Few-shot Domain Adaptation by Causal Mechanism Transfer

Takeshi Teshima [1 2] **Issei Sato** [1 2] **Masashi Sugiyama** [2 1]

Abstract

We study *few-shot supervised domain adaptation* (DA) for regression problems, where only a few labeled target domain data and many labeled source domain data are available. Many of the current DA methods base their *transfer assumptions* on either parametrized distribution shift or apparent distribution similarities, e.g., identical conditionals or small distributional discrepancies. However, these assumptions may preclude the possibility of adaptation from intricately shifted and apparently very different distributions. To overcome this problem, we propose *mechanism transfer*, a meta-distributional scenario in which a data generating *mechanism* is invariant among domains. This transfer assumption can accommodate nonparametric shifts resulting in apparently different distributions while providing a solid statistical basis for DA. We take the structural equations in causal modeling as an example and propose a novel DA method, which is shown to be useful both theoretically and experimentally. Our method can be seen as the first attempt to fully leverage the structural causal models for DA.

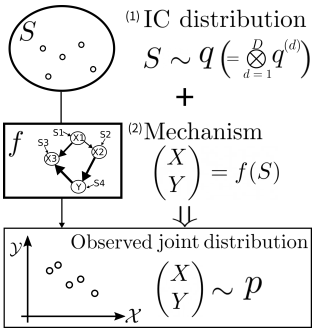


Figure 1: Nonparametric generative model of nonlinear independent component analysis. Our meta-distributional transfer assumption is built on the model, where there exists an invertible function f representing the mechanism to generate the labeled data (X, Y) from the independent components (ICs), S, sampled from q. As a result, each pair (f, q) defines a joint distribution p.

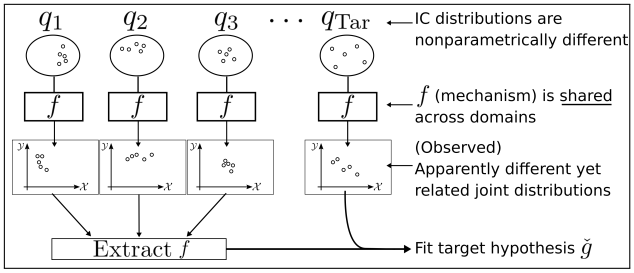


Figure 2: Our assumption of common generative mechanism. By capturing the common data generation mechanism, we enable domain adaptation among seemingly very different distributions without relying on parametric assumptions.

1. Introduction

Learning from a limited amount of data is a long-standing yet actively studied problem of machine learning. Domain adaptation (DA) (Ben-David et al., 2010) tackles this problem by leveraging auxiliary data sampled from related but different domains. In particular, we consider *few-shot supervised* DA for regression problems, where only a few labeled target domain data and many labeled source domain data are available.

A key component of DA methods is the *transfer assumption* (TA) to relate the source and the target distributions. Many of the previously explored TAs have relied on certain direct distributional similarities, e.g., identical conditionals (Shimodaira, 2000) or small distributional discrepancies (Ben-David et al., 2007). However, these TAs may preclude the possibility of adaptation from apparently very different distributions. Many others assume parametric forms of the distribution shift (Zhang et al., 2013) or the distribution family (Storkey & Sugiyama, 2007) which can highly limit the considered set of distributions. (we further review related work in Section 5.1).

To alleviate the intrinsic limitation of previous TAs due to relying on apparent distribution similarities or parametric assumptions, we focus on a meta-distributional scenario where there exists a common generative *mechanism* behind the data distributions (Figures 1,2). Such a common mechanism may be more conceivable in applications involving structured table data such as medical records (Yadav et al.,

[1]The University of Tokyo, Tokyo, Japan [2]RIKEN, Tokyo, Japan. Correspondence to: Takeshi Teshima <teshima@ms.k.u-tokyo.ac.jp>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

2018). For example, in medical record analysis for disease risk prediction, it can be reasonable to assume that there is a pathological mechanism that is common across regions or generations, but the data distributions may vary due to the difference in cultures or lifestyles. Such a hidden structure (pathological mechanism, in this case), once estimated, may provide portable knowledge to enable DA, allowing one to obtain accurate predictors for under-investigated regions or new generations.

Concretely, our assumption relies on the generative model of nonlinear independent component analysis (nonlinear ICA; Figure 1), where the observed labeled data are generated by first sampling latent independent components (ICs) S and later transforming them by a nonlinear invertible *mixing function* denoted by f (Hyvärinen et al., 2019). Under this generative model, our TA is that f representing the mechanism is identical across domains (Figure 2). This TA allows us to formally relate the domain distributions and develop a novel DA method without assuming their apparent similarities or making parametric assumptions.

Our contributions. Our key contributions can be summarized in three points as follows.

1. We formulate the flexible yet intuitively accessible TA of shared generative mechanism and develop a few-shot regression DA method (Section 3). The idea is as follows. First, from the source domain data, we estimate the mixing function f by nonlinear ICA (Hyvärinen et al., 2019) because f is the only assumed relation of the domains. Then, to transfer the knowledge, we perform data augmentation using the estimated f on the target domain data using the independence of the IC distributions. In the end, the augmented data is used to fit a target predictor (Figure 3).

2. We theoretically justify the augmentation procedure by invoking the theory of generalized U-statistics (Lee, 1990). The theory shows that the proposed data augmentation procedure yields the uniformly minimum variance unbiased risk estimator in an ideal case. We also provide an excess risk bound (Mohri et al., 2012) to cover a more realistic case (Section 4).

3. We experimentally demonstrate the effectiveness of the proposed algorithm (Section 6). The real-world data we use is taken from the field of *econometrics*, for which structural equation models have been applied in previous studies (Greene, 2012).

A salient example of the generative model we consider is the structural equations of causal modeling (Section 2). In this context, our method can be seen as the first attempt to fully leverage the structural causal models for DA (Section 5.2).

2. Problem Setup

In this section, we describe the problem setup and the notation. To summarize, our problem setup is *homogeneous*, *multi-source*, and *few-shot supervised* domain adapting regression. That is, respectively, all data distributions are defined on the same data space, there are multiple source domains, and a limited number of labeled data is available from the target distribution (and we do *not* assume the availability of unlabeled data). In this paper, we use the terms *domain* and *distribution* interchangeably.

Notation. Let us denote the set of real (resp. natural) numbers by $\mathbb{R}$ (resp. $\mathbb{N}$). For $N \in \mathbb{N}$, we define $[N] := \{1, 2, \dots, N\}$. Throughout the paper, we fix $D(\in \mathbb{N}) > 1$ and suppose that the input space $\mathcal{X}$ is a subset of $\mathbb{R}^{D-1}$ and the label space $\mathcal{Y}$ is a subset of $\mathbb{R}$. As a result, the overall data space $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ is a subset of $\mathbb{R}^D$. We generally denote a labeled data point by $Z = (X, Y)$. We denote by $\mathcal{Q}$ the set of independent distributions on $\mathbb{R}^D$ with absolutely continuous marginals. For a distribution p, we denote its induced expectation operator by $\mathbb{E}_p$. Table 3 in Supplementary Material provides a summary of notation.

