因果メカニズム転移による小標本ドメイン適応

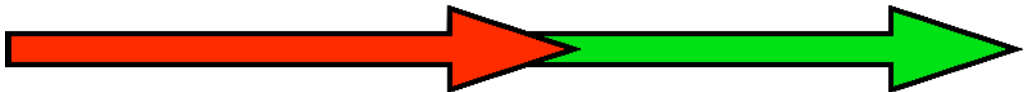
手嶋 毅志 [12]

[1] 東京大学 [2] 理研 AIP

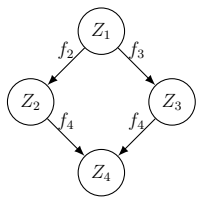
情報論的学習理論と機械学習研究会（IBISML）
オーガナイズドセッション『異なるタスクを活用する機械学習：転移学習，メタ学習』

（佐藤一誠 先生 [12]，杉山将 先生 [21] との共同研究に基づく内容です。
また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。）

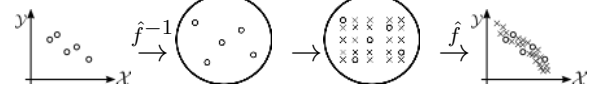
本トークの構成



Part 1. **因果モデリングの紹介**



Part 2. **因果メカニズム転移による小標本ドメイン適応**



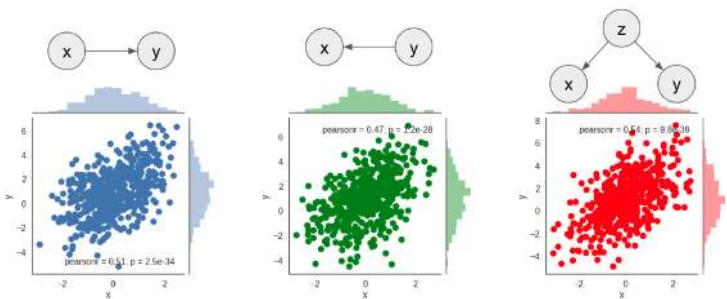
Take-home message

因果モデルによって捉えられるデータ生成過程の情報が転移学習・メタ学習の手がかりになる可能性がある

Part 1.

導入・因果モデリング

- 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
- 今回は Pearl 流の構造的因果モデルを紹介



なぜ「因果モデル」を考えるか

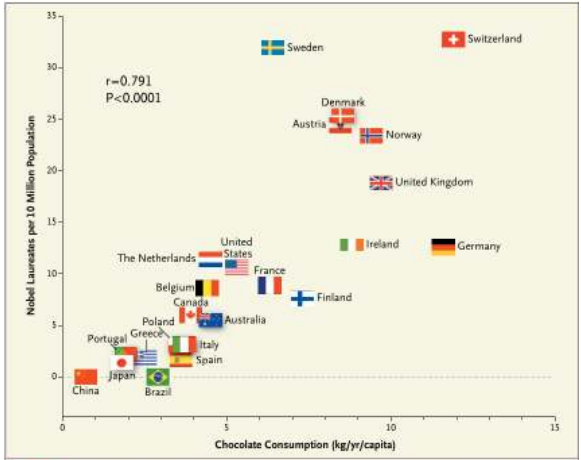


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Figure: [1]

なぜ「因果モデル」を考えるか

？？？「チョコレートを食べさせよう！」

我々🤔「因果関係があるかは怪しい」

……チョコを食べるだけでノーベル賞が穫れるとは思えない

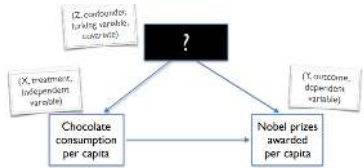


Figure: [2]

