因果的知識は予測モデルの役に立つか？

Takeshi Teshima

[1]The University of Tokyo [2]RIKEN

手嶋 毅志

Junior Research Associate @ RIKEN AIP

Ph.D. candidate (3rd–year) @ UTokyo

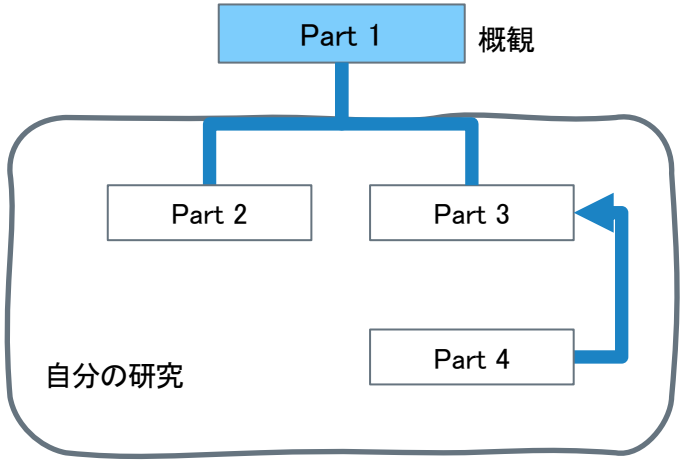
Masason Foundation (3rd)

最近の関心：PeaceTech・AI for Good
（お声掛けください）

https://takeshi-teshima.github.io

Twitter @DiadochosT

Part 1:「機械学習のための因果」概観



因果関係への好奇心・探究心

I would rather discover one causal law than be King of Persia.
Democritus (460–370 B.C.)

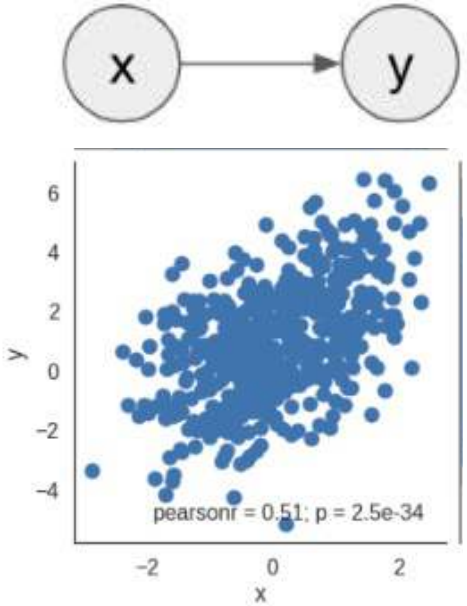
私はペルシャの王になるより
因果の法則を一つでも発見したい

なぜ……？

出典

出典

（統計的）因果とはデータ生成過程の構造



統計的因果関係

- ▷ 確率変数の間の（確率的な）決定関係
- ▷ Xの値を変えると、Yの値が変わる、など

- ▷ 同時確率分布を考えるだけでは捉えられない、「ある変数が異なる変数に影響を与える」という背景構造

※このような構造がいつでもあるとは当然限らないが、このような構造があるという信念を許容できるとき使われる。

データ生成過程を考える理由

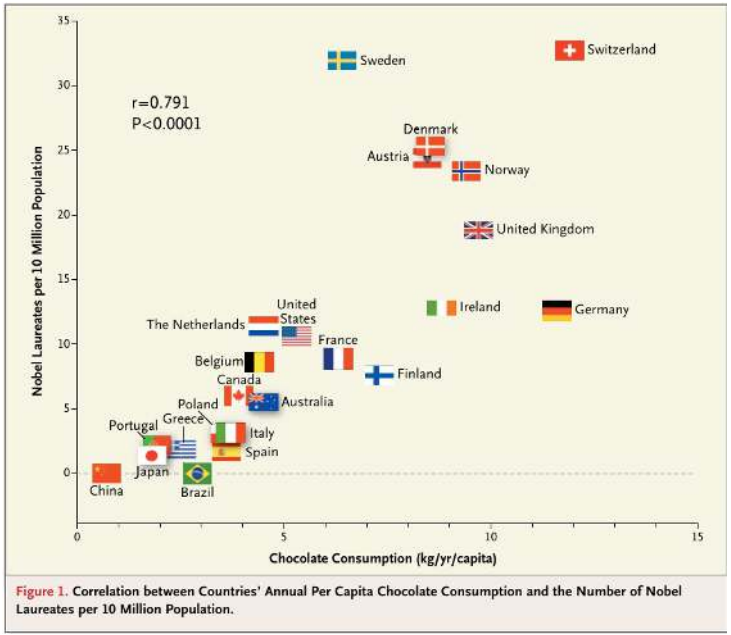


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

(Messerli, 2012)

横軸：年間一人当たりチョコレート消費量
縦軸：人口1000万人あたりのノーベル賞受賞者数
明確な正の相関が見られる

チョコ消費量とノーベル賞数の関係

1. チョコ → ノーベル賞

2. チョコ ← ノーベル賞

3. 経済水準etc. → チョコ, 経済水準etc. → ノーベル賞

（それともそれ以外か？）

なぜデータ生成過程が重要か

同じ分布を誘導する、異なる3つの生成過程

```
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
```

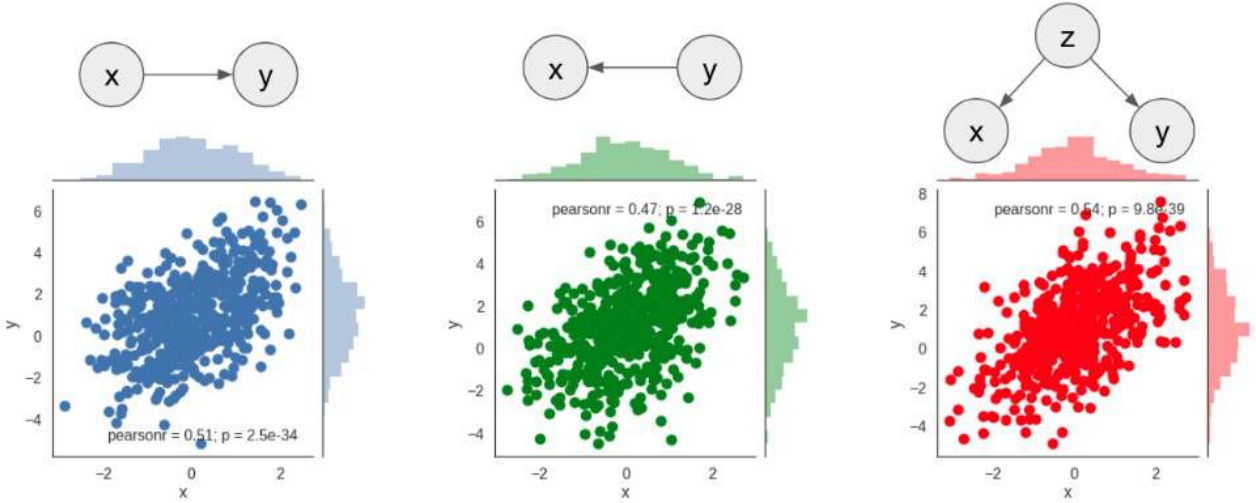


図: (Huszár, 2019)

なぜデータ生成過程が重要か

介入 $\mathrm{do}(X = 3)$ のもとでは異なる振る舞いをする $p(Y|\mathrm{do}(X = 3))$

```
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
```

