Domain Adaptation Under Structural Causal Models

at MSRI workshop on
Foundations Of Stable, Generalizable And Transferable Statistical Learning

---

Yuansi Chen

Joint work with Peter Bühlmann

Department of Statistical Science
Duke University

Many machine learning systems aren't good at domain adaptation (yet)

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

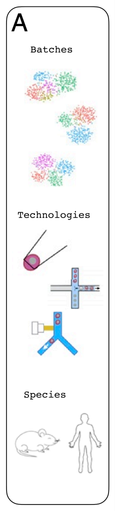
(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

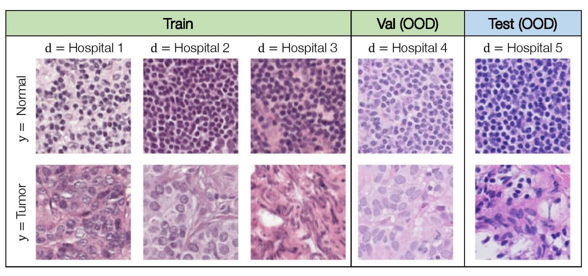
(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

ClarifAI.com, Beery et al. 2018

- Cows in "common" contexts (e.g. Alpine pastures) are detected and classified correctly (A)
- Cows in uncommon contexts (beach, waves and boat) are not detected (B) or classified poorly (C)

In modern scientific problems, we also wish to adapt across multiple domains





- Learn a tumor prediction model that generalize to a new hospital [Bandi et al. 2018]
- scRNA-seq datasets from different batches, technologies, and across species [Peng et al. 2020]

Outline

1. Empirical success of domain adaptation (DA)
   and reflections on its general validity
2. Analysis of popular DA methods
   under structural causal models (SCMs)
3. A new DA method (CIRM)
4. Numerical validation

Empirical success of DA

Domain adaptation problem setup

We observe

- Source data: M separate labeled datasets ($M \geq 1$)

$$S^{Ⓜ} = ((x_1^{Ⓜ}, y_1^{Ⓜ}), \cdots, (x_n^{Ⓜ}, y_n^{Ⓜ})) \text{ from } \mathcal{P}^{Ⓜ}$$

- Target data: unlabeled dataset (red is unobserved)

$$\tilde{S} = ((\tilde{x}_1, \tilde{y}_1), \cdots, (\tilde{x}_n, \tilde{y}_n)) \text{ from } \widetilde{\mathcal{P}}$$

Goal of domain adaptation:

to predict target labels so that the following population target risk is small

$$\tilde{R}(f) = \mathbb{E}_{(X,Y)\sim\widetilde{\mathcal{P}}}[\ell(f(X), Y)]$$

A natural baseline method

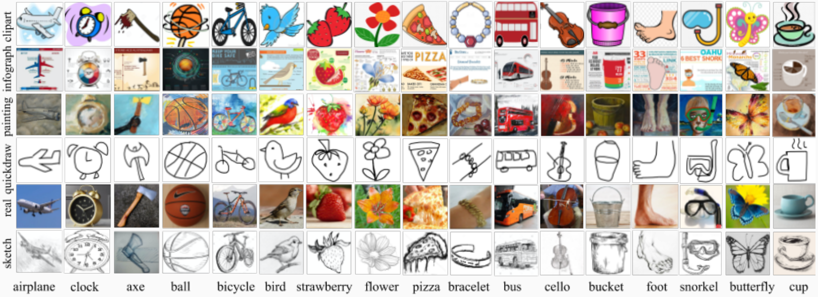
The less ambitious goal is to know if we can outperform SrcPool

- **SrcPool**: Combining all the source data and train a model
  Do not use the target covariates at all

If we don’t assume any relationship between the source and target data distribution, the DA problem is ill-posed

In general, there is no free lunch
But ...

