JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 00, NO. 0, 1–13: Theory and Methods
https://doi.org/10.1080/01621459.2024.2321653

Taylor & Francis
Taylor & Francis Group

Check for updates

Doubly Flexible Estimation under Label Shift

Seong-ho Lee[a], Yanyuan Ma[b], and Jiwei Zhao[c]

[a]Department of Statistics, University of Seoul, Seoul, South Korea; [b]Department of Statistics, Pennsylvania State University, University Park, PA; [c]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI

ABSTRACT

In studies ranging from clinical medicine to policy research, complete data are usually available from a population $\mathcal{P}$, but the quantity of interest is often sought for a related but different population $\mathcal{Q}$ which only has partial data. We consider the setting when both outcome Y and covariate $\mathbf{X}$ are available from $\mathcal{P}$ but only $\mathbf{X}$ is available from $\mathcal{Q}$, under the label shift assumption; that is, the conditional distribution of $\mathbf{X}$ given Y is the same in the two populations. To estimate the parameter of interest in $\mathcal{Q}$ by leveraging information from $\mathcal{P}$, three ingredients are essential: (a) the common conditional distribution of $\mathbf{X}$ given Y, (b) the regression model of Y given $\mathbf{X}$ in $\mathcal{P}$, and (c) the density ratio of the outcome Y between the two populations. We propose an estimation procedure that only needs some standard nonparametric technique to approximate the conditional expectations with respect to (a), while by no means needs an estimate or model for (b) or (c); that is, doubly flexible to the model misspecifications of both (b) and (c). This is conceptually different from the well-known doubly robust estimation in that, double robustness allows at most one model to be misspecified whereas our proposal can allow both (b) and (c) to be misspecified. This is of particular interest in label shift because estimating (c) is difficult, if not impossible, by virtue of the absence of the Y-data from $\mathcal{Q}$. While estimating (b) is occasionally off-the-shelf, it may encounter issues related to the curse of dimensionality or computational challenges. We develop the large sample theory for the proposed estimator, and examine its finite-sample performance through simulation studies as well as an application to the MIMIC-III database. Supplementary materials for this article are available online including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received December 2022
Accepted February 2024

KEYWORDS
Distribution shift; Doubly flexible; Efficient influence function; Label shift; Model misspecification; Semiparametric statistics

1. Introduction

In studies ranging from clinical medicine to policy research, there often exist situations where data and information are collected from a population $\mathcal{P}$, while the study interest lies in a target population $\mathcal{Q}$ which is related to but different from $\mathcal{P}$. For instance, in a clinical trial setting, physicians may need to use findings from a randomized trial on a set of patients ($\mathcal{P}$) whose demographics and comorbidities are different from those of their own patients ($\mathcal{Q}$). As another example, to build a predictive model on pneumonia outbreak for the flu season ($\mathcal{Q}$), researchers might find a similar model during the non-flu season ($\mathcal{P}$) relevant and useful. In these scenarios, there is a discrepancy between the distributions of $\mathcal{P}$ and $\mathcal{Q}$; that is, distribution shift. Distribution shift also refers to the scenario where the distribution of the training sample is different from that of the testing sample. In all these situations, it is of vital interest to develop methods that can appropriately leverage the information from $\mathcal{P}$ to the statistical tasks for $\mathcal{Q}$. Often, both outcome (output, response, label) Y and covariate (input, predictor, feature) $\mathbf{X}$ are available from $\mathcal{P}$ while only $\mathbf{X}$ is available from $\mathcal{Q}$. This setting is also named unsupervised domain adaptation (Quinonero-Candela et al. 2008; Moreno-Torres et al. 2012; Kouw and Loog 2021).

Without any assumptions on the shift, it is certainly impossible to leverage information between two heterogeneous populations. Two types of distribution shifts have been defined in the literature. The first is called covariate shift, where the shift happens between the marginal distributions of $\mathbf{X}$ while the conditional distribution of Y given $\mathbf{X}$ does not change; that is, $p_{\mathbf{X}}(\mathbf{x}) \neq q_{\mathbf{X}}(\mathbf{x})$ and $p_{Y|\mathbf{X}}(y,\mathbf{x}) = q_{Y|\mathbf{X}}(y,\mathbf{x})$. The difference between $\mathcal{P}$ and $\mathcal{Q}$ can be summarized as a density ratio $q_{\mathbf{X}}(\mathbf{x})/p_{\mathbf{X}}(\mathbf{x})$, which is fortunately estimable since covariate $\mathbf{X}$ is available from both populations. Covariate shift aligns with the causal learning setting (Schölkopf et al. 2012) where $\mathbf{X}$ is the cause and Y is the effect. Covariate shift has attracted much attention and has been investigated in many literatures, such as Shimodaira (2000), Huang et al. (2006), Sugiyama et al. (2008), Gretton et al. (2009), Sugiyama and Kawanabe (2012), Kpotufe and Martinet (2021) and the references therein.

The second type, which is the focus of this article, is named label shift, where it assumes that the shift is in the marginal distributions of Y while the process of generating $\mathbf{X}$ given Y is identical in both populations. Specifically, it assumes

$$p_Y(y) \neq q_Y(y), \text{ and } p_{\mathbf{X}|Y}(\mathbf{x},y) = q_{\mathbf{X}|Y}(\mathbf{x},y) \equiv g(\mathbf{x},y).$$

CONTACT Jiwei Zhao jiwei.zhao@wisc.edu Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53726.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.
© 2024 American Statistical Association

Label shift is also called prior probability shift (Storkey 2009; Tasche 2017), target shift (Zhang et al. 2013; Nguyen, Christoffel, and Sugiyama 2016), or class prior change (Du Plessis and Sugiyama 2014; Iyer, Nath, and Sarawagi 2014). Label shift aligns with the anticausal learning setting in which the outcome Y causes the covariate $\mathbf{X}$; for example, diseases cause symptoms or objects cause sensory observations. Consider the situation that one fits a model to predict whether a patient has pneumonia based on observed symptoms, and this model predicts reliably when deployed in the clinic during the non-flu season. When the flu season starts, there is a sudden surge of pneumonia cases and the probability of developing pneumonia rises, while the mechanism of showing symptoms determined by the pneumonia status is rather stable. Label shift also exists in computer vision applications, such as predicting object locations and directions, and human poses; see Martinez et al. (2017), Yang et al. (2018), and Guo et al. (2020).

In the label shift framework, one fundamental problem (Garg et al. 2020) is determining whether the shift has occurred and estimating the label distribution $q_Y(y)$, or equivalently, assessing the density ratio $q_Y(y)/p_Y(y) \equiv \rho(y)$. In contrast to estimating the density ratio $q_{\mathbf{X}}(\mathbf{x})/p_{\mathbf{X}}(\mathbf{x})$ under covariate shift, estimating $\rho(y)$ is a daunting task due to the absence of Y-observations in population $\mathcal{Q}$. Works under label shift are mainly limited to classification problems in the machine learning literature; see detailed review in Section 2. In this article, we take a unique approach by devising a methodology that can be applied to both discrete and continuous Y. We consider estimating a characteristic of the population $\mathcal{Q}$. Specifically, we estimate $\boldsymbol{\theta}$ such that $\mathrm{E}_q\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})\} = \mathbf{0}$ where $\mathbf{U}(\cdot)$ is a user-specified function and $\mathrm{E}_q(\cdot)$ stands for the expectation with respect to $q_Y(y)g(\mathbf{x}, y)$ or equivalently to $q_{Y|\mathbf{X}}(y, \mathbf{x})q_{\mathbf{X}}(\mathbf{x})$. This is a general framework including estimating the mean of Y and the tth quantile of Y as special cases. According to how the nuisance components are estimated, we propose various estimators for $\boldsymbol{\theta}$ and develop large sample theory for these estimators to quantify the estimation uncertainties and to conduct statistical inference.

To estimate $\boldsymbol{\theta}$, three nuisance components are involved. The first is the density ratio $\rho(y)$ which is infeasible to estimate based on the observed data, due to the lack of Y-observations in $\mathcal{Q}$. Our intention is to bypass the challenging task of estimating $\rho(y)$. This turns out to be achievable through careful manipulation of the influence function. In fact, a unique feature of our work is that we do not need to estimate $\rho(y)$. Instead, only a working model, denoted as $\rho^*(y)$, is needed. The second one is $p_{Y|\mathbf{X}}(y, \mathbf{x})$ or $\mathrm{E}_p(\cdot \mid \mathbf{x})$. In contrast to $\rho(y)$, estimating $\mathrm{E}_p(\cdot \mid \mathbf{x})$ is blessed with the observed data from $\mathcal{P}$. Indeed, we can use off-the-shelf machine learning methods or nonparametric regression methods to obtain the corresponding estimator $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$. Nevertheless, we have the option to abstain from estimating $\mathrm{E}_p(\cdot \mid \mathbf{x})$ despite our capability to do so. This means that we can misspecify the conditional distribution $p_{Y|\mathbf{X}}(y, \mathbf{x})$ while we also misspecify the density ratio $\rho(y)$. We call such an estimator $\widehat{\boldsymbol{\theta}}$ doubly flexible—the working density ratio model $\rho^*(y)$ is flexible, so is the working conditional distribution model $p^{\star}_{Y|\mathbf{X}}(y, \mathbf{x})$. Note that our superscripts here are different: superscript * is for the working model of the density ratio whereas * is for the conditional distribution model. This double flexibility is more favorable than the classic "double robustness" in the literature. The standard double robustness means that one can misspecify either one of two models but not both, while here we can misspecify both models. As an alternative, if one chooses to estimate $\mathrm{E}_p(\cdot \mid \mathbf{x})$, say, $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$, we name the corresponding estimator $\widetilde{\boldsymbol{\theta}}$ singly flexible—only flexible in working model $\rho^*(y)$. The third nuisance is the conditional density function $g(\mathbf{x}, y)$ whose estimation might be subject to the curse of dimensionality. Fortunately, in our estimation procedure, $g(\mathbf{x}, y)$ only affects quantities of the form $\mathrm{E}(\cdot \mid y)$, which are one dimensional regression problems hence can be easily solved via basic nonparametric regression tools such as the Nadaraya-Watson estimator.

Our approach to addressing the label shift problem is based on the idea of minimizing the influence of nuisance components on the target quantity of interest. This is accomplished by scrutinizing influence functions and employing orthogonal projections. Even with these projections in place, there will still be a residual effect inherent to the problem itself that cannot be completely eliminated. Consequently, we conduct a focused analysis of this residual effect and develop procedures to effectively manage its impact. To facilitate mathematical derivations, we typically rely on common assumptions related to smoothness, boundedness, and non-singularity, which are not restrictive in practice. We will explicitly articulate these assumptions as regularity conditions when we present our theoretical results.

The remainder of the article is structured as follows. We first provide a thorough literature review on label shift and relevant semiparametric techniques in Section 2. In Section 3, we outline our strategy of incorporating samples from two heterogeneous populations. The proposed doubly flexible estimator is presented in Section 4, and the alternative singly flexible estimator in Section 5. For easier understanding and improved readability, we present both methodologies and theories for a special parameter $\theta = \mathrm{E}_q(Y)$ in the main text, while defer the results for a general $\boldsymbol{\theta}$ such that $\mathrm{E}_q\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})\} = \mathbf{0}$ to the supplementary materials. Section 6 contains empirical results from extensive simulation studies. We present an application to the MIMIC-III database in Section 7. The article is concluded with discussions in Section 8. All the technical details are included in the supplementary materials.