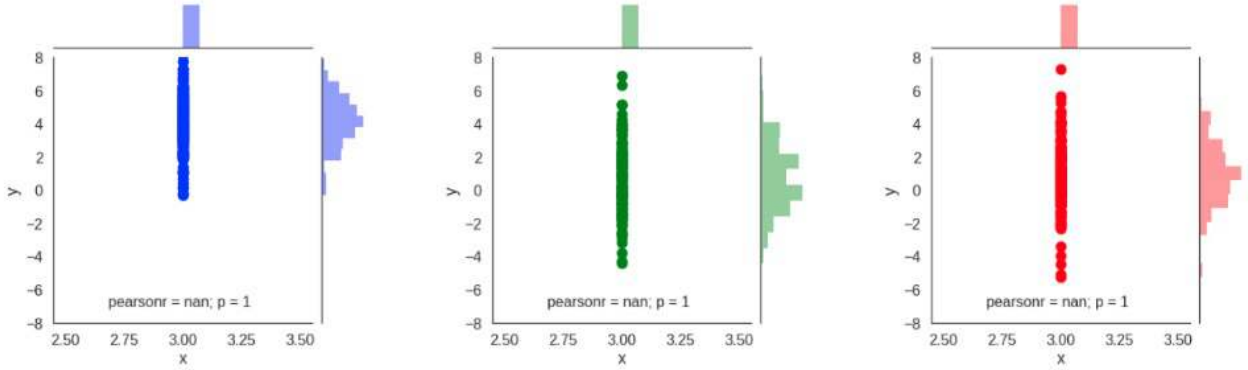


図: (Huszár, 2019)

- 因果を扱う枠組みでは分布の背後のデータ生成過程を考える必要がある

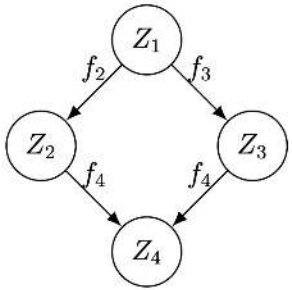
因果モデルの定式化

構造方程式モデル (SEM)/構造的因果モデル (SCM)

(Pearl, 2009)
(Bongers et al., 2021)

「変数間の因果関係は，決定論的な関数関係で捉えられる」という考え方に基づくデータ生成過程のモデル．データ $Z = \{Z_d\}_{d=1}^{D}$ の生成過程を $(\mathcal{F}, q)$ でモデル化

非巡回有向グラフ $\mathcal{G}$



変数間の直接的依存関係を定性的に表現したグラフ

構造方程式 $\mathcal{F} = \{f_d\}_{d=1}^{D}$

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

pa_d は $\mathcal{G}$ における Z_d の親

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

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

S が確率変数→ $\mathcal{F}$ に代入されることで Z が確率変数になる

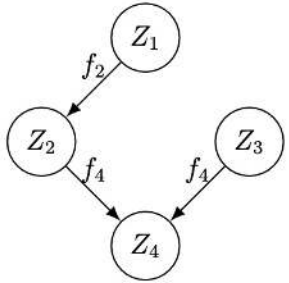
Rubinの潜在反応フレームワークとPearlの構造的因果モデルフレームワークがあるが，
ここでは後者のみ説明．また非巡回，マルコフ的，同次元外生変数という仮定を暗に置いている．

因果モデルの定式化：使用法(☆)

完全介入(Perfect intervention) (Pearl, 2009)

- 構造的因果モデルを考えれば「介入」の定式化が出来る

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



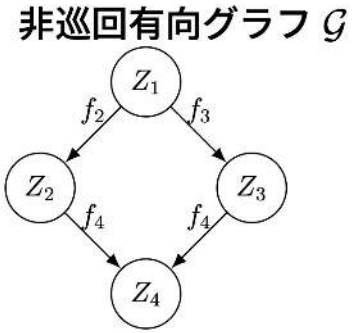
- 介入後分布 $p(Z|\mathrm{do}(Z_I = \zeta_I))$ は $(q, \mathcal{F}')$ により生成される分布

- 実は介入後分布は $\mathcal{G}$ と介入前分布 $p(Z)$ から計算できる ($\mathcal{F}$ が不要)

不完全な介入は例えば (Eaton and Murphy, 2007) を参照

因果モデルの定式化

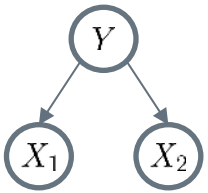
- 因果グラフ (Pearl, 2009) (Bongers et al., 2021)
  構造方程式の引数関係を定性的に表したグラフ



- 因果グラフィカルモデル (Pearl, 2009)
  データの確率分布＋因果グラフ
  ＋介入によってグラフと分布がどう変化するかという仮定

- 因果グラフから条件付き独立性(CI)関係を読み取れる

(Pearl, 2009) (Richardson, 2003)



⇨ $X_1 \perp\!\!\!\perp X_2 \mid Y$

（後のパートで関係）

因果モデルの階層構造

- 因果モデルは階層構造の関係にある

Model	Predict in IID setting	Predict under distr. shift/intervention	Answer counter-factual questions	Obtain physical insight	Learn from data
Mechanistic/physical	yes	yes	yes	yes	?
Structural causal	yes	yes	yes	?	?
Causal graphical	yes	yes	no	?	?
Statistical	yes	no	no	no	yes

(Schölkopf, 2019)

- 上から順に詳細なモデル
  - 物理的因果モデル（微分方程式による） (Mooij et al., 2013)
  - 構造的因果モデル←条件のもとモデルの一部を推定可能
  - 因果グラフィカルモデル←ある程度の条件のもとで推定可能
  - 統計モデル（確率分布モデル）←標準的に推定可能

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 (Hernán and Robins, 2020)

- 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 (Pearl, 2009)

- 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

因果関係への好奇心・探究心

I would rather discover one causal law than be King of Persia.
Democritus (460–370 B.C.)

私はペルシャの王になるより
因果の法則を一つでも発見したい

出典

出典

なぜ……？

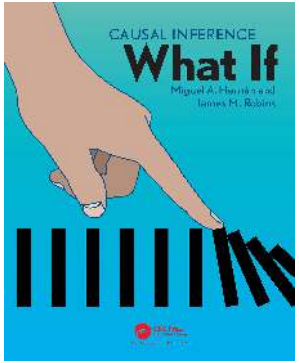
A1. 介入結果の推論ができるから

それだけだろうか……？

なぜ因果を知りたいか？

A1. 反実仮想・介入効果の推論ができるから

- 例：ワクチンの効果(重症率低下)を測りたい
- 使い道は明確……介入の設計



(Hernán and Robins, 2020)

A2. 世界の構造を知りたいから

- 例：ワクチンが効果を発揮する機序を知りたい
- なぜそう思うのか？人間が「因果を知ると役立つ」と思っているなら機械学習に役立つ要素もあるのでは？

因果的知識は予測モデルの役に立つか？

- 因果を考えないことによる不都合の解消
  Amazonでカバンをカートに入れると，パソコンを推薦される
  （同時に買う可能性は高いかもしれないが因果が逆）

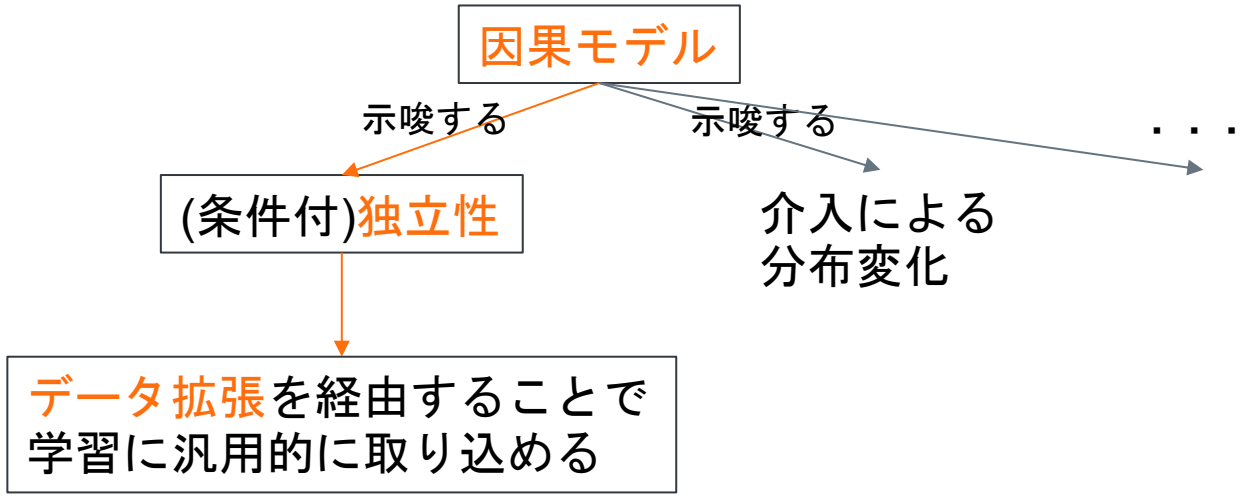
- 機械学習モデルの説明可能性・公平性への応用
  特定の予測が出た「原因」を説明（例．クレジットカード審査）
  特定の特徴量が予測に対し「直接の」影響を与えないようにする
  (Vig et al., 2020) (Wu et al., 2019) (Chikahara et al., 2021)

