統計的因果モデルに基づく転移学習

Takeshi Teshima

Recruit Co., Ltd.

日本統計学会 第17回春季集会
March 4th, 2023

Based on joint work w/ Prof. Sato and Prof. Sugiyama:
Teshima, T., Sato, I., and Sugiyama, M., Few-shot domain adaptation by causal mechanism transfer.
37th International Conference on Machine Learning (ICML 2020).

ICML
International Conference On Machine Learning

Background: Domain Adaptation

Domain adaptation (DA) (e.g., Ben-David et al., 2010)

Learning from the data in "relevant but different" domains.

Question of transfer assumption (TA)

Transfer assumption = how the domains relate to one another.

Q. What transfer assumptions enable domain adaptation?

Idea

Common data-generating mechanism
can be a foundation of domain adaptation.

Intuition

One reason why human beings value causal relations is because causality is portable knowledge that can be valid outside of a system in which we acquired it.

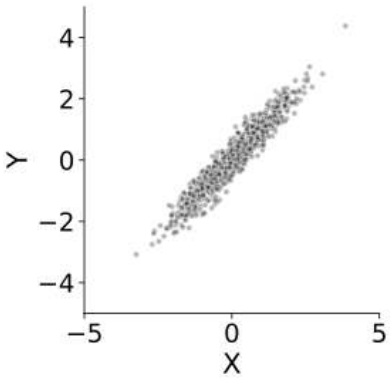
Motivating (Ideal) Example: Regional Disease Prediction

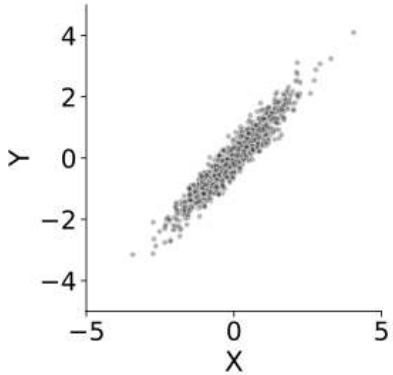
- Predicting disease risk from electronic health records (Yadav et al., 2018).
- Data distribution can vary across regions (diet, demographics, etc.)
- Epidemiological mechanism may be common among regions.

Background: Structural Causal Framework

Two different definitions of (X, Y) with the same distribution:

$\left(e_1, e_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)\right)$





Definitions:

$$X(\cdot) = f_X(e_1(\cdot)),$$
$$Y(\cdot) = f_Y(X(\cdot), e_2(\cdot)),$$

$$Y(\cdot) = g_Y(e_2(\cdot)),$$
$$X(\cdot) = g_X(Y(\cdot), e_1(\cdot)),$$

Graphs: $X \to Y$ $Y \to X$

Scripts:

```
X = normal()
Y = X + a * normal()
```

```
Y = sqrt(1+a^2) * normal()
X = Y / (1+a^2) + a / sqrt(1+a^2) * normal()
```

Under Intervention, Consequences Differ

Distributions after intervention to set $X = 2$:

(a) $X(\cdot) = 2$

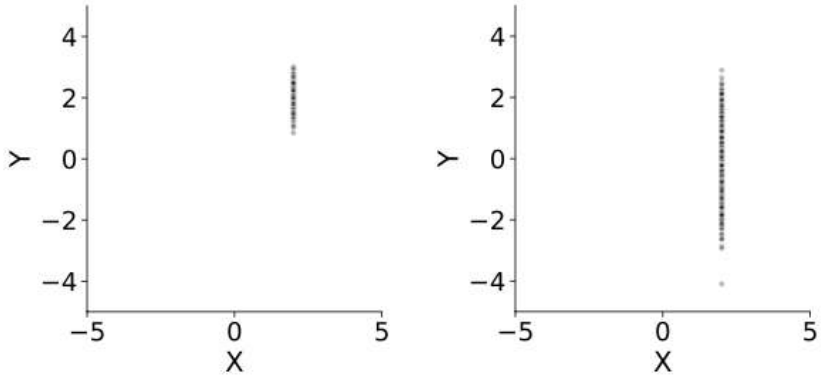
$$Y(\cdot) = f_Y(X(\cdot), e_2(\cdot)).$$

```
X = normal()
X = const
Y = X + a * normal()
X = const
```