Basic setup: Few-shot domain adaptation. Let p_{Tar} be a distribution (the *target distribution*) over $\mathcal{Z}$, and let $\mathcal{G} \subset \{g : \mathbb{R}^{D-1} \to \mathbb{R}\}$ be a hypothesis class. Let $\ell : \mathcal{G} \times \mathbb{R}^D \to [0, B_\ell]$ be a loss function where $B_\ell > 0$ is a constant. Our goal is to find a predictor $g \in \mathcal{G}$ which performs well for p_{Tar}, i.e., the target risk $R(g) := \mathbb{E}_{p_{\mathrm{Tar}}} \ell(g, Z)$ is small. We denote $g^* \in \arg\min_{g \in \mathcal{G}} R(g)$. To this goal, we are given an independent and identically distributed (i.i.d.) sample $\mathcal{D}_{\mathrm{Tar}} := \{Z_i\}_{i=1}^{n_{\mathrm{Tar}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{Tar}}$. In a fully supervised setting where n_{Tar} is large, a standard procedure is to select g by empirical risk minimization (ERM), i.e., $\hat{g} \in \arg\min_{g \in \mathcal{G}} \hat{R}(g)$, where $\hat{R}(g) := \frac{1}{n_{\mathrm{Tar}}} \sum_{i=1}^{n_{\mathrm{Tar}}} \ell(g, Z_i)$. However, when n_{Tar} is not sufficiently large, $\hat{R}(g)$ may not accurately estimate $R(g)$, resulting in a high generalization error of $\hat{g}$. To compensate for the scarcity of data from the target distribution, let us assume that we have data from K distinct *source distributions* $\{p_k\}_{k=1}^K$ over $\mathcal{Z}$, that is, we have independent i.i.d. samples $\mathcal{D}_k := \{Z_{k,i}^{\mathrm{Src}}\}_{i=1}^{n_k} \overset{\text{i.i.d.}}{\sim} p_k (k \in [K], n_k \in N)$ whose relations to p_{Tar} are described shortly. We assume $n_{\mathrm{Tar}}, n_k \geq D$ for simplicity.

Key assumption. In this work, the key transfer assumption is that all domains follow nonlinear ICA models with identical mixing functions (Figure 2). To be precise, we assume that there exists a set of IC distributions $q_{\mathrm{Tar}}, q_k \in \mathcal{Q} (k \in [K])$, and a smooth invertible function $f : \mathbb{R}^D \to \mathbb{R}^D$ (the *transformation* or *mixing*) such that $Z_{k,i}^{\mathrm{Src}} \sim p_k$ is generated by first sampling $S_{k,i}^{\mathrm{Src}} \sim q_k$ and

later transforming it by

$$Z^{\mathrm{Src}}_{k,i} = f(S^{\mathrm{Src}}_{k,i}), \tag{1}$$

and similarly $Z_i = f(S_i), S_i \sim q_{\mathrm{Tar}}$ for p_{Tar}. The above assumption allows us to formally relate p_k and p_{Tar}. It also allows us to estimate f when sufficient identification conditions required by the theory of nonlinear ICA are met. Due to space limitation, we provide a brief review of the nonlinear ICA method used in this paper and the known theoretical conditions in Supplementary Material A. The requirement for multiple source domains comes from the currently known identification condition of nonlinear ICA. Note that complex changes in q are allowed, hence the assumption of invariant f can accommodate intricate shifts in the apparent distribution p. We discuss this further in Section 5.3 by taking a simple example.

Example: Structural equation models A salient example of generative models expressed as Eq. (1) is *structural equation models* (SEMs; Pearl, 2009; Peters et al., 2017), which are used to describe the data generating mechanism involving the causality of random variables (Pearl, 2009). More precisely, the generative model of Eq.(1) corresponds to the *reduced form* (Reiss & Wolak, 2007) of a *Markovian* SEM (Pearl, 2009), i.e., a form where the structural equations to determine Z from (Z, S) are solved so that Z is expressed as a function of S. Such a conversion is always possible because a Markovian SEM induces an *acyclic* causal graph (Pearl, 2009), hence the structural equations can be solved by elimination of variables. This interpretation of reduced-form SEMs as Eq.(1) has been exploited in methods of *causal discovery*, e.g., in the linear non-Gaussian additive-noise models and their successors (Kano & Shimizu, 2003; Shimizu et al., 2006; Monti et al., 2019). In the case of SEMs, the key assumption of this paper translates into the invariance of the structural equations across domains, which enables an intuitive assessment of the assumption based on prior knowledge. For instance, if all domains have the same causal mechanism and are in the same intervention state (including an intervention-free case), the modeling choice is deemed plausible. Note that we do not estimate the original structural equations in the proposed method (Section˜3) but we only require estimating the reduced form, which is an easier problem compared to causal discovery.

3. Proposed Method: Mechanism Transfer

In this section, we detail the proposed method, mechanism transfer (Algorithm 1). The method first estimates the common generative mechanism f from the source domain data and then uses it to perform data augmentation of the target domain data to transfer the knowledge (Figure 3).

Algorithm 1 Proposed method: mechanism transfer

Input: Source domain data sets $\{\mathcal{D}_k\}_{k\in[K]}$, target domain data set $\mathcal{D}_{\mathrm{Tar}}$, nonlinear ICA algorithm ICA, and a learning algorithm $\mathcal{A}_{\mathcal{G}}$ to fit the hypothesis class $\mathcal{G}$ of predictors.

// Step 1. Estimate the shared transformation.

$\hat{f} \leftarrow \mathrm{ICA}(\mathcal{D}_1, \ldots, \mathcal{D}_K)$

// Step 2. Extract and shuffle target independent components

$\hat{s}_i \leftarrow \hat{f}^{-1}(Z_i), \quad (i = 1, \ldots, n_{\mathrm{Tar}})$

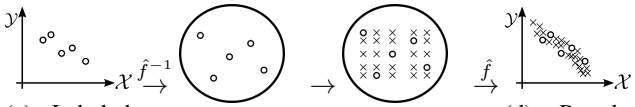
$\{\bar{s}_{\boldsymbol{i}}\}_{\boldsymbol{i}\in[n_{\mathrm{Tar}}]^D} \leftarrow \mathrm{AllCombinations}(\{\hat{s}_i\}_{i=1}^{n_{\mathrm{Tar}}})$

// Step 3. Synthesize target data and fit the predictor.

