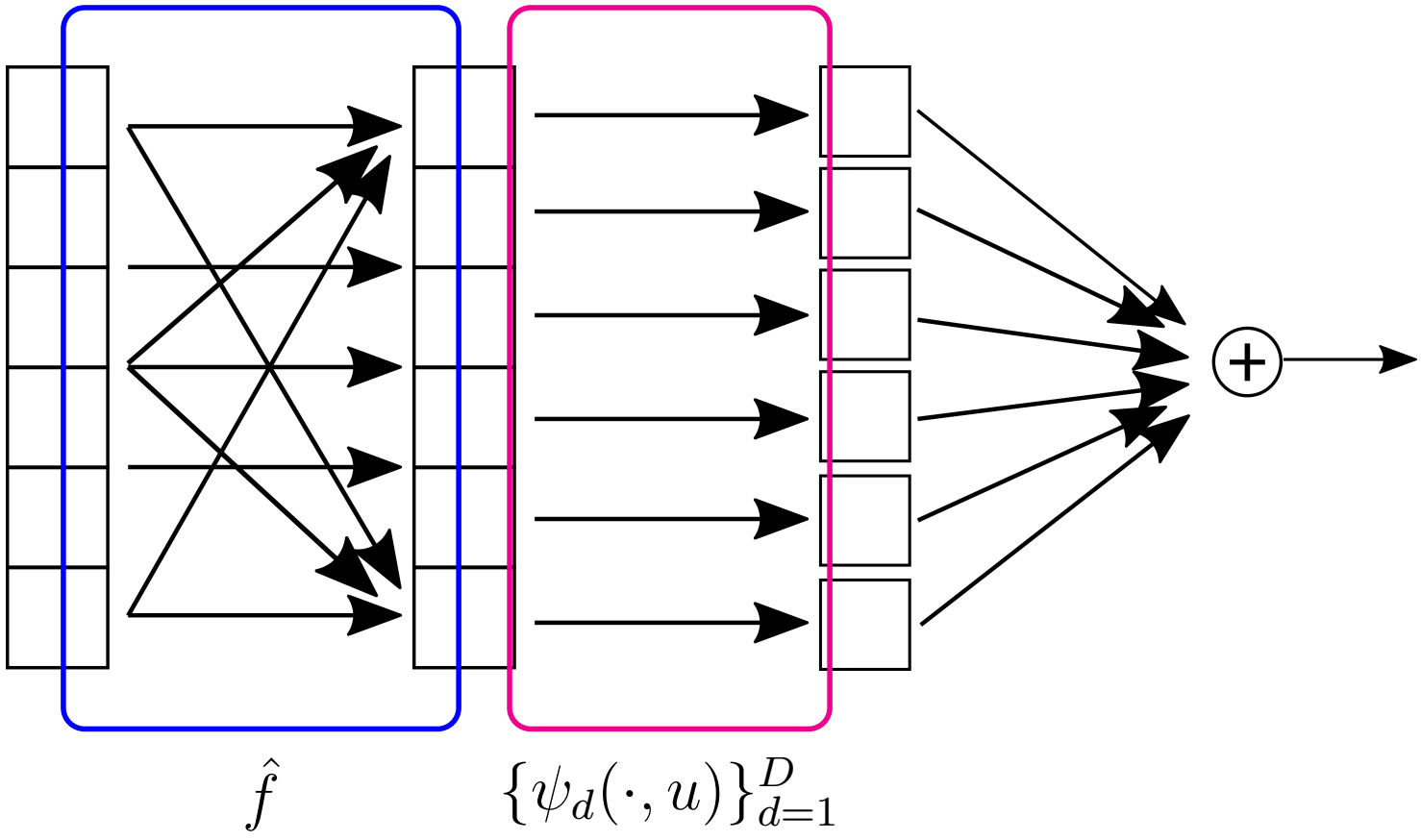
Identification condition of GCL [12]

