Domain Adaptation under Target and Conditional Shift

Kun Zhang KZHANG@TUEBINGEN.MPG.DE
Bernhard Schölkopf BS@TUEBINGEN.MPG.DE
Krikamol Muandet KRIKAMOL@TUEBINGEN.MPG.DE
Zhikun Wang ZHIKUN@TUEBINGEN.MPG.DE
Max Plank Institute for Intelligent Systems, Tübingen, Germany

Abstract

Let X denote the feature and Y the target. We consider domain adaptation under three possible scenarios: (1) the marginal P_Y changes, while the conditional $P_{X|Y}$ stays the same (*target shift*), (2) the marginal P_Y is fixed, while the conditional $P_{X|Y}$ changes with certain constraints (*conditional shift*), and (3) the marginal P_Y changes, and the conditional $P_{X|Y}$ changes with constraints (*generalized target shift*). Using background knowledge, causal interpretations allow us to determine the correct situation for a problem at hand. We exploit importance reweighting or sample transformation to find the learning machine that works well on test data, and propose to estimate the weights or transformations by *reweighting or transforming training data to reproduce the covariate distribution* on the test domain. Thanks to kernel embedding of conditional as well as marginal distributions, the proposed approaches avoid distribution estimation, and are applicable for high-dimensional problems. Numerical evaluations on synthetic and real-world data sets demonstrate the effectiveness of the proposed framework.

1. Introduction

The goal of supervised learning is to infer a function f from a training set $\mathbf{D}^{tr} = \{(x_1^{tr}, y_1^{tr}), ..., (x_m^{tr}, y_m^{tr})\} \subseteq \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ denote the domains of predictors X and target Y, respectively. The estimated f is expected to generalize well on the test set $\mathbf{D}^{te} = \{(x_1^{te}, y_1^{te}), ..., (x_n^{te}, y_n^{te})\} \subseteq \mathcal{X} \times \mathcal{Y}$, where y_i^{te} are unknown. Traditionally, the training set and test set are assumed to follow the same distribution. However, in many real world problems, the training data and test data have different distributions, i.e., $P_{XY}^{tr} \neq P_{XY}^{te}$, and the goal is to find a learning machine that performs well on the test domain. This problem is known as *domain adaptation* in machine learning.

If the data distribution changes arbitrarily, training data would be of no use to make predictions on the test domain. To perform domain adaptation successfully, relevant knowledge in the training (or source) domain should be transferred to the test (or target) domain. For instance, the situation where P_{XY}^{tr} and P_{XY}^{te} only differ in the marginal distribution of the covariate (i.e., $P_X^{tr} \neq P_X^{te}$, while $P_{Y|X}^{tr} = P_{Y|X}^{te}$) is termed covariate shift (Shimodaira, 2000; Sugiyama et al., 2008; Huang et al., 2007) or sample selection bias (Zadrozny, 2004), and has been well studied. For surveys on domain adaptation for classification, see, e.g., Jiang (2008); Pan & Yang (2010); Candela et al. (2009).

In particular, we address the situation where both the marginal distribution P_X and the conditional distribution $P_{Y|X}$ may change across the domains. Clearly, we need to make certain assumptions for the training domain to be adaptable to the test domain. We first consider the case where $P_{X|Y}$ is the same on both domains. As a consequence of Bayes' rule, the changes in P_X and $P_{Y|X}$ are caused by the change in P_Y, the marginal distribution of the target variable. We term this situation *Target Shift* (TarS) which is frequently encountered in practice; for instance, it is known as choice-based or endogenous stratified sampling (Manski & Lerman, 1977) in econometrics, and is sometimes called prior probability shift (Storkey, 2009).

We further discuss the situation where P_Y remains the same, while $P_{X|Y}$ changes, as termed *conditional shift* (ConS). Estimation of $P_{X|Y}^{te}$ under ConS is in general ill-posed; we consider a rather practical yet identifiable case where $P_{X|Y}$ changes under location-scale

Proceedings of the 30^th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

(LS) transformations on X. We show how to transform the training points to mimic the distribution of test data and facilitate learning on the test domain. Finally, the situation in which both P_Y and $P_{X|Y}$ change across domains is termed *generalized target shift* (GeTarS); we focus on LS-GeTarS, i.e., GeTarS with $P_{X|Y}$ changes under LS transformations, and propose practical methods to estimate both changes, making domain adaptation possible.

It has been demostrated that causal information can be derived from changes in data distributions (Tian & Pearl, 2001); on the other hand, knowledge of the data generating process, or causal knowledge, would imply how the data distribution changes across domains and help in domain adaptation. Schölkopf et al. (2012) demonstrated that a number of learning tasks, especially semi-supervised learning, can be understood from the causal point of view. The problems studied here, TarS, ConS, and GeTarS, have clear causal interpretations. Throughout the paper, we assume that Y is a cause of X.[1] If we further know that X depends on the domain (or selection variable) only via Y, we have the $TarS$ situation: the marginal distribution of the cause, P_Y, describes the process which generates Y in the domain, and $P_{X|Y}$ describes the data generating mechanism for X from the cause Y, which is independent of the domain. According to Woodward (2003), the invariance of $P_{X|Y}$ w.r.t. the change in P_Y is one of the features of the causal system $Y \rightarrow X$. Consider the clinical diagnosis as an example. The disease is natually considered as the cause of symptoms; moreover, the marginal distribution of the disease could change across different regions, but the conditional distribution of the symptoms given the disease is expected to be invariant. Furthermore, if both Y and the domain are causes of X while Y is independent of the domain, we have the ConS situation. More generally, the situation where Y is a cause of X and both P_Y and $P_{X|Y}$ depend on the domain corresponds to GeTarS.

In the classification scenario, target shift was referred to the class imbalance problem by Japkowicz & Stephen (2002). To solve it, sometimes it is assumed that P_Y^{te} is known *a priori* (Lin et al., 2002), or that some knowledge about the change in P_Y is known (Yu & Zhou, 2008). However, this is usually not the case in practice. Chan & Ng (2005) proposed to estimate P_Y^{te} with an EM algorithm. Unfortunately, this approach has to estimate $P_{X|Y}^{tr}$, which is a difficult task if the dimensionality of X is high; moreover, it does not apply to regression problems. In fact, lack of information on P_Y^{te} causes the main difficulty in domain adaptation under TarS.