(b) $Y(\cdot) = g_Y(e_2(\cdot)),$

$$X(\cdot) = 2$$

```
Y = sqrt(1+a^2) * normal()
X = const
X = Y / (1+a^2) + a / np.sqrt(1+a^2) * normal()
X = const
```



(a) $X \to Y$ (b) $Y \to X$

After all, distribution is only a footprint of RVs: $\mathbb{P} \circ (X, Y)^{-1}$.

Detailed structure of (the definition of) RVs can be relevant.

Background: Structural Causal Models

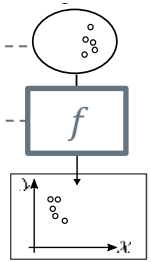
Structural Causal Models (**SCMs**) (Informal)

- Models of data generation processes.
- Capture causal relations by deterministic functions.

Structural Causal Models (**SCMs**)[1] $\langle f, \mathbb{P}_{\mathcal{E}} \rangle \rightsquigarrow \mathbb{P}_{\mathcal{Z}}$ (Pearl, 2009) (Bongers et al., 2021)

- Structural function (**SF**): $f: \mathcal{Z} \times \mathcal{E} \to \mathcal{Z}$
  $Z = f(Z, E)$ (a.s.) (structural equation)
- Exogenous distribution: $\mathbb{P}_{\mathcal{E}}$ (independent)

1. $(E_1, E_2, E_3) \sim \mathbb{P}_{\mathcal{E}}$
2. $\begin{cases} X_1 = f_1(E_1) \\ X_2 = f_2(X_1, E_2) \\ Y = f_Y(X_1, X_2, E_3) \end{cases}$



Induces an (observational) data distribution $\mathbb{P}_{\mathcal{Z}}$.
By replacing some of the SFs, interventional distribution can also be modeled.

1) a.k.a. functional causal models / structural equation models (Pearl, 2009). We only consider acyclic cases in this work.

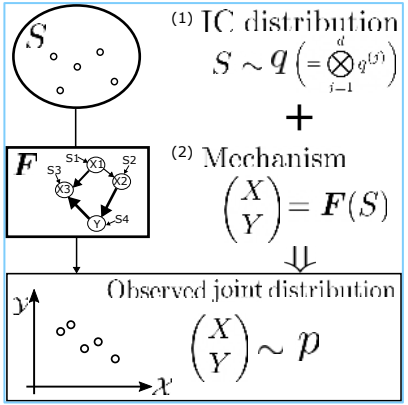
Background: Estimation of Causal Models

Reduced-form Structural Function (RSF) = "Solved" SF (Reiss and Wolak, 2007)

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

Structural-form SE　　Reduced-form SE

SCMs induce ICMs where the RSFs are the mixing maps. (Kano and Shimizu, 2003) (Shimizu et al., 2006)



Independent Component Model (ICM)

Nonlinear Independent Component Analysis (NLICA) (Hyvärinen et al., 2019)

Method to estimate ICMs under "identifiability conditions"
→ Estimate RSF under certain conditions.

(Whether one can recover the original structural form from the RSF is a different question.)

Problem Setup (1/3)

Base Problem: Domain-adapting Regression (i.e., predict continuous Y)

Goal:

Learn $g: X \mapsto Y$ with a small risk $R(g) := \mathbb{E}_{\text{tar}}\ell(g, (X, Y))$.

We assume:

1. Homogeneous (all domains, the same data space) $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^{d-1} \times \mathbb{R}$

2. Few-shot supervised (small sample from target domain is available)

$\mathcal{D}_{\text{tar}} = \{(x_{\text{tar},i}, y_{\text{tar},i})\}_{i=1}^{n_{\text{tar}}}$ from target domain. (n_{tar}: small)

3. Multi-source (samples from multiple source domains exist)

$\mathcal{D}_k = \{(x_{k,i}, y_{k,i})\}_{i=1}^{n_k}$ from domain k ($k \in [K]$). (n_k: large)

ℓ: loss function. $\mathbb{E}_{\text{tar}}$: Expectation w.r.t. target domain distribution.