- データの確率分布の統計モデルだけでは捉えられない議論

なぜ「因果モデル」を考えるか

同じ分布になるような異なる生成過程

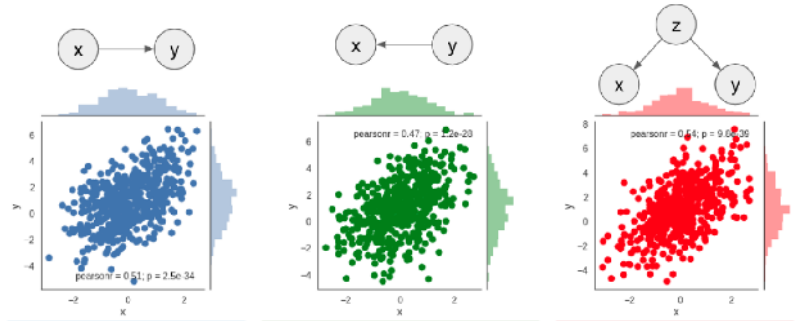


Figure: [3]

なぜ「因果モデル」を考えるか

介入 $\mathrm{do}(X = 3)$ のもとでは異なる振る舞い

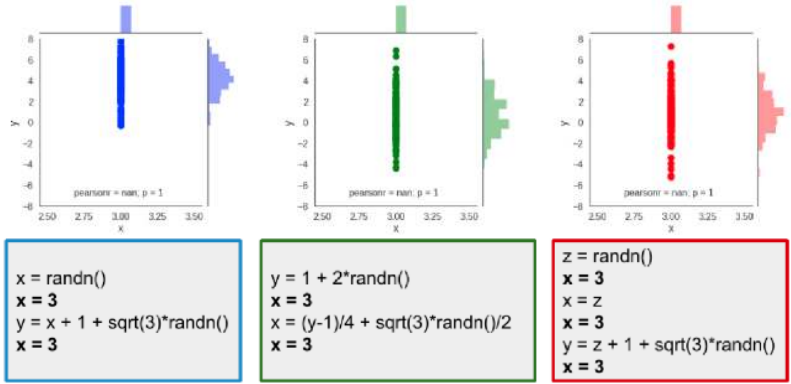


Figure: [3]

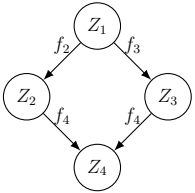
- 因果を扱うには分布の背後のデータ生成過程を考える必要あり

主な因果モデルの具体的定式化

構造方程式モデル (SEM)/構造的因果モデル (SCM) [4–6]

「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 $\boldsymbol{Z} = \{Z_d\}_{d=1}^D$ 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 $\boldsymbol{S} = \{S_i\}_{i=1}^D$ の分布 q

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

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

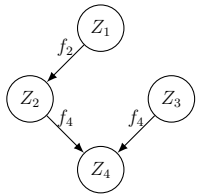
上記は非巡回な Markovian モデルの場合．詳細は [6, 7]．また $\mathcal{G}$ は $\mathcal{F}$ の引数関係から読み取れるので実際はモデルは $(\mathcal{F}, q)$ のみから定まる

主な因果モデルの具体的定式化：使用法 (☆)

完全介入 (Perfect intervention; [5]) $\mathrm{do}(Z_I = \zeta_I)$

- 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 [8] 参照

主な因果モデルの具体的定式化

因果グラフィカルモデル (GCM; [5])

- 観測データの確率分布 $p(Z)$ ＋非巡回有向グラフ $\mathcal{G}$
- ＋ 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

(SCM のうち完全介入の推論に必要な要素だけを抽出したモデル)

- このような「因果的仮定」は（データ分布の背後にある）データの生成過程に関する仮定
- 「データの背後に確率分布がある」よりも強い仮定を置く
  →より強い（因果的）推論が可能になっている

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

Table: [9]