- $\{(z_i, u_i)\}_{i=1}^n$: Data $(u_i \in \mathcal{U})$
- Conditionally independent $q(s|u) = \prod_{d=1}^D q^{(d)}(s^{(d)}|u)$
- Assumption of variability [12]: for any z, distinct $\{u_j\}_{j=0}^{2D} \subset \mathcal{U}$
  $\{w(z|u_j) - w(z|u_0)\}_{j=1}^{2D}$ are linearly independent w/ $w(z|u) :=$
  $\left(\frac{\partial \log q^{(1)}(z_1|u)}{\partial z_1}, \ldots, \frac{\partial \log q^{(D)}(z_D|u)}{\partial z_D}, \frac{\partial^2 \log q^{(1)}(z_1|u)}{\partial z_1^2}, \ldots, \frac{\partial^2 \log q^{(D)}(z_D|u)}{\partial z_D^2}\right)$.
- Some other regularity conditions

Theorem 1 of [12]

Then, upto dimension-wise invertible transformations, GCL outputs a consistent estimator of f.

NLICA via GCL: How it works



- To estimate $p(u|X)$ by this restricted architecture, $\hat{f}$ needs to extract independent components.

Theoretical analysis detail

A1. Theorem: UMVUE if $\hat{f} = f$

- $\hat{R}_{\mathrm{aug}}(g)$ is the (unique) UMVUE of $R(g)$ on $\mathcal{Q}$, i.e.,
- $\boxed{\forall \hat{R}(g) \text{ :unbiased, } \forall q \in \mathcal{Q},\ \mathrm{Var}(\hat{R}_{\mathrm{aug}}(g)) \leq \mathrm{Var}(\hat{R}(g))}$
- $\because$ Rewrite $\hat{R}_{\mathrm{aug}}(g)$ as generalized U-statistic [59].

A2. Theorem: Excess risk bound for $\hat{f} \neq f$

- Under appropriate assumptions, w/ prob. $\geq 1-(\delta+\delta')$,

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

$\mathfrak{R}(\mathcal{G})$: Reduced Rademacher complexity, $\|\cdot\|_{W^{1,1}}$: $(1,1)$-Sobolev norm

Theoretical analysis: How

- Interpret $\hat{R}_{\text{aug}}(g)$ 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,

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

- This is the V-statistic [59] of

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

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

Theoretical analysis detail ①

Q1. What does it mean to exploit independence?

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 $\mathcal{Q}$.
  - $\forall \hat{R}(g)$:unbiased, $\forall q \in \mathcal{Q}$, $\mathrm{Var}(\hat{R}_{\mathrm{aug}}(g)) \leq \mathrm{Var}(\hat{R}(g))$
  - Special case: $\mathrm{Var}(\hat{R}_{\mathrm{aug}}(g)) \leq \mathrm{Var}(\hat{R}_{\mathrm{ERM}}(g))$

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

Proof (Details are skipped)

Rewrite $\hat{R}_{\mathrm{aug}}(g)$ as generalized U-statistic [59] of $R(g)$.

Theoretical analysis detail ②

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

Theorem: generalization error bound

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

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

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

$\mathfrak{R}(\mathcal{G})$: Reduced Rademacher complexity

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

Proof outline $\hat{R}_{\mathrm{aug}}(g)$ is V-statistic [59] of $R(g)$.

Concentration of V-process + Evaluation of $\left\| Q - \check{Q} \right\|_{L^1}$.

Theoretical analysis detail ②

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:

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

  - $\tilde{\ell}(s_1, \ldots, s_D) := \frac{1}{D!} \sum_{\pi \in \mathfrak{S}_D} \ell(g, \hat{f}(s^{(1)}_{\pi(1)}, \ldots, s^{(D)}_{\pi(D)}))$,
  - $\{\sigma_i\}_{i=1}^n$: 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.