Problem Setup (2/3)

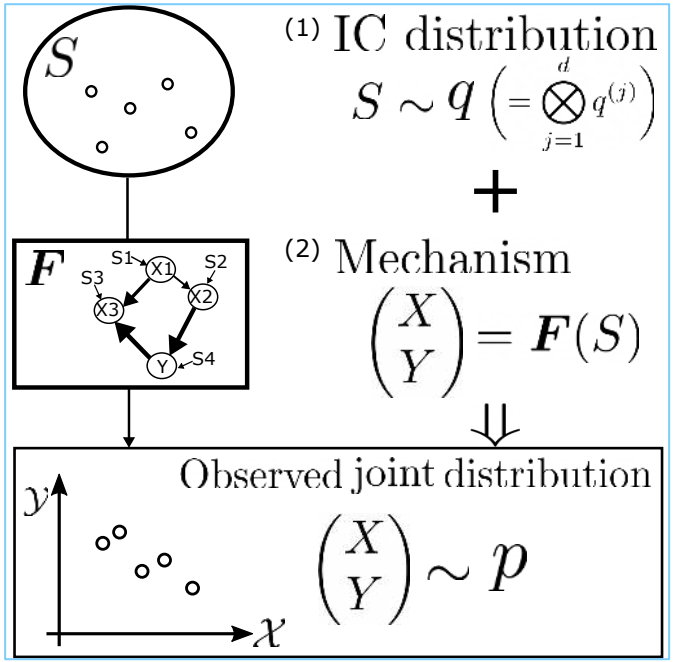
Each domain is an ICM $\mathcal{M}_k = \langle \mathbb{R}^d, \mathbb{R}^d, F, q_k \rangle$

$q_k \in \mathcal{Q}$, $\mathcal{Q}$: set of independent densities.

$F: \mathbb{R}^d \to \mathbb{R}^d$: **invertible**,

and $\mathcal{D}_k \overset{\text{i.i.g.}}{\longleftarrow} \mathcal{M}_k$, i.e.,

1. $S \sim q_k$
2. $Z = F(S)$



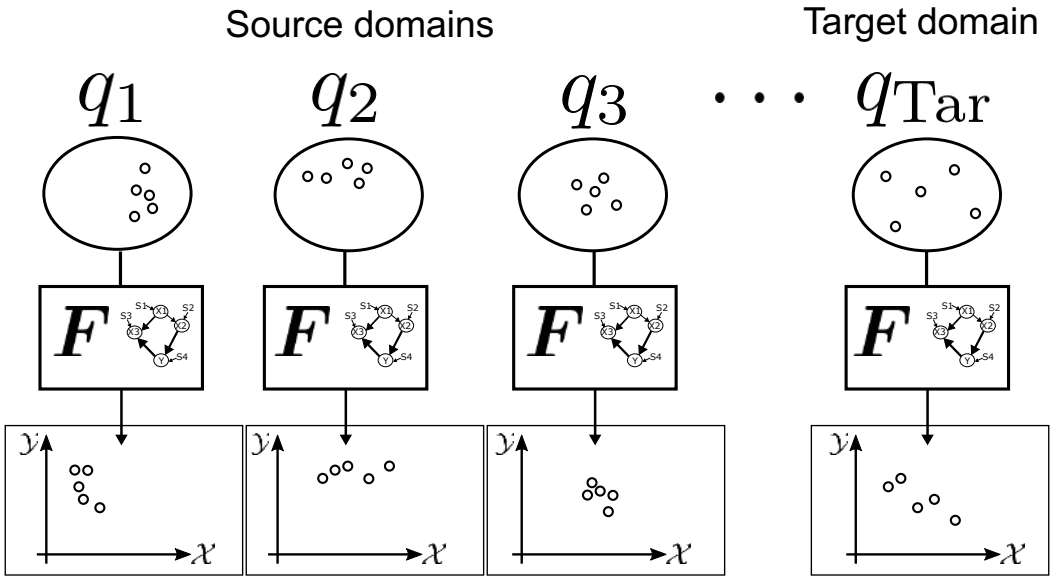
F can be estimated by NLICA (Hyvärinen et al., 2019) under appropriate assumptions.