2. Related Work

Within the label shift framework, the majority of research has been concentrated on classification in machine learning. Saerens, Latinne, and Decaestecker (2002) proposed a simple Expectation-Maximization (EM) (Dempster, Laird, and Rubin 1977) procedure, named maximum likelihood label shift (MLLS), to estimate $q_Y(y)$ assuming access to a classifier that outputs the true conditional probabilities of the population $\mathcal{P}$, $p_{Y|\mathbf{X}}(y, \mathbf{x})$. Later on, Chan and Ng (2005) proposed an EM algorithm that requires the estimation of $g(\mathbf{x}, y)$, which is unfortunately difficult for high-dimensional $\mathbf{X}$ and moreover, it does not apply to regression problems. Alternatively, Lipton, Wang, and Smola (2018) and Azizzadenesheli et al. (2019) proposed moment-matching based estimators, named

black box shift learning (BBSL) and regularized learning under label shift (RLLS), that make use of the invertible confusion matrix of a classifier learned from population $\mathcal{P}$. The connection and comparison of these two lines of research, either empirical or theoretical, remain unclear. To our best knowledge, neither BBSL nor RLLS has been benchmarked against EM. Alexandari, Kundaje, and Shrikumar (2020) showed that, in combination with a calibration named bias-corrected temperature scaling, MLLS outperforms BBSL and RLLS empirically; whereas MLLS underperforms BBSL when applied naively. Under label shift, Maity, Sun, and Banerjee (2020) also studied the minimax rate of convergence for nonparametric classification.

For continuous Y in regression, estimating $q_Y(y)$ becomes the problem of estimating a function instead of a finite number of parameters. Not surprisingly, its literature is quite scarce. Zhang et al. (2013) proposed a nonparametric method to estimate the density ratio by kernel mean matching of distributions; however, it does not scale to large data as the computational cost is quadratic in the sample size. Nguyen, Christoffel, and Sugiyama (2016) considered continuous label shift adaptation and studied an importance weight estimator, but their approach relies on a parametric model for $p_{Y|\mathbf{X}}(y,\mathbf{x})$ hence can only be applied in supervised learning.

Our primary focus in this article is on doubly flexible estimation, taking into consideration the possibility of misspecification in nuisance functions, particularly the density ratio $\rho(y)$. We stress that the superiority of our method compared to the previous literature, such as Lipton, Wang, and Smola (2018) and Azizzadenesheli et al. (2019), can still be demonstrated even in the classification setting where the misspecification of $\rho(y)$ might not be an issue. Existing methods primarily concentrate on developing various estimators for $\rho(y)$ and subsequently constructing consistent estimators for the target quantity, often without a thorough analysis of the associated estimation efficiency. In contrast, our estimator would achieve the semiparametric efficiency bound when the model misspecification issue is absent. Interested readers can find a simulation study in Section 6.2 and further discussion in Section S.11 of the supplementary materials.

The judicious application of semiparametric techniques plays a pivotal role in our approach. Semiparametric theory was established initially in Bickel et al. (1993), and popularized by van der Laan and Robins (2003) and Tsiatis (2006) in causal inference, missing data analysis and related fields. Some recent research on missing data problems that takes advantage of semiparametric theory includes Zhao and Ma (2022) and Li, Miao, and Tchetgen (2021), yet none of those is directly relevant to the problem we address in this article.

3. Model Structure

We consider independent and identically distributed (iid) observations $\{Y_i, \mathbf{X}_i\}, i = 1,\ldots,n_1$ from population $\mathcal{P}$, and iid observations $\mathbf{X}_j, j = n_1+1,\ldots,n_1+n_0$ from population $\mathcal{Q}$. To use the information in population $\mathcal{P}$ under label shift, we stack the two random samples together and assemble a new dataset of size $n = n_1 + n_0$, which represents a random sample for an imaginary population consisting of $100\pi\%$ population $\mathcal{P}$ members and $100(1-\pi)\%$ population $\mathcal{Q}$ members. Here we define $\pi \equiv n_1/n$. Throughout our derivation, other than $\mathrm{E}_p(\cdot)$ and $\mathrm{E}_q(\cdot)$, we also compute $\mathrm{E}(\cdot)$ that is with respect to this imaginary population; however, this imaginary population is only used as an intermediate tool to leverage information from two heterogeneous populations under label shift. Our final conclusion will only be made for the target population $\mathcal{Q}$.

For convenience, we introduce a binary indicator R in this stacked random sample, where $R = 1$ means the subject is from population $\mathcal{P}$ and $R = 0$ population $\mathcal{Q}$. Thus, the likelihood of one observation from the stacked random sample is

$$\{g(\mathbf{x},y)p_Y(y)\}^r \left\{\int g(\mathbf{x},y)q_Y(y)\mathrm{d}y\right\}^{1-r} \pi^r(1-\pi)^{1-r} \quad (1)$$

$$= \{g(\mathbf{x},y)p_Y(y)\}^r \left\{\int g(\mathbf{x},y)\rho(y)p_Y(y)\mathrm{d}y\right\}^{1-r} \pi^r(1-\pi)^{1-r}. \quad (2)$$

Although $g(\mathbf{x},y)$ and $p_Y(y)$ can be identified from (1), unfortunately $q_Y(y)$ may not. Below is a simple example illustrating the possible nonidentifiability of $q_Y(y)$.

Example 1. Consider a discrete Y with three supporting values $0,1,2$ and a discrete X with two supporting values $0,1$. In both populations $\mathcal{P}$ and $\mathcal{Q}$, $g(x,y)$ is given as $\mathrm{pr}(X=0 \mid Y=0) = 1/5$, $\mathrm{pr}(X=0 \mid Y=1) = 1/8$, and $\mathrm{pr}(X=0 \mid Y=2) = 2/3$. In population $\mathcal{P}$, the marginal distribution of Y, $p_Y(y)$, is given as $\mathrm{pr}(Y=0) = 5/16, \mathrm{pr}(Y=1) = 1/2$, and $\mathrm{pr}(Y=2) = 3/16$. In population $\mathcal{Q}$, the marginal distribution of Y, $q_Y(y)$, is given as $\mathrm{pr}(Y=0) = \frac{5(25-416t)}{336}$, $\mathrm{pr}(Y=1) = \frac{32t-1}{6}$, and $\mathrm{pr}(Y=2) = \frac{89+96t}{112}$, where $t \in (1/32, 25/336)$. Clearly this satisfies the label shift assumption, and the marginal distribution of X in population $\mathcal{Q}$ is identifiable since $\mathrm{pr}(X=0) = 7/12$ is free of t; however, the marginal distribution of Y in population $\mathcal{Q}$, $q_Y(y)$, is not identifiable.

The following result demonstrates that the completeness condition on $p_{Y|\mathbf{X}}(y,\mathbf{x})$ would ensure the identifiability. Its proof is contained in Section S.1 of the supplementary materials.

Lemma 1. If the conditional pdf/pmf $p_{Y|\mathbf{X}}(y,\mathbf{x})$ of population $\mathcal{P}$ satisfies the completeness condition in the sense that, for any function $h(Y)$ with finite mean, $\mathrm{E}_p\{h(Y) \mid \mathbf{X}\} = \int h(y)p_{Y|\mathbf{X}}(y,\mathbf{x})\mathrm{d}y = 0$ implies $h(Y) = 0$ almost surely, then all the unknown components in (2), that is, $g(\mathbf{x},y)$, $p_Y(y)$ and $\rho(y)$, are identifiable. Subsequently, $q_Y(y)$ is also identifiable.

The completeness condition in Lemma 1 is mild and has been widely assumed in instrumental variables, measurement error models, and econometrics; see, for example, Newey and Powell (2003), d'Haultfoeuille (2011), and Hu and Shiu (2018). Because the condition is imposed on population $\mathcal{P}$ and we have random observations $(\mathbf{X}_i, Y_i)$'s from population $\mathcal{P}$, it can be examined and verified in empirical studies. One can easily check that many commonly used distributions such as exponential families satisfy the completeness condition. In particular, if the outcome Y is discrete with finitely many supporting values, Newey and Powell (2003) pointed out that the completeness condition only means that the covariate $\mathbf{X}$ has a support whose cardinality is no smaller than that of Y.

In this article, we focus on estimating a characteristic of population $\mathcal{Q}$. For better clarity, we will present the results for $\theta = \mathrm{E}_q(Y) = \int yg(\mathbf{x}, y)q_Y(y)\mathrm{d}\mathbf{x}\mathrm{d}y$ in the main text, then generalize the results to $\boldsymbol{\theta}$ that satisfies $\mathrm{E}_q\{\mathbf{U}(Y, \mathbf{X}, \boldsymbol{\theta})\} = \mathbf{0}$ in the supplementary materials. The main challenge in estimating θ is caused by the lack of knowledge and data on $q_Y(y)$, or equivalently, $\rho(y)$. Nevertheless, we will construct an estimator that bypasses the difficulty of assessing $\rho(y)$. We will show that we only need a working model of $\rho(y)$, denoted as $\rho^*(y)$, that can be flexible. Furthermore, we find that our procedure can also simultaneously avoid estimating $p_{Y|\mathbf{X}}(y, \mathbf{x})$, in that we can insert a possibly misspecified working model $p^{\star}_{Y|\mathbf{X}}(y, \mathbf{x})$. Thus, our procedure is flexible with respect to both $\rho(y)$ and $p_{Y|\mathbf{X}}(y, \mathbf{x})$—doubly flexible. This is a property different from the classic "double robustness" which means that one can only misspecify one of two models but not both. In contrast, here, we can misspecify both.

4. Proposed Doubly Flexible Estimation for $\theta = \mathrm{E}_q(Y)$

If the density ratio function $\rho(y)$ were known, an intuitive estimator of $\theta = \mathrm{E}_q(Y)$ can be created by noticing the relation $\theta = \mathrm{E}_q(Y) = \mathrm{E}_p\{\rho(Y)Y\} = \mathrm{E}\{R\rho(Y)Y\}/\pi$; that is,

$$\breve{\theta} = \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\pi}\rho(y_i)y_i. \tag{3}$$

We call this estimator shift-dependent since it requires the correct specification of $\rho(y)$. Clearly, if a working model $\rho^*(y)$ is adopted, the corresponding estimator $\breve{\theta}^*$ is likely biased.

4.1. General Approach to Estimating θ

The creation of an estimator that is not solely shift-dependent is possible. To motivate our proposed estimator, we first make some simple observations via balancing the samples from populations $\mathcal{P}$ and $\mathcal{Q}$. Recognizing the relation between $\mathrm{E}_p(\cdot)$, $\mathrm{E}_q(\cdot)$ and $\mathrm{E}(\cdot)$, the balancing of Y is $\mathrm{E}\left\{\frac{R}{\pi}\rho(Y)Y\right\} = \mathrm{E}_p\{\rho(Y)Y\} = \mathrm{E}_q(Y) = \mathrm{E}\left(\frac{1-R}{1-\pi}\theta\right)$. Further, replacing the variable Y above by an arbitrary function of $\mathbf{X}$, we obtain another balancing function $\mathrm{E}\left\{\frac{R}{\pi}\rho(Y)b(\mathbf{X})\right\} = \mathrm{E}\left\{\frac{1-R}{1-\pi}b(\mathbf{X})\right\}$. Certainly, we also have $\mathrm{E}\left(\frac{R}{\pi}c\right) = \mathrm{E}\left(\frac{1-R}{1-\pi}c\right)$ for any constant c. Combining the above three, we can obtain a family of mean zero functions

$$\frac{r}{\pi}\{\rho(y)y - b(\mathbf{x})\rho(y) + c\} + \frac{1-r}{1-\pi}\{b(\mathbf{x}) - \theta - c\} : \forall b(\mathbf{x}), \forall c. \tag{4}$$

Note that the model in (2) contains three unknown functions $p_Y(y)$, $g(\mathbf{x}, y)$ and $\rho(y)$. For this model, in Section S.2 of the supplementary materials, we establish that

$$\mathcal{F} \equiv \left[\frac{r}{\pi}\{\rho(y)y - b(\mathbf{x})\rho(y) + c\} + \frac{1-r}{1-\pi}\{b(\mathbf{x}) - \theta - c\} : \mathrm{E}\{b(\mathbf{X}) \mid y\} = y, \forall c\right]$$

is the family of all influence functions (Bickel et al. 1993; Tsiatis 2006) for estimating θ. According to the definition of the influence function, $\mathcal{F}$ is sufficiently comprehensive since it can generate any regular asymptotically linear estimator of θ. The requirement $\mathrm{E}\{b(\mathbf{X}) \mid y\} = y$ in the definition of $\mathcal{F}$ is pivotal. Different from the mean zero function in (4), which critically relies on the correct specification of $\rho(y)$, the element in $\mathcal{F}$ preserves its zero mean even if $\rho(y)$ is misspecified as long as an appropriate $b(\mathbf{x})$ is chosen so that $\mathrm{E}\{b(\mathbf{X}) \mid y\} = y$. To further discover a wise choice of such a $b(\mathbf{x})$, we first derive a special element in $\mathcal{F}$, the efficient influence function $\phi_{\mathrm{eff}}(\mathbf{x}, r, ry)$, that corresponds to the semiparametric efficiency bound and that provides guidance on constructing flexible estimators for θ.