なぜ「因果」を考えるか (その後)

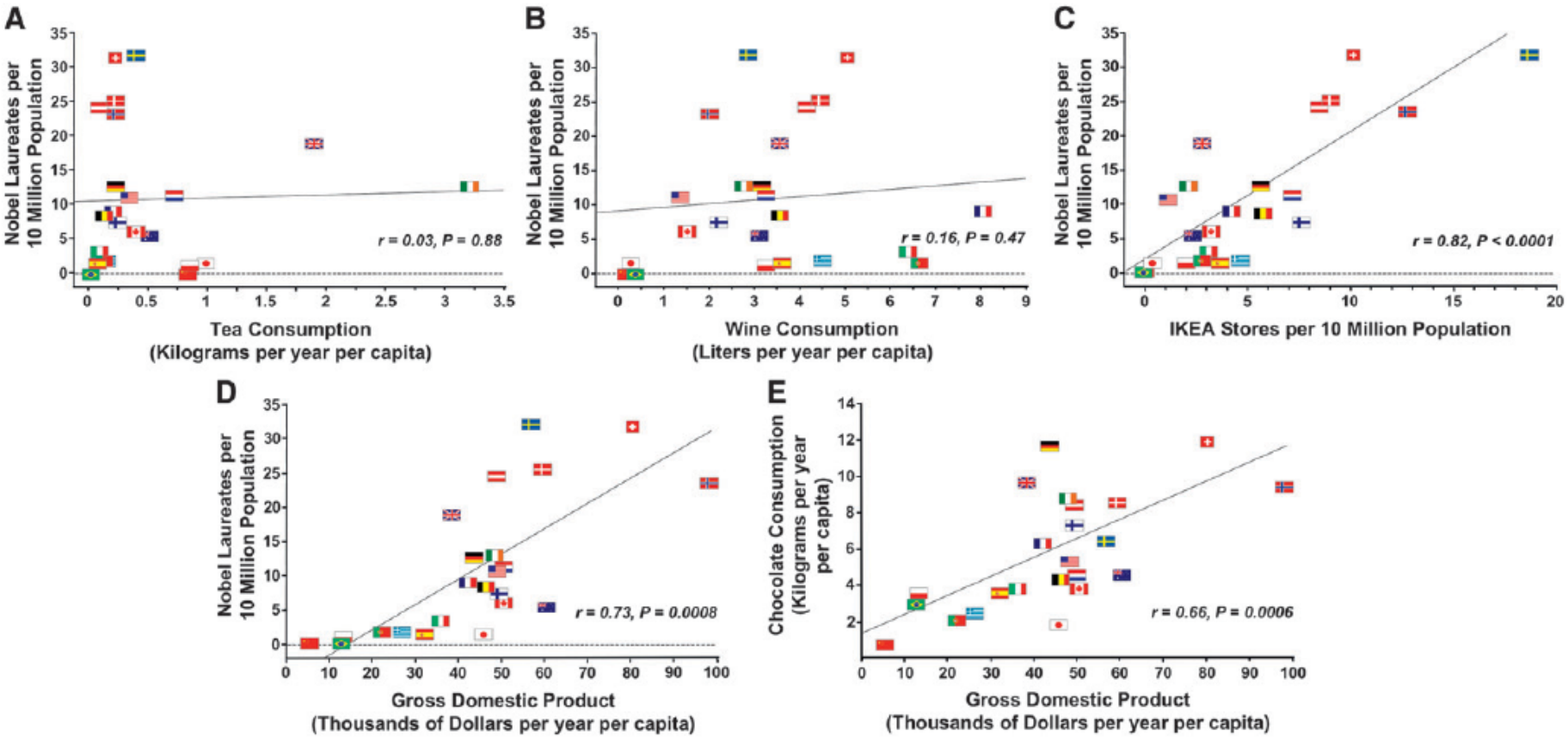


Figure: [60]

- チョコと同じ種類の栄養を豊富に含む他の食材→相関弱い
- 北欧・西欧＋ GDP でだいたい説明がつく

因果の定式化にはデータ生成過程のモデル化が必要

- 通常の機械学習では正確な生成過程モデルは不要

確率モデル (通常の機械学習)