$\bar{z}_{\boldsymbol{i}} \leftarrow \hat{f}(\bar{s}_{\boldsymbol{i}})$

$\check{g} \leftarrow \mathcal{A}_{\mathcal{G}}(\{\bar{z}_{\boldsymbol{i}}\}_{\boldsymbol{i}})$

Output: $\check{g}$: the predictor in the target domain.



(a) Labeled target data (b) Find IC (c) Shuffle (d) Pseudo target data

Figure 3: Schematic illustration of proposed few-shot domain adaptation method after estimating the common mechanism f. With the estimated $\hat{f}$, the method augments the small target domain sample in a few steps to enhance statistical efficiency: (a) The algorithm is given labeled target domain data. (b) From labeled target domain data, extract the ICs. (c) By shuffling the values, synthesize likely values of IC. (d) From the synthesized IC, generate pseudo target data. The generated data is used to fit a predictor for the target domain.

3.1. Step 1: Estimate f using the source domain data

The first step estimates the common transformation f by nonlinear ICA, namely via *generalized contrastive learning* (GCL; Hyvärinen et al., 2019). GCL uses auxiliary information for training a certain binary classification function, $r_{\hat{f},\boldsymbol{\psi}}$, equipped with a parametrized feature extractor $\hat{f} : \mathbb{R}^D \to \mathbb{R}^D$. The trained feature extractor $\hat{f}$ is used as an estimator of f. The auxiliary information we use in our problem setup is the domain indices $[K]$. The classification function to be trained in GCL is $r_{\hat{f},\boldsymbol{\psi}}(z, u) := \sum_{d=1}^{D} \psi_d(\hat{f}^{-1}(z)_d, u)$ consisting of $(\hat{f}, \{\psi_d\}_{d=1}^D)$, and the classification task of GCL is logistic regression to classify (Z^{Src}_k, k) as positive and $(Z^{\mathrm{Src}}_k, k')(k' \neq k)$ as negative. This yields the following domain-contrastive learning criterion to estimate f:

$$\underset{\substack{\hat{f}\in\mathcal{F},\\ \{\psi_d\}_{d=1}^D\subset\Psi}}{\mathrm{argmin}} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \Bigg(\phi\left(r_{\hat{f},\boldsymbol{\psi}}(Z^{\mathrm{Src}}_{k,i}, k)\right) + \mathbb{E}_{k'\neq k}\phi\left(-r_{\hat{f},\boldsymbol{\psi}}(Z^{\mathrm{Src}}_{k,i}, k')\right)\Bigg),$$

where $\mathcal{F}$ and Ψ are sets of parametrized functions, $\mathbb{E}_{k' \neq k}$ denotes the expectation with respect to $k' \sim \mathrm{U}([K] \setminus \{k\})$ (U denotes the uniform distribution), and ϕ is the logistic loss $\phi(m) := \log(1 + \exp(-m))$. We use the solution $\hat{f}$ as an estimator of f. In experiments, $\mathcal{F}$ is implemented by invertible neural networks (Kingma & Dhariwal, 2018), Ψ by multi-layer perceptron, and $\mathbb{E}_{k' \neq k}$ is replaced by a random sampling renewed for every mini-batch.

3.2. Step 2: Extract and inflate the target ICs using $\hat{f}$

The second step extracts and inflates the target domain ICs using the estimated $\hat{f}$. We first extract the ICs of the target domain data by applying the inverse of $\hat{f}$ as

$$\hat{s}_i = \hat{f}^{-1}(Z_i).$$

After the extraction, we inflate the set of IC values by taking all dimension-wise combinations of the estimated IC:

$$\bar{s}_{\boldsymbol{i}} = (\hat{s}_{i_1}^{(1)}, \ldots, \hat{s}_{i_D}^{(D)}), \quad \boldsymbol{i} = (i_1, \ldots, i_D) \in [n_{\mathrm{Tar}}]^D,$$

to obtain new plausible IC values $\bar{s}_{\boldsymbol{i}}$. The intuitive motivation of this procedure stems from the independence of the IC distributions. Theoretical justifications are provided in Section 4. In our implementation, we use invertible neural networks (Kingma & Dhariwal, 2018) to model the function $\hat{f}$ to enable the computation of the inverse $\hat{f}^{-1}$.

3.3. Step 3: Synthesize target data from the inflated ICs

The third step estimates the target risk R by the empirical distribution of the augmented data:

$$\check{R}(g) := \frac{1}{n_{\mathrm{Tar}}^D} \sum_{\boldsymbol{i} \in [n_{\mathrm{Tar}}]^D} \ell(g, \hat{f}(\bar{s}_{\boldsymbol{i}})), \tag{2}$$

and performs empirical risk minimization. In experiments, we use a regularization term $\Omega(\cdot)$ to control the complexity of $\mathcal{G}$ and select

$$\check{g} \in \operatorname*{argmin}_{g \in \mathcal{G}} \left\{ \check{R}(g) + \Omega(g) \right\}.$$

The generated hypothesis $\check{g}$ is then used to make predictions in the target domain. In our experiments, we use $\Omega(g) = \lambda \|g\|^2$, where $\lambda > 0$ and the norm is that of the reproducing kernel Hilbert space (RKHS) which we take the subset $\mathcal{G}$ from. Note that we may well subsample only a subset of combinations in Eq. (2) to mitigate the computational cost similarly to Clémençon et al. (2016) and Papa et al. (2015).

4. Theoretical Insights

In this section, we state two theorems to investigate the statistical properties of the method proposed in Section 3 and provide plausibility beyond the intuition that we take advantage of the independence of the IC distributions.

4.1. Minimum variance property: Idealized case

The first theorem provides an insight into the statistical advantage of the proposed method: in the ideal case, the method attains the minimum variance among all possible unbiased risk estimators.

Theorem 1 (Minimum variance property of $\check{R}$)**.** *Assume that $\hat{f} = f$. Then, for each $g \in \mathcal{G}$, the proposed risk estimator $\check{R}(g)$ is the uniformly minimum variance unbiased estimator of $R(g)$, i.e., for any unbiased estimator $\tilde{R}(g)$ of $R(g)$,*

$$\forall q \in \mathcal{Q}, \quad \mathrm{Var}(\check{R}(g)) \leq \mathrm{Var}(\tilde{R}(g))$$

as well as $\mathbb{E}_{p_{\mathrm{Tar}}} \check{R}(g) = R(g)$ holds.

The proof of Theorem 1 is immediate once we rewrite $R(g)$ as a D-variate regular statistical functional and $\check{R}(g)$ as its corresponding generalized U-statistic (Lee, 1990). Details can be found in Supplementary Material D. Theorem 1 implies that the proposed risk estimator can have superior statistical efficiency in terms of the variance over the ordinary empirical risk.

