Multi-Source Domain Adaptation: A Causal View

Kun Zhang
MPI for Intelligent Systems
72076, Tübingen, Germany
kzhang@tuebingen.mpg.de

Mingming Gong
QCIS, University of Technology Sydney
Ultimo, NSW 2007, Australia
gongmingnju@gmail.com

Bernhard Schölkopf
MPI for Intelligent Systems
72076, Tübingen, Germany
bs@tuebingen.mpg.de

Abstract

This paper is concerned with the problem of domain adaptation with multiple sources from a causal point of view. In particular, we use causal models to represent the relationship between the features X and class label Y, and consider possible situations where different modules of the causal model change with the domain. In each situation, we investigate what knowledge is appropriate to transfer and find the optimal target-domain hypothesis. This gives an intuitive interpretation of the assumptions underlying certain previous methods and motivates new ones. We finally focus on the case where Y is the cause for X with changing P_Y and $P_{X|Y}$, that is, P_Y and $P_{X|Y}$ change independently across domains. Under appropriate assumptions, the availability of multiple source domains allows a natural way to reconstruct the conditional distribution on the target domain; we propose to model $P_{X|Y}$ (the process to generate effect X from cause Y) on the target domain as a linear mixture of those on source domains, and estimate all involved parameters by matching the target-domain feature distribution. Experimental results on both synthetic and real-world data verify our theoretical results.

Traditional machine learning relies on the assumption that both training and test data are from the same distribution. In practice, however, training and test data are probably sampled under different conditions, thus violating this assumption, and the problem of domain adaptation (DA) arises. Consider remote sensing image classification as an example. Suppose we already have several data sets on which the class labels are known; they are called source domains here. For a new data set, or a target domain, it is usually difficult to find the ground truth reference labels, and we aim to determine the labels by making use of the information from the source domains. Note that those domains are usually obtained in different areas and time periods, and that the corresponding data distribution various due to the change in illumination conditions, physical factors related to ground (e.g., different soil moisture or composition), vegetation, and atmospheric conditions. Other well-known instances of this situation include sentiment data analysis (Blitzer, Dredze, and Pereira 2007) and flow cytometry data analysis (Blanchard, Lee, and Scott 2011). DA approaches have many applications in varies areas including natural language processing, computer vision, and biology. For surveys on DA, see, e.g., (Jiang 2008; Pan and Yang 2010; Candela et al. 2009).

In this paper, we consider the situation with n source domains on which both the features X and label Y are given, i.e., we are given $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) = (x_k^{(i)}, y_k^{(i)})_{k=1}^{m_i}$, where $i = 1, ..., n$, and m_i is the sample size of the ith source domain. Our goal is to find the classifier for the target domain, on which only the features $\mathbf{x}^t = (x_k^t)_{k=1}^m$ are available. Here we are concerned with a difficult scenario where no labeled point is available in the target domain, known as unsupervised domain adaptation. Since P_{XY} changes across domains, we have to find what knowledge in the source domains should be transferred to the target one. Previous work in domain adaptation has usually assumed that P_X changes but $P_{Y|X}$ remain the same, i.e., the covariate shift situation; see, e.g., (Shimodaira 2000; Huang et al. 2007; Sugiyama et al. 2008; Ben-David, Shalev-Shwartz, and Urner 2012). It is also known as sample selection bias (particularly on the features X) in (Zadrozny 2004).

In practice it is very often that both P_X and $P_{Y|X}$ change simultaneously across domains. For instance, both of them are likely to change over time and location for a satellite image classification system. If the data distribution changes arbitrarily across domains, clearly knowledge from the sources may not help in predicting Y on the target domain (Rosenstein et al. 2005). One has to find what type of information should be transferred from sources to the target. One possibility is to assume the change in both P_X and $P_{Y|X}$ is due to the change in P_Y, while $P_{X|Y}$ remains the same, as known as prior probability shift (Storkey 2009; Plessis and Sugiyama 2012) or target shift (Zhang et al. 2013). The latter further models the change in $P_{X|Y}$ caused by a location-scale (LS) transformation of the features for each class. The constraint of the LS transformation renders $P_{X|Y}$ on the target domain, denoted by $P^t_{X|Y}$, identifiable; however, it might be too restrictive.

Fortunately, the availability of multiple source domains provides more hints as to find $P^t_{X|Y}$, as well as $P^t_{Y|X}$. Several algorithms have been proposed to combine knowledge from multiple source domains. For instance, (Mansour, Mohri, and Rostamizadeh 2008) proposed to form the target hypothesis by combining source hypotheses with a distribu-

tion weighted rule. (Gao et al. 2008), (Duan et al. 2009), and (Chattopadhyay et al. 2011) combine the predictions made by the source hypotheses, with the weights determined in different ways.

An intuitive interpretation of the assumptions underlying those algorithms would facilitate choosing or developing DA methods for the problem at hand. To the best of our knowledge, however, it is still missing in the literature. One of our contributions in this paper is to provide such an interpretation. This paper studies the multi-source DA problem from a causal point of view where we consider the underlying data generating process behind the observed domains. We are particularly interested in what types of information stay the same, what types of information change, and how they change across domains. This enables us to construct the optimal hypothesis for the target domain in various situations. To this end, we use causal models to represent the relationship between X and Y, because they provide a compact description of the properties of the change in the data distribution.[1] They, for instance, help characterize transportability of experimental findings (Pearl and Bareinboim 2011) or recoverability from selection bias (Bareinboim, Tian, and Pearl 2014).

As another contribution, we further focus on a typical DA scenario where both P_Y and $P_{X|Y}$ (or the causal mechanism to generate effect X from cause Y) change across domains, but their changes are independent from each other, as implied by the causal model $Y \to X$. We assume that the source domains contains rich information such that for each class, $P^t_{X|Y}$ can be approximated by a linear mixture of $P_{X|Y}$ on source domains. Together with other mild conditions on $P_{X|Y}$, we then show that $P^t_{X|Y}$, as well as P^t_Y, is identifiable (or can be uniquely recovered). We present a computationally efficient method to estimate the involved parameters based on kernel mean distribution embedding (Smola et al. 2007; Gretton et al. 2007), followed by several approaches to constructing the target classifier using those parameters.