- 受動的予測にフォーカス. $p(\mathrm{Nobel}|\mathrm{choco} = 10kg)$
- データの確率分布をモデル化すれば大体足りる

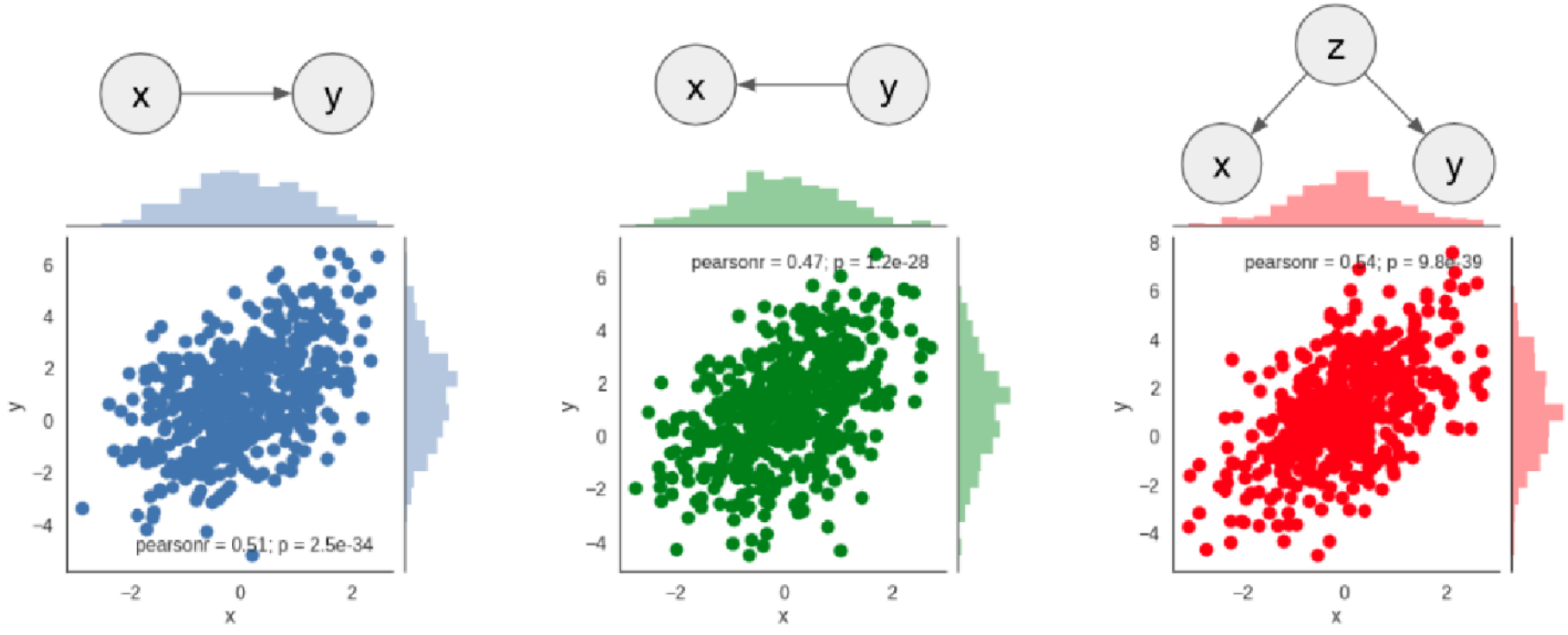
因果モデル [4]

- データの生成過程をモデル化する (分布も含意).
- 介入の結果も予測. $p(\mathrm{Nobel}|\mathrm{do}(\mathrm{choco} = 10kg))$

- より強い仮定→確率モデルより強いことが出来る.