In this paper we provide practical approaches for domain adaptation under TarS, LS-ConS, and LS-GeTarS, by sample importance reweighting or sample transformation. The approach for TarS also applies to regression. Kernel embedding of both conditional and marginal distributions provides a convenient tool to estimate the importance weights or the sample transformations. With it, we are able to avoid estimating any distribution explicitly, and the proposed approaches apply to high-dimenional problems without any difficulty. We note that kernel distribution embedding has been used to correct for covariate shift in Huang et al. (2007); Gretton et al. (2008), but the studied problems are inherently different: they used the kernel mean matching to estimate the ratio P_X^{te}/P_X^{tr}, avoiding estimating P_X^{te} and P_X^{tr} explicitly from data; in our problems we are interested in how P_Y^{te} is different from P_Y^{tr} (for TarS and GeTarS) or how $P_{X|Y}^{tr}$ changes to $P_{X|Y}^{te}$ (for ConS and GeTarS), but there are no data points available to estimate P_Y^{te} or $P_{X|Y}^{te}$, making the problems much more difficult to solve.

2. Distribution Shift Correction

In this section, we outline two scenarios for distribution shift correction, namely, *importance reweighting* and *sample transformation.*

Importance Reweighting We aim to find the function $f(x)$ that minimizes the expected loss on test data. Assume *the support of* P_{XY}^{te} *is contained by that of* P_{XY}^{tr}. The expected loss is $[P^{te}, \theta, l(x,y;\theta)] = \mathbb{E}_{(X,Y)\sim P^{te}}[l(x,y;\theta)] = \int P_{XY}^{tr} \cdot \frac{P_{XY}^{te}}{P_{XY}^{tr}} \cdot l(x,y,\theta)dxdy = \mathbb{E}_{(X,Y)\sim P^{tr}}[\beta^*(y)\cdot\gamma^*(x,y)\cdot l(x,y;\theta)]$, where θ denotes the parameters in the loss function $l(x,y;\theta)$, $\beta^*(y) \triangleq P_Y^{te}/P_Y^{tr}$ and $\gamma^*(x,y) \triangleq P_{X|Y}^{te}/P_{X|Y}^{tr}$. Here we *factorize* P_{XY} *as* $P_Y P_{X|Y}$ *instead of* $P_X P_{Y|X}$ because it provides a more convenient way to handle the change in P_{XY}, according to our assumptions given later. In practice, we minimize the empirical loss,

$$\widehat{R} = \frac{1}{m}\sum_{i=1}^{m} \beta^*(y_i^{tr})\gamma^*(x_i^{tr}, y_i^{tr}) l(x_i^{tr}, y_i^{tr};\theta), \qquad (1)$$

to find the supervised learning machine which is expected to work well on test data, if $\beta^*(y_i^{tr})\gamma^*(x_i^{tr}, y_i^{tr})$ are given. Readers who are interested in how to reduce the variance of the empirical expected loss may refer to, e.g., Shimodaira (2000); Robert & Casella (2004).

Sample Transformation and Reweighting Sample reweighting only applies when the suport of P_{XY}^{te}

[1] This is usually the case, especially for classification: in many cases features were generated from classes; for instance, see the handwriting digit recognition problem.

is contained in that of P^{tr}_{XY}; even under this condition, it is usually very difficult to estimate $\gamma^*(x,y)$ without prior knowledge on how $P_{X|Y}$ changes. Therefore, in the case where both P_Y and $P_{X|Y}$ change, the application of the sample reweighting scheme is rather limited. Instead, if we can find the transformation from $P^{tr}_{X|Y}$ to $P^{te}_{X|Y}$, i.e., find the transformation $\mathcal{T}$ such that the conditional distribution of $X^{new} = \mathcal{T}(X^{tr}, Y^{tr})$ satisfies $P^{new}_{X|Y} = P^{te}_{X|Y}$, we can calculate the expected loss on the test domain: $R[P^{te}, \theta, l(x,y;\theta)] = \mathbb{E}_{(X,Y)\sim P^{te}}[l(x,y;\theta)] = \int P^{tr}_{Y} \cdot \beta^*(y) \cdot P^{te}_{X|Y} \cdot l(x,y;\theta)dxdy = \mathbb{E}_{(X,Y)\sim P^{tr}_Y P^{new}_{X|Y}}[\beta^*(y)\cdot l(x,y;\theta)]$. Note that Y^{tr} is an argument of the transformation $\mathcal{T}$, i.e., $\mathcal{T}$ might be different at different Y values. This empirical loss can be calculated on the transformed training points $(\mathbf{x}^{new}, \mathbf{y}^{tr})$ with weights β^*:

$$\widehat{R}[P^{te}, \theta, l(x,y;\theta)] = \frac{1}{m}\sum_{i=1}^{m}\beta^*(y_i^{tr})l(x_i^{new}, y_i^{tr};\theta). \quad (2)$$

Classification and Regression Machines In this paper, we use support vector machine (SVM) and kernel ridge regression (KRR) for classification and regression problems, respectively. The standard formulation of both SVM and KRR can be straightforwardly modified to incorporate the importance weights according to (1) and (2). Details are skipped.

3. Correction for Target Shift

Unfortunately, unlike the covariate shift, the weights $\beta^*(y_i)\gamma^*(x_i,y_i)$ cannot be directly estimated because P^{te}_Y and $P^{te}_{X|Y}$ are unknown on the test data. Below we first consider the situation where $P^{te}_{X|Y} = P^{tr}_{X|Y}$, i.e., $\gamma^*(x,y) \equiv 1$, and propose a practical method to estimate $\beta^*(\mathbf{y}^{tr})$ as well as P^{te}_Y based on kernel embedding of conditional and marginal distributions.

3.1. Assumptions

We first consider Target Shift (TarS):

$\mathbf{A_1^{TarS}}$**:** $P^{te}_{X|Y} = P^{tr}_{X|Y}$ and $P^{te}_{Y} \neq P^{tr}_{Y}$.

That is, the difference between P^{tr}_{XY} and P^{te}_{XY} is caused by a shift in target distribution P_Y.

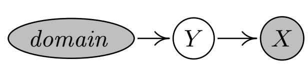


Figure 1. A causal model for TarS.

Fig. 1 shows a causal interpretation of TarS. For classification problems, it is possible to estimate P^{te}_Y in an iterative way by maximizing the likelihood on $\mathbf{x}^{te}$, for instance, with the EM algorithm (Chan & Ng, 2005); however, such approaches involve estimation of $P^{tr}_{X|Y}$ explicitly, which is difficult for high-dimensional problems. They are also not practical for regression.

We make the following assumptions on P^{te}_Y and $P^{tr}_{X|Y}$.

$\mathbf{A_2^{TarS}}$**:** The support of P^{te}_Y is contained in the support of P^{tr}_Y (i.e., roughly speaking, the training set is richer than the test set).