DA has huge empirical successes I



DomainNet dataset [Peng et al. 2019]

- Used VisDA-2019 Challenge: predict unlabelled clipart images from other datasets
- Top accuracy 76.0% compared to less than 10% for SrcPool

DA has huge empirical successes II

Sentiment analysis for Amazon product review data

Text data from: books, DVDs, electronics and kitchen appliances

(Multi-Domain Sentiment Dataset, Blitzer et al. 2007)

Image classification from various domains

(Office-Caltech dataset, Hoffman et al. 2012)

DA has huge empirical successes III

Digit classification under perturbations

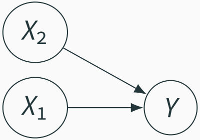
METHOD (SOURCE / TARGET)	MNIST / MNIST-M	SYN NUMBERS / SVHN	SVHN / MNIST	SYN SIGNS / GTSRB
SOURCE ONLY	.5225	.8674	.5490	.7900
SA (Fernando et al., 2013)	.5690 (4.1%)	.8644 (−5.5%)	.5932 (9.9%)	.8165 (12.7%)
DANN	**.7666** (52.9%)	**.9109** (79.7%)	**.7385** (42.6%)	**.8865** (46.4%)
TRAIN ON TARGET	.9596	.9220	.9942	.9980

(Domain-Adversarial Training of Neural Networks (DANN), Ganin et al. 2016)

[PDF] **Domain-adversarial** training of **neural networks**
Y Ganin, E Ustinova, H Ajakan, P Germain… - The journal of machine …, 2016 - jmlr.org
We introduce a new representation learning approach for **domain** adaptation, in which data at training and test time come from similar but different distributions. Our approach is directly inspired by the theory on **domain** adaptation suggesting that, for effective **domain** transfer to …
Cited by 3599 Related articles All 33 versions

Are there negative results on DA?

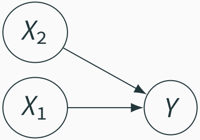
At least in causal prediction setting, DA is no better than SrcPool



We don't gain additional information for $Y \mid X$ from the target X

(On Causal and Anticausal learning, Scholkopf et al. 2012)

At least in causal prediction setting, DA is no better than SrcPool



We don't gain additional information for $Y \mid X$ from the target X

(On Causal and Anticausal learning, Scholkopf et al. 2012)

Examples of causal prediction:

- Predict housing values based on nitric oxides concentration
- Predict fish weight from fish length, fish width and fish type etc.

Main questions

Q: what properties do these success stories share?

Q: can we identify the assumptions needed for popular DA algorithms to have low target risk?

Main questions

Q: what properties do these success stories share?

Q: can we identify the assumptions needed for popular DA algorithms to have low target risk?

You may wonder: maybe domain knowledge is applied in success DA?

True in some cases, but the troubling trend is that many popular DA algorithms are advertised as generic methods

The intriguing success of DIP

Domain invariant projection (DIP) is becoming one of the most popular DA methods

(Pan et al. 2010, Baktashmotlagh et al. 2013, Ganin et al. 2016, etc.)

- **Intuition**: Assumes the existence of a common subspace between the source and target data

The intriguing success of DIP

Domain invariant projection (DIP) is becoming one of the most popular DA methods

(Pan et al. 2010, Baktashmotlagh et al. 2013, Ganin et al. 2016, etc.)

- **Intuition**: Assumes the existence of a common subspace between the source and target data
- **Generic formulation**:

$$f_{\text{DIP}}(x) := u_{\text{DIP}} \circ v_{\text{DIP}}(x)$$

$$u_{\text{DIP}}, v_{\text{DIP}} := \operatorname*{arg\,min}_{u \in \mathcal{U}, v \in \mathcal{V}} \mathbb{E}\ell(u \circ v(X), Y) + \lambda \cdot \mathcal{D}(v(X), v(\tilde{X}))$$

where $\mathcal{D}$ is a distributional distance

Popular implementations of DIP

$$\min_{u \in \mathcal{U}, v \in \mathcal{V}} \mathbb{E}\ell(u \circ v(X), Y) + \lambda \cdot \mathcal{D}(v(X), v(\tilde{X}))$$