因果の定式化にはデータ生成過程のモデル化が必要

- 確率モデルだけでは十分ではないというデモ




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

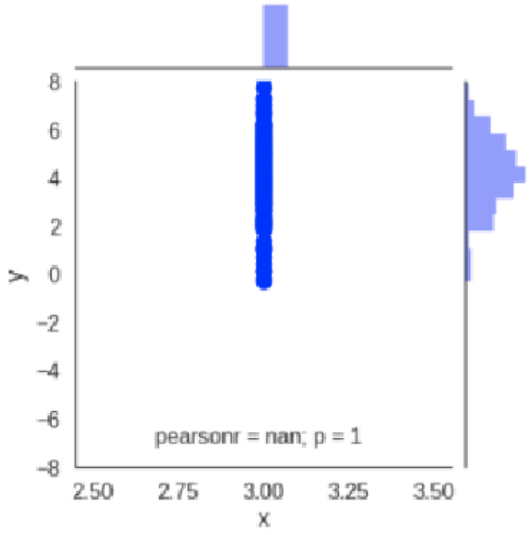
```
y = 1 + 2*randn()
x = (y-1)/4 + sqrt(3)*randn()/2
```

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

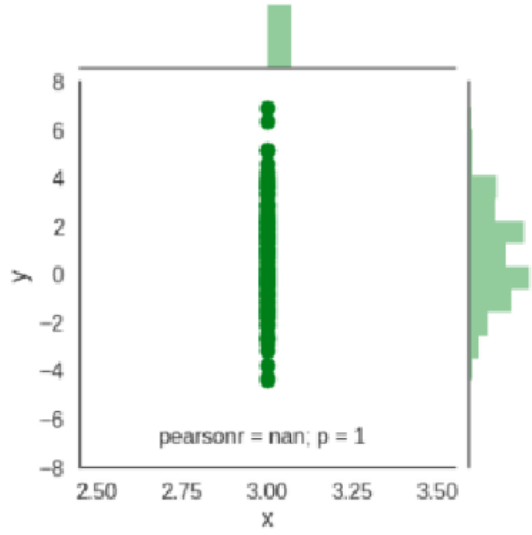
Figure: [61]

- 同じ分布になるような異なる生成過程

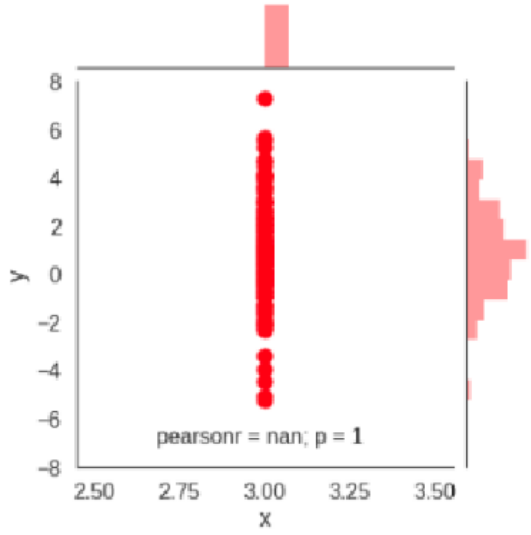
因果の定式化にはデータ生成過程のモデル化が必要



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

Figure: [61]

- 介入 $\mathrm{do}(X = 3)$ のもとでは異なる振る舞い
- 因果を扱うには分布の背後のデータ生成過程を考える必要あり

構造的因果モデル定式化の発展 (☆)

More generic definition

- Allowing cycles: acyclic $\rightarrow$ simple SCMs [6].

Constraint causal models

- Expressing constraints such as $pV = nRT$ [62].

Micro-foundations for SCMs

- From ODEs to SCMs [63].

Algorithmic causal models

- Kolmogorov 複雑度で「条件付き独立」を置き換える (データ 1 点ごとの議論が出来る) [64]

By-ML: 因果探索 [8]

Approach	Example	Ref.
(1) Constraint based	PC, FCI	[8]
(2) Score based	GES	[8, 9]
(3) Restricted SEM based	ANM, PNL	[8, 11]
(4) ICA-based	LiNGAM	[8, 10, 12]
Others	JCI	[13–15, 65]

- 最適化も進展 [66]
- サーベイも [8]

* This is only an incomplete list. Many methods estimate only the graph.