$\mathbf{A_3^{TarS}}$**:** There exists only one possible distribution of Y that, together with $P^{tr}_{X|Y}$, leads to P^{te}_X.

Imagine that we can draw a biased sample from the training data; here the selection variable depends only on Y, i.e., it is independent of X given Y. Denote by $P^{new}(\cdot)$ the distribution on this sample. Note that $P^{new}_{X|Y} = P^{tr}_{X|Y} = P^{te}_{X|Y}$. Thus, we can make P^{new}_X identical to P^{te}_X by adjusting P^{new}_Y.

Let $\beta(y)$ be the ratio of the P^{new}_Y to P^{tr}_Y, i.e., $P^{new}_Y = \beta(y)\cdot P^{tr}_Y$. To make P^{new}_X identical to P^{te}_X, we can adjust $\beta(y)$ to minimize $\mathcal{D}(P^{te}_X, P^{new}_X) = \mathcal{D}\big(P^{te}_X, \int P^{tr}_Y \beta(y) P^{tr}_{X|Y} dy\big)$, where $\mathcal{D}$ measures the difference between two distributions; it can be the mean square error or the Kullback-Leibler distance. To solve this problem, we have to estimate $P^{tr}_{X|Y}$ and P^{tr}_X from the training set, and moreover, the integral makes optimization very difficult.

3.2. A Kernel Mean Matching Approach

Instead, we solve this problem by making use of the kernel mean embedding of the marginal and conditional distributions; see Table 1 for the notation we use. The kernel mean embedding of P_X (Smola et al., 2007; Gretton et al., 2007) is a point in the Reproducing Kernel Hilbert Space (RKHS) given by $\mu[P_X] = \mathbb{E}_{X\sim P_X}[\psi(X)]$, and its empirical estimate is $\hat{\mu}[P_X] = \frac{1}{m}\sum_{i=1}^{m}\psi(x_i)$. The embedding of the conditional distribution has been studied by Song et al. (2009; 2010). The embedding of $P_{X|Y}$ can be considered as an operator mapping from $\mathcal{G}$ to $\mathcal{F}$, defined as $\mathcal{U}[P_{X|Y}] = \mathcal{C}_{XY}C^{-1}_{YY}$, where $\mathcal{C}_{XY}$ and $\mathcal{C}_{YY}$ denote the (uncentered) cross-covariance and covariance operators, respectively (Fukumizu et al., 2004). Furthermore, we have $\mu[P_X] = \mathcal{U}[P_{X|Y}]\mu[P_Y]$.

We make the following assumption on the kernels:

$\mathbf{A_4^{TarS}}$**:** Product kernel kl on $\mathcal{X}\times\mathcal{Y}$ is characteristic.

For characteristic kernels, the kernel mean map μ from the space of the distribution to the RKHS is injective, meaning that all information of the distribution is preserved (Fukumizu et al., 2008; Sriperumbudur et al., 2011). In this paper we use the Gaussian kernel, i.e., $k(x_i,x_j) = \exp\big(-\frac{||x_i-x_j||^2}{2\sigma^2}\big)$, where σ is the kernel width. Note that under assumptions A_3^{TarS} and A_4^{TarS}, for the embedding $\mathcal{U}[P^{tr}_{X|Y}]$, which is a mapping from $\mathcal{G}$ to $\mathcal{F}$, the pre-image of $\mu[P^{te}_X]$ is unique.

Table 1. Notation used in this paper.

random variable	X	Y
domain	$\mathcal{X}$	$\mathcal{Y}$
observation	x	y
data matrix	$\mathbf{x}$	$\mathbf{y}$
kernel	$k(x,x')$	$l(y,y')$
kernel matrix on training set	K	L
feature map	$\psi(x)$	$\phi(y)$
feature matrix on training set	Ψ	Φ
RKHS	$\mathcal{F}$	$\mathcal{G}$

The kernel mean embedding of P_Y^{new} is

$$\mu[P_Y^{new}] = \mathbb{E}_{Y\sim P_Y^{new}}[\phi(Y)] = \mathbb{E}_{Y\sim P_Y^{tr}}[\beta(y)\phi(Y)]. \quad (3)$$

The embedding of P_X^{new} is then given by $\mu[P_X^{new}] = \mathcal{U}[P_{X|Y}^{tr}]\mu[P_Y^{new}]$. Consequently, in the population version, we can find $\beta(y)$ by minimizing the maximum mean discrepancy:

$$\begin{aligned}&\big|\big|\mu[P_X^{new}] - \mu[P_X^{te}]\big|\big| = \big|\big|\mathcal{U}[P_{X|Y}^{tr}]\mu[P_Y^{new}] - \mu[P_X^{te}]\big|\big|\\ =&\big|\big|\mathcal{U}[P_{X|Y}^{tr}]\mathbb{E}_{Y\sim P_Y^{tr}}[\beta(y)\phi(y)] - \mu[P_X^{te}]\big|\big|, \quad (4)\end{aligned}$$

subject to $\beta(y) \geq 0$ and $\mathbb{E}_{P_Y^{tr}}[\beta(y)] = 1$, which guarantees that $P_Y^{new} = \beta(y)P_Y^{tr}$ is a valid distribution.

Theorem 1 *Under assumptions A_2^{TarS}, A_3^{TarS}, and A_4^{TarS}, the minimization problem (4) is convex in β. Further suppose A_1^{TarS} holds. Then the solution to (4) is $\beta(y) = \frac{P_Y^{te}(y)}{P_Y^{tr}(y)}$.*

For a proof see the supplementary material. In practice we have to use an empirical version. The empirical estimate of $\mathcal{U}_{X|Y}$ is $\hat{\mathcal{U}}_{X|Y} = \Psi(L+\lambda I)^{-1}\Phi^\intercal$. Recall that m and n are the sizes of the training and test sets. Denote by $\mathbf{1}_n$ the vector of 1's of length n, and by K^c the "cross" kernel matrix between $\mathbf{y}^{te}$ and $\mathbf{y}^{tr}$, i.e., $K_{ij}^c = k(x_i^{te}, x_j^{tr})$. Let β stand for $\beta(\mathbf{y}^{tr})$ and β_i for $\beta(y_i^{tr})$. The empirical version of the square of (4) is