One might wonder how to find the causal information underlying the data to facilitate domain adaptation. We note that in practice, background causal knowledge is usually available, helping formulating how to transfer the knowledge from source domains to the target. Even if this is not the case, multiple source domains with different data distributions may allow one to identify the causal structure, since the causal knowledge can be seen from the change in data distributions; see e.g., (Tian and Pearl 2001).

1 Possible DA Situations and Their Solutions

DA can be considered as a learning problem in nonstationary environments (Sugiyama and Kawanabe 2012). It is helpful to find how the data distribution changes; it provides the clues as to find the learning machine for the target domain.

[1]The causal model also describes how the components of the joint distribution are related to each other, which, for instance, gives a causal explanation of the behavior of semi-supervised learning (Schölkopf et al. 2012).

Table 1: Notation used in this paper.

X, Y	random variables
$\mathcal{X}, \mathcal{Y}$	domains
$P^{(i)}_{XY}$	distribution in the ith source domain
P^t_{XY}	distribution in the target domain
$(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) = (x^{(i)}_k, y^{(i)}_k)^{m_i}_{k=1}$	sample in the ith source domain
$\mathbf{x}^{(i)}_j = (x^{(i)}_{jk})^{m_{ij}}_{k=1}$	X values with $Y = c_j$ in the ith source domain
$\mathbf{x}^t = (x^t_k)^m_{k=1}$	X values in the target domain
K^t	kernel matrix on $\mathbf{x}^t$
K^{it}	"cross" kernel matrix between $\mathbf{x}^{(i)}$ and $\mathbf{x}^t$
$\psi(X)$	feature map of X

We focus on how causality, which provides a compact and intuitive description about distribution changes, helps us in DA. Generally speaking, in the unconfounded case, the process that generates the effect from the cause does not depend on that generating the cause (Pearl 2000). We can represent such knowledge with graphical models, or selection diagrams defined in (Pearl and Bareinboim 2011). In particular, let us consider four situations which are often the case in practice; see Fig. 1. Here W_s and V_s are represent domain-specific selection variables, and they are hidden variables.[2]

Below we discuss what knowledge to transfer from source domains to target, and how to construct the optimal target-domain hypothesis in each situation. For clarity and simplicity of the presentation, the causal models in the figure are simplified—we do not consider the existence of possible confounders underlying X and Y or the relationship between the components of X. We would like to remark that in many supervised tasks, Y is the cause of X, e.g., in clinic diagnosis and handwritten digit recognition problems. The analysis in this section applies to both classification and regression.

Situation 1 (Fig. 1.a): $X \to Y$ with changing P_X and fixed $P_{Y|X}$ (covariate shift). Theoretically speaking, in this case P_X is irrelevant for modeling $P_{Y|X}$; however, if one uses a simple model to predict Y, which is usually the case, under-fit of the conditional model causes the predicted Y to depend on the input distribution P_X; importance reweighting according to the difference in P_X between the source and target domains is widely used to correct covariate shift (Shimodaira 2000; Sugiyama et al. 2008).

[2]Such variable are graphically depicted as square nodes in (Pearl and Bareinboim 2011). We would like to distinguish between the domain-specific selection diagram and the sample selection bias procedure used in (Bareinboim, Tian, and Pearl 2014). In the former, the selection variables W_s and V_s are root variables and encode the information that they change the corresponding data-generating process across domains. In the latter, the selection variable is a sink node and encodes the property of the final sampling procedure.

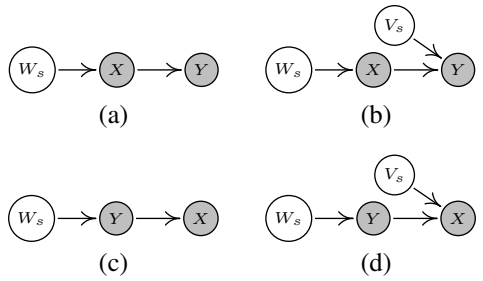

Figure 1: Possible situations of DA. X denote the feature vector, and Y is the target to be predicted. W_s and V_s are domain-specific selection variables assumed to be independent, leading to changing P_{XY} across domains. (a) Covariate shift: P_X is changed by W_s, but $P_{Y|X}$ does not change. (b) W_s and V_s change P_X and $P_{Y|X}$, respectively. (c) Target shift: W_s changes P_Y, with $P_{X|Y}$ unchanged. (d) W_s and V_s change P_Y and $P_{X|Y}$, respectively. In the first two situations, we consider X as a cause for Y, whilst in the last two situations, Y is a cause of X.

Situation 2 (Fig. 1.b): $X \to Y$ with changing $P_{Y|X}$ (and possibly changing P_X). Below we derive the optimal hypothesis for the target domain. Let $P^{t*}_{Y|X}$ be the underlying optimal posterior of Y on the target domain; see Table 1 for the notation used in this paper. Since V_s is unknown, we can estimate the optimal hypothesis by minimizing the expected Kullback-Leibler divergence between $P^t_{XY|W_s,V_s} = P^t_{X|W_s} P^t_{Y|X,V_s} = P^t_X P^t_{Y|X,V_s}$ and $P^t_X P^{t*}_{Y|X}$ (or maximizing the expected likelihood), which is given below, and the following position gives the solution.

$$\begin{aligned}&\mathbb{E}_{V_s} KL(P^t_{XY|W_s,V_s} || P^t_X P^{t*}_{Y|X}) \\ =&\mathbb{E}_{X,Y,V_s} \log\Big(\frac{P^t_X P^t_{Y|X,V_s}}{P^t_X P^{t*}_{Y|X}}\Big) = \mathbb{E}_{X,Y,V_s} \log\Big(\frac{P^t_{Y|X,V_s}}{P^{t*}_{Y|X}}\Big). \quad (1)\end{aligned}$$

Proposition 1. *Minimizing (1) w.r.t. a valid conditional distribution $P^{t*}_{Y|X}$ has the solution $P^{t*}_{Y|X} = \int P_{Y|X,V_s} dP_{V_s} = \mathbb{E}_{V_s}[P_{Y|X,V_s}]$*

In practice, the constructed optimal hypothesis would be $\hat{P}^{t*}_{Y|X} = \frac{1}{n}\sum_{i=1}^n P^{(i)}_{Y|X}$. That is, the learned target hypothesis is a convex combination (or more specifically, the average) of the source hypotheses. In (Mansour, Mohri, and Rostamizadeh 2008) this is known as the convex combination rule.