Proposition 1. The efficient influence function $\phi_{\mathrm{eff}}(\mathbf{x}, r, ry)$ for θ is

$$\begin{aligned}&\phi_{\mathrm{eff}}(\mathbf{x}, r, ry)\\ &= \frac{r}{\pi}\rho(y)\left[y - \frac{\mathrm{E}_p\{a(Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y) \mid \mathbf{x}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho(Y) \mid \mathbf{x}\}}\right]\\ &\quad + \frac{1-r}{1-\pi}\left[\frac{\mathrm{E}_p\{a(Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y) \mid \mathbf{x}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho(Y) \mid \mathbf{x}\}} - \theta\right],\end{aligned}$$

where $a(y)$ satisfies

$$\mathrm{E}\left[\frac{\mathrm{E}_p\{a(Y)\rho(Y) \mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^2(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho(Y) \mid \mathbf{X}\}} \mid y\right] = y. \tag{5}$$

The detailed derivation of Proposition 1 is provided in Section S.3 of the supplementary materials. Clearly, the unique $b(\mathbf{x})$ that leads to the efficient influence function is

$$\begin{aligned}b(\mathbf{x}) &\equiv \frac{\mathrm{E}_p\{a(Y)\rho(Y) \mid \mathbf{x}\}}{\mathrm{E}_p\{\rho^2(Y) \mid \mathbf{x}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho(Y) \mid \mathbf{x}\}}\\ &= \frac{\mathrm{E}_q\{a(Y) \mid \mathbf{X}\}}{\mathrm{E}_q\{\rho(Y) \mid \mathbf{X}\} + \pi/(1-\pi)}.\end{aligned}$$

In principle, if both $\rho(y)$ and $b(\mathbf{x})$ were known, we can estimate θ by solving the estimating equation $\sum_{i=1}^{n}\phi_{\mathrm{eff}}(\mathbf{x}_i, r_i, r_iy_i) = 0$, which leads to

$$\breve{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{r_i}{\pi}\rho(y_i)\{y_i - b(\mathbf{x}_i)\} + \frac{1-r_i}{1-\pi}b(\mathbf{x}_i)\right]. \tag{6}$$

However, the estimator $\breve{\theta}$ is impractical because of the following three obstacles. First, as we pointed out, $\rho(y)$ is almost infeasible to estimate based on the observed data. Second, $\mathrm{E}_p(\cdot \mid \mathbf{x})$ is unknown and needs to be estimated. Though various off-the-shelf machine learning or nonparametric regression methods are available, when the dimension of $\mathbf{x}$ is high, their performances are not always satisfactory and their computation can be expensive. The third obstacle lies in solving $a(y)$ from the integral equation (5), which requires $g(\mathbf{x}, y)$ to evaluate its left hand side. Estimating conditional density $g(\mathbf{x}, y)$ could be even more difficult than estimating $\mathrm{E}_p(\cdot \mid \mathbf{x})$, due to the curse of dimensionality.

Our proposed estimator will bypass the challenging task of estimating $\rho(y)$. Throughout the estimation procedure, only a working model $\rho^*(y)$ is needed, which can be arbitrarily misspecified hence is flexible. This turns out achievable through

careful manipulation of other components of the efficient influence function. Our proposed estimator can also avoid estimating $\mathrm{E}_p(\cdot \mid \mathbf{x})$ even though we can do it if we decide to. This means that we can misspecify the conditional density model $p_{Y|\mathbf{X}}(y, \mathbf{x})$, encoded as $p^\star_{Y|\mathbf{X}}(y, \mathbf{x})$, while we also misspecify the density ratio $\rho(y)$. We call such an estimation procedure doubly flexible. To overcome the third obstacle, we recognize that $g(\mathbf{x}, y)$ only affects quantities of the form $\mathrm{E}(\cdot \mid y)$, which are one dimensional regression problems hence can be easily solved via the most basic nonparametric regression procedure such as the Nadaraya-Watson estimator.

In a nutshell, a unique feature of our work is the tolerance of both $\rho^*(y)$ and $p^\star_{Y|\mathbf{X}}(y, \mathbf{x})$, which can be simultaneously misspecified. We thus name the procedure doubly flexible.

4.2. Proposed Estimator $\widehat{\theta}$: Doubly Flexible in $\rho^*(y)$ and $p^\star_{Y|\mathbf{X}}(y, \mathbf{x})$

Interestingly and critically, we discover that, even when both $\rho^*(y)$ and $p^\star_{Y|\mathbf{X}}(y, \mathbf{x})$ are misspecified, the corresponding estimator following the implementation of $\breve{\theta}$ in (6) is still consistent for θ. We summarize this result in Proposition 2 and give its proof in Section S.4 of the supplementary materials. Below, we use superscripts * and * to indicate that the corresponding quantities are calculated based on the working models $\rho^*(y)$ and $p^\star_{Y|\mathbf{X}}(y, \mathbf{x})$, respectively.

Proposition 2. Define $\widehat{\theta}_t = \frac{1}{n}\sum_{i=1}^n \left[\frac{r_i}{\pi}\rho^*(y_i)\{y_i - b^{*\star}(\mathbf{x}_i)\} + \frac{1-r_i}{1-\pi}b^{*\star}(\mathbf{x}_i)\right]$, where

$$b^{*\star}(\mathbf{x}) \equiv \frac{\mathrm{E}^\star_p\{a^{*\star}(Y)\rho^*(Y) \mid \mathbf{x}\}}{\mathrm{E}^\star_p\{\rho^{*2}(Y) \mid \mathbf{x}\} + \pi/(1-\pi)\mathrm{E}^\star_p\{\rho^*(Y) \mid \mathbf{x}\}}, \text{ and}$$

$a^{*\star}(y)$ is a solution to

$$\mathrm{E}\left[\frac{\mathrm{E}^\star_p\{a^{*\star}(Y)\rho^*(Y) \mid \mathbf{X}\}}{\mathrm{E}^\star_p\{\rho^{*2}(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\mathrm{E}^\star_p\{\rho^*(Y) \mid \mathbf{X}\}} \mid y\right] = y. \quad (7)$$

Then $\widehat{\theta}_t$ is a consistent estimator of θ.

In Proposition 2, the subscript $_t$ in $\widehat{\theta}_t$ indicates the conditional density $g(\mathbf{x}, y)$ in (7) is the truth. In reality, note that $g(\mathbf{x}, y)$ only involves in the evaluation of the conditional expectation $\mathrm{E}(\cdot \mid y)$ on the left hand side of (7). This is a one dimensional regression problem and can be easily estimated by many basic nonparametric regression procedures such as the Nadaraya-Watson estimator. Specifically, we approximate the integral equation (7) by

$$\begin{aligned} y &= \widehat{\mathrm{E}}\left[\frac{\mathrm{E}^\star_p\{a^{*\star}(Y)\rho^*(Y) \mid \mathbf{X}\}}{\mathrm{E}^\star_p\{\rho^{*2}(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\mathrm{E}^\star_p\{\rho^*(Y) \mid \mathbf{X}\}} \mid y\right] \\ &= \sum_{i=1}^n \frac{\mathrm{E}^\star_p\{a^{*\star}(Y)\rho^*(Y) \mid \mathbf{x}_i\}}{\mathrm{E}^\star_p\{\rho^{*2}(Y) \mid \mathbf{x}_i\} + \pi/(1-\pi)\mathrm{E}^\star_p\{\rho^*(Y) \mid \mathbf{x}_i\}} \\ &\quad \times \frac{r_i K_h(y - y_i)}{\sum_{j=1}^n r_j K_h(y - y_j)} \\ &= \int a^{*\star}(t)\rho^*(t)\sum_{i=1}^n \\ &\quad \times \frac{p^\star_{Y|\mathbf{X}}(t, \mathbf{x}_i)}{\mathrm{E}^\star_p\{\rho^{*2}(Y) \mid \mathbf{x}_i\} + \pi/(1-\pi)\mathrm{E}^\star_p\{\rho^*(Y) \mid \mathbf{x}_i\}} \\ &\quad \times \frac{r_i K_h(y - y_i)}{\sum_{j=1}^n r_j K_h(y - y_j)}\mathrm{d}t, \end{aligned} \quad (8)$$

where $K_h(\cdot) \equiv K(\cdot/h)/h$, $K(\cdot)$ is a kernel function and h is a bandwidth, with conditions imposed later in our theoretical investigation. Equation (8) is a Fredholm integral equation of the first type, which is ill-posed. Numerical methods to provide stable and reliable solutions have been well studied in the literature (Hansen 1992). In our numerical implementations in Sections 6 and 7, we use Landweber's iterative method (Landweber 1951) that is well-known to produce a convergent solution. We provide those technical details in Section S.5 of the supplementary materials. We summarize the complete estimation procedure in Algorithm 1.

Algorithm 1 Proposed Estimator $\widehat{\theta}$: Doubly Flexible in $\rho^*(y)$ and $p^\star_{Y|\mathbf{X}}(y, \mathbf{x})$

Input: data from population $\mathcal{P}$: $(y_i, \mathbf{x}_i, r_i = 1)$, $i = 1, \ldots, n_1$, data from population $\mathcal{Q}$: $(\mathbf{x}_j, r_j = 0)$, $j = n_1 + 1, \ldots, n$, and value $\pi = n_1/n$.

do

(a) adopt a working model for $\rho(y)$, denoted as $\rho^*(y)$;

(b) adopt a working model for $p_{Y|\mathbf{X}}(y, \mathbf{x})$, denoted as $p^\star_{Y|\mathbf{X}}(y, \mathbf{x})$ or $p^\star_{Y|\mathbf{X}}(y, \mathbf{x}, \widehat{\boldsymbol{\zeta}})$;

(c) compute $w_i = [\mathrm{E}^\star_p\{\rho^{*2}(Y) \mid \mathbf{x}_i\} + \pi/(1-\pi)\mathrm{E}^\star_p\{\rho^*(Y) \mid \mathbf{x}_i\}]^{-1}$ for $i = 1, \ldots, n$;

(d) obtain $\widehat{a}^{*\star}(\cdot)$ by solving the integral equation (8);

(e) compute $\widehat{b}^{*\star}(\mathbf{x}_i) = w_i\mathrm{E}^\star_p\{\widehat{a}^{*\star}(Y)\rho^*(Y) \mid \mathbf{x}_i\}$ for $i = 1, \ldots, n$;

(f) obtain $\widehat{\theta}$ as

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n \left[\frac{r_i}{\pi}\rho^*(y_i)\{y_i - \widehat{b}^{*\star}(\mathbf{x}_i)\} + \frac{1-r_i}{1-\pi}\widehat{b}^{*\star}(\mathbf{x}_i)\right]. \quad (9)$$

Output: $\widehat{\theta}$.