- 因果モデルの存在を（仮定して）事前知識として使う
  直観：「因果」は「異なる系でも有効」という特性を持つ
  （そうでないと我々は因果として認識しない）
  →使い回せる知識として少数データ学習に活かせる可能性

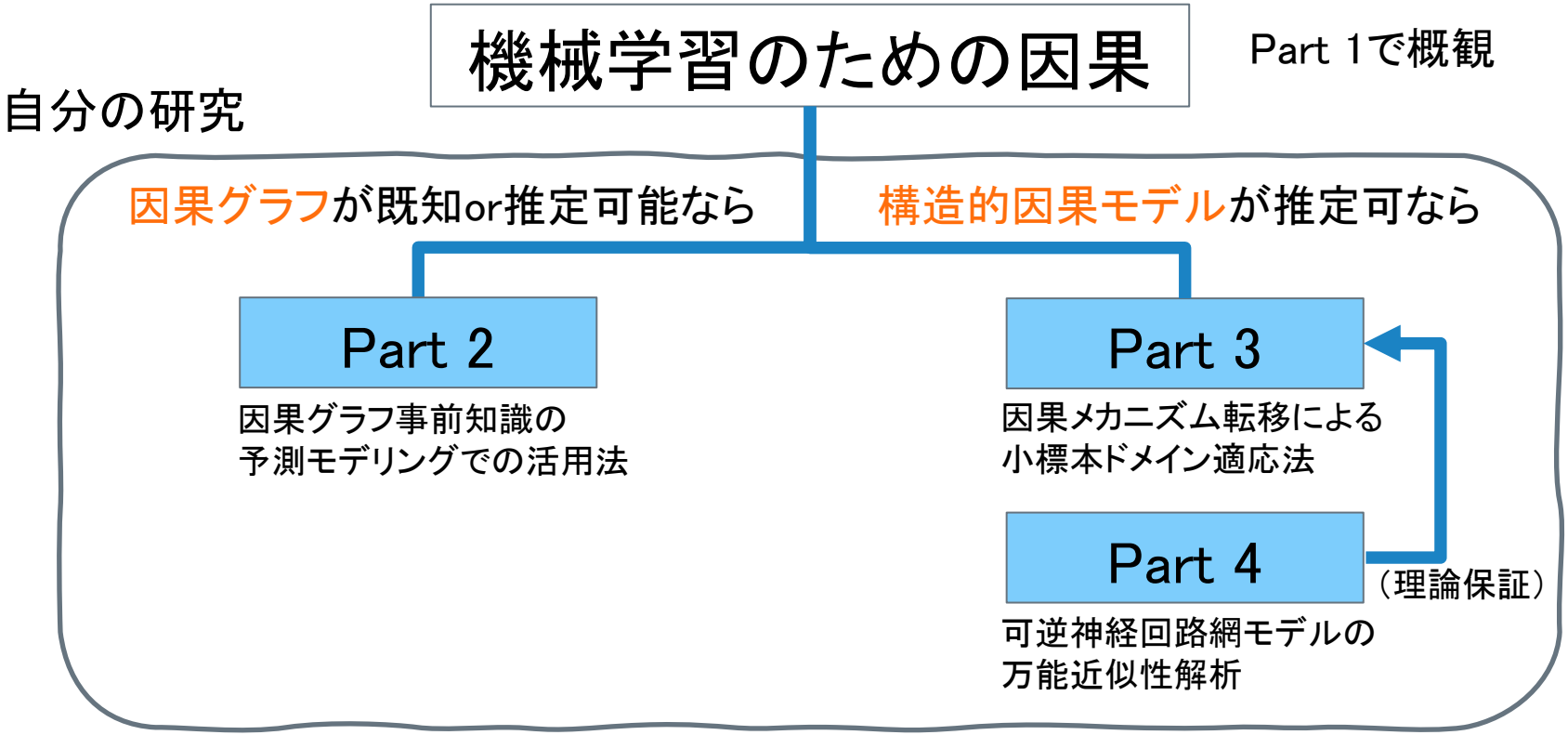
ここからの試み

基本のアイデア

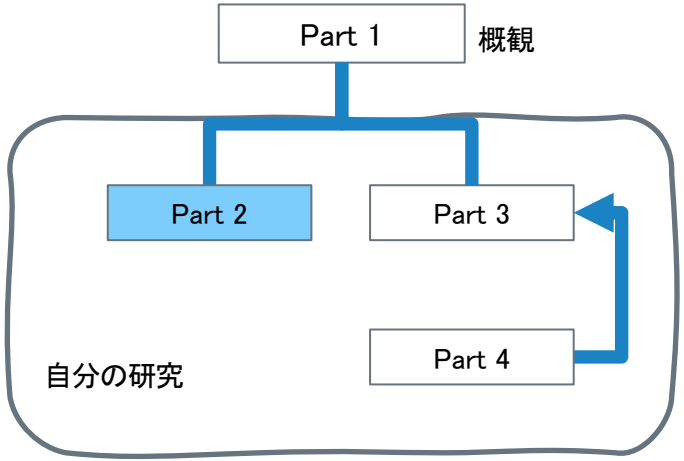
- 因果モデルが示唆する「統計的独立性」はデータ拡張を通して活用できる



本日の内容


機械学習のための因果
Part 1で概観
自分の研究
因果グラフが既知or推定可能なら
構造的因果モデルが推定可なら
Part 2
因果グラフ事前知識の
予測モデリングでの活用法
Part 3
因果メカニズム転移による
小標本ドメイン適応法
Part 4
（理論保証）
可逆神経回路網モデルの
万能近似性解析

Part 2: 因果グラフの事前知識があるとき



因果グラフの事前知識を簡単なデータ拡張で予測モデリングに組み込む方法

Teshima, T. and Sugiyama, M.,
Incorporating causal graphical prior knowledge into predictive modeling via simple data augmentation.
37th Conference on Uncertainty in Artificial Intelligence, accepted.

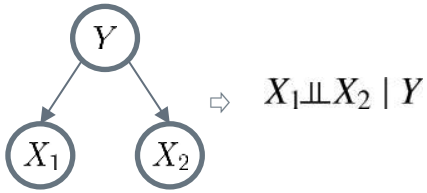
37th Conference on Uncertainty in Artificial Intelligence
July 27-30, 2021
Online
uai2021

背景：因果グラフ

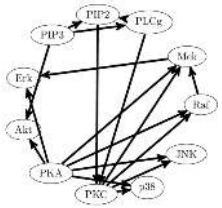
因果グラフ（CGs） (Pearl, 2009)

データの生成過程に関する我々の知識/仮定を表現するグラフ

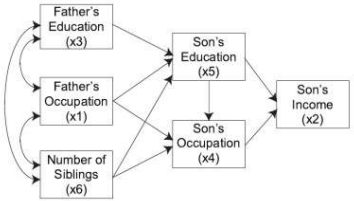
因果グラフからは条件付き独立性(CI)関係を読み取れる (Pearl, 2009) (Richardson, 2003)



例：



Biology (Sachs et al., 2005)



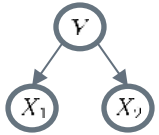
Sociology (Shimizu et al., 2011) (Duncan et al., 1972)

リサーチクエスチョン

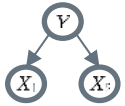
因果グラフの形で得られた事前知識は、どうすれば予測に活用できるか？

アイデア：データ拡張で条件付き独立性を取り込む

例（3変数のケース）

因果グラフ  が与えられたもとで Y を (X_1, X_2) から予測する

アイデア：データ拡張

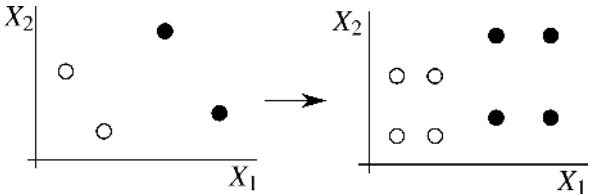
因果グラフ  からは $X_1 \perp\!\!\!\perp X_2 \mid Y$ が従う