Situation 3 (Fig. 1.c): $Y \to X$, with changing P_Y and fixed $P_{X|Y}$. This is called prior probability shift (Storkey 2009) or target shift (Zhang et al. 2013). (Plessis and Sugiyama 2012) and (Zhang et al. 2013) studied how to estimate the change in P_Y in this situation, and the latter also applies for regression problems (i.e., with continuous Y).

Here we consider multiple source domains. Suppose P^t_Y can be represented as $P^t_Y = \sum_{i=1}^n \tilde{\alpha}_i P^{(i)}_Y$; we can derive the posterior of Y on the target domain:

$$\begin{aligned}P^t_{Y|X} &= \frac{P_{X|Y} P^t_Y}{P^t_X} = \frac{P_{X|Y}\sum_{i=1}^n \tilde{\alpha}_i P^{(i)}_Y}{P^t_X} \\ &= \frac{\sum_{i=1}^n \tilde{\alpha}_i P^{(i)}_{XY}}{\sum_{i=1}^n \tilde{\alpha}_i P^{(i)}_X} = \sum_{i=1}^n \frac{\tilde{\alpha}_i P^{(i)}_X}{\sum_{q=1}^n \tilde{\alpha}_q P^{(q)}_X} P^{(i)}_{Y|X}. \quad (2)\end{aligned}$$

The hypothesis for the target domain is then a distribution weighted combination of the individual hypotheses on source domains. This combination rule has been discussed in (Mansour, Mohri, and Rostamizadeh 2008), and here we have shown that in Situation 3 it is actually optimal. (Mansour, Mohri, and Rostamizadeh 2008) also compared this combination rule against the convex combination rule (see Situation 2), and the former was shown to be superior. This is consistent with the fact that in most classification problems Y is the cause for X; one can think of handwritten digit recognition and medical diagnosis as typical examples.

Situation 4 (Fig. 1.d): $Y \to X$ with changing $P_{X|Y}$ (and possibly changing P_Y). This is known as generalized target shift in (Zhang et al. 2013), where only a single source domain was considered. In Situation 4 we have to make certain assumptions on how $P_{X|Y}$ changes; with them, fortunately, P^t_X might provide additional knowledge to find the optimal classifier. This case will be further discussed in detail in the next section.

2 DA with Independently Changing P_Y & $P_{X|Y}$

Here we consider Situation 4, where P_Y and $P_{X|Y}$ both change across domains, as shown in Fig. 1.d. According to the graphical model or the causal explanation $Y \to X$, we know that P_Y and $P_{X|Y}$ change independent from each other. In this section we restrict our attention to classification problems. Generally speaking, without further conditions on the data generating process, it is not possible to recover $P^t_{X|Y}$, the conditional distribution on the target domain. It is possible to solve the problem under rather restrictive assumptions. For instance, (Zhang et al. 2013) considers DA with a single source domain, and assumes that the change in $P_{X|Y}$ follows the location-scale (LS) transformation; $P^t_{X|Y}$ is then generally identifiable. They have reported that LS generalized target shift produces a much better performance on remote sensing image classification then all alternatives, which demonstrates that Situation 4 is practically relevant for some rather complex DA problems.

Compared to a single source domain, multiple source domains contain much richer information as to how to determine $P_{X|Y}$ on the target domain, and we can avoid the constraint of the LS transformation.

2.1 Model: Target Conditional as a Linear Mixture of Source Conditionals

Motivation One can consider $P_{X|Y,V_s}$ (which is the conditional $P_{X|Y}$ in the domain associated with V_s; see Fig. 1.d)

as the mechanism to generate features from the class label given the domain. Imagine that there exist L elementary "sub-mechanisms", or class conditional feature distributions, $\tilde{P}^{(l)}_{X|Y}$, $l = 1, ..., L$, so that the mechanism in each domain, $P_{X|Y,V_s}$, is a mixture of those sub-mechanisms, i.e., $P_{X|Y=c_j,V_s} = \sum_{l=1}^{L} \tilde{\alpha}_{V_s,j,l} \tilde{P}^{(l)}_{X|Y=c_j}$, where $\mathring{\alpha}_{V_s,j,l}$ depend on both V_s and j, $\mathring{\alpha}_{V_s,l} \geq 0$, and $\sum_{l=1}^{L} \mathring{\alpha}_{V_s,l} = 1$. Consequently, in the multi-source DA scenario, if for each j, the rank of $\{P^{(i)}_{X|Y=c_j} \mid i = 1, ..., n\}$ is equal to L, $P^{t}_{X|Y=c_j}$ can always be represented as a linear mixture of $P^{(i)}_{X|Y=c_j}$.

More generally speaking, the proposed approach was also inspired by latent variable modeling. According to Fig. 1.d, we know that $P_{X|Y=c_j}$, or computationally more easily, its kernel embedding (Smola et al. 2007; Gretton et al. 2007), is actually a function of V_s:

$$\mu[P_{X|Y=c_j,V_s}] = \int \psi(x) P_{X|Y=c_j,V_s} dx = F_j(V_s), \quad (3)$$

where F_j are infinite-dimensional vector functions, which might vary for different values of j. Here V_s contains domain-specific conditions. For instance, for object recognition, it may contain the illumination condition, the angle from which the image was taken, etc.

One can see that the intrinsic dimensionality of $\{\mu[P^{(i)}_{X|Y=c_j}] \mid i = 1, ..., n\}$, is upper bounded by the intrinsic dimensionality of V_s, denoted by d_f. They are equal if F_j is non-degenerate, i.e., if there is no loss of degree of freedom in the transformation (3). We define d_f as the *degree-of-freedom in the conditional distribution change*. Generally speaking, the higher d_f, the more complex the change in $P_{X|Y=c_j}$ across domains. Since on source domains we only know that V_s might change across domains but cannot access its values, we cannot directly find d_f.