Remark 1. In step (a) of Algorithm 1, one may adopt an arbitrary $\rho^*(y)$ as long as it is a ratio of two pdfs/pmfs and satisfies $\mathrm{E}\{R\rho^*(Y)\} = \pi$. The most convenient choice is $\rho^*(y) = 1$ on the support of $p_Y(y)$. If one desires to explore more complex alternatives to numerically investigate the impact of the choice on the resulting estimator, we suggest to first adopt some positive function $\widetilde{\rho}(y)$, for example, $\widetilde{\rho}(y) = \exp(a + by)$ on the support of $p_Y(y)$, then normalize it by multiplying $c^* \equiv \pi/\{n^{-1}\sum_{i=1}^n r_i\widetilde{\rho}(y_i)\}$, that is, $\rho^*(y) = c^*\widetilde{\rho}(y)$. This is a valid choice and was used in our simulation experiments.

Remark 2. In step (b) of Algorithm 1, one may adopt a completely specified $p^\star_{Y|\mathbf{X}}(y, \mathbf{x})$ or a partially specified model $p^\star_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\zeta})$ with an unknown parameter $\boldsymbol{\zeta}$. If the latter case, a natural strategy is to estimate $\boldsymbol{\zeta}$ first based on the observed samples from $\mathcal{P}$ via, say MLE, to obtain $\widehat{\boldsymbol{\zeta}}$, then use $p^\star_{Y|\mathbf{X}}(y, \mathbf{x}, \widehat{\boldsymbol{\zeta}})$ to replace the completely fixed $p^\star_{Y|\mathbf{X}}(y, \mathbf{x})$. In fact, we will show that the action of estimating $\boldsymbol{\zeta}$ has no consequence in terms of estimating θ. This is an important discovery, because this means

one can always include a reasonably flexible model $p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\zeta})$ so that it has a good chance of approximating the true $p_{Y|\mathbf{X}}(y,\mathbf{x})$. If $p_{Y|\mathbf{X}}(y,\mathbf{x}) = p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\zeta}_0)$ for certain $\boldsymbol{\zeta}_0$, then even though the additional parameter $\boldsymbol{\zeta}$ causes extra work, the reward is that θ can be estimated as efficiently as if we knew $p_{Y|\mathbf{X}}(y,\mathbf{x})$ completely. In all the subsequent steps, we replace $p^\star_{Y|\mathbf{X}}(y,\mathbf{x})$ by $p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$ for its generality, bearing in mind that $p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$ degenerates to $p^\star_{Y|\mathbf{X}}(y,\mathbf{x})$ when the parameter $\boldsymbol{\zeta}$ vanishes.

We now study the theoretical properties of $\widehat{\theta}$ defined in (9). The main technical challenge is quantifying the gap between the solutions for the integral equations (7) and (8), encoded as $a^{*\star}(y)$ and $\widehat{a}^{*\star}(y)$, respectively. To facilitate the derivation, we define the linear operator

$$\begin{aligned}
&\mathcal{L}^{*\star}(a)(y)\\
&\equiv p_Y(y)\mathrm{E}\left[\frac{\mathrm{E}^\star_p\{a(Y)\rho^*(Y)\mid\mathbf{X}\}}{\mathrm{E}^\star_p\{\rho^{*2}(Y)\mid\mathbf{X}\}+\pi/(1-\pi)\mathrm{E}^\star_p\{\rho^*(Y)\mid\mathbf{X}\}}\mid y\right]\\
&=\int a(t)u^{*\star}(t,y)\mathrm{d}t,\ \text{where}\\
&u^{*\star}(t,y)\equiv p_Y(y)\\
&\int\frac{\rho^*(t)p^\star_{Y|\mathbf{X}}(t,\mathbf{x},\boldsymbol{\zeta})}{\mathrm{E}^\star_p\{\rho^{*2}(Y)\mid\mathbf{x}\}+\pi/(1-\pi)\mathrm{E}^\star_p\{\rho^*(Y)\mid\mathbf{x}\}}g(\mathbf{x},y)\mathrm{d}\mathbf{x}.
\end{aligned}$$

Apparently, $a^{*\star}(y)$ satisfies $\mathcal{L}^{*\star}(a^{*\star})(y) = v(y)$, where $v(y)\equiv p_Y(y)y$. Similarly, $\widehat{a}^{*\star}(y)$ satisfies $\widehat{\mathcal{L}}^{*\star}(\widehat{a}^{*\star})(y) = \widehat{v}(y)$, where

$$\begin{aligned}
&\widehat{\mathcal{L}}^{*\star}(a)(y)\equiv n_1^{-1}\sum_{i=1}^n r_iK_h(y-y_i)\\
&\qquad\frac{\mathrm{E}^\star_p\{a(Y)\rho^*(Y)\mid\mathbf{x}_i,\widehat{\boldsymbol{\zeta}}\}}{\mathrm{E}^\star_p\{\rho^{*2}(Y)\mid\mathbf{x}_i,\widehat{\boldsymbol{\zeta}}\}+\pi/(1-\pi)\mathrm{E}^\star_p\{\rho^*(Y)\mid\mathbf{x}_i,\widehat{\boldsymbol{\zeta}}\}},\ \text{and}\\
&\widehat{v}(y)\equiv n_1^{-1}\sum_{i=1}^n v_{i,h}(y)\equiv n_1^{-1}\sum_{i=1}^n r_iK_h(y-y_i)y.
\end{aligned}$$

We first establish in Lemma 2 that given regularity conditions (A1)-(A4), the linear operator $\mathcal{L}^{*\star}$, as well as its inverse, is well behaved.

(A1) The working model $p^\star_{Y|\mathbf{X}}(y,\mathbf{x})$ or $p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$ satisfies the completeness condition stated in Lemma 1.

(A2) $\rho^*(y) > \delta$ for all y on the support of $p_Y(y)$ where δ is a positive constant, and $\rho^*(y)$ is twice differentiable and its derivative is bounded.

(A3) The function $u^{*\star}(t,y)$ is bounded and has bounded derivatives with respect to t and y on its support. The function $a^{*\star}(y)$ in (7) is bounded.

(A4) The support sets of $g(\mathbf{x},y)$, $p_Y(y)$, $\rho^*(y)$ are compact.

Lemma 2. Let $\|a\|_\infty \equiv \sup_y|a(y)|$. Under Conditions (A1)-(A4), the linear operator $\mathcal{L}^{*\star}: L^\infty(R)\to L^\infty(R)$ is invertible. In addition, there exist positive finite constants c_1, c_2 such that for all $a(y)\in L^\infty(R)$, (i) $c_1\|a\|_\infty \le \|\mathcal{L}^{*\star}(a)\|_\infty \le c_2\|a\|_\infty$, and (ii) $\|\mathcal{L}^{*\star-1}(a)\|_\infty \le c_1^{-1}\|a\|_\infty$.

The proof of Lemma 2 is in Section S.6 of the supplementary materials. To analyze the asymptotic normality of the estimator $\widehat{\theta}$, we add two more regularity conditions on the kernel function and the bandwidth h.

(A5) The kernel function $K(\cdot)\ge 0$ is symmetric, bounded, and twice differentiable with bounded first derivative. It has support on $(-1,1)$ and satisfies $\int_{-1}^1 K(t)\mathrm{d}t = 1$.

(A6) The bandwidth h satisfies $n_1(\log n_1)^{-4}h^2 \to \infty$ and $n^2n_1^{-1}h^4\to 0$.

Condition (A5) is standard for kernel functions. Note that we only need a one-dimensional kernel function $K(\cdot)$ in our estimation procedure. Condition (A6) specifies the requirement of the bandwidth h associated with kernel function $K(\cdot)$. In general we need both $h^{-1} = o\{n_1^{1/2}(\log n_1)^{-2}\}$ and $h = o(n^{-1/2}n_1^{1/4})$. If π is further assumed to be bounded away from zero, the second requirement becomes $h = o(n_1^{-1/4})$ and one can simply choose $h = n_1^{-1/3}$ to meet both requirements. We are now ready to present the asymptotic normality of the estimator $\widehat{\theta}$ below. Its proof is contained in Section S.7 of the supplementary materials.

Theorem 1. Assume $\widehat{\boldsymbol{\zeta}}$ satisfies $\|\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$ and $\mathrm{E}^\star_p\{\|\mathbf{S}^\star_{\boldsymbol{\zeta}}(Y,\mathbf{x},\boldsymbol{\zeta})\|_2 \mid \mathbf{x}\}$ is bounded, where $\mathbf{S}^\star_{\boldsymbol{\zeta}}(y,\mathbf{x},\boldsymbol{\zeta})\equiv\partial\log p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\zeta})/\partial\boldsymbol{\zeta}$. For any choice of $p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\zeta})$ and $\rho^*(y)$, under Conditions (A1)-(A6),

$$\sqrt{n_1}(\widehat{\theta}-\theta)\to N(0,\sigma^2_\theta)$$

in distribution as $n_1\to\infty$, where σ^2_θ equals

$$\begin{aligned}
&\mathrm{var}\Bigg(\sqrt{\pi}\phi^{*\star}_{\mathrm{eff}}(\mathbf{X},R,RY)+\frac{R}{\sqrt{\pi}}\\
&\left[\frac{\mathrm{E}^\star_p\{a^{*\star}(Y)\rho^*(Y)\mid\mathbf{X}\}}{\mathrm{E}^\star_p\{\rho^{*2}(Y)\mid\mathbf{X}\}+\pi/(1-\pi)\mathrm{E}^\star_p\{\rho^*(Y)\mid\mathbf{X}\}}-Y\right]\\
&\{\rho^*(Y)-\rho(Y)\}\Bigg).
\end{aligned}$$

In Theorem 1, the only requirement on $\widehat{\boldsymbol{\zeta}}$ is $\|\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$. Thus, the asymptotic variance of $\sqrt{n_1}(\widehat{\boldsymbol{\zeta}}-\boldsymbol{\zeta})$ does not affect the result in Theorem 1 as long as $\widehat{\boldsymbol{\zeta}}$ is $\sqrt{n_1}$-consistent for $\boldsymbol{\zeta}$. This is easily achievable by constructing a standard MLE or moment based estimator for $\boldsymbol{\zeta}$ in the regression model of Y given $\mathbf{X}$, $p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\zeta})$, based on the n_1 observations from population $\mathcal{P}$. Also, Theorem 1 indicates that, instead of solving the exact equation (7), solving the approximate equation (8) does not affect $\widehat{\theta}$ in terms of its leading order asymptotic property. In other words, if one could solve (7) for $a^{*\star}(\cdot)$ and construct the corresponding estimator, it would have exactly the same asymptotic distribution as $\widehat{\theta}$ here, as long as the kernel function $K(\cdot)$ and the bandwidth h are appropriately chosen.

In addition, it is clear that the estimator $\widehat{\theta}$ is $\sqrt{n_1}$-consistent, even if n goes to infinity much faster than n_1 does. The intuition is that when we only have n_1 complete observations in this problem, although a much larger n_0 can help us better understand the label shift mechanism, it cannot improve the convergence rate of $\widehat{\theta}$.

Finally, Theorem 1 also indicates that, when $\widehat{\boldsymbol{\zeta}}$ is a $\sqrt{n_1}$-consistent estimator for $\boldsymbol{\zeta}$ such that $p^\star_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\zeta}) = p_{Y|\mathbf{X}}(y,\mathbf{x})$, and $\rho^*(y)$ is correctly specified as $\rho(y)$, the corresponding estimator $\widehat{\theta}$ achieves the semiparametric efficiency bound and is the efficient estimator. We state this result formally as Corollary 1. Since it is a special case of Theorem 1, its proof is omitted.