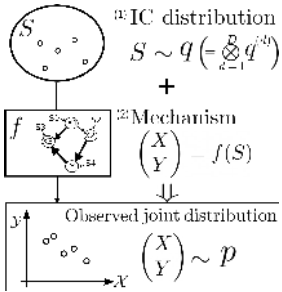
- 上から順に詳細なモデル
  - 機械論的/物理的因果モデル (微分方程式による; [10, 11])
  - 構造的因果モデル←条件のもとモデルの一部を推定可能 [12–15]
  - 因果グラフィカルモデル←ある程度の条件のもとで推定可能 [5, 16–18]
  - 統計モデル (確率分布モデル) ←標準的に推定可能

非線形 ICA による誘導形 SEM の推定

誘導形の構造方程式: $(\mathbf{X}, Y)$ **について「解いた」構造方程式** [19]

$$\begin{cases} Z_1 &= f'_1(S_1) \\ Z_2 &= f'_2(Z_1, S_2) \\ Z_3 &= f'_3(Z_1, S_3) \\ Z_4 &= f'_4(Z_2, Z_3, S_4) \end{cases} \Rightarrow \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{pmatrix} = f \begin{pmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{pmatrix}$$

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 ICA** により f **を推定可能** [15](今回の提案手法でもこれを用いる．f' まで復元出来るか否かは別問題)

ICA = 独立成分分析 (Independent component analysis)

機械学習との接点

機械学習の為の因果 (Causality for ML; **主に Pearl 系の枠組み**)

- **「因果はデータ生成過程の知識」→「学習にも何かしら役に立つはず」** [20–28]

機械学習による因果 (Causality by ML; **主に Rubin，Robins 系の枠組み**)

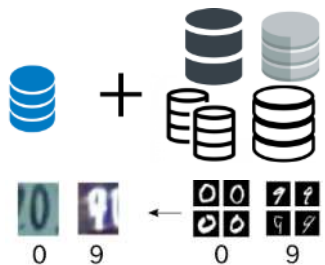
- **因果的知識発見のために機械学習を使う**
- **因果効果推定** [29, 30]**，因果表現学習** [31–33]**，……**

機械学習における因果 (Causality in ML)

- **機械学習システムの中の因果構造 (データ→予測→施策)**
- **解釈性** [34]**，公平性** [35]**，……**

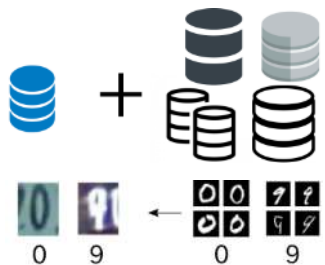
モチベーション: ドメイン適応

- ラベル付きデータが少数しか得られない場合，事前知識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」確率分布のデータを学習に活用



モチベーション: ドメイン適応

- 任意の分布から学習できるわけはなく，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（と思われる）から

動機付けとなる例（仮想的）：地域別の疾病予測器

ゴール：疾病リスクを医療記録から予測する [36]
データ分布は地域ごとに異なりうる（生活習慣の違いなど）
疫学的メカニズムは地域によらず共通

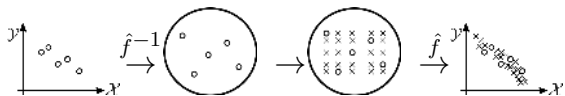
Part 2.

因果メカニズム転移による小標本ドメイン適応

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

Few-shot domain adaptation by causal mechanism transfer.

ICML
International Conference
On Machine Learning



定式化 1/3: データと目標

問題設定: ドメイン適応回帰 (実ベクトル X から実数値 Y を予測)

期待損失 $R(g) := \mathbb{E}_{\mathrm{tar}}\ell(g, X, Y)$ が小さい $g : \mathbb{R}^{D-1} \to \mathbb{R}$ を学習