(F corresponds to the RSF of the SCM of each domain. Slight generalization to assuming an SCM.)

(NLICA: NonLinear Independent Component Analysis. F: mixing map/function) (Hyvärinen et al., 2019)

Problem Setup (3/3)

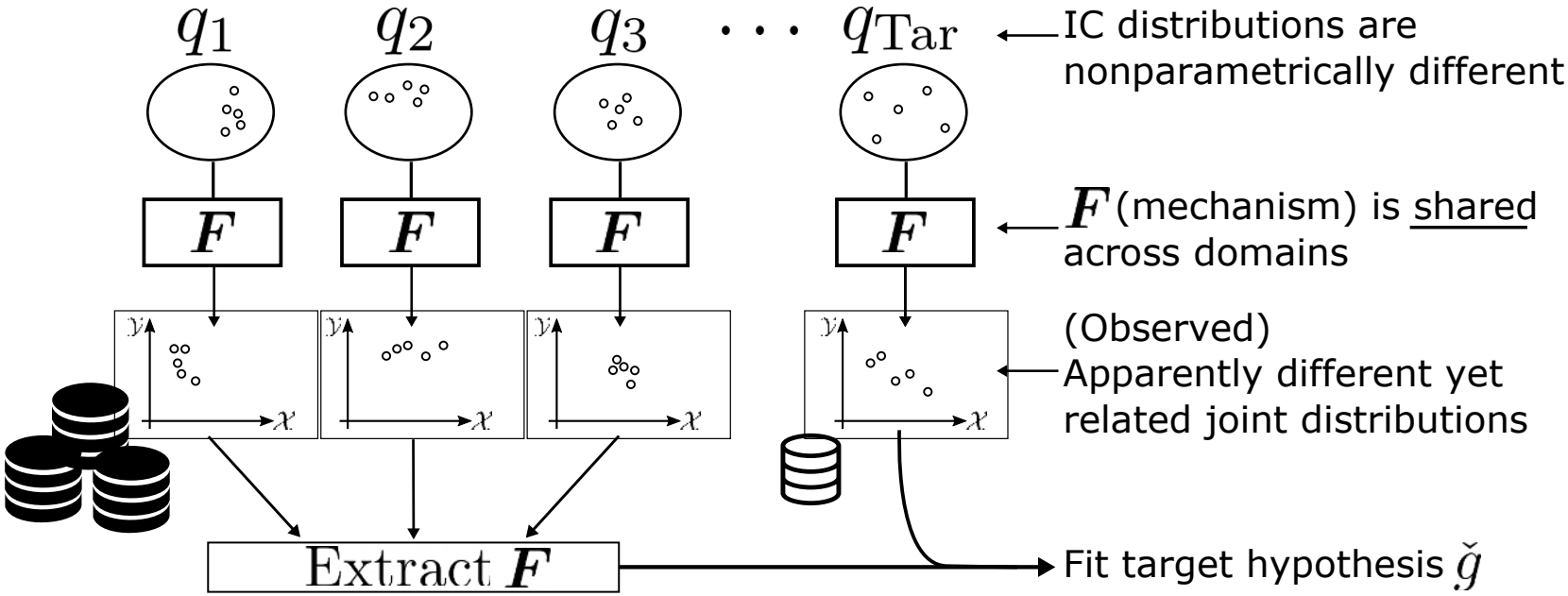
Main Assumption: Generative Mechanism F is common.