References

[1] F. H. Messerli, 'Chocolate Consumption, Cognitive Function, and Nobel Laureates', *New England Journal of Medicine*, vol. 367, no. 16, pp. 1562–1564, Oct. 2012.

[2] A. Eggers, *Multivariate relationships*, Feb. 2016.

[3] M. A. Hernn and J. M. Robins, *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

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

[5] S. Wright, 'Correlation and causation', *Journal of Agricultural Research*, vol. 20, no. 7, pp. 557–585, 1921.

[6] S. Bongers, P. Forr, J. Peters, B. Schlkopf, and J. M. Mooij, 'Foundations of structural causal models with cycles and latent variables', *arXiv:1611.06221 [cs, stat]*, May 2020. arXiv: 1611.06221 [cs, stat].

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

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

References (cont.)

[9] B. Huang, K. Zhang, Y. Lin, B. Schlkopf, and C. Glymour, 'Generalized Score Functions for Causal Discovery',, ACM Press, 2018, pp. 1551–1560.

[10] S. Shimizu, P. O. Hoyer, A. Hyvrinen, and A. J. Kerminen, 'A linear non-Gaussian acyclic model for causal discovery', *The Journal of Machine Learning Research*, vol. 7, no. Oct, pp. 2003–2030, 2006.

[11] J. Peters, J. Mooij, D. Janzing, and B. Schoelkopf, 'Identifiability of Causal Graphs using Functional Models', *arXiv:1202.3757 [cs, stat]*, Feb. 2012. arXiv: 1202.3757 [cs, stat].

[12] A. Hyvrinen, 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.

[13] 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].

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

References (cont.)

[15] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniuis, B. Steudel, and B. Schlkopf, 'Information-geometric approach to inferring causal directions', *Artificial Intelligence*, vol. 182, pp. 1–31, May 2012.

[16] K. Chalupka, P. Perona, and F. Eberhardt, 'Visual causal feature learning', in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, M. Meila and T. Heskes, Eds., Amsterdam, the Netherlands: AUAI Press, 2015, pp. 181–190.

[17] K. Chalupka, T. Bischoff, F. Eberhardt, and P. Perona, 'Unsupervised discovery of El Nio using causal feature learning on microlevel climate data', in *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, A. T. Ihler and D. Janzing, Eds., new york city, NY, USA: AUAI Press, 2016.

[18] K. Chalupka, F. Eberhardt, and P. Perona, 'Causal feature learning: An overview', *Behaviormetrika*, vol. 44, no. 1, pp. 137–164, Jan. 2017.

[19] J. Peters, D. Janzing, and B. Schlkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*, ser. Adaptive Computation and Machine Learning Series. Cambridge, Massachuestts: The MIT Press, 2017.

References (cont.)

[20] B. Schlkopf, 'Causality for Machine Learning', *arXiv:1911.10500 [cs, stat]*, Dec. 2019. arXiv: 1911.10500 [cs, stat].

[21] B. Schlkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij, 'On causal and anticausal learning', in *Proceedings of the 29th International Coference on Machine Learning*, Omnipress, 2012, pp. 459–466.

[22] K. Zhang, B. Schlkopf, 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.

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

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

References (cont.)

[25] B. Schlkopf, D. Hogg, D. Wang, D. Foreman-Mackey, D. Janzing, C.-J. Simon-Gabriel, and J. Peters, 'Removing systematic errors for exoplanet search via latent causes', in *International Conference on Machine Learning*, Jun. 2015, pp. 2218–2226.

[26] 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.

[27] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, 'Invariant Risk Minimization', *arXiv:1907.02893 [cs, stat]*, Mar. 2020. arXiv: 1907.02893 [cs, stat].

[28] 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.

[29] W. H. Greene, *Econometric Analysis*, Seventh. Boston: Prentice Hall, 2012.