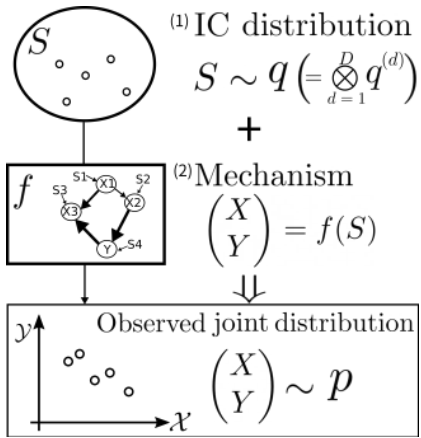
1. 等質性 (全ドメインは同じデータ空間を持つ) $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^{D-1} \times \mathbb{R}$
2. マルチソース (複数の転移元ドメインがある)
   $\mathcal{D}_k = \{(x_{k,i}, y_{k,i})\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{src}(k)}$ $(k = 1, \ldots, K)$ (n_k:大)
3. 小標本教師付き (転移先分布のラベル付きデータが少数ある)
   $\{(x_{\mathrm{tar},i}, y_{\mathrm{tar},i})\}_{i=1}^{n_{\mathrm{tar}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{tar}}$ (n_{tar} :小)

を仮定

ℓ: 損失関数

定式化 2/3: 因果モデルの存在仮定

- 各ドメインは次の非線形 ICA モデルに従う



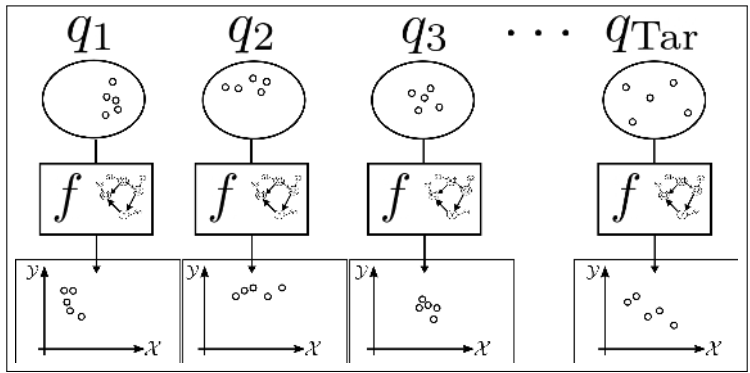
分布 p は (f, q) から生成

1. D 次元独立成分 S が q からサンプル
2. 可逆な f が S を $(X, Y) = f(S)$ に変換

- f は適当な仮定のもと非線形 ICA により推定可能
- f は SEM の「誘導形」に相当

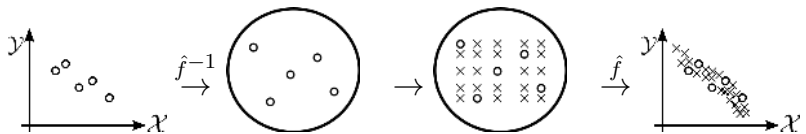
定式化 3/3: 主仮定

- 主仮定: 生成メカニズム (から定まる) f が共通



- q には柔軟な変化を許容 $\rightsquigarrow$ 見かけ上大きく異なる分布間での転移学習も可能に

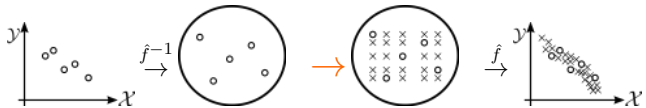
手法のアイデア



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) $\hat{f} \leftarrow \mathrm{ICA}(\mathcal{D}_1, \ldots, \mathcal{D}_K)$
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** $\hat{R}_{\mathrm{aug}}(g) := \frac{1}{n_{\mathrm{tar}}^D} \sum_{j=1}^{n_{\mathrm{tar}}^D} \ell(g, \bar{x}_j, \bar{y}_j)$