Allow nonparametric variation in $q \in \mathcal{Q}$

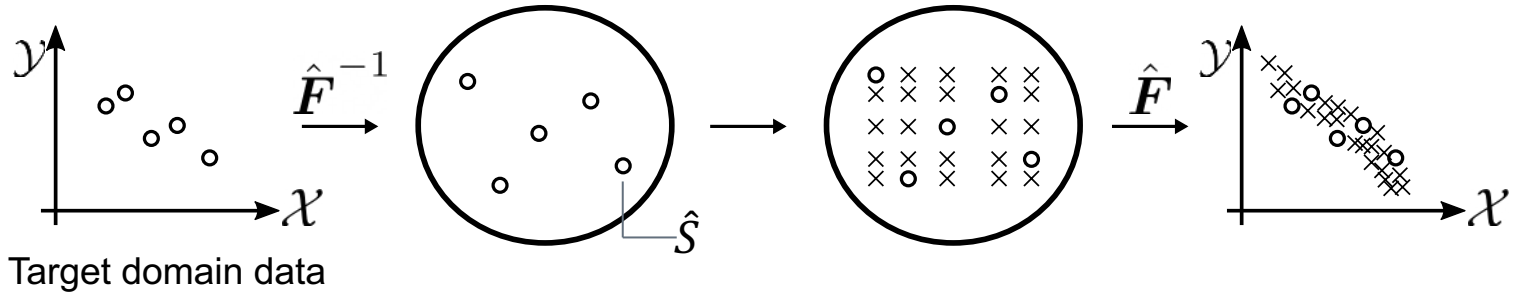
⇝ Accommodate apparently different distributions for DA.

Problem Setup Overview



Proposed Method

Proposed Method (Causal Mechanism Transfer; CMT)

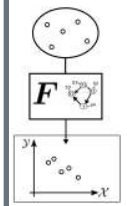


Estimation 1. Estimate F from source-domain data (ICA) (Hyvärinen et al., 2019).

Inflation 2. Estimate ICs of target-domain data using $\hat{F}^{-1}$.

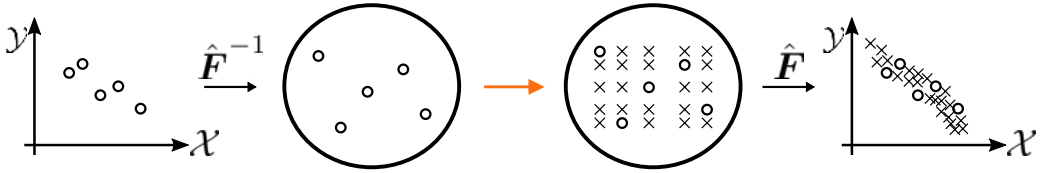
Inflation 3. Apply "IC element exchange" to generate "IC candidates"

Synthesis 4. Use $\hat{F}$ on IC candidates to get target-domain pseudo data



We used invertible neural networks to implement $\hat{F}$ (Kingma and Dhariwal, 2018).

Proposed algorithm: Inflation step (IC element exchange)



Select one data point for each dimension (with replacement)
= sampling from empirical marginal distribution.

$$
\begin{array}{c|cccccc}
 & \hat{S}_1 & \hat{S}_2 & \cdots & \hat{S}_{n-1} & \hat{S}_n & \\
1 & \hat{s}_{11} & \hat{s}_{12} & \cdots & \hat{s}_{1,n-1} & \hat{s}_{1n} & \to \hat{s}_{1,n-1} \\
2 & \hat{s}_{21} & \hat{s}_{22} & \cdots & \hat{s}_{2,n-1} & \hat{s}_{2n} & \to \hat{s}_{22} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \to \vdots \\
d-1 & \hat{s}_{d-1,1} & \hat{s}_{d-1,2} & \cdots & \hat{s}_{d-1,n-1} & \hat{s}_{d-1,n} & \to \hat{s}_{d-1,1} \\
d & \hat{s}_{d1} & \hat{s}_{d2} & \cdots & \hat{s}_{d,n-1} & \hat{s}_{dn} & \to \hat{s}_{d2}
\end{array}
$$

Theoretical Analysis

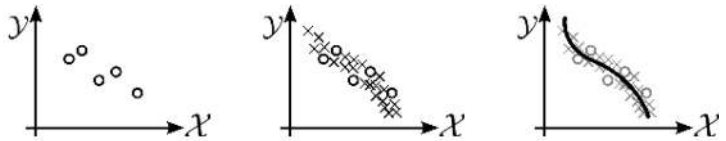
Q1. How does the method help statistically?