$$\begin{aligned}&\Big|\Big|\hat{\mathcal{U}}_{X|Y}\cdot\frac{1}{m}\sum_{i=1}^m \beta_i\phi(y_i^{tr}) - \frac{1}{n}\sum_{i=1}^n \psi(x_i^{te})\Big|\Big|^2\\ =&\frac{1}{m^2}\beta^\intercal\phi^\intercal(\mathbf{y}^{tr})\hat{\mathcal{U}}_{X|Y}^\intercal\hat{\mathcal{U}}_{X|Y}\phi(\mathbf{y}^{tr})\beta\\ &\qquad - \frac{2}{mn}\mathbf{1}_n^\intercal\psi^\intercal(\mathbf{x}^{te})\hat{\mathcal{U}}_{X|Y}\phi(\mathbf{y}^{tr})\beta + \text{const}\\ =&\frac{1}{m^2}\beta^\intercal\underbrace{\Omega K\Omega^T}_{\triangleq A}\beta - \frac{2}{mn}\underbrace{\mathbf{1}_n^\intercal K^c\Omega^T}_{\triangleq M}\beta + \text{const}, \quad (5)\end{aligned}$$

where we use short-hand notation $\Omega \triangleq L(L+\lambda I)^{-1}$. As shown by Huang et al. (2007, Lemma 3), if $\beta_i \in [0, B_\beta]$, i.e., B_β is the upper bound of β, given that β_i has finite mean and non-zero variance, the sample mean $\frac{1}{m}\sum_{i=1}^m \beta_i$ converges in distribution to a Gaussian variable with mean $\mathbb{E}_{P_Y^{tr}}[\beta(y)]$ and standard deviation bounded by $\frac{B_\beta}{2\sqrt{m}}$. As $\mathbb{E}_{P_Y^{tr}}[\beta(y)] = 1$, we have the following constrained quadratic programming (QP) problem:

$$\begin{aligned}\underset{\boldsymbol{\beta}}{\text{minimize}} \quad & \frac{1}{2}\beta^\intercal A\beta - \frac{m}{n}M\beta,\\ \text{s.t.} \quad & \beta_i \in [0, B_\beta] \ \text{ and } \ \Big|\sum_{i=1}^m \beta_i - m\Big| \leq m\epsilon,\end{aligned}$$

where a good choice of ϵ is $\mathcal{O}\left(\frac{B}{2\sqrt{m}}\right)$.

Note that β estimated this way is not necessarily a function of y: different data points in the training set with the same y value could correspond to different β values. We also found that the β values estimated by solving the above optimization problem usually change dramatically along with y. We can improve the estimation quality of β by making use of reparameterization. First consider the case where Y is discrete. Let C be the cardinality of Y and denote by $v_1, ..., v_C$ its possible values. We can define a matrix $R^{(d)}$ where $R_{ik}^{(d)}$ is 1 if $y_i = v_k$ and is zero everywhere else. β can then be reparameterized as $\beta = R^{(d)}\alpha$,where the C-dimensional vector α is the new parameter.

We then consider the case where Y is continuous. Usually both distributions P_Y^{tr} and P_Y^{te} are smooth, and so is $\beta(y)$. Therefore, we would like to enforce the smoothness of $\beta(y)$ w.r.t. y. Let $R^{(c)} \triangleq L_\beta(L_\beta + \lambda_\beta I)^{-1}$, where L_β is a kernel matrix of $\mathbf{y}$ with the Gaussian kernel and λ_β is the regularization parameter.[2] Inspired by KRR (Saunders et al., 1998), we parameterize $\beta(\mathbf{y}^{tr})$ as $\beta = R^{(c)}\alpha$ with new parameter α. One can consider β as a smoothed version of α.

Finally, we find α (and β) in both cases by solving:

$$\begin{aligned}\underset{\boldsymbol{\alpha}}{\text{minimize}} \quad & \frac{1}{2}\alpha^\intercal\big[R^\intercal AR\big]\alpha - \frac{m}{n}\big[MR\big]\alpha,\\ \text{s.t.} \quad & 0 \leq \big[R\alpha\big]_i \leq B_\beta \text{ and } \big|\mathbf{1}_m^\intercal R\alpha - m\big| \leq m\epsilon, \quad (6)\end{aligned}$$

where R stands for $R^{(d)}$ or $R^{(c)}$, depending on whether Y is discrete or continuous. In all our experiments, we set $B_\beta = 10$ and $\epsilon = \frac{B_\beta}{4\sqrt{m}}$. We then set $\boldsymbol{\beta}^*$ in (1) to the estimated $\boldsymbol{\beta}$ and $\gamma^*(x_i, y_i) \equiv 1$. Minimizing (1) produces the classifier or regression model after correction for TarS.

4. Location-Scale Conditional Shift

[2]Note that although L_β and L are both kernel matrices of $\mathbf{y}$, they have different purposes and might have different hyperparameters, so we use different notations.

In practice $P^{tr}_{X|Y}$ and $P^{te}_{X|Y}$ might differ to some extent. It is certainly not possible to transfer useful knowledge from the training domain to the test domain if $P_{X|Y}$ changes arbitrarily. However, under certain assumptions on the change in $P_{X|Y}$, one could estimate $P^{te}_{X|Y}$ without knowing Y on test data. In this section we assume that $P_{X|Y}$ changes across domains and that $P^{tr}_Y = P^{te}_Y$. We term this situation Conditional Shift (ConS); Fig. 2 gives its causal interpretation. This situation might be less realistic in practice and will not be considered in our experiments; however, it serves as a foundation of a more general situation, GeTarS, which will be studied in Sec. 5. When considering ConS and GeTarS, we focus on classification problems.

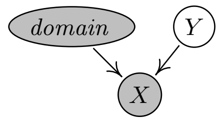


Figure 2. A causal model for ConS.

4.1. Assumptions and Identifiability

In some situations, we can formulate how the conditional distribution changes. For instance, for the same image, features such as intensities and colors are influenced by illumination, viewing angles, etc., which might change across domains. Modeling such a change enables distribution matching between the training domain and test domain, and consequently improves the performance on the test domain. Here we use the approach of *transforming training data to reproduce the covariate distribution on the test domain*; see Sec. 2. Since we can model the transformation from $P^{tr}_{X|Y}$ to $P^{te}_{X|Y}$, we do not need the condition that the support of $P^{te}_{X|Y}$ is contained in that of $P^{tr}_{X|Y}$, making the approach more practical.

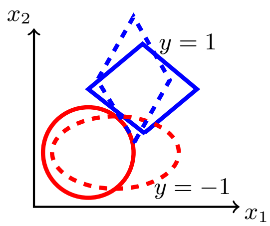


Figure 3. An illustration of LS-ConS where Y is binary and X is two-dimensional. Red lines are contours of $P_{X|Y}(x|y = -1)$, and blues ones are those of $P_{X|Y}(x|y = 1)$. Solid and dashed lines represent the contours on the training and test domains, respectively.

