Correcting for Nonignorable Selection Bias and Missing Response in Regression using Privileged Information

Abstract

When estimating a regression model, we might have data where some labels are missing, or our data might be biased by a selection mechanism. When the response or selection mechanism is *ignorable* (i.e., independent of the response variable given the features) one can use off-the-shelf regression methods; in the *nonignorable* case one typically has to adjust for bias. We observe that *privileged data* (i.e. data that is only available during training) might render a nonignorable selection mechanism ignorable. We refer to this scenario as *Privilegedly Missing at Random* (PMAR). We propose a novel imputation-based regression method, named *repeated regression*, that is suitable for PMAR. We also consider an importance weighted regression method, and a doubly robust combination of the two. The proposed methods are easy to implement with most popular out-of-the-box regression algorithms. We empirically assess the performance of the proposed methods with extensive simulated experiments and on a synthetically augmented real-world dataset. We conclude that repeated regression can appropriately correct for bias, and can have considerable advantage over weighted regression, especially when extrapolating to regions of the feature space where response is never observed.

1 INTRODUCTION

Regression is a primary technique in data science, machine learning and statistics. When presented with data $(X_1, Y_1), ..., (X_n, Y_n)$ from the distribution $\mathbb{P}(X, Y)$, the goal is to find the conditional expectation of Y given X, i.e. estimate the function $\mu(x) = \mathbb{E}[Y|X = x]$. Practitioners are often presented with either incomplete data (where some values are missing) or data that is not representative of the population (drawn from some other $\tilde{\mathbb{P}}(X, Y)$ with $\tilde{\mathbb{P}} \neq \mathbb{P}$), and hence have to correct for the bias that is present in their training data. The discrepancy between $\tilde{\mathbb{P}}$ and $\mathbb{P}$ can for example be induced by a selection mechanism.

Consider the following example of a follow-up study of HIV-positive individuals, where we are interested in the prediction of the 3-year risk of death (Y) after a certain antiretroviral drug dosage (X) has been administered [Hernán and Robins, 2021]. Severe side effects of the drug (Z) might cause the patient to drop out of the study, censoring the measurement of Y. When other factors than drug dosage influence the side effects Z and the outcome Y, one must include measurements of Z in the dataset for correct estimation of the model $\mathbb{E}[Y|X]$. As these side effects cannot be known before the drug has been administered, including Z in the features of the model is undesirable when the model is used for e.g. optimising drug dosage. In this work, we demonstrate how one can incorporate such *privileged information* in a regression model.

An important line of work on selection bias is by Pearl [2012] and Bareinboim et al. [2014], who consider the problem of estimating $\mathbb{P}(Y|X)$ from a potentially biased dataset by leveraging knowledge of the underlying causal graph. They derive an expression of $\mathbb{P}(Y|X)$ as an integral of quantities that can be estimated from readily available biased and additional ('external') unbiased data. Although identification of a conditional distribution might be useful in certain scenarios, this does not tell the practitioner how to estimate a regression model, especially when dealing with continuous variables. In this work we address this problem, keeping in mind the applicability of the proposed methods.

Missingness problems are often characterised by the presence of certain conditional independencies in the data. Typically these independencies are untestable, making it unclear whether these conditional independence assumptions are appropriate for the data at hand. Drawing a causal graph of the data generating process is often helpful for gaining a

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

better understanding of the problem. Moreover, the causal graph identifies conditional independencies in the data, and thereby allows the practitioner to motivate these conditional independence assumptions. As posed by Rubin [1976]:

"The inescapable conclusion seems to be that when dealing with real data, the practising statistician should explicitly consider the process that causes missing data far more often than he does. However, to do so, he needs models for this process and these have not received much attention in the statistical literature."[1]

Contrary to what is often seen in the literature we do not need a full probabilistic model of the missingness mechanism, as the identification of certain conditional independencies can already be sufficient for estimating a regression model.

Our contributions are as follows. First, this paper serves a pedagogical purpose by reviewing literature of missingness and selection bias, as addressed in Section 2. We address what the consequences of certain conditional independencies are for the practitioner. While doing so, we motivate the use of *privileged information* [Vapnik and Vashist, 2009] when missingness or selection is nonignorable, and introduce the *Privilegedly Missing at Random* setting (PMAR). Secondly, in Section 3 we formulate a novel imputation-based estimator for the PMAR setting, which we refer to as *repeated regression*. We point out that importance weighted regression is also suitable for the PMAR setting, and combine the two methods into a doubly robust estimator. These estimators are formulated such that they are easy to implement using out-of-the box regression algorithms. We then address the intricacies of evaluation under selection bias and missingness in Section 4. We warn the practitioner that, to our knowledge, there is no appropriate way of evaluating a regression method on a finite, biased dataset without relying on auxiliary models. Lastly, in Section 5 we assess the performance of the formulated methods on extensive simulated experiments and on synthetically augmented real-world data.[2] We observe that repeated regression has considerable advantage over importance weighting, especially when extrapolating.