DIP Variants	Func class $\mathcal{U}, \mathcal{V}$	Distance $\mathcal{D}$	When is better than SrcPool
TCA (Pan et al.) 2009	linear	mean diff	WiFi localization
DIP (Baktashmotlagh et al.) 2013	linear	MMD Gaussian kernel	Office-Caltech
DANN (Ganin et al.) 2016	conv nets	Generative adversarial nets	MNIST-M
M3SDA (Peng et al.) 2016	conv nets	Moment matching	DomainNet

Two precise motivating questions

- What assumptions are needed for DIP to outperform SrcPool?
- Can we design datasets that make DIP fail drastically?

Previous ways to formulate DA

- DA = classic VC theory + divergence between source and target
  Ben-David et al. 2007, 2010; Mansour et al. 2009; Cortes and Mohri 2011, 2014; Hoffman et al. 2018; Redko et al. 2020 ...
- Missing data y imputation via expectation maximization Amini and Gallinari 2003; McLachlan and Krishnan 2007 ...
- Distributional robustness Huber, 1964; Gao et al., 2017; Sinha et al., 2018; Duchi and Namkoong, 2018; Yuan et al., 2019 ...
- $Y \mid X$ invariant, but covariate shift Quionero-Candela et al. 2009; Storkey 2009; Sugiyama and Kawanabe 2012 ...
- $X \mid Y$ invariant, but label shift Lipton et al., 2018; Aziz- zadenesheli et al., 2019; Garg et al., 2020 ...
- Full structural causal model (SCM) Pearl and Bareinboim 2014

Analysis of DA methods under structural causal models

Structural causal models (SCMs)

Introduced and polished by Pearl (2000) as mathematical models to describe causal relationships between variables

It combines

- structural equations used in economics and social science
- causal framework of Neyman and Rubin
- graphical models for probabilistic reasoning

SCMs are needed to prove guarantees but the DA algorithms do not need SCMs to run

DA under linear SCM with noise interventions

Source data generation $\mathcal{P}^{ⓜ}$

$$\begin{bmatrix} X^{ⓜ} \\ Y^{ⓜ} \end{bmatrix} = \begin{bmatrix} \mathbf{B} & b \\ \omega^\top & 0 \end{bmatrix} \begin{bmatrix} X^{ⓜ} \\ Y^{ⓜ} \end{bmatrix} + g(a^{ⓜ}, \varepsilon^{ⓜ})$$

Target data generation $\widetilde{\mathcal{P}}$

$$\begin{bmatrix} \widetilde{X} \\ \widetilde{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{B} & b \\ \omega^\top & 0 \end{bmatrix} \begin{bmatrix} \widetilde{X} \\ \widetilde{Y} \end{bmatrix} + g(\widetilde{a}, \widetilde{\varepsilon})$$

DA methods in consideration (1)

Oracle and baseline methods

- **OLSTar:** OLS on target data only
- **Causal:** $x \mapsto x^\top \omega$
- **OLSSrc①:** OLS on source dataset 1 only

DA methods in consideration (2)

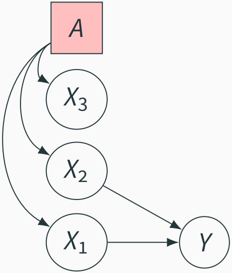
DA methods

- **DIP:** $x \mapsto x^\top \beta_{\text{DIP}}^{①} + \beta_{\text{DIP},0}^{①}$

$$\beta_{\text{DIP}}^{①}, \beta_{\text{DIP},0}^{①} := \underset{\beta,\beta_0}{\arg\min}\ \mathbb{E}_{(X,Y)\sim\mathcal{P}^{①}} \left(Y - X^\top \beta - \beta_0\right)^2$$
$$\text{s.t. } \mathbb{E}_{X\sim\mathcal{P}_X^{①}} \left[X^\top \beta\right] = \mathbb{E}_{X\sim\widetilde{\mathcal{P}}_X} \left[X^\top \beta\right]$$