⇨ X_1 と X_2 を、Y でグループ分けした学習データ内で交換する

Y	X_1	X_2
○	a	c
○	b	d
●	α	γ
●	β	δ

Y	X_1	X_2
○	a	c
○	a	d
○	b	c
○	b	d

Y	X_1	X_2
●	α	γ
●	α	δ
●	β	γ
●	β	δ



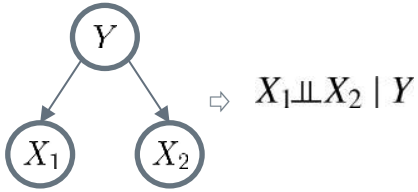
Q. 一般のグラフの場合はどうするか？

背景：因果グラフ

因果グラフ（CGs） (Pearl, 2009)

データの生成過程に関する我々の知識/仮定を表現するグラフ

因果グラフからは条件付き独立性(CI)関係を読み取れる (Pearl, 2009) (Richardson, 2003)

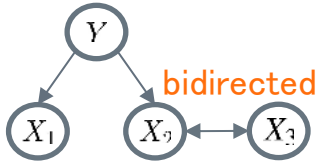


非巡回有向混合グラフ（ADMGs） (Richardson, 2003) (Richardson et al., 2017)

非巡回有向グラフに、双方向辺も許したグラフ $\mathcal{G} = ([D], \mathcal{E}, \mathcal{B})$

未観測変数がある因果モデルに対応して現れる

（準マルコフモデル; cf. Latent projection）. (Tian and Pearl, 2002)



Topological ADMG Factorization (Tian and Pearl, 2002) (Bhattacharya et al., 2020)

ADMG $\mathcal{G}$ を誘導する準マルコフモデルについて、$p(\mathbf{Z}) = \prod_{j=1}^{D} p_{j|\mathrm{mp}(j)}(Z^j|\mathbf{Z}^{\mathrm{mp}(j)})$ が成立する

$\mathrm{mp}(j)$：“Markov pillow” of variable Z^j (Generalization of “parents” in ADMGs.)

問題設定とゴール

$\mathbf{Z} = (Z^1, \ldots, Z^D) \sim p$: X と Y のデータ対（各 Z^j は連続か離散か）

主仮定

- $p(\mathbf{Z})$ が $\mathcal{G}$ に関してtopological ADMG factorizationを満たす

$$p(\mathbf{Z}) = \prod_{j=1}^{D} p_{j|\mathrm{mp}(j)}(Z^j|\mathbf{Z}^{\mathrm{mp}(j)})$$ (Bhattacharya et al., 2020)

所与のもの：

- ラベル付きデータ $\mathcal{D} = \{\mathbf{Z}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} p$
- 真のADMG $\mathcal{G}$ の推定値 $\hat{\mathcal{G}}$

ゴール（教師付き学習）

予測器 $f : X \mapsto Y$ のうち $R(f) = \mathbb{E}[\ell(f, \mathbf{Z})]$ が小さいものを見つける

一般のADMGの場合の提案手法

- Topological ADMG factorization: $p(\mathbf{Z}) = \prod_{j=1}^{D} p_{j|\mathrm{mp}(j)}(Z^j|\mathbf{Z}^{\mathrm{mp}(j)})$.

- それぞれの条件付き確率密度関数を、カーネルに基づく推定量で置き換える. $K^j : \mathcal{Z}^{\mathrm{mp}(j)} \to \mathbb{R}_{\geq 0}$ として,

$$p(\mathbf{Z}) \simeq \prod_{j=1}^{D} \hat{p}_{j|\mathrm{mp}(j)}(Z^j|\mathbf{Z}^{\mathrm{mp}(j)}) := \frac{\sum_{i=1}^{n} \delta_{Z_i^j}(Z^j) K^j(\mathbf{Z}^{\mathrm{mp}(j)} - \mathbf{Z}_i^{\mathrm{mp}(j)})}{\sum_{k=1}^{n} K^j(\mathbf{Z}^{\mathrm{mp}(j)} - \mathbf{Z}_k^{\mathrm{mp}(j)})}$$

経験条件付き密度 (Stute, 1986)

- プラグインリスク推定量

$$\hat{R}_{\mathrm{aug}}(f) = \int_{\mathcal{Z}} \ell(f, \mathbf{Z}) \prod_{j=1}^{D} \hat{p}_{j|\mathrm{mp}(j)}(Z^j|\mathbf{Z}^{\mathrm{mp}(j)}) \mathrm{d}\mathbf{Z} = \sum_{\boldsymbol{i} \in [n]^D} \hat{w}_{\boldsymbol{i}} \cdot \ell(f, \mathbf{Z}_{\boldsymbol{i}})$$

拡張データ + 重み付け

一般のADMGの場合の提案手法

- プラグイン推定量は次のように書き換えられる

$$\hat{R}_{\mathrm{aug}}(f) = \sum_{\boldsymbol{i} \in [n]^D} \hat{w}_{\boldsymbol{i}} \cdot \ell(f, \mathbf{Z}_{\boldsymbol{i}})$$

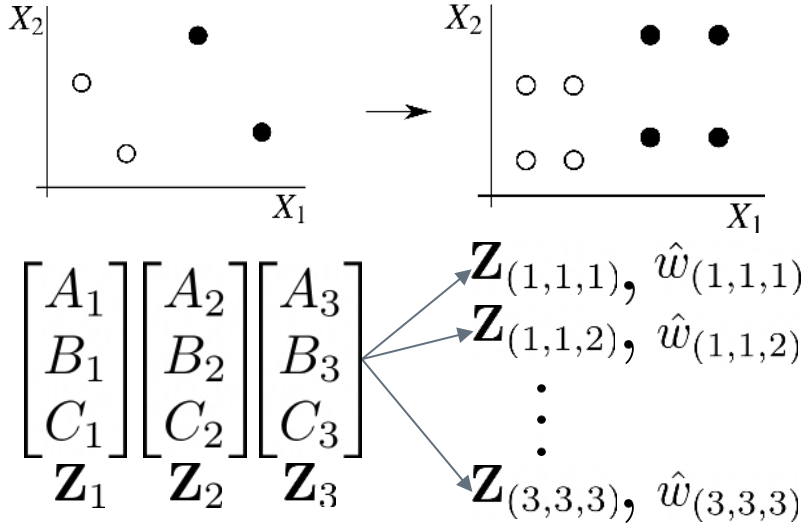
where

$$\mathbf{Z}_{\boldsymbol{i}} = \begin{bmatrix} Z^1_{i_1} \\ \vdots \\ Z^D_{i_D} \end{bmatrix} \qquad \hat{w}_{\boldsymbol{i}} = \prod_{j=1}^{D} \frac{K^j(\mathbf{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}} - \mathbf{Z}^{\mathrm{mp}(j)}_{i})}{\sum_{k=1}^{n} K^j(\mathbf{Z}^{\mathrm{mp}(j)}_{\boldsymbol{i}_{1:j-1}} - \mathbf{Z}^{\mathrm{mp}(j)}_{k})}$$

$$\boldsymbol{i} = (i_1, \ldots, i_D)$$

- これはデータ拡張によって計算できる:

$$\underset{\mathbf{Z}_1}{\begin{bmatrix} A_1 \\ B_1 \\ C_1 \end{bmatrix}} \underset{\mathbf{Z}_2}{\begin{bmatrix} A_2 \\ B_2 \\ C_2 \end{bmatrix}} \underset{\mathbf{Z}_3}{\begin{bmatrix} A_3 \\ B_3 \\ C_3 \end{bmatrix}} \longrightarrow \mathbf{Z}_{(1,1,2)} = \begin{bmatrix} A_1 \\ B_1 \\ C_2 \end{bmatrix}$$



理論解析

Q. 提案手法は統計的にはどのように役立つのか？

設定および鍵となる仮定

- 真の因果グラフが存在し，これが完全に推定できている：$\hat{\mathcal{G}} = \mathcal{G}$
- 真の密度関数とカーネル関数が，十分な滑らかさと有界性の仮定を満たす