Corollary 1. Assume $\widehat{\boldsymbol{\zeta}}$ satisfies $\|\widehat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}\|_2 = O_p(n_1^{-1/2})$ and $\mathrm{E}_p^\star\{\|\mathbf{S}_{\boldsymbol{\zeta}}^\star(Y, \mathbf{x}, \boldsymbol{\zeta})\|_2 \mid \mathbf{x}\}$ is bounded. If $p_{Y|\mathbf{X}}^\star(y, \mathbf{x}, \boldsymbol{\zeta}) = p_{Y|\mathbf{X}}(y, \mathbf{x})$ and $\rho^*(y) = \rho(y)$, under Conditions (A1)-(A6),

$$\sqrt{n_1}(\widehat{\theta}_{\mathrm{eff}} - \theta) \to N[0, \mathrm{var}\{\sqrt{\pi}\phi_{\mathrm{eff}}(\mathbf{X}, R, RY)\}]$$

in distribution as $n_1 \to \infty$.

5. Alternative Estimator $\widetilde{\theta}$: Singly Flexible in $\rho^*(y)$

Because the assessment of $\mathrm{E}_p(\cdot \mid \mathbf{x})$ only relies on the observed data, instead of adopting an arbitrary known model $\mathrm{E}_p^\star(\cdot \mid \mathbf{x})$ or parametric model $\mathrm{E}_p^\star(\cdot \mid \mathbf{x}, \boldsymbol{\zeta})$, one might be willing to estimate $\mathrm{E}_p(\cdot \mid \mathbf{x})$ in a model free fashion and replace $\mathrm{E}_p^\star(\cdot \mid \mathbf{x})$ in the estimation procedure presented in Section 4.2 by a well-behaved estimator $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$. Here we consider a general estimator $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$ which has convergence rate faster than $n_1^{-1/4}$. This rate is achievable for many nonparametric regression or machine learning algorithms (Chernozhukov et al. 2018), see for example, Chen and White (1999) for a class of neural network models, Wager and Walther (2015) for a class of regression trees and random forests, and Bickel, Ritov, and Tsybakov (2009), Bühlmann and Van De Geer (2011), Belloni and Chernozhukov (2011), and Belloni and Chernozhukov (2013) for a variety of sparse models. Meanwhile, we still do not aim to estimate $\rho(y)$ since we do not have the Y-data in population $\mathcal{Q}$. We denote the corresponding estimator $\widetilde{\theta}$ and call it singly flexible because of its flexibility in using a working model $\rho^*(y)$.

The idea behind the estimator $\widetilde{\theta}$ is similar to $\widehat{\theta}$, therefore, we only emphasize the difference from Section 4.2. Similar to (7), we define $a^*(y)$ as the solution of

$$\mathrm{E}\left[\frac{\mathrm{E}_p\{a^*(Y)\rho^*(Y) \mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^{*2}(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho^*(Y) \mid \mathbf{X}\}} \mid y\right] = y. \quad (10)$$

Equivalently, $a^*(y)$ satisfies $\mathcal{L}^*(a^*)(y) = v(y)$, where

$$\mathcal{L}^*(a)(y) \equiv p_Y(y)\mathrm{E}\left[\frac{\mathrm{E}_p\{a(Y)\rho^*(Y) \mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^{*2}(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho^*(Y) \mid \mathbf{X}\}} \mid y\right]$$
$$= \int a(t)u^*(t, y)\mathrm{d}t, \text{ and}$$
$$u^*(t, y) \equiv p_Y(y) \int \frac{\rho^*(t)p_{Y|\mathbf{X}}(t, \mathbf{x})}{\mathrm{E}_p\{\rho^{*2}(Y) \mid \mathbf{x}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho^*(Y) \mid \mathbf{x}\}} g(\mathbf{x}, y)\mathrm{d}\mathbf{x}.$$

Using the estimator $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$, we approximate the integral equation (10) as

$$y = \widehat{\mathrm{E}}\left[\frac{\widehat{\mathrm{E}}_p\{a^*(Y)\rho^*(Y) \mid \mathbf{X}\}}{\widehat{\mathrm{E}}_p\{\rho^{*2}(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\widehat{\mathrm{E}}_p\{\rho^*(Y) \mid \mathbf{X}\}} \mid y\right]$$
$$= \sum_{i=1}^n \frac{\widehat{\mathrm{E}}_p\{a^*(Y)\rho^*(Y) \mid \mathbf{x}_i\}}{\widehat{\mathrm{E}}_p\{\rho^{*2}(Y) \mid \mathbf{x}_i\} + \pi/(1-\pi)\widehat{\mathrm{E}}_p\{\rho^*(Y) \mid \mathbf{x}_i\}} \times \frac{r_i K_h(y - y_i)}{\sum_{j=1}^n r_j K_h(y - y_j)}, \quad (11)$$

and we write $\widehat{a}^*(y)$ as the solution to $\widehat{\mathcal{L}}^*(\widehat{a}^*)(y) = \widehat{v}(y)$, where

$$\widehat{\mathcal{L}}^*(a)(y) \equiv n_1^{-1}\sum_{i=1}^n r_i K_h(y - y_i) \times \frac{\widehat{\mathrm{E}}_p\{a(Y)\rho^*(Y) \mid \mathbf{x}_i\}}{\widehat{\mathrm{E}}_p\{\rho^{*2}(Y) \mid \mathbf{x}_i\} + \pi/(1-\pi)\widehat{\mathrm{E}}_p\{\rho^*(Y) \mid \mathbf{x}_i\}}.$$

We summarize the algorithm for computing the estimator $\widetilde{\theta}$ below.

Algorithm 2 Alternative Estimator $\widetilde{\theta}$: Single Flexible in $\rho^*(y)$

Input: data from population $\mathcal{P}$: $(y_i, \mathbf{x}_i, r_i = 1)$, $i = 1, \ldots, n_1$, data from population $\mathcal{Q}$: $(\mathbf{x}_j, r_j = 0)$, $j = n_1 + 1, \ldots, n$, and value $\pi = n_1/n$.

do

(a) adopt a working model for $\rho(y)$, denoted as $\rho^*(y)$;

(b) adopt a nonparametric or machine learning algorithm for estimating $\mathrm{E}_p(\cdot \mid \mathbf{x})$, denoted as $\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$;

(c) compute $\widehat{w}_i = [\widehat{\mathrm{E}}_p\{\rho^{*2}(Y) \mid \mathbf{x}_i\} + \pi/(1-\pi)\widehat{\mathrm{E}}_p\{\rho^*(Y) \mid \mathbf{x}_i\}]^{-1}$ for $i = 1, \ldots, n$;

(d) obtain $\widehat{a}^*(\cdot)$ by solving the integral equation (11);

(e) compute $\widehat{b}^*(\mathbf{x}_i) = \widehat{w}_i\widehat{\mathrm{E}}_p\{\widehat{a}^*(Y)\rho^*(Y) \mid \mathbf{x}_i\}$ for $i = 1, \ldots, n$;

(f) obtain $\widetilde{\theta}$ as

$$\widetilde{\theta} = \frac{1}{n}\sum_{i=1}^n \left[\frac{r_i}{\pi}\rho^*(y_i)\{y_i - \widehat{b}^*(\mathbf{x}_i)\} + \frac{1 - r_i}{1-\pi}\widehat{b}^*(\mathbf{x}_i)\right]. \quad (12)$$

Output: $\widetilde{\theta}$.

To develop the asymptotic normality of the estimator $\widetilde{\theta}$, instead of Condition (A3), we need

(A7) The function $u^*(t, y)$ is bounded and has bounded derivatives with respect to t and y on its support. The function $a^*(y)$ in (10) is bounded.

We present Theorem 2, with its proof contained in Section S.8 of the supplementary materials.

Theorem 2. Assume $\widehat{\mathrm{E}}_p$ satisfies $|\widehat{\mathrm{E}}_p\{a(Y) \mid \mathbf{x}\} - \mathrm{E}_p\{a(Y) \mid \mathbf{x}\}| = o_p(n_1^{-1/4})$ for any bounded function $a(y)$. For any choice of $\rho^*(y)$, under Conditions (A2), (A4)-(A7), $\sqrt{n_1}(\widetilde{\theta} - \theta) \to N(0, \sigma_\theta^2)$ in distribution as $n_1 \to \infty$, where σ_θ^2 equals

$$\mathrm{var}\left(\sqrt{\pi}\phi_{\mathrm{eff}}^*(\mathbf{X}, R, RY) + \frac{R}{\sqrt{\pi}}\left[\frac{\mathrm{E}_p\{a^*(Y)\rho^*(Y) \mid \mathbf{X}\}}{\mathrm{E}_p\{\rho^{*2}(Y) \mid \mathbf{X}\} + \pi/(1-\pi)\mathrm{E}_p\{\rho^*(Y) \mid \mathbf{X}\}} - Y\right]\{\rho^*(Y) - \rho(Y)\}\right).$$

It is direct from Theorem 2 that when the posited model $\rho^*(y)$ is correctly specified, the estimator $\widetilde{\theta}$ becomes the efficient estimator for θ. We point out this consequence as Corollary 2.

Corollary 2. Assume $\widehat{\mathrm{E}}_p$ satisfies $|\widehat{\mathrm{E}}_p\{a(Y) \mid \mathbf{x}\} - \mathrm{E}_p\{a(Y) \mid \mathbf{x}\}| = o_p(n_1^{-1/4})$ for any bounded function $a(y)$. If $\rho^*(y) = \rho(y)$, under

Conditions (A2), (A4)-(A7),

$$\sqrt{n_1}(\widetilde{\theta}_{\text{eff}} - \theta) \to N[0, \text{var}\{\sqrt{\pi}\phi_{\text{eff}}(\mathbf{X}, R, RY)\}]$$

in distribution as $n_1 \to \infty$.

Last but not least, Sections 4 and 5 here only present the results for estimating $\theta = \mathrm{E}_q(Y)$. The whole story can be extended to a general parameter $\boldsymbol{\theta}$ such that $\mathrm{E}_q\{\mathbf{U}(\mathbf{X}, Y, \boldsymbol{\theta})\} = \mathbf{0}$, and the results are stated in Sections S.9 and S.10 of the supplementary materials. In our numerical studies in Sections 6 and 7, we analyze both $\mathrm{E}_q(Y)$ and the tth quantile of population $\mathcal{Q}$, defined as $\tau_{q,t}(Y) = \inf\left[y : \mathrm{E}_q\{I(Y \le y)\} \ge t\right]$, where $0 < t < 1$. This corresponds to $\mathrm{E}_q\left[\eta_t\left\{Y - \tau_{q,t}(Y)\right\}\right] = 0$ where $\eta_t(r) = t - I(r < 0)$.

6. Simulation Studies

6.1. Continuous Outcome

We conduct simulation studies to assess the finite sample performance of our proposed methods. We report the results for the mean $\mathrm{E}_q(Y)$ and the median $\tau_{q,0.5}(Y)$ of the outcome Y in population $\mathcal{Q}$.