4.2. Excess risk bound: More realistic case

In real situations, one has to estimate f. The following theorem characterizes the statistical gain and loss arising from the estimation error $f - \hat{f}$. The intuition is that the increased number of points suppresses the possibility of overfitting because the hypothesis has to fit the majority of the inflated data, but the estimator $\hat{f}$ has to be accurate so that fitting to the inflated data is meaningful. Note that the theorem is agnostic to how $\hat{f}$ is obtained, hence it applies to more general problem setup as long as f can be estimated.

Theorem 2 (Excess risk bound)**.** *Let $\check{g}$ be a minimizer of Eq. (2). Under appropriate assumptions (see Theorem 3 in Supplementary Material), for arbitrary $\delta, \delta' \in (0, 1)$, we have with probability at least $1 - (\delta + \delta')$,*

$$\begin{aligned} &R(\check{g}) - R(g^*) \\ &\leq \underbrace{C \sum_{j=1}^{D} \left\| f_j - \hat{f}_j \right\|_{W^{1,1}}}_{\textit{Approximation error}} + \underbrace{4D\mathfrak{R}(\mathcal{G}) + 2DB_\ell \sqrt{\frac{\log 2/\delta}{2n}}}_{\textit{Estimation error}} \\ &\quad + \underbrace{\kappa_1(\delta', n) + DB_\ell B_q \kappa_2 (f - \hat{f})}_{\textit{Higher order terms}}. \end{aligned}$$

Here, $\|\cdot\|_{W^{1,1}}$ is the $(1,1)$-Sobolev norm, and we define the effective Rademacher complexity $\mathfrak{R}(\mathcal{G})$ by

$$\frac{1}{n} \mathbb{E}_{\hat{S}} \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \left| \sum_{i=1}^{n} \sigma_i \mathbb{E}_{S'_2, \ldots, S'_D} [\tilde{\ell}(\hat{s}_i, S'_2, \ldots, S'_D)] \right| \right], \tag{3}$$

where $\{\sigma_i\}_{i=1}^n$ are independent sign variables, $\mathbb{E}_{\hat{S}}$ is the expectation with respect to $\{\hat{s}_i\}_{i=1}^{n_{\mathrm{Tar}}}$, the dummy variables $S'_2, \ldots, S'_D$ are i.i.d. copies of $\hat{s}_1$, and $\tilde{\ell}$ is defined by using the degree-D symmetric group $\mathfrak{S}_D$ as

$$\tilde{\ell}(s_1, \ldots, s_D) := \frac{1}{D!} \sum_{\pi \in \mathfrak{S}_D} \ell(g, \hat{f}(s^{(1)}_{\pi(1)}, \ldots, s^{(D)}_{\pi(D)})),$$

and $\kappa_1(\delta', n)$ and $\kappa_2(f - \hat{f})$ are higher order terms. The constants B_q and B_ℓ depend only on q and ℓ, respectively, while C' depends only on f, q, ℓ, and D.

Details of the statement and the proof can be found in Supplementary Material C. The Sobolev norm (Adams & Fournier, 2003) emerges from the evaluation of the difference between the estimated IC distribution and the ground-truth IC distribution. In Theorem 2, the utility of the proposed method appears in the effective complexity measure. The complexity is defined by a set of functions which are marginalized over all but one argument, resulting in mitigated dependence on the input dimensionality from exponential to linear (Supplementary Material C, Remark 3).

5. Related Work and Discussion

In this section, we review some existing TAs for DA to clarify the relative position of the paper. We also clarify the relation to the literature of causality-related transfer learning.

5.1. Existing transfer assumptions

Here, we review some of the existing work and TAs. See Table 1 for a summary.

(1) Parametric assumptions. Some TAs assume parametric distribution families, e.g., Gaussian mixture model in covariate shift (Storkey & Sugiyama, 2007). Some others assume parametric distribution shift, i.e., parametric representations of the target distribution given the source distributions. Examples include location-scale transform of class conditionals (Zhang et al., 2013; Gong et al., 2016), linearly dependent class conditionals (Zhang et al., 2015), and low-dimensional representation of the class conditionals after kernel embedding (Stojanov et al., 2019). In some applications, e.g., remote sensing, some parametric assumptions have proven useful (Zhang et al., 2013).

(2) Invariant conditionals and marginals. Some methods assume invariance of certain conditionals or marginals (Qui, 2009), e.g., $p(Y|X)$ in the covariate shift scenario (Shimodaira, 2000), $p(Y|\mathcal{T}(X))$ for an appropriate feature transformation $\mathcal{T}$ in transfer component analysis (Pan et al., 2011), $p(Y|\mathcal{T}(X))$ for a feature selector $\mathcal{T}$ (Rojas-Carulla et al., 2018; Magliacane et al., 2018), $p(X|Y)$ in the target shift (TarS) scenario (Zhang et al., 2013; Nguyen et al., 2016), and few components of regular-vine copulas and marginals in Lopez-paz et al. (2012). For example, the covariate shift scenario has been shown to fit well to brain computer interface data (Sugiyama et al., 2007).

Table 1: Comparison of TAs for DA (*Parametric*: parametric distribution family or distribution shift, *Invariant dist.*: invariant distribution components such as conditionals, marginals, or copulas. *Disc.* / *IPM*: small discrepancy or integral probability metric, *Param-transfer*: existence of transferable parameter, *Mechanism*: invariant mechanism). AD: adaptation among Apparently Different distributions is accommodated. NP: Non-Parametrically flexible. BCI: Brain computer interface. The numbers indicate the paragraphs of Section 5.1.

TA	AD	NP	Suited app. example
(1) Parametric	✓	-	Remote sensing
(2) Invariant dist.	-	✓	BCI
(3) Disc. / IPM	-	✓	Computer vision
(4) Param-transfer	✓	✓	Computer vision
(Ours) Mechanism	✓	✓	Medical records

(3) Small discrepancy or integral probability metric. Another line of work relies on certain distributional similarities, e.g., integral probability metric (Courty et al., 2017) or hypothesis-class dependent discrepancies (Ben-David et al., 2007; Blitzer et al., 2008; Ben-David et al., 2010; Kuroki et al., 2019; Zhang et al., 2019; Cortes et al., 2019). These methods assume the existence of the *ideal joint hypothesis* (Ben-David et al., 2010), corresponding to a relaxation of the covariate shift assumption. These TA are suited for unsupervised or semi-supervised DA in computer vision applications (Courty et al., 2017).