We assume that the shape of the distribution of each feature X_i, as well as the dependence structure between features, is preserved across the domains. More precisely, we assume that given any y value, $P^{te}_{X_i|Y}$ and $P^{tr}_{X_i|Y}$ only differs in the location and scale:

$\mathbf{A^{ConS}}$: There exists $\mathbf{w}(Y^{tr}) = \text{diag}[w_1(Y^{tr}), ..., w_d(Y^{tr})]$ and $\mathbf{b}(Y^{tr}) = [b_1(Y^{tr}), ..., b_d(Y^{tr})]^\intercal$, where d is the dimensionality of X, such that the conditional distribution of $X^{new} \triangleq \mathbf{w}(Y^{tr})X^{tr} + \mathbf{b}(Y^{tr})$ given Y^{tr} is the same as that of X^{te} given Y^{te}.

We term this situation location-scale ConS (LS-ConS). In matrix form, the transformed training points

$$\mathbf{x}^{new} \triangleq \mathbf{x}^{tr} \odot \mathbf{W} + \mathbf{B}, \tag{7}$$

where the ith columns of $\mathbf{W}$ and $\mathbf{B}$ are $[w_1(y_i), ..., w_d(y_i)]^\intercal$ and $[b_1(y_i), ..., b_d(y_i)]^\intercal$, respectively, are expected to have the same distribution as the test data. Fig. 3 illustrates on how the contours of $P_{X|Y}$ change across domains under LS-ConS.

The following theorem states that $P^{new}_{X|Y}$ is identifiable under some conditions on $P^{tr}_{X|Y}(x|y_i)$.

Theorem 2 *Let $P^{(\mathbf{w}_i,\mathbf{b}_i)}_{X|Y}(x|y_i)$ be the LS transformed version of $P^{tr}_{X|Y}(x|y_i)$ with parameters $(\mathbf{w}_i, \mathbf{b}_i)$ and $P^{te}_Y = P^{tr}_Y$. Suppose A^{ConS} holds, i.e., $\forall i$, $\exists(\mathbf{w}^*_i, \mathbf{b}^*_i)$ such that $P^{(\mathbf{w}^*_i,\mathbf{b}^*_i)}_{X|Y}(x|y_i) = P^{te}_{X|Y}(x|y_i)$. Further assume*

$\mathbf{A_2^{ConS}}$: *Set $\{c_{i1}P^{(\mathbf{w}_i,\mathbf{b}_i)}_{X|Y}(x|y_i) + c_{i2}P^{(\mathbf{w}'_i,\mathbf{b}'_i)}_{X|Y}(x|y_i)\,;\, i = 1, ..., C\}$ is linearly independent $\forall\, c_{i1}, c_{i2}$ ($c_{i1}^2 + c_{i2}^2 \neq 0$), $\mathbf{w}_i$, $\mathbf{w}'_i$ ($||\mathbf{w}_i||^2_F + ||\mathbf{w}'_i||^2_F \neq \mathbf{0}$), and $\mathbf{b}_i$, $\mathbf{b}'_i$.*

If $\exists\, (\mathbf{w}_i, \mathbf{b}_i)$ such that $P^{te}_X = \sum_i P^{tr}_Y(y_i)P^{(\mathbf{w}_i,\mathbf{b}_i)}_{X|Y}(x|y_i)$, then we have $\forall\, i$, $P^{(\mathbf{w}_i,\mathbf{b}_i)}_{X|Y}(x|y_i) = P^{te}_{X|Y}(x|y_i)$.

A necessary condition for A_2^{ConS} is that $P^{tr}_{X|Y}(x|y_i)$, $i = 1, ..., C$, are linearly independent after any LS transformations. Rougly speaking, the higher d, the less likely for this assumption to be violated.

4.2. A Kernel Approach

As in Sec. 3.2, we parameterize $\mathbf{W}$ and $\mathbf{B}$ as $\mathbf{W} = R\mathbf{G}$ and $\mathbf{B} = R\mathbf{H}$, where $\mathbf{G}$ and $\mathbf{H}$ are the parameters to be estimated, and R is $R^{(c)}$ or $R^{(d)}$, depending on whether Y is discrete or continuous. In this way $\mathbf{W}$ and $\mathbf{B}$ are guaranteed to be functions of y, and the number of parameters is greatly reduced.

Noting the relationship between X^{new} and X^{tr}, and using the substitution rule, we have

$$\begin{aligned}
&\mathcal{U}[P^{new}_{X|Y}] = \mathcal{C}_{X^{new}Y}C^{-1}_{YY} \\
&= \mathbb{E}_{(X^{new},Y)\sim P^{new}_{XY}}[\psi(X^{new}) \otimes \phi^\intercal(Y)]\, \mathbb{E}^{-1}_{Y\sim P^{tr}_Y}[\phi(Y) \otimes \phi^\intercal(Y)] \\
&= \mathbb{E}_{(X^{tr},Y)\sim P^{tr}_{XY}}[\psi(X^{new}) \otimes \phi^\intercal(Y)] \cdot \mathbb{E}^{-1}_{Y\sim P^{tr}_Y}[\phi(Y) \otimes \phi^\intercal(Y)].
\end{aligned}$$

The empirical estimate of $\mathcal{U}[P^{new}_{X|Y}]$ is consequently

$$\begin{aligned}
\hat{\mathcal{U}}[P^{new}_{X|Y}] =& \frac{1}{m}\psi(\mathbf{x}^{new}) \cdot \phi^\intercal(\mathbf{y}^{tr}) \cdot \left[\frac{1}{m}\phi(\mathbf{y}^{tr})\phi^\intercal(\mathbf{y}^{tr}) + \tilde{\lambda}I\right]^{-1} \\
=& \tilde{\Psi}(L + \lambda I)^{-1}\Phi^\intercal,
\end{aligned} \tag{8}$$

where $\tilde{\Psi} = \psi(\mathbf{x}^{new})$.