For simplicity, let us assume that F_j in (3) can be approximated by a linear function,[3] i.e., $\mu[P_{X|Y=c_j,V_s}] = L_{F_j} V_s$, where L_{F_j} has an infinite number of rows and d_f columns. That is,

$$\left(\mu[P^{(1)}_{X|Y=c_j}],\ \cdots,\ \mu[P^{(n)}_{X|Y=c_j}]\right) = L_{F_j} \cdot \left(V_s^{(1)},\ \cdots,\ V_s^{(n)}\right).$$

If we further assume that V^t_s can be constructed as a linear mixture of V_s on source domains, then $P^{t}_{X|Y=c_j}$ *is a linear mixture of* $P_{X|Y=c_j}$ *on source domains*. This tends to be the case if d_f is small: in this case, the rank of V_s is small, and then the class conditional feature distributions are likely to be linearly dependent, that is, the target-domain conditional distribution is likely to be represented as a linear mixture of those on source domains. If needed, in such situations we can directly estimate d_f from source domains by finding the rank of the estimated $\mu[P^{(i)}_{X|Y=c_j}]$, $i = 1, ..., n$, under the condition that we have enough source domains which are diverse enough. More specifically, let $\hat{\overrightarrow{\mu}}_j = \left(\hat{\mu}[P^{(1)}_{X|Y=c_j}], ..., \hat{\mu}[P^{(n)}_{X|Y=c_j}]\right) = \left(\frac{1}{m_{1j}}\psi(\mathbf{x}^{(1)}_j)\mathbf{1}, ..., \frac{1}{m_{1j}}\psi(\mathbf{x}^{(n)}_j)\mathbf{1}\right)$, where $\mathbf{1}$ denotes the vector of all 1's of an appropriate size; under this condition, d_f can be estimated as the maximum of the following quantity for all j:

$$\text{rank}(\hat{\overrightarrow{\mu}}_j) = \text{rank}(\hat{\overrightarrow{\mu}}_j^{\intercal} \hat{\overrightarrow{\mu}}_j) = \text{rank}(Q_j), \quad (4)$$

where the (i, i')th entry of Q_j is $\frac{1}{m_{ij} m_{i'j}} \mathbf{1}^{\intercal} K(\mathbf{x}^{(i)}_j), \mathbf{x}^{(i')}_j)) \mathbf{1}$. In practice, an appropriately chosen threshold is needed to determine the rank, due to the estimation error in the kernel mean embedding.

Formulation Motivated by this, we make the following assumption on $P_{X|Y}$ on the target domain.

$\mathbf{A_1}$. For each y, $P^{t}_{X|Y=y}$ is a mixture of $P_{X|Y=y}$ on the source domains, i.e., there exist α_{ij}, which satisfy the constraint $\sum_{i=1}^{n} \alpha_{ij} = 1$ for all j, such that

$$P^{new}_{X|Y=c_j} = \sum_{i=1}^{n} \alpha_{ij} P^{(i)}_{X|Y=c_j} \quad (5)$$

is equal to $P^{t}_{X|Y=c_j}$, where c_j is the jth possible value of Y.[4]

Denote by P^{new}_Y a marginal distribution of Y, and use $P^{new}_Y(c_j)$ as shorthand for $P^{new}_Y(Y = c_j)$. The corresponding joint distribution is

$$P^{new}_{X,Y=c_j} = P^{new}_Y(c_j) P^{new}_{X|Y=c_j}, \quad (6)$$

and the marginal distribution of X is then

$$P^{new}_X = \sum_{j=1}^{C} P^{new}_Y(c_j) \sum_{i=1}^{n} \alpha_{ij} P^{(i)}_{X|Y=c_j}. \quad (7)$$

We aim to match P^{new}_X with P^{t}_X by tuning the parameters α_{ij} and $P^{new}_Y(c_j)$. Here we have the constraints $P^{new}_Y(c_j) \geq 0$, and $\sum_{j=1}^{C} P^{new}_Y(c_j) = 1$. Let $\beta_{ij} \triangleq P^{new}_Y(c_j)\alpha_{ij}$, which satisfy the condition

$$\sum_{j=1}^{C} \sum_{i=1}^{n} \beta_{ij} = 1. \quad (8)$$

Once we find the values of β_{ij}, we can reconstruct p^{new}_Y and α_{ij} by $P^{new}_Y(c_j) = \sum_{i=1}^{n} \beta_{ij}$, and $\alpha_{ij} = \frac{\beta_{ij}}{P^{new}_Y(c_j)}$. The following theorem states that under mild conditions, $P^{t}_{X|Y}$ can be uniquely recovered.

[3] This holds if F_j is essentially linear, or if V_s does not change too much so that one can use linear approximation for F_j on all observed domains.

[4] We have two remarks here. First, for the domains with $P^{(i)}_Y(c_j) = 0$, $P^{(i)}_{X|Y=c_j}$ is undefined, and one can simply set $\alpha_{ij} = 0$. Second, usually the weights α_{ij} in a distribution mixture model are assumed to be nonnegative; however, this is not necessary to guarantee that the constructed $P^{t}_{X|Y=c_j}$ is a valid distribution. For flexibility of the mixture model, we allow α_{ij} to be negative, as long as $P^{new}_{X|Y=c_j}$ is a valid distribution, which, under appropriate assumptions, is achieved by matching P^{new}_X with P^{t}_X, as implied by Theorem 1.

Theorem 1. *Let Assumption A_1 hold. Further make the following assumption:*

$\mathbf{A_2}$. *For any constants d_{ij} that satisfy $\sum_{i=1}^{n} d_{ij}^2 \neq 0$, it holds that $\sum_{i=1}^{n} d_{ij} P^{(i)}_{X|Y=c_j}$, $j = 1, ..., C$, are always linearly independent, if they are not zero.*

Then if $P_X^{new} = P_X^t$, we have $P_Y^{new} = P_Y^t$ and $P^{new}_{X|Y} = P^t_{X|Y}$, i.e, P^{new}_{XY} is identical to P^t_{XY}.

To get an idea how strong (or weak) Assumption A_2 is, note that it is an assumption of linear independence of probability measures, or of densities (as functions of x). For continuous x, those are objects in infinite-dimensional spaces, and linear independence is the generic case rather than a special situation.

