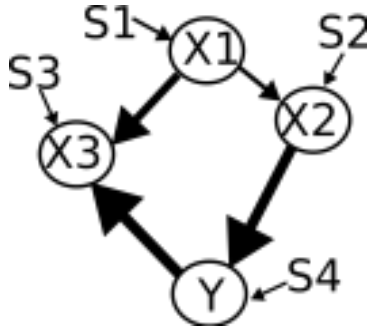
Few-shot Domain Adaptation by Causal Mechanism Transfer

Domain adaptation

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

Causal mechanism



A. Common causal mechanism as the relation.

Summary

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

What's the intuition?

Intuition

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

Motivating example: Regional disease prediction

- Predict disease risk from medical records. [1]
- Data distributions may vary for different lifestyles.
- Common pathological mechanism across regions.

Few-shot Domain Adaptation by Causal Mechanism Transfer

Takeshi Teshima[12], Issei Sato[12], and Masashi Sugiyama[21]

[1] The University of Tokyo [2] RIKEN

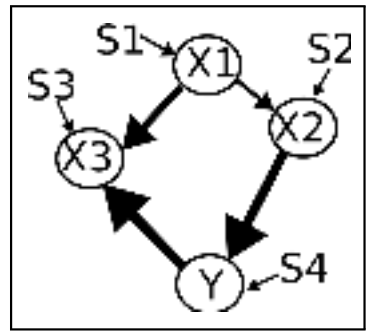
(This work was supported by RIKEN Junior Research Associate Program.)

Preliminary: Causal model (SEMs)

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

- Generative model for the joint distribution of data.
- Consists of deterministic functions of the form:

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



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

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

[2] Acyclicity is assumed.

Nonlinear-ICA and reduced-form SEM

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

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

Structural equations

$\Longrightarrow$

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

Reduced form

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

[3] ICA = Independent component analysis.

Formulation 1/3: Data and goal

Basic setup: regression domain adaptation

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

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

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

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

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

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

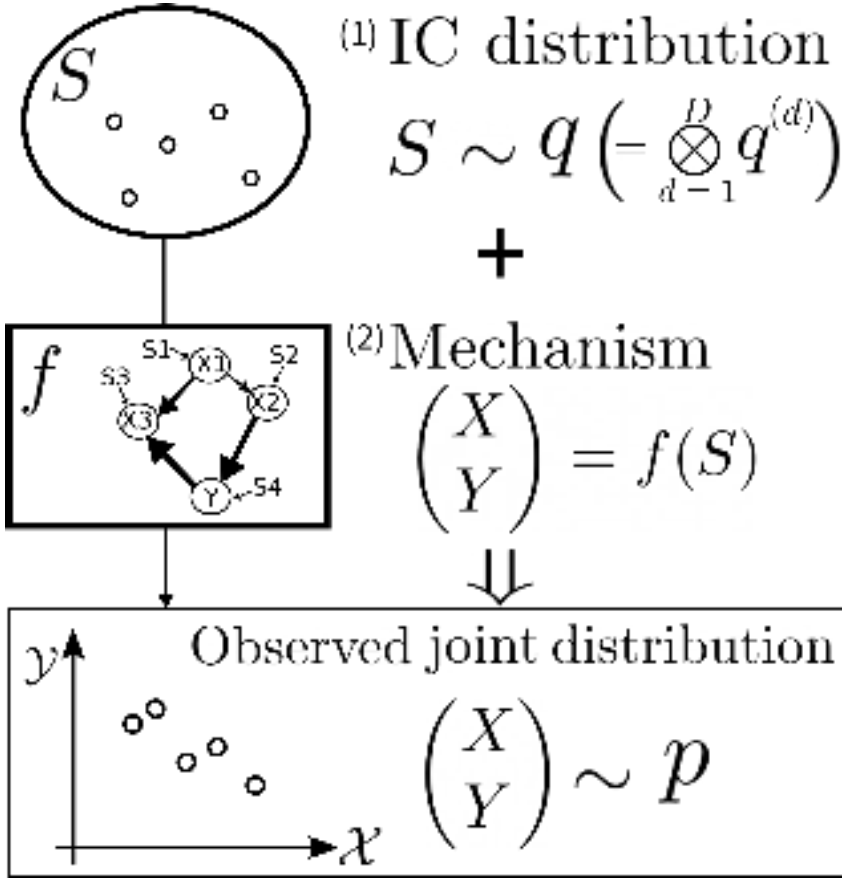
Goal: accurate predictor for the target distribution

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

(ℓ: loss function)

Formulation 2/3: Each domain

- Each domain follows a nonlinear-ICA model.