(4) Transferable parameter. Some others consider parameter transfer (Kumagai, 2016), where the TA is the existence of a parameterized feature extractor that performs well in the target domain for linear-in-parameter hypotheses and its learnability from the source domain data. For example, such a TA has been known to be useful in natural language processing or image recognition (Lee et al., 2009; Kumagai, 2016).

5.2. Causality for transfer learning

Our method can be seen as the first attempt to fully leverage structural causal models for DA. Most of the causality-inspired DA methods express their assumptions in the level of *graphical causal models* (GCMs), which only has much coarser information than *structural causal models* (SCMs) (Peters et al., 2017, Table 1.1) exploited in this paper. Compared to previous work, our method takes one step further

to assume and exploit the invariance of SCMs. Specifically, many studies assume the GCM $X \leftarrow Y$ (the *anticausal* scenario) following the seminal meta-analysis of Schölkopf et al. (2012) and use it to motivate their parametric distribution shift assumptions or the parameter estimation procedure (Zhang et al., 2013; 2015; Gong et al., 2016; 2018). Although such assumptions on the GCM have the virtue of being more robust to misspecification, they tend to require parametric assumptions to obtain theoretical justifications. On the other hand, our assumption enjoys a theoretical guarantee without relying on parametric assumptions.

One notable work in the existing literature is Magliacane et al. (2018) that considered the domain adaptation among *different intervention states*, a problem setup that complements ours that considers an intervention-free (or identical intervention across domains) case. To model intervention states, Magliacane et al. (2018) also formulated the problem setup using SCMs, similarly to the present paper. Therefore, we clarify a few key differences between Magliacane et al. (2018) and our work here. In terms of the methodology, Magliacane et al. (2018) takes a variable selection approach to select a set of predictor variables with an invariant conditional distribution across different intervention states. On the other hand, our method estimates the SEMs (in the reduced form) and applies a data augmentation procedure to transfer the knowledge. To the best of our knowledge, the present paper is the first to propose a way to directly use the estimated SEMs for domain adaptation, and the fine-grained use of the estimated SEMs enables us to derive an excess risk bound. In terms of the plausible applications, their problem setup may be more suitable for application fields with interventional experiments such as genomics, whereas ours may be more suited for fields where observational studies are more common such as health record analysis or economics. In Appendix E, we provide a more detailed comparison.

5.3. Plausibility of the assumptions

Checking the validity of the assumption. As is often the case in DA, the scarcity of data disables data-driven testing of the TAs, and we need domain knowledge to judge the validity. For our TA, the intuitive interpretation as invariance of causal models (Section˜2) can be used.

Invariant causal mechanisms. The invariance of causal mechanisms has been exploited in recent work of causal discovery such as Xu et al. (2014) and Monti et al. (2019), or under the name of the *multi-environment setting* in Ghassami et al. (2017). Moreover, the SEMs are normally assumed to remain invariant unless explicitly intervened in (Hünermund & Bareinboim, 2019). However, the invariance assumption presumes that the intervention states do not vary across domains (allowing for the intervention-free case), which can be limiting for some applications where different interventions are likely to be present, e.g., different treatment policies being put in place in different hospitals. Nevertheless, the present work can already be of practical interest if it is combined with the effort to find suitable data or situations. For instance, one may find medical records in group hospitals where the same treatment criteria is put in place or local surveys in the same district enforcing identical regulations. In future work, relaxing the requirement to facilitate the data-gathering process is an important area. For such future extensions, the present theoretical analyses can also serve as a landmark to establish what can be guaranteed in the basic case without mechanism alterations.

Fully observed variables. As the first algorithm in the approach to fully exploit SCMs for DA, we also consider the case where all variables are observable. Although it is often assumed in a causal inference problem that there are some unobserved confounding variables, we leave further extension to such a case for future work.

Required number of source domains. A potential drawback of the proposed method is that it requires a number of source domains in order to satisfy the identification condition of the nonlinear ICA, namely GCL in this paper (Supplementary Material A). The requirement solely arises from the identification condition of the ICA method and therefore has the possibility to be made less stringent by the future development of nonlinear ICA methods. Moreover, if one can accept other identification conditions, one-sample ICA methods (e.g., linear ICA) can be also used in the proposed approach in a straightforward manner, and our theoretical analyses still hold regardless of the method chosen.

Flexibility of the model. The relation between X and Y can drastically change while f is invariant. For example, even in a simple additive noise model $(X, Y) = f(S_1, S_2) = (S_1, S_1 + S_2)$, the conditional $p(Y|X)$ can shift drastically if the distribution of the independent noise S_2 changes in a complex manner, e.g., becoming multimodal from unimodal.

6. Experiment

In this section, we provide proof-of-concept experiments to demonstrate the effectiveness of the proposed approach. Note that the primary purpose of the experiments is to confirm whether the proposed method can properly perform DA in real-world data, and it is not to determine which DA method and TA are the most suited for the specific dataset.

6.1. Implementation details of the proposed method

Estimation of f (Step 1). We model $\hat{f}$ by an 8-layer Glow neural network (Supplementary Material B.2). We model ψ_d by a 1-hidden-layer neural network with a varied number of hidden units, K output units, and the rectified linear unit activation (LeCun et al., 2015). We use its k-th output ($k \in [K]$) as the value for $\psi_d(\cdot, k)$. For training, we use the Adam optimizer (Kingma & Ba, 2017) with fixed parameters $(\beta_1, \beta_2, \epsilon) = (0.9, 0.999, 10^{-8})$, fixed initial learning rate 10^{-3}, and the maximum number of epochs 300. The other fixed hyperparameters of $\hat{f}$ and its training process are described in Supplementary Material B.

Augmentation of target data (Step 3). For each evaluation step, we take all combinations (with replacement) of the estimated ICs to synthesize target domain data. After we synthesize the data, we filter them by applying a novelty detection technique with respect to the union of source domain data. Namely, we use one-class support vector machine (Schölkopf et al., 2000) with the fixed parameter $\nu = 0.1$ and radial basis function (RBF) kernel $k(x, y) = \exp(-\|x - y\|^2/\gamma)$ with $\gamma = D$. This is because the estimated transform $\hat{f}$ is not expected to be trained well outside the union of the supports of the source distributions. After performing the filtration, we combined the original target training data with the augmented data to ensure the original data points to be always included.

Predictor hypothesis class $\mathcal{G}$. As the predictor model, we use the kernel ridge regression (KRR) with RBF kernel. The bandwidth γ is chosen by the median heuristic similarly to Yamada et al. (2011) for simplicity. Note that the choice of the predictor model is for the sake of comparison with the other methods tailored for KRR (Cortes et al., 2019), and that an arbitrary predictor hypothesis class and learning algorithm can be easily combined with the proposed approach.