A sufficient condition for Assumption A_2 is that $P^{(i)}_{X|Y=c_j}$, $i = 1, ..., n$, $j = 1, ..., C$, are linearly independent. Note that this conditional is much stronger: Assumption A_2 allows $P^{(i)}_{X|Y=c_j}$, $i = 1, ..., n$, to be linear dependent for the same j. In fact, here we do not care about the the identifiability of the parameters β_{ij} (or α_{ij} and P_Y), but the identifiability of $P^{new}_{X|Y}$.

2.2 Parameter Estimation by Reproducing the Target Feature Distribution

We can estimate the parameters β_{ij}, and hence α_{ij} and P_Y^{new}, by minimizing the maximum mean discrepancy (MMD; see (Gretton et al. 2007)):

$$\Big|\Big|\mu[P_X^{new}] - \mu[P_X^t]]\Big|\Big| = \Big|\Big|\mathbb{E}_{P_X^{new}}[\psi(X)] - \mu[P_X^t]\Big|\Big|$$
$$= \Big|\Big| \sum_{j=1}^{C} P_Y^{new}(c_j) \sum_{i=1}^{n} \alpha_{ij} \mu[P^{(i)}_{X|Y=c_j}] - \mu[P_X^t]\Big|\Big|. \quad (9)$$

Let $x_{jk}^{(i)}$, $k = 1, ..., m_{ij}$ denote the data points of X in the ith source domain for which $Y = c_j$, where m_{ij} is the total number of points in the ith source domain for which $Y = c_j$. Similarly, x_k^t denotes the kth point of X in the target domain. In practice, we minimize the square of the empirical version of (9):

$$J_0 = \Big|\Big| \sum_{j=1}^{C} P_Y^{new}(c_j) \sum_{i=1}^{n} \frac{\alpha_{ij}}{m_{ij}} \sum_{k=1}^{m_{ij}} \psi(x_{jk}^{(i)}) - \frac{1}{m} \sum_{k=1}^{m} \psi(x_k^t)\Big|\Big|^2$$
$$= \sum_{j=1}^{C}\sum_{i=1}^{n}\sum_{j'=1}^{C}\sum_{i'=1}^{n} \frac{\beta_{ij}\beta_{i'j'}}{m_{ij}m_{i'j'}} \sum_{k=1}^{m_{ij}}\sum_{k'=1}^{m_{i'j'}} k(x_{jk}^{(i)}, x_{j'k'}^{(i')}) -$$
$$2\sum_{j=1}^{C}\sum_{i=1}^{n} \frac{\beta_{ij}}{mm_{ij}} \sum_{k=1}^{m_{ij}}\sum_{k'=1}^{m} k(x_{jk}^{(i)}, x_{k'}^t) + \text{const.} \quad (10)$$

Let $\vec{\beta} \triangleq (\beta_{11}, ..., \beta_{1C}, \beta_{21}, ..., \beta_{2C}, ..., \beta_{n1}, ..., \beta_{nC})^\intercal$, $\mathbf{A}$ be a $nC \times nC$ matrix with $\mathbf{A}_{(i-1)C+j,(i'-1)C+j'} = \frac{1}{m_{ij}m_{i'j'}} \sum_k \sum_k k(x_{jk}^{(i)}, x_{j'k'}^{(i')}) = \frac{1}{m_{ij}m_{i'j'}} \mathbf{1}^\intercal K(\mathbf{x}_j^{(i)}, \mathbf{x}_{j'}^{(i')})\mathbf{1}$ for $i \in \{1, 2, ..., n\}$, $i' \in \{1, 2, ..., n\}$, $j \in \{1, 2, ..., C\}$, and $j' \in \{1, 2, ..., C\}$, and $\mathbf{b}$ be a nC-dimensional vector with its entries $\mathbf{b}_{(i-1)C+j} = -\frac{1}{mm_{ij}} \sum_{k=1}^{m_{ij}} \sum_{k'=1}^{m} k(x_{jk}^{(i)}, x_{k'}^t) = -\frac{1}{mm_{ij}} \mathbf{1}^\intercal K(\mathbf{x}_j^{(i)}, \mathbf{x}_{k'}^t)$ for $i \in \{1, 2, ..., n\}$, $i' \in \{1, 2, ..., n\}$, $j \in \{1, 2, ..., C\}$. $\vec{\beta}$ can then be estimated by minimizing J_0:

$$J_0 = \vec{\beta}^\intercal \mathbf{A} \vec{\beta} + 2\mathbf{b}^\intercal \vec{\beta} + \text{const}, \quad (11)$$

subject to the constraint (8).[5] This is a quadratic programming (QP) problem. After finding the values of $\vec{\beta}$, we can then construct α_{ij} and $P_Y^{new}(c_j)$. For some practical issues in this optimization procedure, including enforcing the sparsity constraint on α_{ij}; see Supplementary Material.

In our experiments, we use the Gauss kernel, which is known to be characteristic; unless specified otherwise, we adopt the median heuristic to set the kernel width.

2.3 Construction of Target Classifiers

Given the estimated parameters β_{ij} (or α_{ij}), we then present several natural ways to construct the target-domain classifier or directly determine the class labels on the target domain.