We first generate a binary indicator $R_i, i = 1, \ldots, n$ from the Bernoulli distribution with probability 0.5, and record $n_1 = \sum_{i=1}^n r_i$, $\pi = n_1/n$. Then we generate Y_i from $N(0,1)$ if $R_i = 1$ and from $N(1,1)$ if $R_i = 0$. The $\mathbf{X} \mid Y$ distribution is generated from a three-dimensional normal with mean $(-0.5, 0.5, 1)^{\mathrm{T}} Y_i$ and covariance $\mathbf{I}$, the identity matrix. This implies, $\mathrm{E}_q(Y) = 1$, $\tau_{q,0.5} = 1$ and the true density ratio model $\rho(y) = \exp\big(-0.5 + y\big)$. One can derive that $p_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\zeta})$ follows normal with mean $(1, \mathbf{x}^{\mathrm{T}})\boldsymbol{\beta}$ where $\boldsymbol{\beta} = (0, -0.2, 0.2, 0.4)^{\mathrm{T}}$ and variance $\sigma^2 = 0.4$. Here we denote $\boldsymbol{\zeta} \equiv (\boldsymbol{\beta}^{\mathrm{T}}, \sigma^2)^{\mathrm{T}}$.

We use the following misspecified working models. We define $\rho^*(y) \equiv c^* \exp\big(-0.7 + 1.2y\big)$, where $c^* \equiv \pi / \{n^{-1} \sum_{i=1}^n r_i \exp\big(-0.7 + 1.2y_i\big)\}$ in order to satisfy $\mathrm{E}\{R\rho^*(Y)\} = \pi$. For the working model $p^\star_{Y|\mathbf{X}}$, we define $\mathbf{x}^\star \equiv [x_1, \exp(x_2/2), x_3/\{1 + \exp(x_2)\} + 10]^{\mathrm{T}}$ and define $p^\star_{Y|\mathbf{X}}(y, \mathbf{x}^\star, \boldsymbol{\zeta}^\star)$ as the normal distribution with mean $(1, \mathbf{x}^{\star\mathrm{T}})\boldsymbol{\beta}^\star$ where $\boldsymbol{\beta}^\star = (-7.000, -0.223, 0.363, 0.664)^{\mathrm{T}}$ and variance $\sigma^{\star 2} = 0.449$. The parameter $\boldsymbol{\zeta}^\star \equiv (\boldsymbol{\beta}^{\star\mathrm{T}}, \sigma^{\star 2})^{\mathrm{T}}$ is obtained by minimizing the Kullback-Leibler distance $D_{kl}(p_{Y|\mathbf{X}} \| p^\star_{Y|\mathbf{X}})$.

We implement the following seven estimators:

1. `shift-dependent`*: $\breve{\theta}$ in (3) with $\rho^*(y)$;
2. `doubly-flexible`**: $\widehat{\theta}$ in (9) with $\rho^*(y)$ and $p^\star_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\zeta}^\star)$, theoretically analyzed in Theorem 1;
3. `singly-flexible`*: $\widetilde{\theta}$ in (12) with $\rho^*(y)$, theoretically analyzed in Theorem 2;
4. `shift-dependent`^{0}: $\breve{\theta}$ in (3) with correct $\rho(y)$;
5. `doubly-flexible`^{0}: $\widehat{\theta}_{\text{eff}}$ with correct $\rho(y)$ and $p_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\zeta})$, theoretically analyzed in Corollary 1;
6. `singly-flexible`^{0}: $\widetilde{\theta}_{\text{eff}}$ with correct $\rho(y)$, theoretically analyzed in Corollary 2;
7. `oracle`: the $\sqrt{n_0}$-consistent estimator $\frac{1}{n}\sum_{i=1}^n \frac{1-r_i}{1-\pi} y_i$.

Note that the last four estimators (shown as "gray" in Figure 1) are unrealistic since they either use the unknown models $\rho(y)$ and $p_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\zeta})$ or the Y-data in population $\mathcal{Q}$.

In implementing estimators `doubly-flexible`**, `singly-flexible`*, `doubly-flexible`^{0} and `singly-flexible`^{0}, we solve the integral equations (8) and (11) using the Nadaraya-Watson estimator for $\mathrm{E}(\cdot \mid y)$ with Gaussian kernel and bandwidth $h = n_1^{-1/3}$ that is discussed in Condition (A6). Numerically, the integrations are approximated by the Gauss-Legendre quadrature with 50 points on the interval $[-5, 5]$ and the integral equations are evaluated at $y_i, i = 1, \ldots, n_1$. In addition, for estimators `singly-flexible`* and `singly-flexible`^{0}, we estimate $\mathrm{E}_p(\cdot \mid \mathbf{x})$ using the Nadaraya-Watson estimator based on the product Gaussian kernel with bandwidth $2.5 n_1^{-1/7}$, where the order comes from the optimal bandwidth $n_1^{-1/(4+d)}$ with d the dimensionality of covariate $\mathbf{X}$. See Section S.5 of the supplementary materials for technical details on the numerical implementation.

Based on 1000 simulation replicates, Figure 1 illustrates the boxplots of the estimates for the mean and the median. With the misspecified working model $\rho^*(y)$, the estimator `shift-dependent`* is biased; in contrast, the proposed estimators `doubly-flexible`** and `singly-flexible`* are both unbiased. When the correct model $\rho(y)$ is used, not surprisingly, all of the estimators `shift-dependent`^{0}, `doubly-flexible`^{0} and `singly-flexible`^{0} are unbiased. It is also clear that the two proposed flexible estimators are always more efficient than the shift-dependent estimator no matter the correct model $\rho(y)$ is used or not.

To further demonstrate the efficiency comparison and the inference results, in Table 1, we report the mean squared error (MSE), the empirical bias (Bias), the empirical standard error (SE), the average of estimated standard error ($\widehat{\text{SE}}$), and the empirical coverage at 95% confidence level (CI), for each of the estimators. The estimator `shift-dependent`* has an incorrect coverage (over-coverage for mean estimation and under-coverage for median estimation) because of its severe bias. This issue is not mitigated at all or becomes even worse in Table 1 when we increase the size of the stacked random sample from 500 to 1000. On the contrary, the estimators `doubly-flexible`** and `singly-flexible`* are correctly covered. Though there is no theoretical justification, `doubly-flexible`** is slightly less efficient than `singly-flexible`* in this setting. This indicates that the effort of correctly estimating $\mathrm{E}_p(\cdot \mid \mathbf{x})$ pays off in the sense of improving the estimation efficiency. With the correct $\rho(y)$ model used, each of the estimators `doubly-flexible`^{0} and `singly-flexible`^{0} is more efficient than its counterpart `doubly-flexible`** and `singly-flexible`*. The $\sqrt{n_0}$-consistent estimator `oracle` is the most efficient one in this simulation setting.

6.2. Binary Outcome

We conduct an additional simulation study to compare our proposed methods to existing methods in binary classification case under the correct specification of ρ and $p_{Y|\mathbf{X}}$. We report the results for estimation of $\mathrm{E}_q(Y)$, the mean of the outcome Y in population $\mathcal{Q}$. For comparision, we implement the following six estimators:

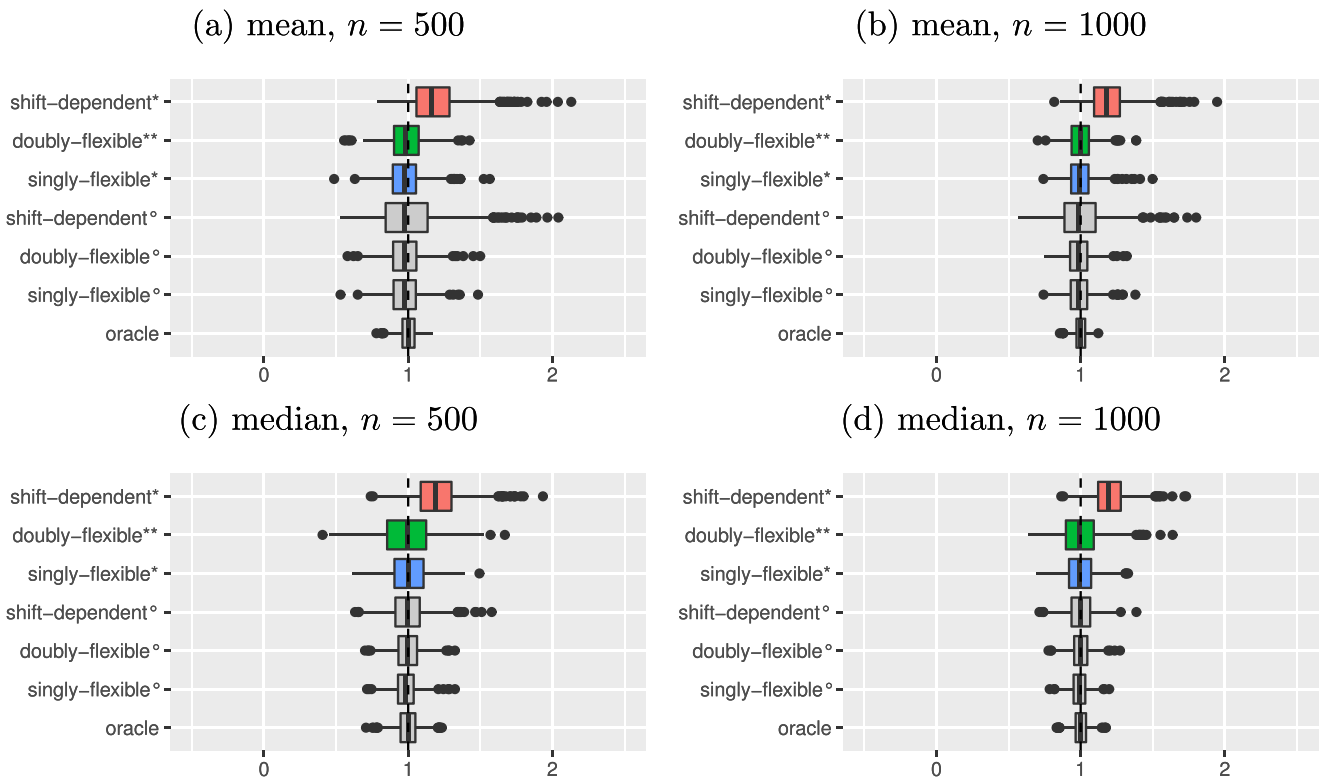

Figure 1. Boxplots of estimates in the simulation study in Section 6.1. Dashed line: the true estimand.

1. `shift-dependent`^{0}: $\check{\theta}$ in (3) with correct $\rho(y)$;
2. `BBSE`: Black Box Shift Estimation by Lipton, Wang, and Smola (2018);
3. `RLLS`: Regularized Learning under Label Shifts by Azizzadenesheli et al. (2019);
4. `doubly-flexible`^{0}: $\widehat{\theta}_{\text{eff}}$ with correct $\rho(y)$ and $p_{Y|\mathbf{X}}(y, \mathbf{x}, \boldsymbol{\zeta})$, theoretically analyzed in Corollary 1;
5. `singly-flexible`^{0}: $\widetilde{\theta}_{\text{eff}}$ with correct $\rho(y)$, theoretically analyzed in Corollary 2;
6. `oracle`: the $\sqrt{n_0}$-consistent estimator $\frac{1}{n}\sum_{i=1}^{n}\frac{1-r_i}{1-\pi}y_i$.

We refer the readers to Section S.11 of the supplementary materials for details on the data generating process and the implementation of the estimators.

To demonstrate the efficiency comparison, in Table 2, we report the mean squared error (MSE), the empirical bias (Bias), the empirical standard error (SE), and the asymptotic relative efficiency (ARE, the ratio of SE of each estimator to `doubly-flexible`^{0}) from 1000 simulation replicates. Agreeing with the results in Corollaries 1 and 2, our proposed estimators `doubly-flexible`^{0} and `singly-flexible`^{0} performed more efficiently than the existing methods `BBSE` and `RLLS` in terms of MSE. ARE further indicates that `doubly-flexible`^{0} was about 1.5 times more efficient than `BBSE` and `RLLS` in terms of standard error. It is also worthwhile to note that `doubly-flexible`^{0} and `singly-flexible`^{0} performed similarly to each other, suggesting that both estimators achieved the semiparametric efficiency. In addition, in this simulation setting, we can also observe that `doubly-flexible`^{0} and `singly-flexible`^{0} were almost as efficient as `oracle`, while the performances of `BBSE` and `RLLS` are similar to or worse than `shift-dependent`^{0}.