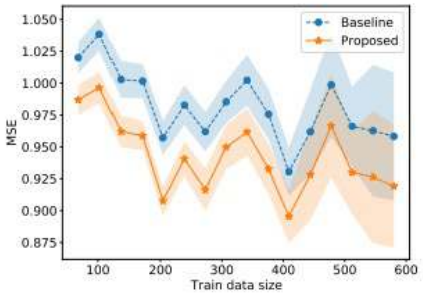
Theorem (Excess Risk Bound; インフォーマル) $\hat{f} \in \arg\min_{f \in \mathcal{F}} \{\hat{R}_{\mathrm{aug}}(f)\},\ f^* \in \arg\min_{f \in \mathcal{F}} \{R(f)\}$

$$R(\hat{f}) - R(f^*) \leq \underbrace{C_1 R_{\mathbf{H}} + C_p}_{\text{Kernel Bias}} + \underbrace{C_2 R_K + C_3 R_{\mathcal{F},K}}_{\text{Complexity terms}} + \underbrace{C_4 \sqrt{\frac{\log(4D/\delta)}{2n}}}_{\text{Uncertainty}}$$

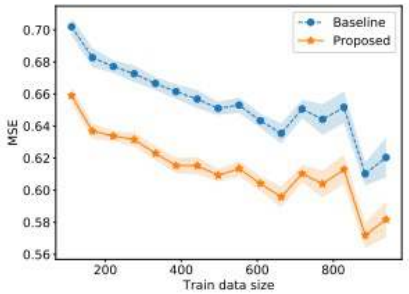
w/ high probability.

- 複雑度の項は，通常の経験リスク最小化に現れるRademacher複雑度と比べて改善されたサンプルサイズ依存性を持つ：過学習の抑制効果が示唆される
  （直観：合成データの増加 ⇨ 過学習がより困難に）
- しかしカーネル関数による近似に由来するバイアスが導入される

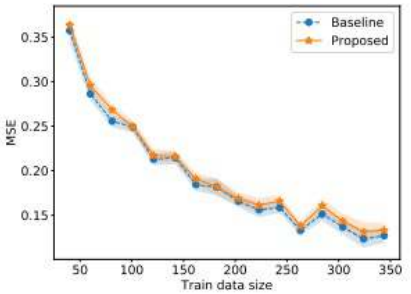
実験結果



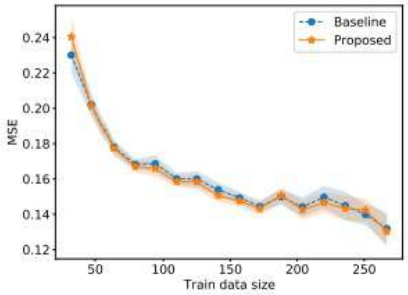
(a) Sachs data.



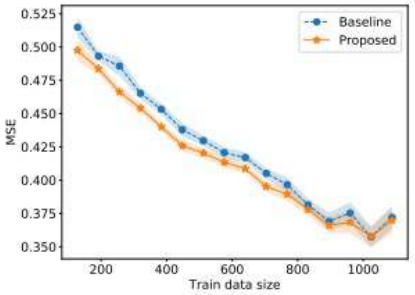
(b) GSS data.



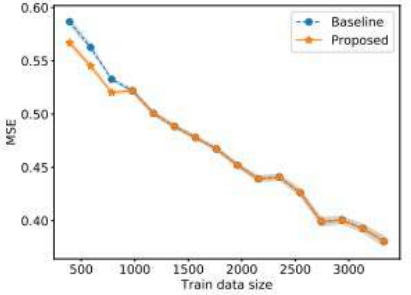
(c) Boston Housing data.



(d) Auto MPG data.



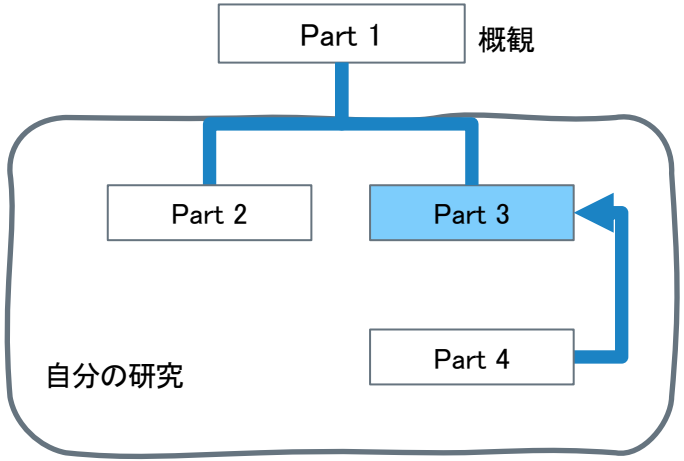
(e) Red Wine data.



(f) White Wine data.

- 小標本なときに性能が改善

Part 3：構造的因果モデルが推定可能なとき

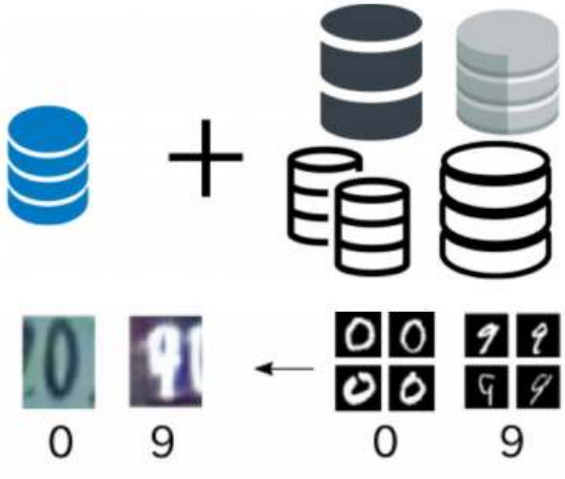


因果メカニズム転移による少数データドメイン適応

Teshima, T., Sato, I., and Sugiyama, M.,
Few-shot domain adaptation by causal mechanism transfer.
Thirty-seventh International Conference on Machine Learning (ICML 2020).

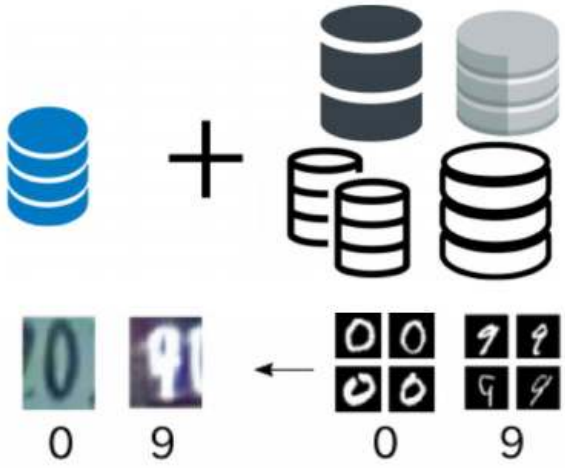
目標:ドメイン適応

- ラベル付きデータが少数しか得られない場合，
  事前情報の利用が性能向上のために重要.
- ドメイン適応:「関連するが異なる」確率分布から得られたデータを学習に活用するという方策



目標:ドメイン適応

- 任意の分布から学習できるわけではなく，分布間の関係性を表現する「転移仮定」を置く必要がある.
- 中心的な問い:どのような関係性があればドメイン適応できるか？



アイデア

共通のデータ生成（因果）メカニズムは
ドメイン適応の土台となりうる

直観

- 人間が因果関係を重視するのは，
  因果は一度発見できれば異なる系にも適用できる知識だから．

動機付けとなる例（仮想的）：地域別の疾病予測器

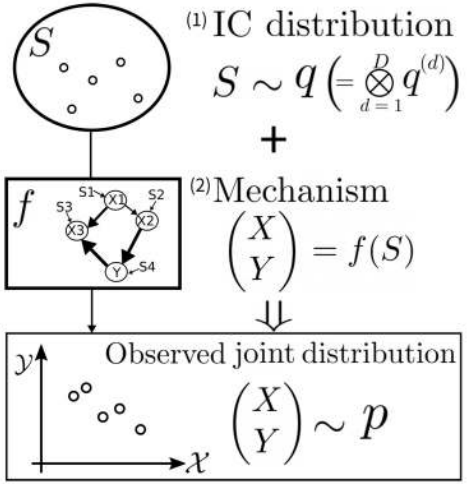
- 疾病リスクを医療記録から予測する (Yadav et al., 2018)
  データ分布は地域毎に異なりうる（生活習慣等）
  疫学的メカニズムは地域によらず共通