By importance reweighting on source samples (denoted `weigh_sample`) The first approach is to train the classifier on the original data points in source domains with appropriate importance weights. Once we find α_{ij} and $P_Y^{new}(c_j)$, we can construct P_{XY}^{new}, which mimics P_{XY}^t. According to (5), since an empirical estimator of $P^{(i)}_{X|Y}(x|y = c_j)$ is $\hat{P}^{(i)}_{X|Y}(x|y = c_j) = \frac{1}{m_{ij}} \sum_{k=1}^{m_{ij}} \delta\big(x - x_{jk}^{(i)}\big)$, where $\delta(\cdot)$ is the Dirac delta function, an empirical estimator of $P_{XY}^{new}(x, y = c_j)$ is $\hat{P}_{XY}^{new} = P_Y^{new}(c_j) \sum_{i=1}^{n} \frac{\alpha_{ij}}{m_{ij}} \sum_{k=1}^{m_{ij}} \delta\big(x - x_{jk}^{(i)}\big)$. We aim to find the function $f(x)$ which minimizes the expected loss on the target domain. Denoted by $l(x, y; \theta)$ the loss function, where θ denotes the involved parameters, the expected loss is $R[P_{XY}^t, \theta, l(x, y; \theta)] = \mathbb{E}_{P_{XY}^t}[l(x, y; \theta)]$. Its empirical estimator is $R_{emp}[\hat{P}_{XY}^{new}, \theta, l(x, y; \theta)] = \int \hat{P}_{XY}^{new} l(x, y; \theta) dxdy = \sum_{j=1}^{C} \sum_{i=1}^{n} \sum_{k=1}^{m_{ij}} \frac{\alpha_{ij} P_Y^{new}(c_j)}{m_{ij}} l(x_{jk}^{(i)}, c_j; \theta)$. We can then train the classifier on all source data points with the reweighting coefficients $\frac{\alpha_{ij} P_Y^{new}(c_j)}{m_{ij}}$.

By generative modeling (denoted `genar_model`) The second approach is purely generative. Let $\eta_j(x) \triangleq P^t_{Y=c_j|X}(x) = \frac{P_Y^t(c_j) P^t_{X|Y=c_j}}{P_X^t}$. For any value of x, if $\eta_j(x)$ is known, one can directly find the class label for x by comparing $\eta_j(x), j = 1, ..., C$. We propose a method to estimate $\eta_j(x)$ without explicitly estimating those involved distributions. Again, we make use of the kernel mean embedding of distributions. For details see Supplementary Material.

By weighted combination of source classifiers (denoted `combn_classf`) Alternatively, we can combine the individual source classifiers to form the one for the target do-

[5] Here we use a hard constraint on β_{ij}. Note that in (Huang et al. 2007; Gretton et al. 2008), a slightly different constraint was used for importance weights to correct for covariate shift.

main:

$$P^t_{Y|X}(y = c_j|x) = \frac{P^t_Y(c_j)\sum_{i=1}^n \alpha_{ij} P^{(i)}_{X|Y}(x|y = c_j)}{P^t_X}$$

$$= \sum_{i=1}^n \gamma_j^{(i)}(x) P^{(i)}_{Y|X}(y = c_j|x), \qquad (12)$$

where $\gamma_j^{(i)}(x) \triangleq \frac{\alpha_{ij} P^t_Y(c_j) P^{(i)}_X(x)}{P^{(i)}_Y(c_j) P^t_X(x)}$. Note that under Assumption A_1, we have $\sum_{i=1}^n \gamma_j^{(i)}(x) = 1$. The weights $\gamma_j^{(i)}(x)$ can be estimated in a similar way to η_j in approach `genar_model`. This method involves construction of n classifiers and combines them with weights $\gamma_j^{(i)}(x)$, which depend on all of the test point x, domain i, and class j.

Comparisons of those approaches involve theoretical studies of discriminative and generative classifiers and the behavior of importance reweighting and weighted combination of classifiers. Generally speaking, as a generative approach, `genar_model` might not work well when X is high-dimensional. We will next compare them empirically.

2.4 Special Case: Distribution Weighted Hypothesis Combination

The distribution weighted hypothesis combination rule (Mansour, Mohri, and Rostamizadeh 2008) is actually a special case of the proposed `combn_classf` under additional constraints; as stated in the following theorem.

Theorem 2. *Suppose the conditions in Theorem 1 hold. The source hypothesis combination rule (12) reduces to the distribution weighted combination rule in the form of (2) under any of the following conditions:*

1. *$P_{X|Y}$ does not change across domains, and P^t_Y is a linear mixture of P_Y on source domains, or*

2. *P_Y does not change, and α_{ij} in (5) are the same for all classes $j = 1, ..., C$, or*

3. *both $P_{X|Y}$ and P_Y change, but $\alpha_{ij} P^{(i)}_Y(c_j)/P^t_Y(c_j)$ are the same for all j.*

The three conditions in the above theorem all constrain how P_Y or $P_{X|Y}$ change. For the distribution weighted rule, the same coefficient, $1/n$, was used in (Mansour, Mohri, and Rostamizadeh 2008) for all sources; here we denote this method by `simple_adapt`. We propose to use kernel mean matching (KMM; see (Huang et al. 2007)) to estimate $\tilde{\alpha}_i$ in the distribution weighted rule (2) from data such that $\sum_i \tilde{\alpha}_i P^{(i)}_X$ is as close to P^t_X as possible, and the resulting *hybrid* method is denoted by `dstr_wgh (H)`. Moreover, note that in our `dstr_wgh (H)`, the weights can be negative, while in (Mansour, Mohri, and Rostamizadeh 2008) all coefficients have to be nonnegative.

3 Experiments

3.1 Simulations

We first test the performance of the multi-source DA methods proposed in Section 2.3 for classification on simulated data. We generated the data according to Assumption A_1 in Sec. 2.1: on each domain, we generated the data points belonging to each class as a mixture of three fixed Gaussians, which have different means or variances, with random coefficients, and P_Y was also randomly chosen on each domain. We used three source domains, and in each domain the number of points in each class is a random number between 50 and 600. Fig. 2 shows the simulated data in one replication.

We compare the three classification approaches proposed in Section 2.3 against a number of alternatives. We include the following representative hypothesis combination methods for comparison: LWE (Gao et al. 2008), convex hypothesis combination (Mansour, Mohri, and Rostamizadeh 2008), denoted `convex`, `simple_adapt` (Mansour, Mohri, and Rostamizadeh 2008), and `dstr_wgh (H)`, which adopts the distribution weighted combination rule (2) with the weights $\tilde{\alpha}_i$ estimated from data. KMM for correcting covariate shift (Huang et al. 2007), the pooling SVM (denoted `pool_SVM`), which merge all source data to train the SVM, domain-invariant component analysis (DICA) (Muandet, Balduzzi, and Schölkopf 2013), and Learning marginal predictors (LMP) proposed by (Blanchard, Lee, and Scott 2011) are also included.