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

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

理論的正当化

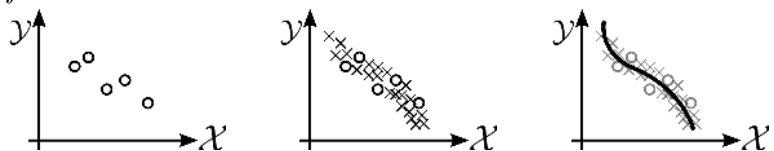
Q1. この手法は統計的にはどう役立つ？

定理: もし $\hat{f} = f$ なら提案リスク推定量は一様最小分散不偏推定量

💬 この手法は分散に関して役立つはず

Q2. もし $\hat{f} \neq f$ ならどうなる？

定理: $\hat{f} \neq f$ の場合の汎化誤差バウンド



💬 😍 過学習を抑制 😬 バイアスが発生

提案手法は，いつ，何を，どのように転移するか

いつ転移するか (When to transfer)

「背後のデータ生成過程が共通のときに」転移する

何を転移するか (What to transfer)

- 「因果モデルで捉えられるデータ生成過程の知識を」転移する
- 具体的には「独立成分の取り出し方+再合成の仕方を」転移する

どのように転移するか (How to transfer)

転移元ドメインで因果構造を学習し，少ない転移先データを水増しする

実験設定

- データ: ガソリン消費データ・セット [37].
  - 計量経済学のパネルデータ (過去にも構造方程式が適用)
  - 18 ヵ国 (=ドメイン)， 19 年， $D = 4$
- ドメイン適応回帰の比較手法

名称	比較手法　　(予測器仮説集合: カーネルリッジ回帰)
TarOnly	転移先データのみで学習
SrcOnly	転移元データのみで学習
S&TV	転移元データで学習し転移先データで交差検証
TrAdaBoost	回帰転移学習のブースティング法 [38].
IW	RuLSIF による重要度重み付き学習 [39].
GDM	一般化乖離度最小化 [40].
Copula	ノンパラメトリック R-vine コピュラ法 [41].
LOO (参考)	一個抜き交差検証による誤差推定値

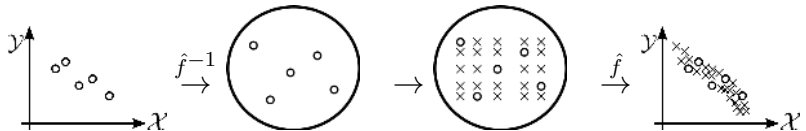
実験結果

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)	226.96 (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

まとめ

1. 共通の生成過程という転移仮定 $\rightsquigarrow$ 小標本ドメイン適応手法を開発
2. 提案法は推定した因果モデルをデータ拡張に用いて過学習を抑制
3. 実データ実験により有効性を検証



Take-home message

因果モデルによって捉えられるデータ生成過程の情報が転移学習・メタ学習の手がかりになる可能性がある

Appendix

なぜ「因果」を考えるか (その後)

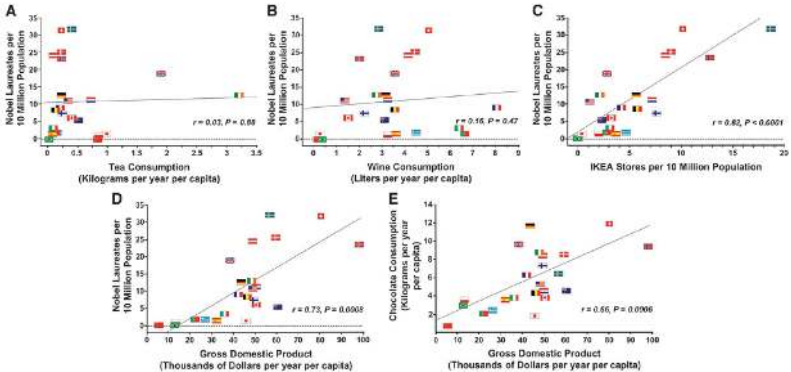


Figure: [42]

- 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

理論の詳細

A1. **定理: もし** $\hat{f} = f$ **なら最小分散不偏推定量** (UMVUE)