Let $\tilde{K}$ be the kernel matrix corresponding to the feature matrix $\tilde{\Psi}$, i.e., $\tilde{K}_{i,j} = k(x_i^{new}, x_j^{new})$, and $\tilde{K}^c$ the cross kernel matrix between $\mathbf{x}^{te}$ and $\mathbf{x}^{new}$, i.e., $\tilde{K}^c_{ij} = k(x_i^{te}, x_j^{new})$. We aim to minimize $\left|\left|\mu[P_X^{new}] - \mu[P_X^{te}]\right|\right|^2$, whose empirical version is

$$
\begin{aligned}
J^{ConS} &\triangleq \left|\left|\hat{\mu}[P_X^{new}] - \hat{\mu}[P_X^{te}]\right|\right|^2 = \left|\left|\hat{\mathcal{U}}[P_{X|Y}^{new}]\hat{\mu}[P_Y^{tr}] - \hat{\mu}[P_Y^{te}]\right|\right|^2 \\
&= \frac{1}{m^2}\mathbf{1}_m^\intercal \phi^\intercal(\mathbf{y}^{tr})\hat{\mathcal{U}}^\intercal[P_{X|Y}^{new}]\hat{\mathcal{U}}[P_{X|Y}^{new}]\phi(\mathbf{y}^{tr})\mathbf{1}_m \\
&\qquad - \frac{2}{mn}\mathbf{1}_n^\intercal \psi^\intercal(\mathbf{x}^{te})\hat{\mathcal{U}}[P_{X|Y}^{new}]\phi(\mathbf{y}^{tr})\mathbf{1}_m \\
&= \frac{1}{m^2}\mathbf{1}_m^\intercal \Omega \tilde{K} \Omega^T \mathbf{1}_m - \frac{2}{mn}\mathbf{1}_n^\intercal \tilde{K}^c \Omega^T \mathbf{1}_m. \qquad (9)
\end{aligned}
$$

We then estimate $\mathbf{W}$ (or $\mathbf{G}$) together with $\mathbf{B}$ (or $\mathbf{H}$) by minimizing J^{ConS}. In practice we also regularize (9) to prefer the change in $P_{X|Y}$ to be as little as possible, i.e., to make entries of $\mathbf{W}$ close to one and those of $\mathbf{B}$ close to zero. This is particularly useful in case assumption A_2^{ConS} is violated; we then prefer the slightest change in the conditional, among all possibilities. The regularization term is

$$J^{reg} = \frac{\lambda_{LS}}{m} \cdot ||\mathbf{W} - \mathbf{1}_m \mathbf{1}_d^\intercal||_F^2 + \frac{\lambda_{LS}}{m} \cdot ||\mathbf{B}||_F^2. \qquad (10)$$

One can find the derivarive of J^{ConS} and J^{reg} w.r.t. $\mathbf{G}$ and $\mathbf{H}$, and use the scaled conjugate gradient (SCG) to minimize $J^{ConS} + J^{reg}$. After estimating $\mathbf{W}$ and $\mathbf{B}$, we transform $\mathbf{x}^{tr}$ to $\mathbf{x}^{new}$ according to (7), and $(\mathbf{x}^{new}, \mathbf{y}^{tr})$ would have the same distribution as the test data, under assumption A^{ConS}. Consequently, the classifier or regressor trained on $(\mathbf{x}^{new}, \mathbf{y}^{tr})$ is expected to generalize well to the test domain.

5. LS Generalized Target Shift

We then consider a more general situation where both P_Y and $P_{X|Y}$ change, called Generalized Target Shift (GeTarS). Fig. 4 gives the causal model underlying the GeTarS situation.

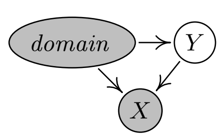


Figure 4. A causal model for GeTarS.

In this setting, we assume that $P_Y^{te} \neq P_Y^{tr}$ and that assumption A^{ConS} holds, i.e., we consider LS-GeTarS, and aim to estimate the importance weights $\beta^*(y_i) \triangleq \frac{P_Y^{te}(y_i)}{P_Y^{tr}(y_i)}$ and the matrices $\mathbf{W}$ and $\mathbf{B}$ in (7). They would transform the training data to mimic the distribution of the test data, and the learning machine learned on the reweighted transformed data is expected to work well on the test data. Parameters can be estimated by reweighting and transforming the training data to reproduce P_X^{te}, i.e., by minimizing $||\mu[P_X^{new}] - \mu[P_X^{te}]||$, where $P_X^{new} = \int P_Y^{new} P_{X|Y}^{new} dy$, $P_Y^{new} = \beta P_Y^{tr}$, and $P_{X|Y}^{new}(x|y_i) = P_{X|Y}^{(\mathbf{w}_i, \mathbf{b}_i)}(x|y_i)$. The following theorem provides the identifiability of p_Y^{new} and $P_{X|Y}^{new}$.

Theorem 3 *Suppose A^{ConS} holds. Under assumption A_2^{ConS}, if there exist $(\mathbf{w}_i, \mathbf{b}_i)$ such that $P_X^{te} = \sum_i P_Y^{new}(y_i) P_{X|Y}^{(\mathbf{w}_i, \mathbf{b}_i)}(x|y_i)$, then we have $P_Y^{new} = P_Y^{te}$, and $\forall$ i, $P_{X|Y}^{(\mathbf{w}_i, \mathbf{b}_i)}(x|y_i) = P_{X|Y}^{te}(x|y_i)$.*

Combining (3) and (8), we can find the empirical version of $||\mu[P_X^{new}] - \mu[P_X^{te}]||^2$:

$$
\begin{aligned}
J =& \left|\left|\hat{\mu}[P_X^{new}] - \hat{\mu}[P_X^{te}]\right|\right|^2 = \left|\left|\hat{\mathcal{U}}[P_{X|Y}^{new}]\hat{\mu}[P_Y^{te}] - \hat{\mu}[P_X^{te}]\right|\right|^2 \\
=& \left|\left|\frac{1}{m}\hat{\mathcal{U}}[P_{X|Y}^{new}]\phi(\mathbf{y}^{tr})\beta - \frac{1}{n}\psi(\mathbf{x}^{te})\mathbf{1}_n\right|\right|^2 \\
=& \frac{1}{m^2}\beta^\intercal \Omega \tilde{K} \Omega^T \beta - \frac{2}{mn}\mathbf{1}_n^\intercal \tilde{K}^c \Omega^T \beta. \qquad (11)
\end{aligned}
$$

When minimizing J, we would also like the difference between $P_{X|Y}^{te}$ and $P_{X|Y}^{tr}$, as measured by J^{reg} given in (10), to be as little as possible. Combining both constraints, we estimate the involved parameters β, $\mathbf{W}$, and $\mathbf{B}$ by minimizing

$$J^{GeTarS} = J + \lambda_{LS} J^{reg}. \qquad (12)$$

Finally, for parameter estimation, we iteratively alternate between the QP to minimize (11) w.r.t β and the SCG optimization procedure w.r.t. $\{\mathbf{W}, \mathbf{B}\}$. For details of the two optimization sub-procedures, see Sections 3 and 4, respectively. After estimating the parameters, we train the learning machine by minimizing the weighted loss (2) on $(\mathbf{x}^{new}, \mathbf{y}^{tr})$.