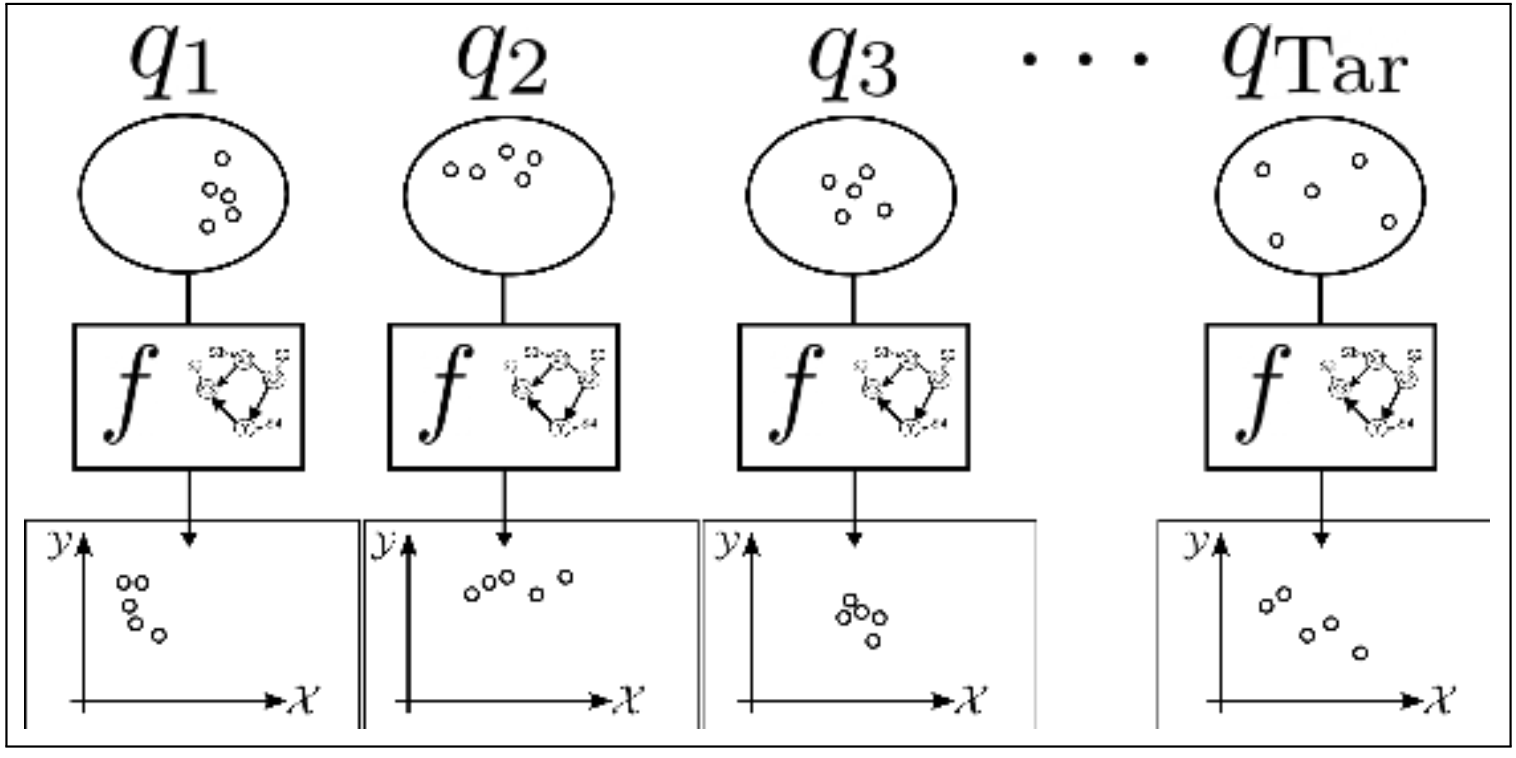
Dist. p consists of (f, q)