7. Data Application

We now illustrate the numerical performance of our proposed method through analyzing the Medical Information Mart for Intensive Care III (MIMIC-III), an openly available electronic health records database, developed by the MIT Lab for Computational Physiology (Johnson et al. 2016). It comprises deidentified health-related records including demographics, vital signs, laboratory test and medications, for 46,520 patients who admitted to the intensive care unit of the Beth Israel Deaconess Medical Center between 2001 and 2012.

The outcome of interest Y in our analysis is the sequential organ failure assessment (SOFA) score (Singer et al. 2016), used to track a patient's status during the stay in an intensive care unit to determine the extent of a patient's organ function or rate of failure. The score is based on six different sub-scores, with one of each for the respiratory, cardiovascular, hepatic, coagulation, renal and neurological systems. The SOFA score ranges from 0 (best) to 24 (worst). We include 16 covariates from either chart events (6 variables, diastolic blood pressure, systolic blood pressure, blood glucose, respiratory rate per minute, and two measures from body temperature) or laboratory tests (10 variables, peripheral caillary oxygen saturation, two measures from each of hematocrit level, platelets count and red blood cell count, and three measures from blood urea nitrogen). We choose these covariates through assessing whether the absolute correlation with the outcome Y is greater than 0.2 and whether the missing rate is less than 1%. In our analysis, we only include the first

Table 1. Summary of estimation results in the simulation study in Section 6.1.

(a) mean

n	Estimator	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x})$	MSE	Bias	SE	$\widehat{\text{SE}}$	CI
500	shift-dependent*	$\rho^*(y)$	–	0.0699	0.1840	0.1899	0.2791	1.000
	doubly-flexible*⋆	$\rho^*(y)$	$p^\star_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.0173	−0.0120	0.1311	0.1287	0.943
	singly-flexible*	$\rho^*(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	0.0162	−0.0201	0.1258	0.1230	0.941
	shift-dependent^0	$\rho(y)$	–	0.0533	0.0049	0.2309	0.2119	0.899
	doubly-flexible^0	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.0153	−0.0231	0.1214	0.1212	0.941
	singly-flexible^0	$\rho(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	0.0138	−0.0221	0.1155	0.1169	0.939
	oracle	-	–	0.0040	0.0006	0.0636	0.0633	0.943
1000	shift-dependent*	$\rho^*(y)$	–	0.0561	0.1906	0.1406	0.2077	0.999
	doubly-flexible*⋆	$\rho^*(y)$	$p^\star_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.0085	0.0013	0.0922	0.0912	0.955
	singly-flexible*	$\rho^*(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	0.0081	−0.0031	0.0899	0.0850	0.952
	shift-dependent^0	$\rho(y)$	–	0.0275	0.0024	0.1660	0.1533	0.927
	doubly-flexible^0	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.0075	−0.0125	0.0856	0.0861	0.958
	singly-flexible^0	$\rho(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	0.0069	−0.0094	0.0824	0.0827	0.955
	oracle	–	–	0.0020	0.0008	0.0451	0.0447	0.945

(b) median

n	Estimator	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x})$	MSE	*Bias*	SE	$\widehat{\text{SE}}$	CI
500	shift-dependent*	$\rho^*(y)$	–	0.0695	0.2025	0.1687	0.1547	0.802
	doubly-flexible*⋆	$\rho^*(y)$	$p^\star_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.0368	−0.0072	0.1918	0.1764	0.940
	singly-flexible*	$\rho^*(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	0.0211	0.0024	0.1453	0.1390	0.941
	shift-dependent^0	$\rho(y)$	–	0.0178	0.0011	0.1336	0.1286	0.947
	doubly-flexible^0	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.0093	−0.0033	0.0964	0.0950	0.951
	singly-flexible^0	$\rho(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	0.0071	−0.0162	0.0827	0.0881	0.956
	oracle	–	–	0.0064	−0.0020	0.0799	0.0821	0.946
1000	shift-dependent*	$\rho^*(y)$	–	0.0554	0.2018	0.1210	0.1104	0.560
	doubly-flexible*⋆	$\rho^*(y)$	$p^\star_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.0229	−0.0013	0.1515	0.1444	0.943
	singly-flexible*	$\rho^*(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	0.0128	−0.0023	0.1130	0.1124	0.954
	shift-dependent^0	$\rho(y)$	–	0.0091	0.0005	0.0955	0.0915	0.934
	doubly-flexible^0	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.0047	0.0000	0.0688	0.0669	0.954
	singly-flexible^0	$\rho(y)$	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	0.0036	−0.0098	0.0594	0.0625	0.948
	oracle	–	–	0.0031	0.0004	0.0558	0.0578	0.959

Table 2. Summary of estimation results in the simulation study in Section 6.2.

n	Estimator	ρ	$p_{Y\|\mathbf{X}}$	MSE	Bias	SE	ARE
200	shift-dependent^0	ρ	-	0.0059	−0.0007	0.0769	1.5135
	BBSE	–	$p_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.0060	−0.0048	0.0775	1.5250
	RLLS	–	$p_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.0060	−0.0048	0.0775	1.5250
	doubly-flexible^0	ρ	$p_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.0026	−0.0039	0.0508	1
	singly-flexible^0	ρ	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	0.0025	−0.0001	0.0503	0.9899
	oracle	–	–	0.0025	0.0014	0.0499	0.9824
400	shift-dependent^0	ρ	–	0.0026	−0.0011	0.0514	1.4920
	BBSE	–	$p_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.0033	−0.0041	0.0573	1.6635
	RLLS	–	$p_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.0033	−0.0041	0.0573	1.6635
	doubly-flexible^0	ρ	$p_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.0012	−0.0026	0.0345	1
	singly-flexible^0	ρ	$\widehat{\mathrm{E}}_p(\cdot \mid \mathbf{x})$	0.0014	−0.0013	0.0372	1.0795
	oracle	–	–	0.0012	−0.0005	0.0347	1.0062

admission if the patient was admitted to the intensive care unit more than once. We also exclude patients whose outcome Y is greater than or equal to 20, whose age is greater than or equal to 65, and who has missing values in any of the covariates. This results in a total of $n = 16{,}691$ records.

In our analysis, we define the population $\mathcal{P}$ as patients with private, goverment, and self-pay insurances ($R = 1, n_1 = 11{,}695$), and population $\mathcal{Q}$ as patients whose insurance type is either Medicaid or Medicare ($R = 0, n_0 = 4{,}996$). The label shift assumption that the conditional distribution of $\mathbf{X}$ given Y remains the same can be tested via the conditional independence of $\mathbf{X}$ and R given Y. In our analysis, we test the conditional independence between R and each of covariates by the invariant environment prediction test (Heinze-Deml, Peters, and Meinshausen 2018) in R package `CondIndTests`, and the p-values range from 0.460 to 0.628. This indicates the label shift assumption is indeed sensible in our analysis. We first compute the sample mean (3.7409) and sample tth quantiles (1,1,3,5,8 for $t = (10, 25, 50, 75, 90)\%$) of SOFA scores among patients whose insurance type is either Medicaid or Medicare. We regard these estimates as `oracle` in order to compare with our proposed methods.

Table 3. Estimation results in the data application.

(a) mean

	Estimator	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x})$	Estimate	*diff.withoracle*	$\widehat{\text{SE}}$	CI
	shift-dependent*	$\rho^*(y)$	–	4.0529	0.3120	0.0593	[3.9367, 4.1691]
	doubly-flexible*⋆	$\rho^*(y)$	$p^{\star}_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	3.7579	0.0170	0.0803	[3.6005, 3.9153]
	singly-flexible*	$\rho^*(y)$	$\widehat{\text{E}}_p(\cdot \mid \mathbf{x})$	3.7542	0.0133	0.0496	[3.6570, 3.8514]
	oracle	–	–	3.7409	–	0.0405	[3.6616, 3.8202]

(b) quantiles

τ	Estimator	$\rho(y)$	$p_{Y\|\mathbf{X}}(y,\mathbf{x})$	Estimate	*diff.withoracle*	$\widehat{\text{SE}}$	CI
10%	shift-dependent*	$\rho^*(y)$	-	1.0000	0.0000	0.0102	[0.9801, 1.0199]
	doubly-flexible*⋆	$\rho^*(y)$	$p^{\star}_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	0.9998	−0.0002	0.0095	[0.9812, 1.0185]
	singly-flexible*	$\rho^*(y)$	$\widehat{\text{E}}_p(\cdot \mid \mathbf{x})$	0.9998	−0.0002	0.0099	[0.9803, 1.0192]
	oracle	-	-	1.0000	–	0.0218	[0.9572, 1.0428]
25%	shift-dependent*	$\rho^*(y)$	-	1.9995	0.9995	0.0249	[1.9508, 2.0483]
	doubly-flexible*⋆	$\rho^*(y)$	$p^{\star}_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	1.0002	0.0002	0.0276	[0.9460, 1.0543]
	singly-flexible*	$\rho^*(y)$	$\widehat{\text{E}}_p(\cdot \mid \mathbf{x})$	1.0004	0.0004	0.0196	[0.9620, 1.0389]
	oracle	–	-	1.0000	–	0.0315	[0.9383, 1.0617]
50%	shift-dependent*	$\rho^*(y)$	-	3.0002	0.0002	0.0408	[2.9203, 3.0801]
	doubly-flexible*⋆	$\rho^*(y)$	$p^{\star}_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	3.0001	0.0001	0.0333	[2.9348, 3.0654]
	singly-flexible*	$\rho^*(y)$	$\widehat{\text{E}}_p(\cdot \mid \mathbf{x})$	3.0005	0.0005	0.0334	[2.9350, 3.0659]
	oracle	–	-	3.0000	–	0.0550	[2.8923, 3.1077]
75%	shift-dependent*	$\rho^*(y)$	-	5.9999	0.9999	0.0742	[5.8544, 6.1454]
	doubly-flexible*⋆	$\rho^*(y)$	$p^{\star}_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	5.0001	0.0001	0.0759	[4.8512, 5.1489]
	singly-flexible*	$\rho^*(y)$	$\widehat{\text{E}}_p(\cdot \mid \mathbf{x})$	5.0001	0.0001	0.0381	[4.9254, 5.0748]
	oracle	–	–	5.0000	–	0.0591	[4.8842, 5.1158]
90%	shift-dependent*	$\rho^*(y)$	–	8.9999	0.9999	0.1380	[8.7295, 9.2703]
	doubly-flexible*⋆	$\rho^*(y)$	$p^{\star}_{Y\|\mathbf{X}}(y,\mathbf{x},\widehat{\boldsymbol{\zeta}})$	8.0004	0.0004	0.1003	[7.8039, 8.1969]
	singly-flexible*	$\rho^*(y)$	$\widehat{\text{E}}_p(\cdot \mid \mathbf{x})$	8.0004	0.0004	0.0638	[7.8754, 8.1253]
	oracle	–	–	8.0000	–	0.1093	[7.7858, 8.2142]

To identify a reasonable working model $\rho^*(y)$, we model the data $Y + 0.001$ from population $\mathcal{P}$ as a parametric gamma distribution $f(y,\alpha,\beta) = \Gamma(\alpha)^{-1}\beta^{-\alpha}y^{\alpha-1}\exp\big(-y/\beta\big)$ with $\Gamma(\cdot)$ the Γ-function, the shape parameter $\alpha > 0$ and the scale parameter $\beta > 0$. We estimate the unknown parameters α and β as $\widehat{\alpha}$ and $\widehat{\beta}$ using the MLE. For the Y-data in population $\mathcal{Q}$, we assume $Y + 0.001$ follows a similar gamma distribution with shape parameter $\widehat{\alpha}+1$ and scale parameter $\widehat{\beta}$. Hence, we use the working model

$$\rho^*(y) = \frac{f(y+0.001,\widehat{\alpha}+1,\widehat{\beta})}{f(y+0.001,\widehat{\alpha},\widehat{\beta})}.$$

To implement the estimator `doubly-flexible*⋆`, we impose a parametric model $p^{\star}_{Y|\mathbf{X}}(y,\mathbf{x},\boldsymbol{\zeta})$ by regressing $Y+0.001$ on $\mathbf{X}$ as a generalized linear model with gamma distribution, and estimate $\boldsymbol{\zeta}$ using the MLE. To implement the estimator `singly-flexible*`, we identify the first three principal components from the 16 covariates, and then estimate $\text{E}_p(\cdot \mid \mathbf{x})$ as a function of those three principal components using the Nadaraya-Watson estimator based on the product Gaussian kernel with bandwidth $0.5n_1^{-1/7}$. To solve the corresponding integral equations, similar to Section 6.1, we approximate $\text{E}(\cdot \mid y)$ by its Nadaraya-Watson estimator with the Gaussian kernel and bandwidth $h = n_1^{-1/3}$. In addition, for numerical implementation, the integration is approximated at 50 equally-spaced points on the interval [0, 19], and the integral equations are evaluated at each supporting point of $\{y_i : i = 1,\ldots,n_1\}$. See Section S.5 of the supplementary materials for technical details on the implementation.