For how to select the hyperparameters involved in our methods, please refer to the supplementary material or the approach used for kernel-based conditional independence test (Zhang et al., 2011). The MATLAB source code for correcting TarS and LS-GeTarS is available at

`http://people.tuebingen.mpg.de/kzhang/Code-TarS.zip` .

6. Simulations

We use simulations to study the performance of the proposed approach for TarS and LS-GeTarS in four scenarios. They are *(a)* a nonlinear regression problem under TarS, *(b)* a classification problem under TarS, *(c)* a classification problem approximately following LS-GeTarS, and *(d)* a classification problem under non-LS-GeTarS with slight changes in the conditionals. See

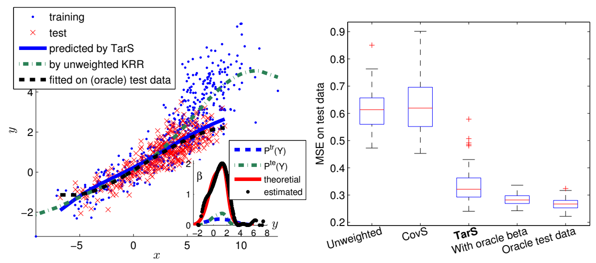
Fig. 5 (left) for the training and test points generated in one random replication. The training and test sets consist of 500 and 400 data points, respectively.

We compare our approaches to correction for TarS (Section 3) and for LS-GeTarS (Section 5) with the baseline (unweighted) least squares KRR or SVM, the importance weighting approach to correction for covariate shift (CovS) proposed in Huang et al. (2007); Gretton et al. (2008), as well as two "oracle" approaches: one uses the theoretical values of $\beta^*(y) = P_Y^{te}/P_Y^{te}$, and the other trains the learning machine directly on the test set. Note that the result learned on the test set certainly has the best performance, but in practice it cannot be applied; it is given to show the limit of the performance that any domain-adaptation approach can achieve. Since in the considered classification problems X is low-dimensional, it is possible to apply the EM algorithm proposed by Chan & Ng (2005) to estimate P_Y^{te}, so it is also included for comparison. We repeated the simulations for 100 times.

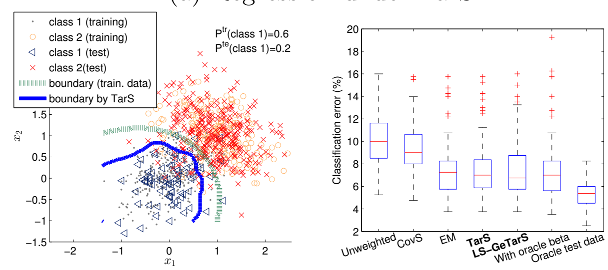
Fig. 5 (right) shows the boxplot of the performances of all approaches, measured by the mean square error (MSE) or classification error on the test set; for illustrative purposes, the left panels show the data points generated in one replication as well as the regression lines or decision boundaries learned by selected approaches. Under TarS, *(a, b)*, and non-LS-GeTarS with slightly changing conditionals, *(d)*, compared to the baseline unweighted method, clearly our approaches for TarS and LS-GeTarS improve the performance significantly. For regression under TarS, the estimated β values are very close to the theoretical ones, as seen from the lower-right corner of Fig. 5 (a, left). EM achieves a similar performance as TarS, since $P_{X|Y}$ can be modeled well in this simple case. In *(c)* the conditional $P_{X|Y}$ changes significantly, such that none of the approaches correcting for CovS or TarS helps, but since the change approximately follows LS-GeTarS, our approach for LS-GeTarS greatly improves the classification performance. Compared to the unweighted method, the important reweighting approach for CovS slightly improves the performance in settings *(b)* and *(d)*, and make it worse in *(a)* and *(c)*.

7. Real-world Data Sets

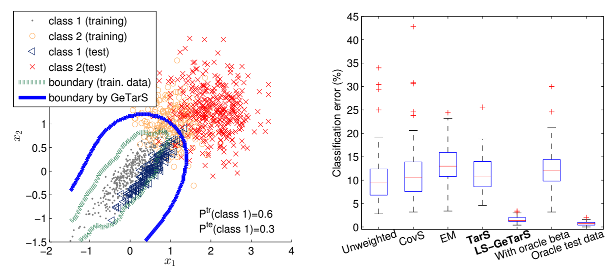
We evaluate the performance of the proposed approaches for regression and classification on real data. We first consider prediction of nonstationary processes, and then tackle the remote sensing image classification problem, with images obtained on different areas.



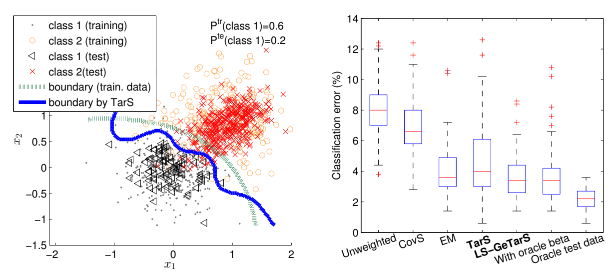
(a) Regression under TarS



(b) Classification under TarS



(c) Classification under LS-GeTarS



(d) Classification under non-location-scale GeTarS

Figure 5. Four simulation settings together with the performances of different approaches. Left panels show the data points together with the decision boundaries (or regression lines) obtained by selected approaches in one replication, and right panels give the boxplot of the performances of different approaches for 100 random replications. (a) For a regression problem with X depending on Y nonlinearly. (b) For a classification problem under TarS. (c) For a classification problem under shape-preserving GeTarS. (d) For a classification problem under GeTarS but the shape of the conditional distribution changes. Note that y-values of the test data were not given in the training phase, and they are plotted for illustrative purposes.

7.1. Regression under TarS

We first applied our approach for prediction on suitable data selected from the cause-effect pairs.[3] We selected data set No. 68, since 1) the data are non-stationary time series, 2) there is a strong dependence between the two variables so that one can be predicted non-trivially by the other, and 3) the variables are believed to have a direct causal relation, so that the invariance of the conditional distribution of one variable (effect) given the other (cause) is likely to hold approximately. Fig. 6 (top) showing the time series as well as the joint distribution. Here X and Y stand for the number of bytes sent by a computer at the tth minute and the number of open http connections at the same time, respectively. It is natural to have the causal relation $Y \to X$, and we aim to predict Y from X without making use of temporal dependence in the data. One subsample was always used for training, because on it Y has large values. The remaining data were divided into four subsets, and each time one of them was used for test and the others included for training.