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

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

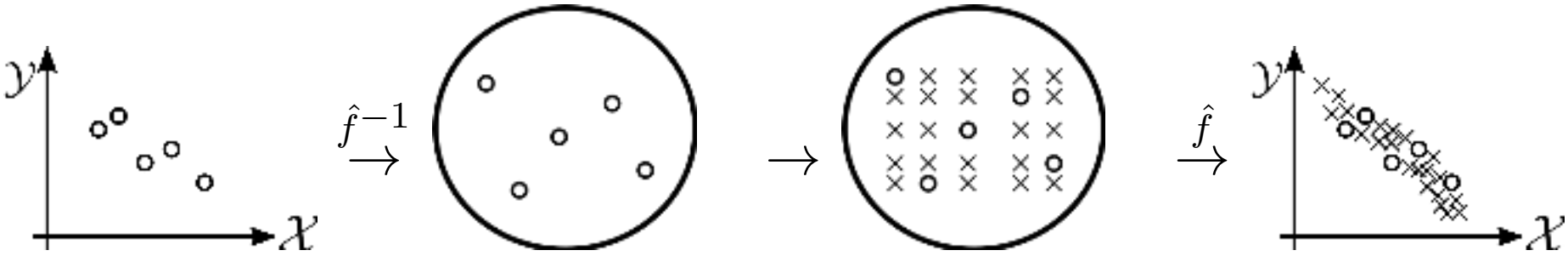
Formulation 3/3: Key assumption

- Key Assumption: generative mechanism f is common.



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

Proposed method: Strategy



Idea: How to exploit the assumption

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

Theoretical analyses

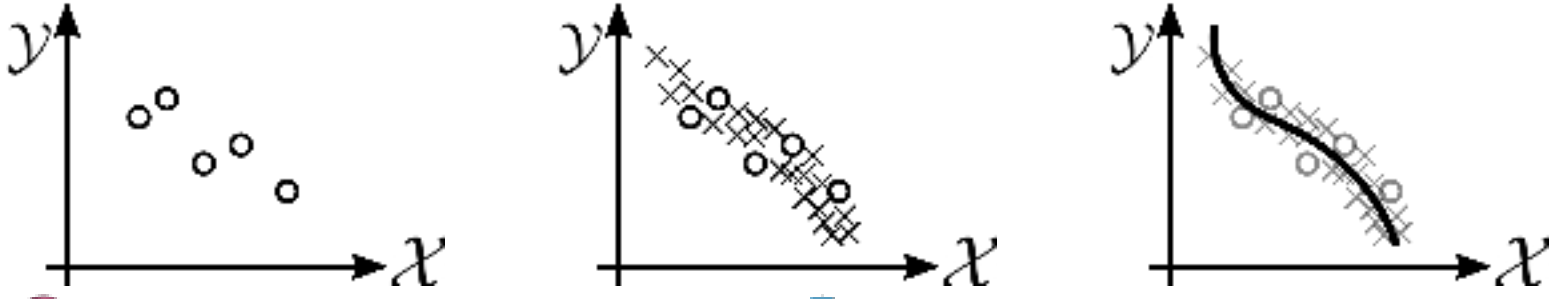
Q1. How does the method statistically help?

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

The method should help in terms of variance.

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

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



Mitigate overfitting. Introduce bias.

Experiment: Setup

- Dataset: Gasoline consumption dataset [3].
  - Panel data from econometrics (SEMs have been applied).
  - 18 countries (=domains), 19 years, $D = 4$.
- Baselines for regression domain adaptation.

Name	Compared method (predictor: KRR)
TarOnly	Train on target.
SrcOnly	Train on source.
S&TV	Train on source, CV on target.
TrAdaBoost	Boosting for few-shot regression transfer [4].
IW	Joint importance weight using RuLSIF [5].
GDM	Generalized discrepancy minimization [6].
Copula	Non-parametric R-vine copula method [7].
LOO (reference)	LOOCV error estimate.

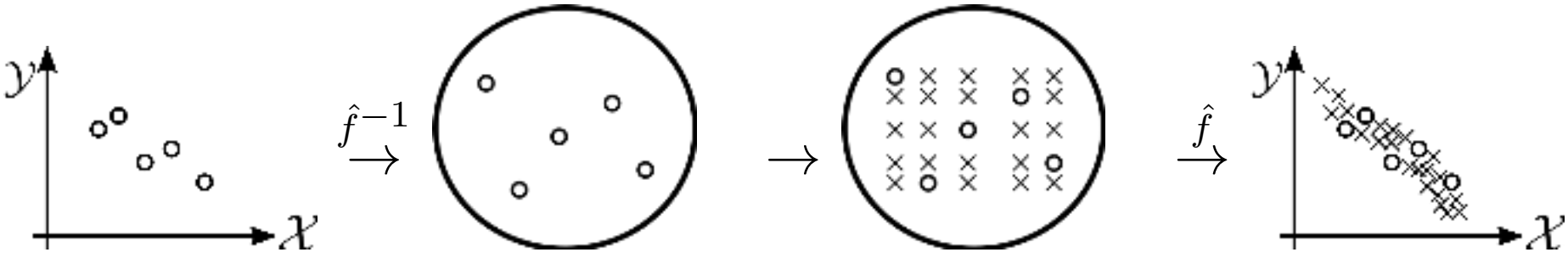
Experiment: Result

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

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

Take-home message

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



References

[1] 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.

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

[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 (cont.)

[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.