構造的因果モデルの(部分的)推定

誘導形の構造方程式: (X,Y)について「解いた」構造方程式 (Reiss and Wolak, 2007)

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

構造方程式・構造形　　構造方程式・誘導形



識別条件のもとで非線形独立成分分析(NLICA)により f を推定可能 (Hyvärinen et al., 2019)
(f'まで復元出来るか否かは別問題)

問題設定

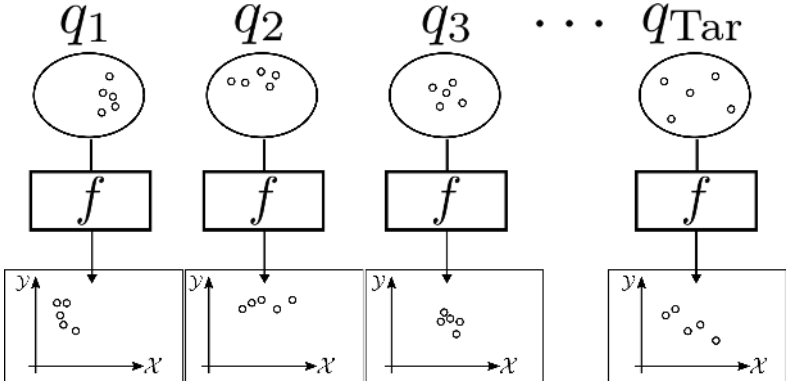
基本設定：回帰（実数値予測）での転移学習

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

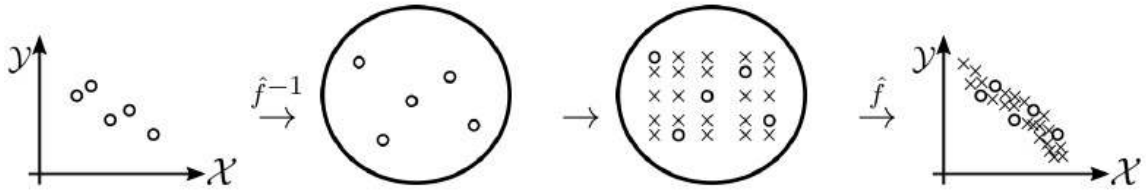
- 期待損失 $R(g) := \mathbb{E}_{\mathrm{tar}} \ell(g, X, Y)$ が小さい予測器 $g : \mathbb{R}^{D-1} \to \mathbb{R}$ を学習
- 複数の転移元分布データ $\mathcal{D}_k = \{(x_{k,i}, y_{k,i})\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{src}(k)} \ (k = 1, \ldots, K)$
- 少数の転移先分布データ $\{(x_{\mathrm{tar},i}, y_{\mathrm{tar},i})\}_{i=1}^{n_{\mathrm{tar}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{tar}}$ （n_k :大，n_{tar} :小）

鍵となる仮定：どの分布も背後に同一の構造方程式を持つ

- q には柔軟な変化を許容
  →表面上大きく異なる分布間での転移も可能
- f は可逆かつNLICA(一般化対照学習)で推定可能と仮定(Hyvärinen et al., 2019)



手法のアイデア

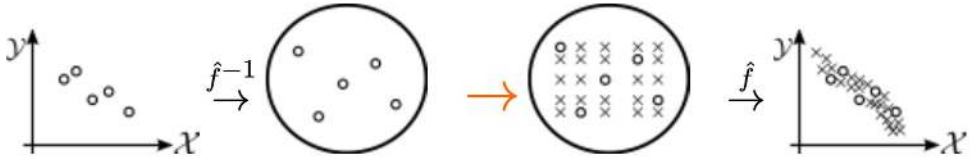


アイデア（仮定をどう利用するか）

1. 転移元ドメインから f を推定 (NLICA) $\hat{f} \leftarrow \mathrm{ICA}(\mathcal{D}_1, \ldots, \mathcal{D}_K)$ (Hyvärinen et al., 2019)
2. $\hat{f}^{-1}$ で 転移先データの独立成分を推定 $\hat{s}_{\mathrm{tar},i} \leftarrow \hat{f}^{-1}(x_{\mathrm{tar},i}, y_{\mathrm{tar},i})$
3. 値の交換により "独立成分候補" を得る $\{\bar{s}_j\}_{j=1}^{n_{\mathrm{tar}}^D} \leftarrow \mathrm{Shuffle}(\{\hat{s}_{\mathrm{tar},i}\}_i)$
4. 独立成分候補から転移先データを生成 $\{(\bar{x}_j, \bar{y}_j)\}_{j=1}^{n_{\mathrm{tar}}^D} \leftarrow \hat{f}(\{\bar{s}_j\}_j)$
5. 生成されたデータで 予測器 g を学習 $\hat{R}_{\mathrm{aug}}(g) := \frac{1}{n_{\mathrm{tar}}^D} \sum_{j=1}^{n_{\mathrm{tar}}^D} \ell(g, \bar{x}_j, \bar{y}_j)$

（上記の $\hat{f}$, $\hat{f}^{-1}$ は可逆ニューラルネットワークモデルで実現(Kingma and Dhariwal, 2018)）

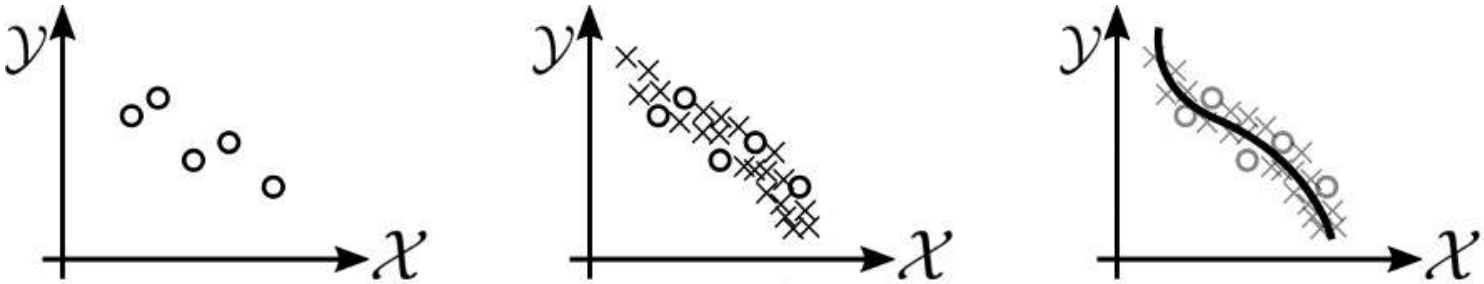
値の交換により独立成分候補を得る，とは?



- 各次元 1 データを選択 (重複可) ＝経験周辺分布からサンプリング

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

データの拡張がどのように役立つか



- 理論的解析（汎化誤差の確率的上界）から分かること:

  データが増加することにより過適合は軽減
  代償として，$\hat{f} \neq f$ の度合いに応じてバイアスが発生

実験設定と結果

- データ (Greene, 2012)
  ガソリン消費量予測
  18ヵ国(=ドメイン)
  19年間分，入力3次元
- 比較手法
  TarOnly: 転移先データのみで学習
  SrcOnly: 転移元データのみで学習