Simplest DIP is considered

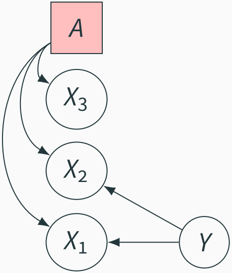
- linear function classes
- mean difference is used as distributional distance



Ex1: Causal prediction

$$X_1 = \varepsilon_{X_1} + a_1$$
$$X_2 = \varepsilon_{X_2} + a_2$$
$$X_3 = \varepsilon_{X_3} + a_3$$
$$Y = X_1 + X_2 + \varepsilon_Y + a_Y, \text{ w/}$$

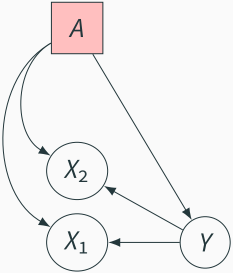
$$a^{①} = \begin{bmatrix} 1 & 1 & 1 & 0 \end{bmatrix}^\top$$
$$\widetilde{a} = \begin{bmatrix} -1 & -1 & -1 & 0 \end{bmatrix}^\top$$



Ex2: Anticausal

$$X_1 = Y + \varepsilon_{X_1} + a_1$$
$$X_2 = Y + \varepsilon_{X_2} + a_2$$
$$X_3 = \varepsilon_{X_3} + a_3$$
$$Y = \varepsilon_Y + a_Y, \text{ w/}$$

$$a^{①} = \begin{bmatrix} 1 & 1 & 1 & 0 \end{bmatrix}^\top$$
$$\widetilde{a} = \begin{bmatrix} -1 & -1 & -1 & 0 \end{bmatrix}^\top$$



Ex3: Anticausal + a_Y

$$X_1 = Y + \varepsilon_{X_1} + a_1$$
$$X_2 = -Y + \varepsilon_{X_2} + a_2$$
$$Y = \varepsilon_Y + a_Y, \text{ w/}$$

$$a^{①} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\top$$
$$\widetilde{a} = \begin{bmatrix} -1 & -1 & -1 \end{bmatrix}^\top$$

Performance of DA methods on three examples

Risk \ Methods	**OLSTar** (oracle)	**Causal**	**OLSSrc**①	**DIP**①
Ex 1, target risk Causal	**0.200**	0.200	0.200	16.333
Ex 2, target risk Anticausal	**0.040**	0.200	2.600	0.086
Ex 3, target risk Anticausal, a_Y	**0.040**	1.200	0.200	4.066

Performance of DIP on real causal datasets

Classification accuracy (the higher the better) on UCI datasets

Methods / Accuracy (%)	OLSSrc①	**DIP①**
DNA Splice-junction Causal	**95.7 ± 1.4**	71.8 ± 7.7
Balance Scale Causal	**92.7 ± 2.4**	69.1 ± 2.5
Chess (King Rook-King) Causal	**57.8 ± 1.1**	56.0 ± 0.7

YC — On-going work with Keru Wu

Sufficient assumptions for DIP target risk guarantees

- Linear SCM
- Anticausal prediction, $\omega = 0$
- No intervention on Y, $a_Y^{①} = \widetilde{a}_Y = 0$
- DIP matching penalty fits the noise intervention type

Sufficient assumptions for DIP target risk guarantees

- Linear SCM
- Anticausal prediction, $\omega = 0$
- No intervention on Y, $a_Y^{①} = \widetilde{a}_Y = 0$
- DIP matching penalty fits the noise intervention type

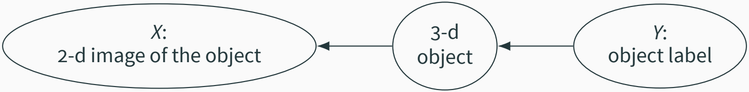
Theorem 1 (DIP, informal)
Under above assumptions

$$\underbrace{\tilde{R}\left(f_{\mathrm{DIP}}^{①}\right)}_{\text{DIP target risk}} = \underbrace{R^{①}\left(f_{\mathrm{DIP}}^{①}\right)}_{\text{DIP source risk}} \approx \underbrace{\tilde{R}\left(f_{\mathrm{OLSTar}}\right)}_{\text{oracle target risk}}$$