Theorem: If $\hat{F} = F$, then the proposed risk estimator is the uniformly minimum variance unbiased risk estimator (UMVUE)

The method can be useful in terms of variance.

Q2. What if $\hat{F} \neq F$?

Theorem: an excess risk bound.



The method suppresses overfitting but may introduces bias proportionally to the degree of $\hat{F} \neq F$.

Experiment Setup and Results

- Data (Greene, 2012)
  - Gasoline consumption prediction
  - 18 countries (=domains)
  - 19 years, 3-dim. input
- Compared methods:
  - TarOnly: learning with only target-domain data
  - SrcOnly: learning with only source-domain data

Even when the compared methods that use source-domain data suffered from the negative transfer, we observed

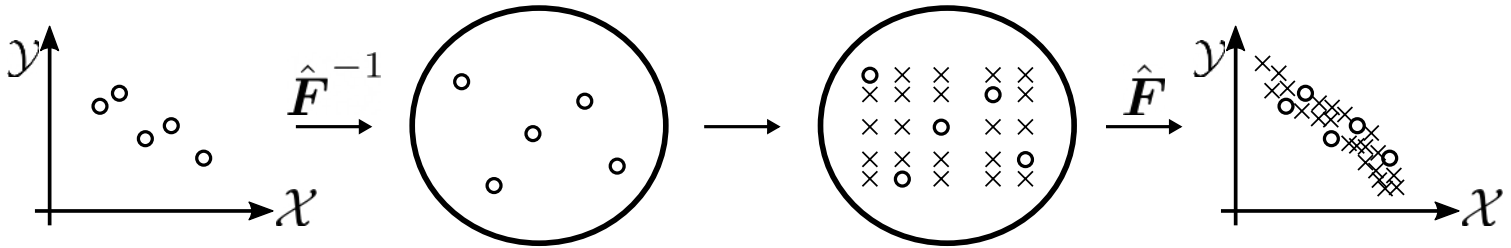
Proposed > *TrgOnly*

Target	(LOO)	TarOnly	**Prop**	SrcOnly	S&TV	TrAda	GDM	Copula	IW(.5)
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)	34.06 (0.67)
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)	105.68 (3.13)
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)	571.33 (1.6)
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)	21.83 (0.93)
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)	113.5 (1.15)
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)	807.88 (4.14)
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)	39.56 (1.7)
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)	5.79 (0.12)
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)	237.15 (6.46)
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)	129.3 (10.47)
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)	33.38 (1.63)
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)	27.09 (0.76)
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)	56.54 (2.16)
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)	113.55 (1.72)
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)	9.33 (0.22)
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)	139.29 (2.41)
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)	69.19 (0.87)
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)	393.45 (1.68)
#Best	-	2	10	2	4	0	0	0	0

Baselines/proposed ← → Compared DA methods

Summary

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 demonstrates the validity.



Take-home message

Info of data generating mechanism captured by causal models may serve as keys to perform transfer/meta learning.

References

References (A–E)

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].

Box, G. E. P. (1976), "Science and statistics", Journal of the American Statistical Association, 71 (356): 791–799, doi:10.1080/01621459.1976.10480949.

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.

References (A–E)

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.

A. Eggers. (Feb. 2016) "Multivariate relationships", [Lecture note]. Available at: http://andy.egge.rs/teaching/qs1/week_6_multivariate_2016_to_distribute.pdf (Accessed: 25 Nov. 2020).

L. Dinh, J. Sohl-Dickstein, and S. Bengio. (2017). 'Density estimation using real NVP', in 5th International Conference on Learning Representations, Conference Track Proceedings.