References (cont.)

[30] D. Pardoe and P. Stone, 'Boosting for regression transfer', in *Proceedings of the Twenty-Seventh International Conference on Machine Learning*, Haifa, Israel, 2010, pp. 863–870.

[31] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama, 'Relative density-ratio estimation for robust distribution comparison', in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2011, pp. 594–602.

[32] 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.

[33] D. Lopez-paz, J. M. Hernndez-lobato, and B. Schlkopf, '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.

[34] S. Clmenon, 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.

References (cont.)

[35] G. Papa, S. Clmenon, and A. Bellet, 'SGD Algorithms based on Incomplete U-statistics: Large-Scale Minimization of Empirical Risk', in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., Curran Associates, Inc., 2015, pp. 1027–1035.

[36] S. Robbiano and J. Tressou, 'Maximal deviations of incomplete U-statistics with applications to empirical risk sampling', in *Proceedings of the 2013 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, May 2013, pp. 19–27.

[37] T. Teshima, I. Sato, and M. Sugiyama, 'Few-shot domain adaptation by causal mechanism transfer', in *Proceedings of Machine Learning and Systems 2020*, 2020, pp. 1820–1831.

[38] M. Gong, K. Zhang, B. Huang, C. Glymour, D. Tao, and K. Batmanghelich, 'Causal generative domain adaptation networks', *arXiv:1804.04333 [cs, stat]*, Apr. 2018. arXiv: 1804.04333 [cs, stat].

References (cont.)

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

[40] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schlkopf, '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.

[41] 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.

[42] J. Quionero-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.

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

References (cont.)

[44] 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.

[45] 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.

[46] 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.

[47] 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. Schlkopf, J. C. Platt, and T. Hoffman, Eds., MIT Press, 2007, pp. 137–144.

[48] 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.

References (cont.)

[49] 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.

[50] 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.

[51] 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.

[52] 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.

References (cont.)

[53] 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.

[54] 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.

[55] D. P. Kingma and P. Dhariwal, 'Glow: Generative flow with invertible 1x1 convolutions', 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. 10 215–10 224.

[56] A. Hyvrinen 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.

References (cont.)

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

[58] I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvrinen, 'Variational autoencoders and nonlinear ICA: A unifying framework', *arXiv:1907.04809 [cs, stat]*, Jul. 2019. arXiv: 1907.04809 [cs, stat].

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

[60] P. Maurage, A. Heeren, and M. Pesenti, 'Does Chocolate Consumption Really Boost Nobel Award Chances? The Peril of Over-Interpreting Correlations in Health Studies', *The Journal of Nutrition*, vol. 143, no. 6, pp. 931–933, Jun. 2013.

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

[62] T. Blom, S. Bongers, and J. M. Mooij, 'Beyond Structural Causal Models: Causal Constraints Models', *arXiv:1805.06539 [cs, stat]*, Aug. 2019. arXiv: 1805.06539 [cs, stat].

References (cont.)

[63] P. K. Rubenstein, S. Bongers, B. Schoelkopf, and J. M. Mooij, 'From Deterministic ODEs to Dynamic Structural Causal Models', *arXiv:1608.08028 [cs]*, Jul. 2018. arXiv: 1608.08028 [cs].

[64] D. Janzing and B. Scholkopf, 'Causal Inference Using the Algorithmic Markov Condition', *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 5168–5194, Oct. 2010.

[65] D. Janzing, B. Steudel, N. Shajarisales, and B. Schlkopf, 'Justifying Information-Geometric Causal Inference', in *Measures of Complexity*, V. Vovk, H. Papadopoulos, and A. Gammerman, Eds., Springer International Publishing, 2015, pp. 253–265.

[66] X. Zheng, C. Dan, B. Aragam, P. Ravikumar, and E. Xing, 'Learning Sparse Nonparametric DAGs', in *International Conference on Artificial Intelligence and Statistics*, Jun. 2020, ch. Machine Learning, pp. 3414–3425.