The results are summarized in Table 3. The estimator `shift-dependent*` that relies on a misspecified model $\rho^*(y)$ severely over-estimates the quantities compared to the `oracle` estimate, in most of the scenarios including estimating the mean, 25%, 75%, and 90% quantiles. As a consequence, the `oracle` estimate cannot be covered by the confidence intervals. In contrast, the proposed estimators `doubly-flexible*⋆` and `singly-flexible*`, although also rely on $\rho^*(y)$, provide almost identical estimates as `oracle`. Accordingly, the confidence intervals from the proposed methods all cover the `oracle` estimate. In these scenarios, `singly-flexible*` is more efficient than `doubly-flexible*⋆`, which echoes our findings in Section 6.

When estimating 10% and 50% quantiles, we find that `shift-dependent*` gives almost the same estimate as `oracle`. It might be plausible that the difference between $\rho^*(y)$ and the true $\rho(y)$ is minor for estimating these two quantities. Nevertheless, the proposed estimators `doubly-flexible*⋆` and `singly-flexible*` are still more efficient than the estimator `shift-dependent*`.

Finally it is interesting to observe that, the estimator `oracle` is even less efficient than the proposed estimators `doubly-flexible*⋆` and `singly-flexible*`, in estimating all of the quantiles. This is because `oracle` is $\sqrt{n_0}$-consistent whereas the two proposed estimators are both $\sqrt{n_1}$-consistent. In this application, n_1 is 2.34 times greater than n_0, which might result in the situation that the `oracle` estimate being less efficient. In Section 6, we also considered situations that n_1 is much larger than n_0 and similar phenomenon was observed as well, with detailed results omitted.

8. Conclusion

In this article, we estimate a characteristic of a target population $\mathcal{Q}$, via exploiting the data and information from a different but relevant population $\mathcal{P}$, under the label shift assumption. Different from most existing literatures, our proposal is devised to accommodate both classification and regression problems. We primarily propose the doubly flexible estimate, whose unique feature is to simultaneously allow both models to be misspecified thus is flexible: the density ratio model $\rho(y)$ that governs the label shift mechanism, and the conditional distribution model $p_{Y|\mathbf{X}}(y, \mathbf{x})$ of population $\mathcal{P}$. While the estimation of the latter can be done via off-the-shelf procedures sometimes, it often faces curse of dimensionality or computational challenges. Further, estimating $\rho(y)$ is even more difficult because the Y-data in population $\mathcal{Q}$ is not accessible in our procedure.

Supplementary Materials

The supplement includes all of the technical details.

Acknowledgments

The authors would like to thank the Editor, an Associate Editor, and the three reviewers for their insightful comments which have helped improve the manuscript substantially.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

The research is supported in part by NSF (DMS 1953526, 2122074, 2310942), NIH (R01DC021431) and the American Family Funding Initiative of UW-Madison.

References

Alexandari, A. M., Kundaje, A., and Shrikumar, A. (2020), "Maximum Likelihood with Bias-Corrected Calibration is Hard-to-Beat at Label Shift Adaptation," in *Proceedings of the 37th International Conference on Machine Learning*. [3]

Azizzadenesheli, K., Liu, A., Yang, F., and Anandkumar, A. (2019), "Regularized Learning for Domain Adaptation Under Label Shifts," arXiv preprint arXiv:1903.09734. [2,3,9]

Belloni, A., and Chernozhukov, V. (2011), "L1-Penalized Quantile Regression in High-Dimensional Sparse Models," *The Annals of Statistics*, 39, 82–130. [7]

——— (2013), "Least Squares after Model Selection in High-Dimensional Sparse Models," *Bernoulli*, 19, 521–547. [7]

Bickel, P. J., Klaassen, J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press. [3,4]

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732. [7]

Bühlmann, P., and Van De Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Berlin: Springer. [7]

Chan, Y. S., and Ng, H. T. (2005), "Word Sense Disambiguation with Distribution Estimation," in *IJCAI* (Vol. 5), Citeseer, pp. 1010–1015. [2]

Chen, X., and White, H. (1999), "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators," *IEEE Transactions on Information Theory*, 45, 682–691. [7]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters: Double/Debiased Machine Learning," *The Econometrics Journal*, 21, C1–C68. [7]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1–22. [2]

d'Haultfoeuille, X. (2011), "On the Completeness Condition in Nonparametric Instrumental Problems," *Econometric Theory*, 27, 460–471. [3]

Du Plessis, M. C., and Sugiyama, M. (2014), "Semi-Supervised Learning of Class Balance under Class-Prior Change by Distribution Matching," *Neural Networks*, 50, 110–119. [2]

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. (2020), "A unified view of label shift estimation," *Advances in Neural Information Processing Systems*, 33, 3290–3300. [2]

Gretton, A., Smola, A. J., Huang, J., Schmittfull, M., Borgwardt, K. M., and Schölkopf, B. (2009), "Covariate Shift by Kernel Mean Matching," in *Dataset Shift in Machine Learning*, pp. 131–160, MIT Press. [1]

Guo, J., Gong, M., Liu, T., Zhang, K., and Tao, D. (2020), "Ltf: A Label Transformation Framework for Correcting Label Shift," in *International Conference on Machine Learning*, pp. 3843–3853, PMLR. [2]

Hansen, P. C. (1992), "Numerical Tools for Analysis and Solution of Fredholm Integral Equations of the First Kind," *Inverse problems*, 8, 849–872. [5]

Heinze-Deml, C., Peters, J., and Meinshausen, N. (2018), "Invariant Causal Prediction for Nonlinear Models," *Journal of Causal Inference*, 6. *https://doi.org/10.1515/jci-2017-0016* [10]

Hu, Y., and Shiu, J.-L. (2018), "Nonparametric Identification Using Instrumental Variables: Sufficient Conditions for Completeness," *Econometric Theory*, 34, 659–693. [3]

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. (2006), "Correcting Sample Selection Bias by Unlabeled Data," in *Advances in Neural Information Processing Systems* (Vol. 19). [1]

Iyer, A., Nath, S., and Sarawagi, S. (2014), "Maximum Mean Discrepancy for Class Ratio Estimation: Convergence Bounds and Kernel Selection," in *International Conference on Machine Learning*, PMLR, pp. 530–538. [2]

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016), "MIMIC-III, a Freely Accessible Critical Care Database," *Scientific Data*, 3, 1–9. [9]

Kouw, W. M., and Loog, M. (2021), "A Review of Domain Adaptation without Target Labels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43, 766–785. [1]

Kpotufe, S., and Martinet, G. (2021), "Marginal Singularity and the Benefits of Labels in Covariate-Shift," *The Annals of Statistics*, 49, 3299–3323. [1]

Landweber, L. (1951), "An Iteration Formula for Fredholm Integral Equations of the First Kind," *American Journal of Mathematics*, 73, 615–624. [5]

Li, W., Miao, W., and Tchetgen, E. T. (2021), "Nonparametric Inference About Mean Functionals of Nonignorable Nonresponse Data Without Identifying the Joint Distribution," arXiv preprint arXiv:2110.05776. [3]

Lipton, Z., Wang, Y.-X., and Smola, A. (2018), "Detecting and Correcting for Label Shift with Black Box Predictors," in *International Conference on Machine Learning*, PMLR, pp. 3122–3130. [2,3,9]

Maity, S., Sun, Y., and Banerjee, M. (2020), "Minimax Optimal Approaches to the Label Shift Problem," arXiv preprint arXiv:2003.10443. [3]

Martinez, J., Hossain, R., Romero, J., and Little, J. J. (2017), "A Simple Yet Effective Baseline for 3D Human Pose Estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2640–2649. [2]

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. (2012), "A Unifying View on Dataset Shift in Classification," *Pattern Recognition*, 45, 521–530. [1]

Newey, W. K., and Powell, J. L. (2003), "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578. [3]

Nguyen, T. D., Christoffel, M., and Sugiyama, M. (2016), "Continuous Target Shift Adaptation in Supervised Learning," in *Asian Conference on Machine Learning*, pp. 285–300, PMLR. [2,3]

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2008), *Dataset Shift in Machine Learning*, Cambridge, MA: MIT Press. [1]

Saerens, M., Latinne, P., and Decaestecker, C. (2002), "Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure," *Neural Computation*, 14, 21–41. [2]

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012), "On Causal and Anticausal Learning," in *29th International Conference on Machine Learning (ICML 2012)*, pp. 1255–1262, Omnipress. [1]

Shimodaira, H. (2000), "Improving Predictive Inference Under Covariate Shift by Weighting the Log-Likelihood Function," *Journal of Statistical Planning and Inference*, 90, 227–244. [1]

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., et al. (2016), "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA*, 315, 801–810. [9]

Storkey, A. (2009), "When Training and Test Sets are Different: Characterizing Learning Transfer," *Dataset Shift in Machine Learning*, 30, 3–28. [2]

Sugiyama, M., and Kawanabe, M. (2012), *Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation*, Cambridge, MA: MIT press. [1]

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. [1]

Tasche, D. (2017), "Fisher Consistency for Prior Probability Shift," *The Journal of Machine Learning Research*, 18, 3338–3369. [2]

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer. [3,4]

van der Laan, M. J., and Robins, J. M. (2003), *Unified Methods for Censored Longitudinal Data and Causality*, Springer. [3]

Wager, S., and Walther, G. (2015), "Adaptive Concentration of Regression Trees, with Application to Random Forests," arXiv preprint arXiv:1503.06388. [7]

Yang, X., Sun, H., Sun, X., Yan, M., Guo, Z., and Fu, K. (2018), "Position Detection and Direction Prediction for Arbitrary-Oriented Ships via Multitask Rotation Region Convolutional Neural Network," *IEEE Access*, 6, 50839–50849. [2]

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. (2013), "Domain Adaptation Under Target and Conditional Shift," in *International Conference on Machine Learning*, PMLR, pp. 819–827. [2,3]

Zhao, J., and Ma, Y. (2022), "A Versatile Estimation Procedure Without Estimating the Nonignorable Missingness Mechanism," *Journal of the American Statistical Association*, 117, 1916–1930. [3]