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

M. Bauer and A. Mnih. (2019). 'Resampled priors for variational autoencoders', in Proceedings of Machine Learning Research, vol. 89, PMLR, Apr. 2019, pp. 66–75.

References (F–J)

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.

F. Huszr. (Jan. 2019), "Causal Inference 2: Illustrating Interventions via a Toy Example". Available at: https://www.inference.vc/causal-inference-2-illustrating-interventions-in-a-toy-example/ (Accessed: 25 Nov. 2020).

W. H. Greene. (2012). "Econometric Analysis", Seventh. Boston: Prentice Hall.

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

Hume, D. (1748). *An Enquiry Concerning Human Understanding*.

References (K–O)

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.

References (K–O)

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.

F. H. Messerli. (2012), “Chocolate Consumption, Cognitive Function, and Nobel Laureates”, New England Journal of Medicine, vol. 367, no. 16, pp. 1562–1564.

A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis. (2018). ‘Parallel WaveNet: Fast high-fidelity speech synthesis’, in Proceedings of the 35th International Conference on Machine Learning, pp. 3918–3926.

D. P. Kingma and P. Dhariwal. (2018). ‘Glow: Generative flow with invertible 1x1 convolutions’, in Advances in Neural Information Processing Systems 31.

S. Kim, S.-G. Lee, J. Song, J. Kim, and S. Yoon. (2019). ‘FloWaveNet : A generative flow for raw audio’, in Proceedings of the 36th International Conference on Machine Learning.

E. T. Nalisnick, A. Matsukawa, Y. W. Teh, D. Grr, and B. Lakshminarayanan. (2019). ‘Hybrid models with deep and invertible features’, in Proceedings of the 36th International Conference on Machine Learning, vol. 97, Long Beach, California, USA: PMLR, pp. 4723–4732.

References (K–O)

Khemakhem, I., Kingma, D. P., Monti, R. P., & Hyvärinen, A. (2019). Variational autoencoders and nonlinear ICA: A unifying framework. ArXiv:1907.04809 [Cs, Stat]. http://arxiv.org/abs/1907.04809

J. M. Mooij, D. Janzing, and B. Schölkopf. (2013). "From ordinary differential equations to structural causal models: The deterministic case." In Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence (UAI), pp. 440–448.

Mitchell, T.M., 1980. The need for biases in learning generalizations.

References (P–T)

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].

References (P–T)

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.

References (P–T)

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.

J. Pearl. (2009), "Causality: Models, Reasoning and Inference", Second. Cambridge, U.K.; New York: Cambridge University Press.

Schölkopf, B. (2019). "Causality for Machine Learning", ArXiv:1911.10500 [Cs, Stat].

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

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

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

S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. (2018). 'Domain adaptation by using causal inference to predict invariant conditional distributions', in Advances in Neural Information Processing Systems 31, Curran Associates, Inc., pp. 10846–10856.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y., 2021. Toward causal representation learning. Proceedings of the IEEE 109, 612–634.

References (P–T)

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.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

References (U–Z)

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.

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

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.

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

References (U-Z)

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

References (Ch4)

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

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

References (Ch4)

[7] 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.

[8] 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.

[9] 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.

[10] 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.

References (Ch4)

[11] 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.

[12] 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].

[13] 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.

[14] 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.

[15] 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 (Ch4)

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

References (Ch4)

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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 (Ch4)

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

References (Ch4)

[31] 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.

[32] 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.

[33] 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.

[34] 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.

References (Ch4)

[35] S. Wright, ‘Correlation and causation’, Journal of Agricultural Research, vol. 20, no. 7, pp. 557–585, 1921.

[36] 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].

[37] J. Mooij, MLSS 2019: Causality, 2019.

[38] 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.

[39] 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 (Ch4)

[40] ——, ‘Nonlinear ICA of temporally dependent stationary sources’, in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017,

pp. 460–469.

[41] 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].

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

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

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

[45] 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].

[46] 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 (Ch4)

[47] 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.

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

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