($\mathcal{Q}$ **のもとで**)$\hat{R}_{\mathrm{aug}}(g)$ **は** $R(g)$ **の** (unique **な**)UMVUE**，即ち**

$\forall \hat{R}(g)$:**不偏**, $\forall q \in \mathcal{Q}$, $\mathrm{Var}(\hat{R}_{\mathrm{aug}}(g)) \leq \mathrm{Var}(\hat{R}(g))$

A2. **定理:** $\hat{f} \neq f$ **のときの過剰リスク上界**

適切な仮定のもとで確率 $1 - (\delta + \delta')$ **以上で**

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

$\mathfrak{R}(\mathcal{G})$: **有効ラデマッハ複雑度，** $\|\cdot\|_{W^{1,1}}$: $(1,1)$-**ソボレフノルム**

何故複数の転移元分布が必要?

- この要請は非線形 ICA 手法に由来する
  - 今回は複数の転移元分布を必要とする一般化対照学習 (GCL) を使用 [15]
- 単一サンプルからの非線形 ICA は不可能性が知られている [43]
  - (1) f の関数クラスを強く制約するか (2) 補助情報を用いる
- 任意の非線形 ICA 手法は今回の提案法と簡単に組み合わせられる

一般化対照学習 (GCL)

- 非線形 ICA が近年実現されている [15, 44–46].
- 補助情報 (例．時系列方向の依存性など) を利用する [1]

GCL による非線形 ICA [15]

- データとともに補助変数が与えられている (u): $\{(X_i, u_i)\}_{i=1}^n$
- 潜在変数分布は u を条件付けたとき独立: $p(s|u) = \prod_d q^{(d)}(s^{(d)}|u)$
- 二値分類器 $r(x,u) = \sigma(\sum_{d=1}^D \psi_d(h(x)_d, u))$ を学習：$(x_i, u_i) : +1$ vs. $(x_i, \tilde{u}) : -1$.

  σ: シグモイド活性化関数
- このとき，十分な理論的条件のもとで $h : \mathcal{X} \to \mathbb{R}^D$ は f の一致推定量 $(n \to \infty)$

[1] 今回のケースでは転移元ドメイン番号 (k) を補助情報として利用した

GCL での識別条件

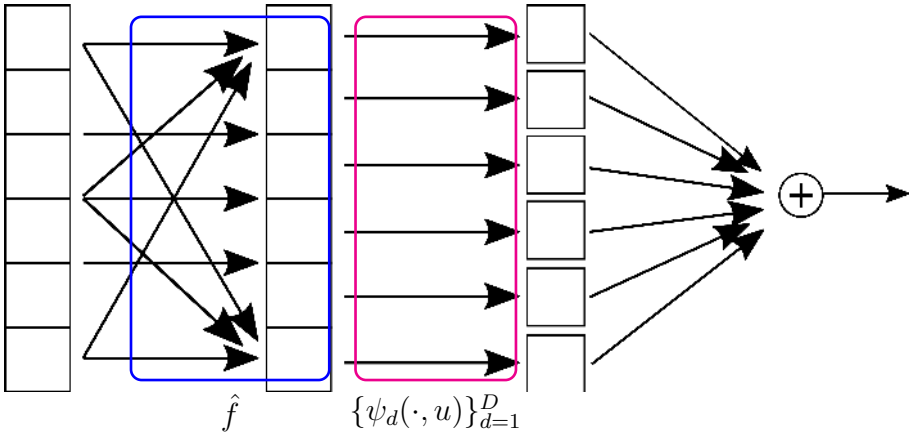
GCL の識別条件 [15]