In our methods, we simply set the kernel width to 0.5, and the SVM parameters were selected by 5-fold cross validation on the parameter grids. Fig. 3 gives the boxplot of the misclassification rate of each method over 50 replications. We use both the Wilcoxon signed ranks test and Friedman test, recommended by (Demšar 2006), for performance comparison. With both tests, we found that on simulated data, the proposed approaches `weigh_sample` and `combn_classf` outperform all alternatives with p values smaller than 0.01, and that `genar_model` outperforms all the remaining methods with the p values smaller than 0.05. `dstr_wgh (H)` and `simple_adapt` are closely behind, verifying the finding that distribution weighted rule outperforms the convex combination of the source hypotheses reported in (Mansour, Mohri, and Rostamizadeh 2008).

Since the data points from each class were drawn from the mixture of three Gaussians with random coefficients, for each class, d_f, the degree-of-freedom in the conditional distribution change, as defined in Section 2.1, is 3. Recall that it indicates how many non-redundant source domains are needed to reconstruct $P^t_{X|Y}$. On the simulated data we found that $\text{rank}(Q_j) = 3$. We also varied the number of source domains from 3 to 5, and the rank of Q_j is still 3, as confirmed by the test of the rank of Hermitian positive semidefinite matrices (Camba-Mendez and Kapetanios 2005).

3.2 Real data: Sentiment analysis & Object recognition

The sentiment data (Blitzer, Dredze, and Pereira 2007) consist of review text and labels for four categories of goods (domains): *book, dvd, electronics,* and *kitchen*; each domain contains 2000 data points (or reviews) with four labels (or classes). We repeated the experiments on this dataset by (Mansour, Mohri, and Rostamizadeh 2008), but with a more general setting. (Mansour, Mohri, and Rostamizadeh 2008)

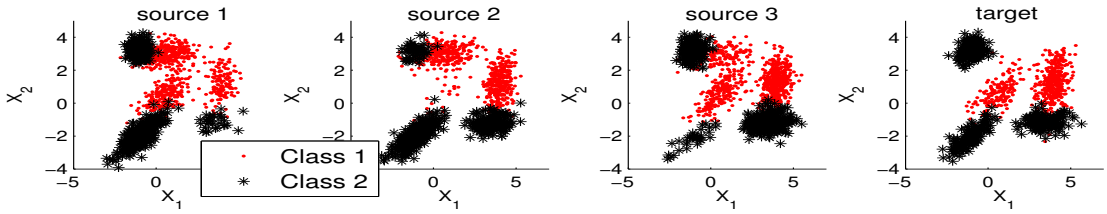

Figure 2: Simulated data in one replication.

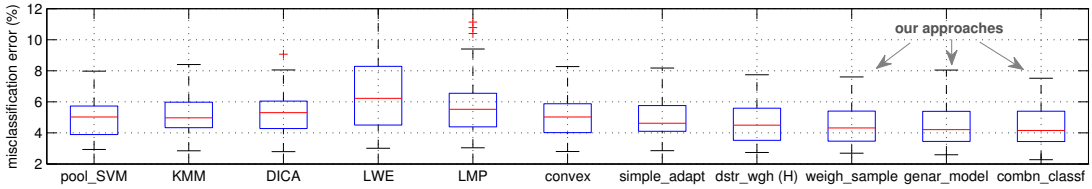

Figure 3: Boxplot of misclassification rate of each method on simulated data (50 replications).

constructed the target domain as a uniform mixture of data points randomly sampled from the four domains; the rest of the data were used as source-domain data. For each class, we sampled $w\%$ (w is a random number between 20 and 50) of the points from each source domain as the target-domain data. Our sampling scheme is more general: in our case P^t_{XY} is not necessary a uniform mixture of $P^{(i)}_{XY}$. We use the the frequency of the unigrams that appear 50 times or more in every domain as the features (in total there are 308 features). Each method was repeated 10 times by randomly sampling the data. The mean and standard deviation of the accuracies on target domains by each method are given in the upper part of Table 2. `combn_classf` and `weigh_sample` give the best accuracies.

We also compared our approaches with alternatives on the object recognition data (Griffin, Holub, and Perona 2007), as done by (Gong et al. 2012). We evaluated different methods on four object recognition datasets (domains): Amazon (images downloaded from Amazon), Webcam (low-resolution images by a web camera), DSLR (high-resolution images by a SLR camera), and Caltech-256 (Griffin, Holub, and Perona 2007). We extracted 10 common categories among all domains. There are 8 to 151 samples per category per domain, and 2533 images in total. We used three domains as sources and the rest one as the target. We followed the feature extraction scheme in (Gong et al. 2012). We used SVM for all the DA methods, and the SVM hyper parameters were selected by 5-fold cross validation on a grid. The results are shown in table 2. The Friedman test gives the p value 0.02, indicating that those approaches give different performances at the significance level 0.05; furthermore, `combn_classf` performs best, closely followed by `dstr_wgh (H)`.

4 Conclusion and discussions

We provided a causal view to domain adaptation with multiple source domains and noted that the background causal knowledge—the data-generating process—helps greatly in domain adaptation. Under different causal assumptions, the knowledge to be transferred from source domains to the target may be different, leading to different algorithms for domain adaptation. We considered several simplified causal models for this task, and accordingly gave the optimal hypothesis for the target domain. In particular, we have focused on a multi-source domain adaptation problem in which P_Y and $P_{X|Y}$ change independently across domains, where X denotes features and Y the target. The proposed methods consist of two steps. One first recovers $P_{X|Y}$ and P_Y on the test domain, by tuning involved parameters to reproduce the corresponding observed feature distribution. The second step constructs the classifier for the target domain or directly determines the target-domain class labels; to this end we presented three natural approaches for target-domain classification, which exploit importance reweighting, use generative learning, or resort to a weighted combination of source hypotheses.

The proposed methods rely on the assumption that for each class, the target-domain conditional distribution $P_{X|Y}$ can be represented as a mixture of those on source domains. We remark that for some real problems, certain features could be highly noisy, and it is worth noting that this assumption might not hold for some features or components of features; therefore it would be beneficial to find appropriate feature representations, as in (Ben-David et al. 2007). Furthermore, another future line of research is to derive convergence bounds and learning guarantees for the proposed domain adaptation approaches, following (Cortes, Mansour, and Mohri 2010; Iyer, Nath, and Sarawagi 2014).