転移元データを用いる他手法が負転移を起こすときも

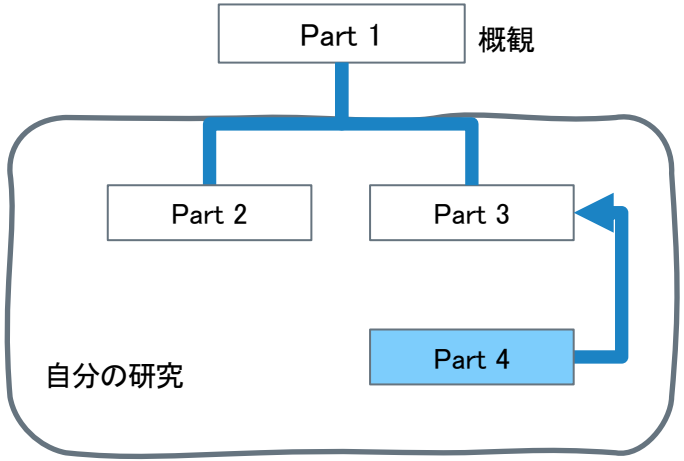
提案法 > *TrgOnly*

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 (0.51)	2.82 (0.73)	1.86 (0.29)	1.63 (0.11)	24.25 (12.50)	23.36 (0.88)	31.37 (1.17)	27.86 (0.94)	27.86 (0.93)	27.52 (0.91)
ESP	1	5.18 (1.05)	6.09 (1.53)	5.17 (1.14)	4.29 (0.72)	14.85 (4.20)	33.16 (6.99)	152.59 (6.19)	53.53 (2.47)	52.56 (2.42)	52.06 (2.40)
SWE	1	6.44 (2.66)	5.47 (2.63)	2.48 (0.23)	2.02 (0.21)	2.18 (0.25)	15.53 (2.59)	2706.85 (17.91)	118.46 (1.64)	118.23 (1.64)	118.27 (1.64)
CHE	1	3.51 (0.46)	2.90 (0.37)	43.59 (1.77)	7.48 (0.49)	38.32 (9.03)	8.43 (0.24)	29.71 (0.53)	9.72 (0.29)	9.71 (0.29)	9.79 (0.28)
TUR	1	1.65 (0.47)	1.06 (0.15)	1.22 (0.18)	0.91 (0.09)	2.19 (0.34)	64.26 (5.71)	142.84 (2.04)	159.79 (2.63)	157.89 (2.63)	157.13 (2.69)
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

提案法の実用化に向けての課題

- 本研究では「可逆ニューラルネットワーク」と呼ばれるモデルを（ f を近似するために）用いた.
- 可逆ニューラルネットワークは新興技術のため，その表現力が十分かどうかは知られていなかった
- 具体的には「万能近似能力」を持つかどうか，という理論的な問い（モデルの基本的な表現能力の性質）

Part 4: 可逆神経回路網モデルの表現力



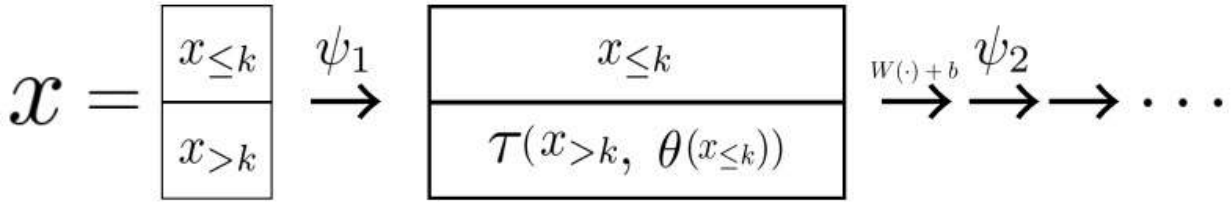
カップリングに基づく可逆ニューラルネットワーク（CF-INN）の万能近似能力

Teshima, T.*, Ishikawa, I.*, Tojo, K., Oono, K., Ikeda, M., and Sugiyama, M., Coupling-based invertible neural networks are universal diffeomorphism approximators. Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS 2020).
(*: Equal contribution. Oral presentation.)

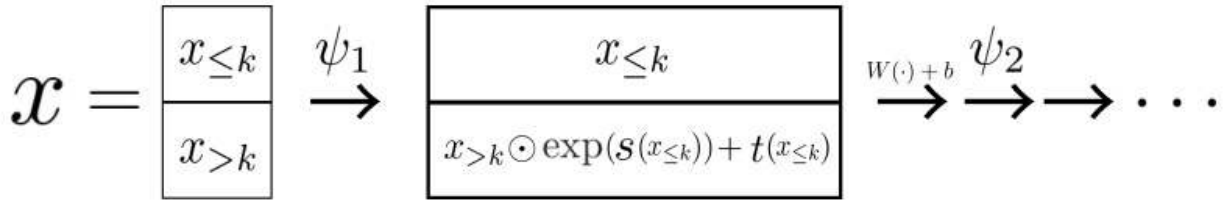
NEURAL INFORMATION PROCESSING SYSTEMS

本研究の理論解析対象

- カップリング層 (Dinh et al., 2017)



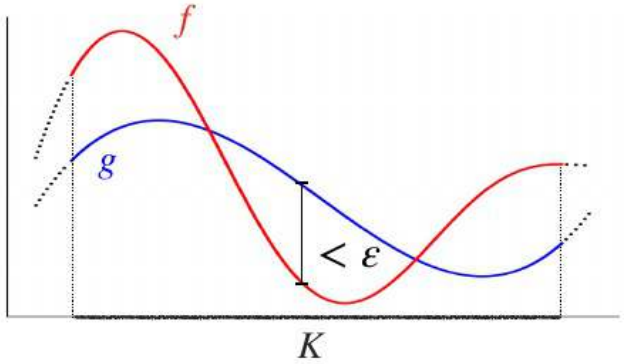
- アフィンカップリング層



生成モデル，確率的推論，特徴抽出，特徴操作といった多様な機械学習タスクに応用を持つ
(Kingma and Dhariwal, 2018), (Oord et al., 2018), (Kim et al., 2019), (Bauer and Mnih, 2019), (Ardizzone et al., 2019), (Nalisnick et al., 2019).

表現力解析の動機

- 機械学習モデルは関数近似器。



- 仮に近似できない関数が存在すると、埋められない性能限界が生じる可能性

  →理論的に（適切に広いクラスの）関数を近似可能であることを確認したい

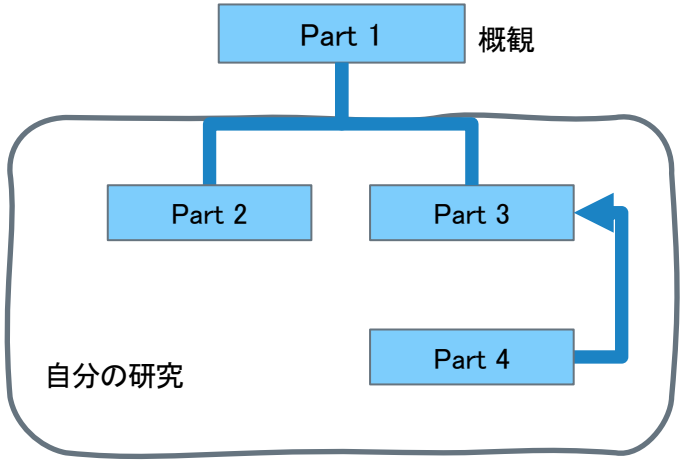
主結果

- 理論の対象となる近似モデル：$\mathrm{INN}_{\mathcal{G}}$
  カップリング層 $\mathcal{G}$ と可逆アフィン変換の有限個の合成

- 近似対象の関数の集合 $\mathscr{D}^2 := \{C^2\text{-diffeo of the form } f : U_f \to f(U_f)\}$
  かなり大きな可逆関数の集合　　$(U_f \subset \mathbb{R}^d : \text{open } C^2\text{-diffeo to } \mathbb{R}^d)$

- 主結果：$\mathrm{INN}_{\mathcal{G}}$ は $\mathscr{D}^2$ に対して万能近似性を持つ
  （適切な仮定のもと，適切な誤差の測り方で）

まとめ



因果関係への好奇心・探究心

I would rather discover one causal law than be King of Persia.
Democritus (460–370 B.C.)

私はペルシャの王になるより
因果の法則を一つでも発見したい

なぜ……？

出典

出典

因果的知識は予測モデルのためにも役に立つ（と思われる）

Appendix

Abadie, A., Cattaneo, M.D., 2018. Econometric methods for program evaluation. Annual Review of Economics 10, 465–503. https://doi.org/10.1146/annurev-economics-080217-053402

Ardizzone, L., Kruse, J., Rother, C., Köthe, U., 2019. Analyzing inverse problems with invertible neural networks, in: 7th International Conference on Learning Representations. OpenReview.net, New Orleans, LA, USA.