Hyperparameter selection. We perform grid-search for hyperparameter selection. The number of hidden units for ψ_d is chosen from $\{10, 20\}$ and the coefficient of weight-decay from $10^{\{-2,-1\}}$. The ℓ^2 regularization coefficient λ of KRR is chosen from $\lambda \in 2^{\{-10,\dots,10\}}$ following Cortes et al. (2019). To perform hyperparameter selection as well as early-stopping, we record the leave-one-out cross-validation (LOOCV) mean-squared error on the target training data every 20 epochs and select its minimizer. The leave-one-out score is computed using the well-known analytic formula instead of training the predictor for each split. Note that we only use the original target domain data and not the synthesized data as the held-out set. In practice, if the target domain data is extremely few, one may well use *percentile-cv* (Ng, 1997) to mitigate overfitting of hyperparameter selection.

Computation environment All experiments were conducted on an Intel Xeon(R) 2.60 GHz CPU with 132 GB memory. They were implemented in Python using the PyTorch library (Paszke et al., 2019) or the R language (R Core Team, 2018).

6.2. Experiment using real-world data

Dataset. We use the gasoline consumption data (Greene, 2012, p.284, Example 9.5), which is a panel data of gasoline usage in 18 of the OECD countries over 19 years. We consider each country as a domain, and we disregard the time-series structure and consider the data as i.i.d. samples for each country in this proof-of-concept experiment. The dataset contains four variables, all of which are log-transformed: motor gasoline consumption per car (the predicted variable), per-capita income, motor gasoline price, and the stock of cars per capita (the predictor variables) (Baltagi & Griffin, 1983). For further details of the data, see Supplementary Material B. We used the dataset because there are very few public datasets for domain adapting regression tasks (Cortes & Mohri, 2014) especially for multi-source DA, and also because the dataset has been used in econometric analyses involving SEMs (Baltagi, 2005), conforming to our approach.

Compared methods. We compare the following DA methods, all of which apply to regression problems. Unless explicitly specified, the predictor class $\mathcal{G}$ is chosen to be KRR with the same hyperparameter candidates as the proposed method (Section 6.1). Further details are described in Supplementary Material B.5.

- Naive baselines (*SrcOnly*, *TarOnly*, and *S&TV*): *SrcOnly* (resp. *TarOnly*) trains a predictor on the source domain data (resp. target training data) without any device. *SrcOnly* can be effective if the source domains and the target domain have highly similar distributions. The *S&TV* baseline trains on both source and target domain data, but the LOOCV score is computed only from the target domain data.

- *TrAdaBoost*: Two-stage TrAdaBoost.R2; a boosting method tailored for few-shot regression transfer proposed in Pardoe & Stone (2010). It is an iterative method with early-stopping (Pardoe & Stone, 2010), for which we use the leave-one-out cross-validation score on the target domain data as the criterion. As suggested in Pardoe & Stone (2010), we set the maximum number of outer loop iterations at 30. The base predictor is the decision tree regressor with the maximum depth 6 (Hastie et al., 2009). Note that although TrAdaBoost does not have a clarified transfer assumption, we compare the performance for reference.

- *IW*: Importance weighted KRR using RuLSIF (Yamada et al., 2011). The method directly estimates a relative joint density ratio function $\frac{p_{\mathrm{Src}}(z)}{\alpha p_{\mathrm{Src}}(z)+(1-\alpha)p_{\mathrm{Tar}}(z)}$ for $\alpha \in [0, 1)$, where p_{Src} is a hypothetical source distribution created by pooling all source domain data. Following Yamada et al. (2011), we experiment on $\alpha \in \{0, 0.5, 0.95\}$ and report the results separately. The regularization coefficient λ' is selected from $\lambda' \in 2^{\{-10,\dots,10\}}$ using importance-weighted cross-validation (Sugiyama et al., 2007).

- *GDM*: Generalized discrepancy minimization (Cortes et al., 2019). This method performs instance-weighted training on the source domain data with the weights that minimize the *generalized discrepancy* (via quadratic programming). We select the hyper-parameters λ_r from $2^{\{-10,\dots,10\}}$ as suggested by Cortes et al. (2019). The selection criterion is the performance of the trained predictor on the target training labels as the method trains on the source domain data and the target unlabeled data.

- *Copula*: Non-parametric regular-vine copula method (Lopez-paz et al., 2012). This method presumes using a specific joint density estimator called regular-vine (R-vine) copulas. Adaptation is realized in two steps: the first step estimates which components of the constructed R-vine model are different by performing two-sample tests based on maximum mean discrepancy (Lopez-paz et al., 2012), and the second step re-estimates the components in which a change is detected using only the target domain data.

- *LOO* (reference score): Leave-one-out cross-validated error estimate is also calculated for reference. It is the average prediction error of predicting for a single held-out test point when the predictor is trained on the rest of the whole target domain data including those in the test set for the other algorithms.

Evaluation procedure. The prediction accuracy was measured by the mean squared error (MSE). For each train-test split, we randomly select one-third (6 points) of the target domain dataset as the training set and use the rest as the test set. All experiments were repeated 10 times with different train-test splits of target domain data.

Results. The results are reported in Table 2. We report the MSE scores normalized by that of *LOO* to facilitate the comparison, similarly to Cortes & Mohri (2014). In many of the target domain choices, the naive baselines (*SrcOnly* and *S&TV*) suffer from negative transfer, i.e., higher average MSE than *TarOnly* (in 12 out of 18 domains). On the other hand, the proposed method successfully performs better than *TarOnly* or is more resistant to negative transfer than the other compared methods. The performances of *GDM*, *Copula*, and *IW* are often inferior even compared to the baseline performance of *SrcAndTarValid*. For *GDM* and *IW*, this can be attributed to the fact that these methods presume the availability of abundant (unlabeled) target domain data, which is unavailable in the current problem setup. For *Copula*, the performance inferior to the naive baselines is possibly due to the restriction of the predictor model to its accompanied probability model (Lopez-paz et al., 2012). *TrAdaBoost* works reasonably well for many but not all domains. For some domains, it suffered from negative transfer similarly to others, possibly because of the very small number of training data points. Note that the transfer assumption of TrAdaBoost has not been stated (Pardoe & Stone, 2010), and it is not understood when the method is reliable. The domains on which the baselines perform better than the proposed method can be explained by the following two cases: (1) easier domains allow naive baselines to perform well and (2) some domains may have deviated f. Case (1) implies that estimating f is unnecessary, hence the proposed method can be suboptimal (more likely for JPN, NLD, NOR, and SWE, where SrcOnly or S&TV improve upon TrgOnly). On the other hand, case (2) implies that an approximation error was induced as in Theorem 2 (more likely for IRL and ITA). In this case, others also perform poorly, implying the difficulty of the problem instance. In either case, in practice, one may well perform cross-validation to fallback to the baselines.