- $\{(z_i, u_i)\}_{i=1}^n$: **データ** $(u_i \in \mathcal{U})$
- **条件付き独立性** $q(s|u) = \prod_{d=1}^D q^{(d)}(s^{(d)}|u)$
- **「変動性の仮定」**[15]: **任意の** z **に対し異なる** $\{u_j\}_{j=0}^{2D} \subset \mathcal{U}$ **が存在し** $\{w(z|u_j) - w(z|u_0)\}_{j=1}^{2D}$ **が線型独立となる．ここで** $w(z|u) := \left(\frac{\partial \log q^{(1)}(z_1|u)}{\partial z_1}, \ldots, \frac{\partial \log q^{(D)}(z_D|u)}{\partial z_D}, \frac{\partial^2 \log q^{(1)}(z_1|u)}{\partial z_1^2}, \ldots, \frac{\partial^2 \log q^{(D)}(z_D|u)}{\partial z_D^2}\right)$
- **その他の正則条件**

定理 1 [15]

このとき，次元ごとの可逆変換の自由度を許して，GCL は f **の一致推定量を与える**

GCL による非線形 ICA の仕組み（直観）



- $p(u|X)$ をこの制約されたアーキテクチャで推定するためには，$\hat{f}$ が独立成分を抽出する必要がある

可逆ニューラルネットワーク (INNs)

- ニューラルネットワークであって設計上可逆なもの
- 今回は Glow を使用 [47]

アファインカップリング層

- カップリング層: いくつかの次元をそのまま出力

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

- 厳密な逆関数が解析的な式で書ける (s と t を $x_{1:d}$ から再計算する)

参考文献

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

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

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

[6] S. Bongers, P. Forré, J. Peters, B. Schölkopf, 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] D. Eaton and K. Murphy, 'Exact Bayesian structure learning from uncertain interventions,' in *Artificial Intelligence and Statistics*, PMLR, Mar. 2007, pp. 107–114.

参考文献 (cont.)

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

[10] J. M. Mooij, D. Janzing, and B. Schölkopf, 'From Ordinary Differential Equations to Structural Causal Models: The deterministic case,' *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2013.

[11] 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].

[12] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen, 'A linear non-Gaussian acyclic model for causal discovery,' *The Journal of Machine Learning Research*, vol. 7, no. 72, pp. 2003–2030, 2006.

[13] 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].

[14] J. Peters and B. Sch, 'Causal Discovery with Continuous Additive Noise Models,' *Journal of Machine Learning Research*, vol. 15, no. June, pp. 2009–2053, 2014.

参考文献 (cont.)

[15] 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.

[16] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, Prediction, and Search*, Second. Cambridge, Massachusetts: MIT Press, 2000.

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

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

[19] P. C. Reiss and F. A. Wolak, 'Structural econometric modeling: Rationales and examples from industrial organization,' in *Handbook of Econometrics*, vol. 6, Elsevier, 2007, pp. 4277–4415.

[20] B. Schölkopf, 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.

参考文献 (cont.)

[21] 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.

[22] 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.

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

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

[25] 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.

[26] B. Schölkopf, 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 *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, Jun. 2015, pp. 2218–2226.

参考文献 (cont.)

[27] 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.

[28] 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].

[29] S. Wager and S. Athey, 'Estimation and inference of heterogeneous treatment effects using random forests,' *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.

[30] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, 'Double/debiased machine learning for treatment and structural parameters,' *The Econometrics Journal*, vol. 21, no. 1, pp. C1–C68, 2018.

参考文献 (cont.)

[31] 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.

[32] K. Chalupka, T. Bischoff, F. Eberhardt, and P. Perona, 'Unsupervised discovery of El Niño 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.

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

[34] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber, 'Investigating gender bias in language models using causal mediation analysis,' in *Advances in Neural Information Processing Systems 33*, 2020.

参考文献 (cont.)

[35] Y. Wu, L. Zhang, X. Wu, and H. Tong, 'PC-Fairness: A unified framework for measuring causality-based fairness,' in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019.

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

参考文献 (cont.)

[41] 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.

[42] 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.

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

[44] 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.

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

参考文献 (cont.)

[46] 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].

[47] 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.