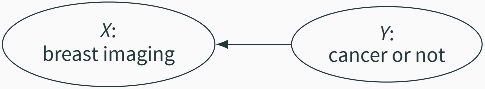
Also, OLSSrc risk is very sensitive to the magnitude of X interventions

Hindsight on the empirical success of DIP

1. Anticausal data generation is plausible for many machine learning datasets
    - Object recognition



    - Breast cancer diagnosis



Hindsight on the empirical success of DIP

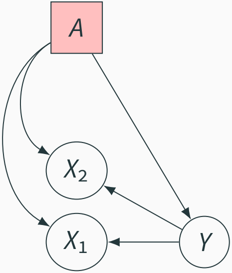
1. Anticausal data generation is plausible for many machine learning datasets
2. Many datasets do not have Y intervention, mainly because many are made-up

Hindsight on the empirical success of DIP

1. Anticausal data generation is plausible for many machine learning datasets
2. Many datasets do not have Y intervention, mainly because many are made-up
3. The use of MMD or CNN-based generative adversarial nets (GANs) for the DIP matching penalty allows to fit a large variety of intervention types

Literature on the empirical failure of DIP
Zhao et al. (2019), Johansson et al. (2019), Li et al. (2019), Tachet des Combes et al. (2020)

Why did DIP fail in simple example 3?



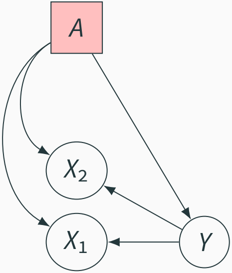
Ex3: Anticausal + a_Y

$$X_1 = Y + \varepsilon_{X_1} + a_1$$
$$X_2 = -Y + \varepsilon_{X_2} + a_2$$
$$Y = \varepsilon_Y + a_Y, \text{ w/}$$

$$a^{①} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\top$$
$$\widetilde{a} = \begin{bmatrix} -1 & -1 & -1 \end{bmatrix}^\top$$

- Matching the distribution of $v(X)$ between source and target data no longer aligns the conditional $X \mid Y$ between source and target

Why did DIP fail in simple example 3?



Ex3: Anticausal + a_Y

$$X_1 = Y + \varepsilon_{X_1} + a_1$$
$$X_2 = -Y + \varepsilon_{X_2} + a_2$$
$$Y = \varepsilon_Y + a_Y, \text{ w/}$$

$$a^{①} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\top$$
$$\widetilde{a} = \begin{bmatrix} -1 & -1 & -1 \end{bmatrix}^\top$$

- Matching the distribution of $v(X)$ between source and target data no longer aligns the conditional $X \mid Y$ between source and target
- Ideally, we want to match the distribution of $v(X \mid Y)$, but we don't have access to Y in target

A new DA method to deal with Y intervention

Conditionally invariant components (CIC) assumption

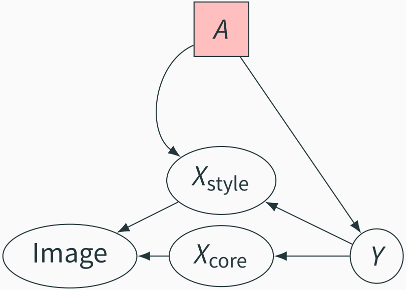
Assumption (CIC, Gong et al. 2016, Heinze-Deml and Meinshausen 2017)

There exists an unknown transformation $\mathcal{T}$ such that the conditional distribution $\mathcal{T}(X) \mid Y$ is invariant across source and target data

If we find such a transformation $\mathcal{T}$,

- If the Y intervention is not too large, then the joint distributiion $(\mathcal{T}(X), Y)$ becomes almost invariant.
- $\mathcal{T}(X)$ can serve as a proxy of Y

Conditionally invariant components (CIC) assumption in Heinze-Deml and Meinshausen 2017