7. Conclusion

In this paper, we proposed a novel few-shot supervised DA method for regression problems based on the assumption of shared generative mechanism. Through theoretical and experimental analysis, we demonstrated the effectiveness of the proposed approach. By considering the latent common structure behind the domain distributions, the proposed method successfully induces positive transfer even when a naive usage of the source domain data can suffer from negative transfer. Our future work includes making an experimental comparison with extensively more datasets and methods as well as an extension to the case where the underlying mechanism are not exactly identical but similar.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful comments and thorough discussions. We would also like to thank Yuko Kuroki and Taira Tsuchiya for proofreading the manuscript. This work was supported by RIKEN Junior Research Associate Program. TT was supported by Masason Foundation. IS was supported by KAKEN 17H04693. MS was supported by JST CREST Grant Number JPMJCR18A2.

Table 2: Results of the real-world data experiments for different choices of the target domain. The evaluation score is MSE normalized by that of *LOO* (the lower the better). All experiments were repeated 10 times with different train-test splits of target domain data, and the average performance is reported with the standard errors in the brackets. The target column indicates abbreviated country names. Bold-face indicates the best score (Prop: proposed method, TrAda: *TrAdaBoost*, the numbers in the brackets of IW indicate the value of α). The proposed method often improves upon the baseline *TarOnly* or is relatively more resistant to negative transfer, with notable improvements in *DEU*, *GBR*, and *USA*.

Target	(LOO)	TarOnly	**Prop**	SrcOnly	S&TV	TrAda	GDM	Copula	IW(.0)	IW(.5)	IW(.95)
AUT	1	5.88 (1.60)	**5.39 (1.86)**	9.67 (0.57)	9.84 (0.62)	5.78 (2.15)	31.56 (1.39)	27.33 (0.77)	39.72 (0.74)	39.45 (0.72)	39.18 (0.76)
BEL	1	10.70 (7.50)	**7.94 (2.19)**	8.19 (0.68)	9.48 (0.91)	8.10 (1.88)	89.10 (4.12)	119.86 (2.64)	105.15 (2.96)	105.28 (2.95)	104.30 (2.95)
CAN	1	5.16 (1.36)	**3.84 (0.98)**	157.74 (8.83)	156.65 (10.69)	51.94 (30.06)	516.90 (4.45)	406.91 (1.59)	592.21 (1.87)	591.21 (1.84)	589.87 (1.91)
DNK	1	3.26 (0.61)	**3.23 (0.63)**	30.79 (0.93)	28.12 (1.67)	25.60 (13.11)	16.84 (0.85)	14.46 (0.79)	22.15 (1.10)	22.11 (1.10)	21.72 (1.07)
FRA	1	2.79 (1.10)	**1.92 (0.66)**	4.67 (0.41)	3.05 (0.11)	52.65 (25.83)	91.69 (1.34)	156.29 (1.96)	116.32 (1.27)	116.54 (1.25)	115.29 (1.28)
DEU	1	16.99 (8.04)	**6.71 (1.23)**	229.65 (9.13)	210.59 (14.99)	341.03 (157.80)	739.29 (11.81)	929.03 (4.85)	817.50 (4.60)	818.13 (4.55)	812.60 (4.57)
GRC	1	3.80 (2.21)	**3.55 (1.79)**	5.30 (0.90)	5.75 (0.68)	11.78 (2.36)	26.90 (1.89)	23.05 (0.53)	47.07 (1.92)	45.50 (1.82)	45.72 (2.00)
IRL	1	**3.05 (0.34)**	4.35 (1.25)	135.57 (5.64)	12.34 (0.58)	23.40 (17.50)	3.84 (0.22)	26.60 (0.59)	6.38 (0.13)	6.31 (0.14)	6.16 (0.13)
ITA	1	**13.00 (4.15)**	14.05 (4.81)	35.29 (1.83)	39.27 (2.52)	87.34 (24.05)	226.95 (11.14)	343.10 (10.04)	244.25 (8.50)	244.84 (8.58)	242.60 (8.46)
JPN	1	10.55 (4.67)	12.32 (4.95)	**8.10 (1.05)**	8.38 (1.07)	18.81 (4.59)	95.58 (7.89)	71.02 (5.08)	135.24 (13.57)	134.89 (13.50)	134.16 (13.43)
NLD	1	3.75 (0.80)	3.87 (0.79)	**0.99 (0.06)**	0.99 (0.05)	9.45 (1.43)	28.35 (1.62)	29.53 (1.58)	33.28 (1.78)	33.23 (1.77)	33.14 (1.77)
NOR	1	2.70 (0.51)	2.82 (0.73)	1.86 (0.29)	**1.63 (0.11)**	24.25 (12.50)	23.36 (0.88)	31.37 (1.17)	27.86 (0.94)	27.86 (0.93)	27.52 (0.91)
ESP	1	5.18 (1.05)	6.09 (1.53)	5.17 (1.14)	**4.29 (0.72)**	14.85 (4.20)	33.16 (6.99)	152.59 (6.19)	53.53 (2.47)	52.56 (2.42)	52.06 (2.40)
SWE	1	6.44 (2.66)	5.47 (2.63)	2.48 (0.23)	**2.02 (0.21)**	2.18 (0.25)	15.53 (2.59)	2706.85 (17.91)	118.46 (1.64)	118.23 (1.64)	118.27 (1.64)
CHE	1	3.51 (0.46)	**2.90 (0.37)**	43.59 (1.77)	7.48 (0.49)	38.32 (9.03)	8.43 (0.24)	29.71 (0.53)	9.72 (0.29)	9.71 (0.29)	9.79 (0.28)
TUR	1	1.65 (0.47)	1.06 (0.15)	1.22 (0.18)	**0.91 (0.09)**	2.19 (0.34)	64.26 (5.71)	142.84 (2.04)	159.79 (2.63)	157.89 (2.63)	157.13 (2.69)
GBR	1	5.95 (1.86)	**2.66 (0.57)**	15.92 (1.02)	10.05 (1.47)	7.57 (5.10)	50.04 (1.75)	68.70 (1.25)	70.98 (1.01)	70.87 (0.99)	69.72 (1.01)
USA	1	4.98 (1.96)	**1.60 (0.42)**	21.53 (3.30)	12.28 (2.52)	2.06 (0.47)	308.69 (5.20)	244.90 (1.82)	462.51 (2.14)	464.75 (2.08)	465.88 (2.16)
#Best	-	2	10	2	4	0	0	0	0	0	0

References

Dataset Shift in Machine Learning. Neural Information Processing Series. MIT Press, Cambridge, Mass, 2009.

Adams, R. A. and Fournier, J. J. *Sobolev Spaces*. Academic press, 2003.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv:1907.02893 [cs, stat]*, March 2020.