2 MISSING DATA, SELECTION BIAS AND PRIVILEGED INFORMATION

2.1 MISSING DATA MECHANISMS

A very general framework for handling missing data is proposed by Rubin [1976]. When modelling the distribution of a set of random variables $X_1, ..., X_m$, this framework takes into account potential missing values of any of the X_i by considering a random vector of *response indicators*: binary variables $S_1, ..., S_m$, where $S_i = 1$ indicates that variable X_i is observed. The process that determines whether we observe X_i can be explicitly modelled as $\mathbb{P}(S_i|X_1, ..., X_m)$. We then define the *observed* random vector $X_1^o, ..., X_m^o$ where $X_i^o = X_i$ if $S_i = 1$ and $X_i =?$ if $S_i = 0$, where '?' denotes the value that is missing in the dataset.

Rubin [1976] proposed the following classification of missing data mechanisms, which approximately categorises the difficulty of many inference problems.

Definition 1. [Rubin, 1976] Given variables X, Y, and response indicator S for Y, we say that Y is

- Missing Completely at Random (MCAR) if the missingness mechanism is independent of all other observed variables, i.e. $S \perp\!\!\!\perp X, Y$;
- Missing at Random (MAR) if the missingness mechanism is independent of the missing variable given all other fully observed variables, i.e. $Y \perp\!\!\!\perp S \mid X$;
- Missing Not at Random (MNAR) if it is neither MCAR nor MAR.

A missingness mechanism is often called *ignorable* if one does not have to correct for bias in the inference problem at hand. Otherwise, it is called *nonignorable*.

A large body of literature has been written about inference under missingness. Important works are the EM algorithm [Dempster et al., 1977], Nobel prize winning work by Heckman [1979] on correcting for selection bias in linear regression, Rosenbaum and Rubin [1984] which laid the foundations of propensity score based methods in causal inference, and a series of generalised estimating equations (GEE) based methods for inference under missingness [Robins et al., 1994, Robins and Rotnitzky, 1995, Rotnitzky et al., 1998, Scharfstein et al., 1999]. For an overview of the field, see Little and Rubin [2019]. In this work we focus on correcting for bias using privileged data in regression, which is not treated in the aforementioned literature.

S-recoverability As proposed by Pearl [2012] and further developed in Bareinboim et al. [2014], *s-recoverability* is a method to deal with selection bias. Consider again the case of modelling the distribution of $X_1, ..., X_m$. Instead of having one response indicator for each X_i as in the missingness framework, Pearl [2012] considers one selection node S where $S = 1$ indicates that all variables are observed (*selected*), and $S = 0$ indicates that no variable is observed, i.e. there is no row for this observation in our dataset. Any data that we observe is drawn from the distribution $\mathbb{P}(X_1, ..., X_m|S = 1)$.

Bareinboim et al. [2014] prove that for discrete variables and under certain positivity assumptions, $\mathbb{P}(Y|X)$ is recoverable from $\mathbb{P}(X, Y|S = 1)$ if and only if $Y \perp\!\!\!\perp S \mid X$. The 'if' part is straightforward, since the conditional independence implies $\mathbb{P}(Y|X) = \mathbb{P}(Y|X, S = 1)$, and the

[1] Quote taken from Mohan and Pearl [2021].

[2] Code for the experiments will be made publicly available.

right-hand-side can be estimated from the data. When this conditional independence is not satisfied, Bareinboim et al. [2014] consider joint measurement of X, Y and some other variable Z in the selection biased dataset (so availability of $\mathbb{P}(X, Y, Z|S = 1)$), satisfying the conditional independence $Y \perp\!\!\!\perp S \mid X, Z$. Additionally, they assume availability of unbiased measurements of (X, Z), i.e. sampled from $\mathbb{P}(X, Z)$, which they refer to as 'external data'. For discrete Z, the quantity $\mathbb{P}(Y|X)$ is then identified with:

$$\begin{aligned}\mathbb{P}(Y|X) &= \sum_z \mathbb{P}(Y|X, Z = z)\mathbb{P}(Z = z|X) \\ &= \sum_z \mathbb{P}(Y|X, Z = z, S = 1)\mathbb{P}(Z = z|X).\end{aligned} \tag{1}$$

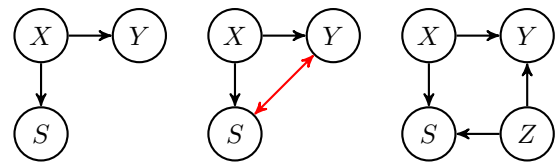
One can straightforwardly replace the sum with an integral when Z is continuous. However, when the domain of Z is countably infinite or continuous, estimating this quantity is not straightforward. The repeated regression estimator proposed in Section 3.1 is a solution to this problem.