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

Bauer, M., Mnih, A., 2019. Resampled priors for variational autoencoders, in: Chaudhuri, K., Sugiyama, M. (Eds.), Proceedings of Machine Learning Research, Proceedings of Machine Learning Research. PMLR, pp. 66–75.

Bhattacharya, R., Nabi, R., Shpitser, I., 2020. Semiparametric inference for causal effects in graphical models with hidden variables. arXiv:2003.12659 [stat.ML].

Bongers, S., Forré, P., Peters, J., Mooij, J.M., 2021. Foundations of structural causal models with cycles and latent variables. arXiv:1611.06221 [cs, stat].

Chikahara, Y., Sakaue, S., Fujino, A., Kashima, H., 2021. Learning individually fair classifier with path-specific causal-effect constraint, in: International Conference on Artificial Intelligence and Statistics. Presented at the International Conference on Artificial Intelligence and Statistics, PMLR, pp. 145–153.

Dinh, L., Sohl-Dickstein, J., Bengio, S., 2017. Density estimation using real NVP, in: 5th International Conference on Learning Representations, Conference Track Proceedings. OpenReview.net, Toulon, France.

Duncan, O.D., Featherman, D.L., Duncan, B., 1972. Socioeconomic Background and Achievement, Socioeconomic background and achievement. Seminar Press, New York.

Eaton, D., Murphy, K., 2007. Exact Bayesian structure learning from uncertain interventions, in: Artificial Intelligence and Statistics. Presented at the Artificial Intelligence and Statistics, PMLR, pp. 107–114.

Greene, W.H., 2012. Econometric Analysis, 7th ed. Prentice Hall, Boston.

Hernán, M.A., Robins, J.M., 2020. Causal Inference: What If. Chapman & Hall/CRC, Boca Raton.

Huszár, F., 2019. Causal Inference 2: Illustrating Interventions via a Toy Example [WWW Document]. inFERENCe. URL https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/ (accessed 2.18.20).

Hyvärinen, A., Sasaki, H., Turner, R., 2019. Nonlinear ICA using auxiliary variables and generalized contrastive learning, in: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics. Presented at the The 22nd International Conference on Artificial Intelligence and Statistics, pp. 859–868.

Kim, S., Lee, S.-G., Song, J., Kim, J., Yoon, S., 2019. FloWaveNet : A generative flow for raw audio, in: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research. PMLR, Long Beach, California, USA, pp. 3370–3378.

Kingma, D.P., Dhariwal, P., 2018. Glow: Generative flow with invertible 1x1 convolutions, in: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 31. Curran Associates, Inc., pp. 10215–10224.

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., Mooij, J.M., 2018. Domain adaptation by using causal inference to predict invariant conditional distributions, in: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 31. Curran Associates, Inc., pp. 10846–10856.

Messerli, F.H., 2012. Chocolate Consumption, Cognitive Function, and Nobel Laureates. New England Journal of Medicine 367, 1562–1564. https://doi.org/10.1056/NEJMon1211064

Mooij, J., 2019. MLSS 2019: Causality.

Mooij, J.M., Janzing, D., Schölkopf, B., 2013. From ordinary differential equations to structural causal models: the deterministic case. Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence.

Nalisnick, E.T., Matsukawa, A., Teh, Y.W., Görür, D., Lakshminarayanan, B., 2019. Hybrid models with deep and invertible features, in: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research. PMLR, Long Beach, California, USA, pp. 4723–4732.

Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G., Lockhart, E., Cobo, L., Stimberg, F., Casagrande, N., Grewe, D., Noury, S., Dieleman, S., Elsen, E., Kalchbrenner, N., Zen, H., Graves, A., King, H., Walters, T., Belov, D., Hassabis, D., 2018. Parallel WaveNet: Fast high-fidelity speech synthesis, in: Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research. PMLR, Stockholmsmässan, Stockholm Sweden, pp. 3918–3926.

Papamakarios, G., Nalisnick, E., Rezende, D.J., Mohamed, S., Lakshminarayanan, B., 2021. Normalizing flows for probabilistic modeling and inference. Journal of Machine Learning Research 22, 1–64.

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

Peters, J., Janzing, D., Schölkopf, B., 2017. Elements of Causal Inference: Foundations and Learning Algorithms, Adaptive Computation and Machine Learning Series. The MIT Press, Cambridge, Massachusetts.

Reiss, P.C., Wolak, F.A., 2007. Structural econometric modeling: rationales and examples from industrial organization, in: Handbook of Econometrics. Elsevier, pp. 4277–4415.

Richardson, T., 2003. Markov properties for acyclic directed mixed graphs. Scandinavian Journal of Statistics 30, 145–157.

Richardson, T.S., Evans, R.J., Robins, J.M., Shpitser, I., 2017. Nested Markov properties for acyclic directed mixed graphs. arXiv:1701.06686 [stat.ME].

Rojas-Carulla, M., Schölkopf, B., Turner, R., Peters, J., 2018. Invariant models for causal transfer learning. Journal of Machine Learning Research 19, 1–34.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P., 2005. Causal protein-signaling networks derived from multiparameter single-cell data. Science 308, 523–529. https://doi.org/10.1126/science.1105809

Schölkopf, B., 2019. Causality for machine learning. arXiv:1911.10500 [cs, stat].

Schölkopf, B., Hogg, D., Wang, D., Foreman-Mackey, D., Janzing, D., Simon-Gabriel, C.-J., Peters, J., 2015. Removing systematic errors for exoplanet search via latent causes, in: Proceedings of the 32nd International Conference on Machine Learning. Presented at the International Conference on Machine Learning, PMLR, pp. 2218–2226.

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

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P.O., Bollen, K., 2011. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. Journal of Machine Learning Research 12, 1225–1248.

Spirtes, P., Glymour, C.N., Scheines, R., 2000. Causation, Prediction, and Search, 2nd ed. MIT Press, Cambridge, Massachusetts.

Stute, W., 1986a. Conditional Empirical Processes. Ann. Statist. 14, 638–647. https://doi.org/10.1214/aos/1176349943

Stute, W., 1986b. On almost sure convergence of conditional empirical distribution functions. Ann. Probab. 14, 891–901. https://doi.org/10.1214/aop/1176992445

Teshima, T., Sato, I., Sugiyama, M., 2020. Few-shot domain adaptation by causal mechanism transfer, in: Proceedings of the 37th International Conference on Machine Learning. Presented at the 37th International Conference on Machine Learning, Online, pp. 9458–9469.

Tian, J., Pearl, J., 2002. A general identification condition for causal effects, in: Proceedings of the Eighteenth National Conference on Artificial Intelligence. AAAI Press/The MIT Press, Menlo Park, CA, pp. 567–573.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., Shieber, S., 2020. Investigating gender bias in language models using causal mediation analysis, in: Advances in Neural Information Processing Systems 33.

Wu, Y., Zhang, L., Wu, X., Tong, H., 2019. PC-Fairness: A unified framework for measuring causality-based fairness, in: Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc.

Yadav, P., Steinbach, M., Kumar, V., Simon, G., 2018. Mining electronic health records (EHRs): a survey. ACM Computing Surveys 50, 1–40.

Yasui S., 2020. 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.

Zhang, K., Gong, M., Schölkopf, B., 2015. Multi-source domain adaptation: a causal view, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press, pp. 3150–3157.

Zhang, K., Schölkopf, B., Muandet, K., Wang, Z., 2013. Domain adaptation under target and conditional shift, in: Proceedings of the 30th International Conference on Machine Learning. Presented at the International Conference on Machine Learning, pp. 819–827.

FAQ

因果推論の入門書・論文

Pearl流の枠組み：

- (Bongers et al., 2021) は定式化が数学的に分かりやすい（動機などを知っているとより読みやすい）
- (Pearl, 2009) は情報量が多い
- 今回のトークは (Peters et al., 2017) の導入の仕方に近い
- (Mooij, 2019) のチュートリアル資料も分かりやすい

Rubin流の枠組み：

- (Hernán and Robins, 2020) の評判が良い
- (Yasui, 2020) は非常に分かりやすい
- (Abadie and Cattaneo, 2018) はコンパクトにまとまっていて読みやすい

可逆ニューラルネットのサーベイ論文

(Papamakarios et al., 2021) が良いサーベイ論文