Baltagi, B. *Econometric Analysis of Panel Data*. New York: John Wiley and Sons, 3rd edition, 2005.

Baltagi, B. H. and Griffin, J. M. Gasoline demand in the OECD: An application of pooling and testing procedures. *European Economic Review*, 22(2):117–137, 1983.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems 19*, pp. 137–144. MIT Press, 2007.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Wortman, J. Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems 20*, pp. 129–136. Curran Associates, Inc., 2008.

Clémençon, S., Colin, I., and Bellet, A. Scaling-up empirical risk minimization: Optimization of incomplete U-statistics. *Journal of Machine Learning Research*, 17 (76):1–36, 2016.

Cortes, C. and Mohri, M. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

Cortes, C., Mohri, M., and Medina, A. M. Adaptation based on generalized discrepancy. *Journal of Machine Learning Research*, 20(1):1–30, 2019.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems 30*, pp. 3730–3739. Curran Associates, Inc., 2017.

Ghassami, A., Salehkaleybar, S., Kiyavash, N., and Zhang, K. Learning causal structures using regression invariance. In *Advances in Neural Information Processing Systems 30*, pp. 3011–3021. Curran Associates, Inc., 2017.

Golub, G. H. and Van Loan, C. F. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. The Johns Hopkins University Press, Baltimore, 4th edition, 2013.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., and Schölkopf, B. Domain adaptation with conditional transferable components. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2839–2848, New York, USA, 2016. PMLR.

Gong, M., Zhang, K., Huang, B., Glymour, C., Tao, D., and Batmanghelich, K. Causal generative domain adaptation networks. *arXiv:1804.04333 [cs, stat]*, April 2018.

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

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

Hayfield, T. and Racine, J. S. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 2008.

Hünermund, P. and Bareinboim, E. Causal inference and data-fusion in econometrics. *arXiv:1912.09104 [econ]*, December 2019.

Hyvärinen, A. and Morioka, H. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems 29*, pp. 3765–3773. Curran Associates, Inc., 2016.

Hyvärinen, A. and Morioka, H. Nonlinear ICA of temporally dependent stationary sources. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 460–469, 2017.

Hyvärinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

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

Ipsen, I. C. F. and Rehman, R. Perturbation bounds for determinants and characteristic polynomials. *SIAM Journal on Matrix Analysis and Applications*, 30(2):762–776, 2008.

Kano, Y. and Shimizu, S. Causal inference using nonnormality. In *Proceedings of the International Symposium on the Science of Modeling, the 30th Anniversary of the Information Criterion,*, pp. 261–270, 2003.

Khemakhem, I., Kingma, D. P., Monti, R. P., and Hyvärinen, A. Variational autoencoders and nonlinear ICA: A unifying framework. *arXiv:1907.04809 [cs, stat]*, July 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv:1412.6980 [cs]*, January 2017.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems 31*, pp. 10215–10224. Curran Associates, Inc., 2018.

Kumagai, W. Learning bound for parameter transfer learning. In *Advances in Neural Information Processing Systems 29*, pp. 2721–2729. Curran Associates, Inc., 2016.

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

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015.

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

Lee, H., Raina, R., Teichman, A., and Ng, A. Y. Exponential family sparse coding with applications to self-taught learning. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence*, pp. 1113–1119, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

Lopez-paz, D., Hernández-lobato, J. M., and Schölkopf, B. Semi-supervised domain adaptation with non-parametric copulas. In *Advances in Neural Information Processing Systems 25*, pp. 665–673. Curran Associates, Inc., 2012.

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems 31*, pp. 10846–10856. Curran Associates, Inc., 2018.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, MA, 2012.

Monti, R. P., Zhang, K., and Hyvärinen, A. Causal discovery with general non-linear relationships using non-linear ICA. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2019.

Ng, A. Y. Preventing "overfitting" of cross-validation data. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 245–253, San Francisco, CA, USA, 1997.

Nguyen, T. D., Christoffel, M., and Sugiyama, M. Continuous Target Shift Adaptation in Supervised Learning. In *Asian Conference on Machine Learning*, volume 45 of *Proceedings of Machine Learning Research*, pp. 285–300. PMLR, 2016.

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

Papa, G., Clémençon, S., and Bellet, A. SGD Algorithms based on Incomplete U-statistics: Large-Scale Minimization of Empirical Risk. In *Advances in Neural Information Processing Systems 28*, pp. 1027–1035. Curran Associates, Inc., 2015.

Pardoe, D. and Stone, P. Boosting for regression transfer. In *Proceedings of the Twenty-Seventh International Conference on Machine Learning*, pp. 863–870, Haifa, Israel, 2010.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

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

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

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018.

Reiss, P. C. and Wolak, F. A. Structural econometric modeling: Rationales and examples from industrial organization. In *Handbook of Econometrics*, volume 6, pp. 4277–4415. Elsevier, 2007.

Rejchel, W. On ranking and generalization bounds. *Journal of Machine Learning Research*, 13(May):1373–1392, 2012.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., and Platt, J. C. Support vector method for novelty detection. In *Advances in Neural Information Processing Systems 12*, pp. 582–588. MIT Press, 2000.

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

Sherman, R. P. Maximal inequalities for degenerate U-processes with applications to optimization estimators. *The Annals of Statistics*, 22(1):439–459, 1994.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. J. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7 (October):2003–2030, 2006.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Stojanov, P., Gong, M., Carbonell, J., and Zhang, K. Data-driven approach to multiple-source domain adaptation. In *Proceedings of Machine Learning Research*, volume 89, pp. 3487–3496. PMLR, 2019.

Storkey, A. J. and Sugiyama, M. Mixture regression for covariate shift. In *Advances in Neural Information Processing Systems 19*, pp. 1337–1344. MIT Press, 2007.

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May): 985–1005, 2007.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 1st edition, 2019.

Xu, L., Fan, T., Wu, X., Chen, K., Guo, X., Zhang, J., and Yao, L. A pooling-LiNGAM algorithm for effective connectivity analysis of fMRI data. *Frontiers in Computational Neuroscience*, 8(October):125, 2014.

Yadav, P., Steinbach, M., Kumar, V., and Simon, G. Mining electronic health records (EHRs): A survey. *ACM Computing Surveys*, 50(6):1–40, 2018.

Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., and Sugiyama, M. Relative density-ratio estimation for robust distribution comparison. In *Advances in Neural Information Processing Systems 24*, pp. 594–602. Curran Associates, Inc., 2011.

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

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

Zhang, Y., Liu, T., Long, M., and Jordan, M. Bridging theory and algorithm for domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 7404–7413, Long Beach, California, USA, 2019. PMLR.