Fig. 6 (bottom) shows the estimated β^* values on the four test sets; they match P_Y^{te} well. Table 2 gives the MSE on the four test sets produced by different approaches. Note that to achieve robustness of the prediction result, we incorporated an exponent q for β^* as the importance weights, as in correction for CovS with importance re-weighting (Shimodaira, 2000). $q = 1$ (i.e., the proposed standard approach) and $q = 0.5$ were used. From Table 2 one can see TarS gives the best results on all four test sets.

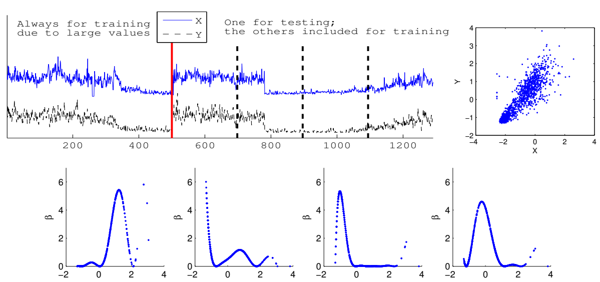


Figure 6. Prediction results on Pair 68 of the cause-effect pairs. Top: time series data of X and Y (left, shifted apart for clarity) and the joint distribution (right). Bottom: estimated β^* values on the four test sets.

Table 2. Prediction performance (MSE) on test sets.

Test set	Unweight.	CovS	CovS ($q = 0.5$)	TarS	TarS (q=0.5)
1	0.3789	0.3844	0.3802	0.3310	**0.3229**
2	0.0969	0.1126	0.1071	0.0937	**0.0887**
3	0.0578	0.0673	0.0659	**0.0466**	0.0489
4	0.2054	0.2126	0.2136	0.2008	**0.1630**

[3] http://webdav.tuebingen.mpg.de/cause-effect/

7.2. Remote Sensing Image Classification

We used a benchmark data set for remote sensing image classification with 14 classes and 145 features; for details of this data set, see (Ham et al., 2005). The labeled samples were collected on two different and spatially disjoint areas, and one would expect that not only P_Y, but also $P_{X|Y}$ changes across them, due to physical factors related to ground, vegetation, and atmospheric conditions. The samples taken on each area were partitioned into a training set TR and a test set TS by random sampling. TR_1, TS_1, TR_2, and TS_2 have sample sizes 1242, 1252, 2621, and 627, respectively. We consider two adaptation problems, $TR_1 \to TS_2$ and $TR_2 \to TS_1$.

After estimating the weights and/or transformed training data (with $\lambda_{LS} = 10^{-4}$), we applied the multi-class classifier with a RBF kernel on the weighted or transformed data. Hyperparameters were selected by cross-validation. Table 3 shows the overall classification error (i.e., the fraction of misclassified points) obtained by different approaches for each domain adaptation problem. We can see that in this experiment, correction for target shift does not significantly improve the performance; in fact, the estimated β values for most classes are rather close to one. However, correction for conditional shift with LS-GeTarS substantially reduces the overall classification error in both cases.

Table 3. A misclassification rate on remote sensing data set under different distribution shift correction schemes.

Problem	Unweight	CovS	TarS	LS-GeTarS
$TR_1 \to TS_2$	20.73%	20.73%	20.41%	**11.96%**
$TR_2 \to TS_1$	26.36%	25.32%	26.28%	**13.56%**

8. Conclusion and Discussions

We have considered domain adaptation where both the distribution of the covariate and the conditional distribution of the target given the covariate change across domains. From the causal point of view, we assume the target causes the covariate, such that the change in the the data distribution can be modeled easily. In particular, we studied three situations, target shift, conditional shift, and generalized target shift which combines the above two situations. We presented practical approaches to handle them based on the kernel mean embedding of conditional and marginal distributions. Simulations were conducted to verify our theoretical claims, and experimental results on diverse real-world problems, showed that (generalized) target shift often happens in domain adaptation, and that the proposed approaches could substantially improve the classification or regression performance.

References

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

Chan, Y. S. and Ng, H. T. Word sense disambiguation with distribution estimation. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pp. 1010–1015, Scotland, 2005.

Fukumizu, K., Bach, F. R., Jordan, M. I., and Williams, C. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *JMLR*, 5:73–99, 2004.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. *NIPS 20*, pp. 489–496, Cambridge, MA, 2008.

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

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. Covariate shift and local learning by distribution matching. In Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., , and Lawrence, N. (eds.), *Dataset shift in machine learning*, pp. 131–160. MIT Press, Cambridge, MA, 2008.

Ham, J., Chen, Y., Crawford, M. M., and Ghosh, J. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Trans. Geosci. Remote Sens.*, 43(3):492–501, 2005.

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

Japkowicz, N. and Stephen, S. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6:429–450, 2002.

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

Lin, Y., Lee, Y., and Wahba, G. Support vector machines for classification in nonstandard situations. *Machine Learning*, 46:191–202, 2002.

Manski, C. and Lerman, S. The estimation of choice probabilities from choice-based samples. *Econometrica*, 45:1977–1988, 1977.

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

Robert, C. P. and Casella, G. *Monte Carlo Statistical Methods*. Springer Press, New York, 2nd edition, 2004.

Saunders, C., Gammerman, A., and Vovk, V. Ridge regression learning algorithm in dual variables. In *Proc. ICML*, pp. 515–521, Madison, WI, 1998.

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

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

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, pp. 13–31. Springer-Verlag, 2007.

Song, L., Huang, J., Smola, A., and Fukumizu, K. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proc. ICML* 2009.

Song, L., Boots, B., Siddiqi, S., Gordon, G., and Smola, A. Hilbert space embeddings of hidden markov models. In *ICML 2010*.

Sriperumbudur, B., Fukumizu, K., and Lanckriet, G. Universality, characteristic kernels and rkhs embedding of measures. *JMLR*, 12:2389–2410, 2011.

Storkey, A. 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*, pp. 3–28. MIT Press, 2009.

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

Tian, J. and Pearl, J. Causal discovery from changes: a bayesian approach. In *UAI2001*, pp. 512–521, 2001.

Woodward, J. *Making things happen: A theory of causal explanation.* Oxford University Press, New York, 2003.

Yu, Y. and Zhou, Z. A framework for modeling positive class expansion with single snapshot. In *PAKDD 2008*.

Zadrozny, B. Learning and evaluating classifiers under sample selection bias. In *Proc. ICML*, pp. 114–121, Banff, Canada, 2004.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. In *UAI 2011*.