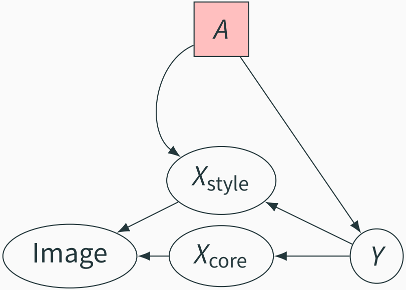


Eyeglass detection in CelebA

(Liu et al. 2015)

Core: eyeglass, **Style:** background, light condition, hairstyle

Conditionally invariant components (CIC) assumption in Heinze-Deml and Meinshausen 2017



Eyeglass detection in CelebA

(Liu et al. 2015)

Core: eyeglass, **Style:** background, light condition, hairstyle

Bias in one source dataset

- people outdoor are more likely to wear glasses
- men are more likely to wear glasses than women

Conditional invariant penalty (CIP) algorithm

Conditional invariance penalty (CIP) minimizes the total source risk by adding the penalty

$$\mathcal{D}\left(\mathcal{T}(X^{①}) \mid Y^{①}, \mathcal{T}(X^{ⓜ}) \mid Y^{ⓜ}\right) \text{ small, for all } 2 \leq m \leq M$$

Logic behind the new method CIRM

- Learn a proxy of Y via CIP across all source environments
- Use the proxy of Y to correct for the Y intervention
- Reduce to the scenario when DIP works

Sufficient assumptions for CIRM target risk guarantees

- Linear SCM
- Anticausal prediction, $\omega = 0$
- ~~no intervention on Y~~
- existence of CICs, enough source envs to learn CICs
- The new matching penalty fits the noise intervention type

Sufficient assumptions for CIRM target risk guarantees

- Linear SCM
- Anticausal prediction, $\omega = 0$
- ~~no intervention on Y~~
- existence of CICs, enough source envs to learn CICs
- The new matching penalty fits the noise intervention type

Theorem 2 (CIRM, informal)
Under above assumptions

$$\underbrace{\tilde{R}\left(f^{①}_{\mathrm{CIRM}}\right)}_{\text{CIRM target risk}} \approx \underbrace{\tilde{R}\left(f_{\mathrm{OLSTar}}\right)}_{\text{oracle target risk}} \leq \underbrace{\tilde{R}\left(f_{\mathrm{CIP}}\right)}_{\text{CIP target risk}}$$

Also, DIP risk is very sensitive to the magnitude of Y intervention

Numerical experiments (take a look at our paper)

Linear SCM simulations

Sim Num	# Src envs	Causal Direction	Interv X type	Interv on Y?	Has CIC?	Better estimator(s)
(i)	single	anticausal	mean shift	N	-	DIP①
(ii)	multiple	anticausal	mean shift	N	-	DIPweigh
(iii)	multiple	anticausal	mean shift	Y	Y	CIRMweigh
(iv)	single	causal	mean shift	N	-	-
(v)	single	mixed	mean shift	N	-	DIP♢①
(vi)	multiple	anticausal	mean shift	Y	N	-
(vii)	multiple	mixed	mean shift	Y	Y	CIRM♢weigh
(viii)	single	anticausal	var shift	N	-	DIP-std+ DIP-MMD
(ix)	multiple	anticausal	var shift	Y	Y	CIRMweigh-std+ CIRMweigh-MMD

Our paper shows that even under linear SCM, can make DA algorithms fail

Summary

- Dangerous to blindly apply DA algorithms domain knowledge matters!
- DIP works under the assumptions
  anticausal prediction & linear SCM & matching penalty fitting the intervention type & no intervention on Y
- DIP can fail!
  - Intervention on Y
  - Too complicated function class $\mathcal{U}$ (not discussed)
- In the presence of Y intervention, conditionally invariant components (CICs) may become a cure. CIRM useful
- The mixed-causal-anticausal DA is chanllenging: is exact causal inference/discovery necessary for DA?

Thank you!