References

Bareinboim, E.; Tian, J.; and Pearl, J. 2014. Recovering from selection bias in causal and statistical inference. In *Proc. 28th AAAI Conference on Artificial Intelligence*, 2410–2416.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2007. Analysis of representations for domain adaptation. In *Proc. of NIPS 2006*.

Ben-David, S.; Shalev-Shwartz, S.; and Urner, R. 2012. Domain adaptation–can quantity compensate for quality? In *ISAIM 2012*.

Blanchard, G.; Lee, G.; and Scott, C. 2011. Generalizing from

Table 2: Classification accuracy of different methods on the sentiment and object recognition data sets.

dataset	pool_SVM	KMM	DICA	LWE	convex	simple_adapt	dstr_wgh (H)	weigh_sample	combn_classf
sentiment	48.37(1.55)	41.75(1.03)	27.69(0.71)	35.39(5.62)	43.95(1.04)	44.76(1.21)	43.85(1.19)	*50.66(1.36)*	**51.72(1.12)**
→amazon	51.93	49.00	51.62	**54.34**	41.38	41.07	*53.50*	52.35	52.14
→caltech	*43.49*	40.37	43.40	40.37	40.55	40.02	**44.92**	*43.49*	43.32
→dslr	49.68	49.68	49.68	17.83	54.14	54.14	*57.96*	50.95	**62.42**
→webcam	56.95	53.90	55.25	37.97	61.02	61.69	*62.37*	60.67	**62.66**

several related classification tasks to a new unlabeled sample. In *NIPS 2011*, 2178–2186.

Blitzer, J.; Dredze, M.; and Pereira, F. 2007. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *In ACL*, 187–205.

Camba-Mendez, G., and Kapetanios, G. 2005. Estimating the rank of the spectral density matrix. *Journal of Time Series Analysis* 26:37–48.

Candela, J.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N., eds. 2009. *Dataset Shift in Machine Learning*. MIT Press.

Chattopadhyay, R.; Ye, J.; Panchanathan, S.; Fan, W.; and Davidson, I. 2011. Multi-source domain adaptation and its application to early detection of fatigue. In *KDD*.

Cortes, C.; Mansour, Y.; and Mohri, M. 2010. Learning bounds for importance weighting. In *NIPS 23*.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7:1–30.

Duan, L.; Tsang, I. W.; Xu, D.; and Chua, T. S. 2009. Domain adaptation from multiple sources via auxiliary classifiers. In *ICML*.

Gao, J.; Fan, W.; Jiang, J.; and Han, J. 2008. Knowledge transfer via multiple model local structure mapping. In *In International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV*.

Gong, B.; Shi, Y.; Sha, F.; and Grauman, K. 2012. Geodesic flow kernel for unsupervised domain adaptation. In *Proc. CVPR 2012*, 2066–2073.

Gretton, A.; Borgwardt, K.; Rasch, M.; Schölkopf, B.; and Smola, A. 2007. A kernel method for the two-sample-problem. In *NIPS 19*, 513–520. Cambridge, MA: MIT Press.

Gretton, A.; Smola, A.; Huang, J.; Schmittfull, M.; Borgwardt, K.; and Schölkopf, B. 2008. Covariate shift and local learning by distribution matching. In *Dataset shift in machine learning*. MIT Press. 131–160.

Griffin, G.; Holub, A.; and Perona, P. 2007. Caltech-256 object category dataset.

Huang, J.; Smola, A.; Gretton, A.; Borgwardt, K.; and Schölkopf, B. 2007. Correcting sample selection bias by unlabeled data. In *NIPS 19*, 601–608.

Iyer, A.; Nath, A.; and Sarawagi, S. 2014. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *Proc. ICML 2014*.

Jiang, J. 2008. *A literature survey on domain adaptation of statistical classifiers*. http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2008. Domain adaptation with multiple sources. In *NIPS 19*, 1041–1048. Cambridge, MA: MIT Press.

Muandet, K.; Balduzzi, D.; and Schölkopf, B. 2013. Domain generalization via invariant feature representation. In *Proc. ICML 2013*.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22:1345–1359.

Pearl, J., and Bareinboim, E. 2011. Transportability of causal and statistical relations: A formal approach. In *Proc. AAAI 2011*, 247–254.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Plessis, M. D., and Sugiyama, M. 2012. Semi-supervised learning of class balance under class-prior change by distribution matching. In *Proc. ICML 2012*.

Rosenstein, M.; Marx, Z.; Kaelbling, L.; and Dietterich, T. 2005. To transfer or not to transfer. In *NIPS 2005 Workshop on Inductive Transfer: 10 Years Later*.

Schölkopf, B.; Janzing, D.; Peters, J.; Sgouritsa, E.; Zhang, K.; and Mooij, J. 2012. On causal and anticausal learning. In *Proc. ICML 2012*.

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

Smola, A.; Gretton, A.; Song, L.; and Schölkopf, B. 2007. A Hilbert space embedding for distributions. In *Proc. ALT 2007*, 13–31. Springer-Verlag.

Storkey, A. 2009. When training and test sets are different: Characterizing learning transfer. In Candela, J.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N., eds., *Dataset Shift in Machine Learning*. MIT Press. 3–28.

Sugiyama, M., and Kawanabe, M. 2012. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. Cambridge, MA, USA: MIT Press.

Sugiyama, M.; Suzuki, T.; Nakajima, S.; Kashima, H.; von Bünau, P.; and Kawanabe, M. 2008. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics* 60:699–746.

Tian, J., and Pearl, J. 2001. Causal discovery from changes: a bayesian approach. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI2001)*, 512–521.

Zadrozny, B. 2004. Learning and evaluating classifiers under sample selection bias. In *Prof. ICML 2004*, 114–121.

Zhang, K.; Schölkopf, B.; Muandet, K.; and Wang, Z. 2013. Domain adaptation under target and conditional shift. In *Proc. ICML 2013*.

Acknowledgments

KZ would like to thank Elias Bareinboim for helpful discussions on selection diagrams and sample selection bias. We are grateful to the anonymous reviewers for their helpful comments and suggestions.