The proposed methods can be applied to the selection bias and missingness settings. In either case, we require availability of i.i.d. observations of $(X, Y, Z) \sim \mathbb{P}(X, Y, Z|S = 1)$ whose index set we denote with $\mathcal{S}$, and additional i.i.d. observations of $(X, Z) \sim \mathbb{P}(X, Z)$ whose index set we denote with $\mathcal{D}$. In the missingness setting, both samples are readily available and we have $\mathcal{S} \subseteq \mathcal{D}$. In the selection bias setting the sample $\mathcal{D}$ consists of 'external' data and we typically have $\mathcal{S} \cap \mathcal{D} = \emptyset$. For the methods of Sections 3.2 and 3.3 we additionally require knowledge of the selection mechanism $\mathbb{P}(S = 1|X, Z)$, which is directly estimable in the missingness setting, and which has to be assumed under selection bias. Throughout this paper, any distinctions between missingness and selection bias will be pointed out when necessary. Otherwise, either setting can be assumed.

2.2 REGRESSION UNDER DIFFERENT MISSINGNESS MECHANISMS

Suppose we are interested in a regression $\mathbb{E}[Y|X]$ with univariate, continuous or binary Y, and arbitrary X.[3] We might be confronted with a dataset with missing values of Y, or we might suspect that some selection mechanism is in play which makes certain (X, Y) pairs unobserved. In this section, we investigate whether there is need for any bias correction, i.e. whether selection/missingness is ignorable. As this investigation is based on conditional independence assumptions, we first elaborate how such assumptions can be motivated.

Causal modelling Missingness mechanisms can be characterised by independencies. This is everything we need: all proposed methods will only require certain conditional independencies in the data, and no causal assumptions. However, the conditional independencies that are assumed are typically untestable. For example, the independence $Y \perp\!\!\!\perp S$ is not testable, as we have not observed any data Y for $S = 0$. To motivate such an independence assumption, one could write down the graphical model, and infer from d-separations in the graph that there must be certain independencies in the data [Pearl, 2009]. Especially when we control the data source, one is likely able to confidently write down the graphical model. In the missingness setting where only Y can be missing (with indicator S), we can draw a more simplified graph and discard the Y^o variable, as in Figure 1a. In the graphical framework of s-recoverability, S is implicitly required to be a sink node (i.e. a node without children) [Bareinboim et al., 2014]. We do not adopt this convention.



(a) Ignorable (b) Nonignorable (c) Privilegedly ignorable

Figure 1: Examples of missingness settings, where S indicates whether Y is observed or not.

Ignorable missingness When the data is MCAR or MAR, we have the conditional independence $Y \perp\!\!\!\perp S \mid X$ and thus

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X, S = 1], \tag{2}$$

in which case we call the missingness/selection mechanism *ignorable* for estimating $\mathbb{E}[Y|X]$. An example of an ignorable (MAR) missingness mechanism is given in Figure 1a. When the missingness mechanism is ignorable, any correctly specified model could directly be learned as *Empirical Risk Minimization* (ERM) is consistent [Sugiyama et al., 2007], and hence, there is no need for bias correction.

Although in this case the missingness or selection mechanism can in principle be ignored, one should beware that under model misspecification ERM is not consistent any more. Also, efficiency of the estimation procedure can be affected by MCAR and MAR. Zadrozny [2004] and Wei Fan et al. [2005] investigate the performance of multiple popular classification algorithms under covariate shift. Robins and Rotnitzky [1995] investigate the asymptotic efficiency of semi-parametric estimation of the regression function via Generalised Estimating Equations (GEE). Note that in the missingness setting, we can use the values of X where Y is unobserved as additional (unlabelled) input of a *semi-supervised learning* algorithm. When doing *anticausal* prediction, these additional samples could benefit performance [Schölkopf et al., 2012].

[3] Note that this setting includes binary classification.

Privilegedly ignorable missingness It might be the case that the missingness mechanism is nonignorable. For example, we may suspect that S and Y are confounded by some variable that is not contained in X, as depicted in Figure 1b. In this case we typically have $\mathbb{E}[Y|X, S = 1] \neq \mathbb{E}[Y|X]$, so naive regression on the available data yields a biased model. However, we might be able to identify the latent confounder, or more generally, any variable (mediator or confounder) that induces a dependence between Y and S after conditioning on X.

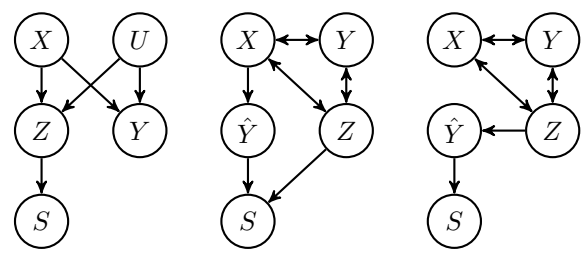
Ideally, we are looking for a set of variables Z such that $Y \perp\!\!\!\perp S \mid X, Z$, such as depicted in Figure 1c. Having identified Z, one might include it into the features of the model and deploy the unbiased model $\hat{\mathbb{E}}[Y|X, Z, S = 1]$. However, in practice it might be costly or impossible to measure Z during deployment, e.g. when Z can only be measured after making a prediction with the regression model. In these settings, it is undesirable to incorporate Z into the features of the deployed model. When Z is only available during training it is referred to as *privileged information*, following Vapnik and Vashist [2009]. The specific setting that we consider in this work is formally defined as follows:

Definition 2 (PMAR)**.** Given variables X and Y, response indicator S for Y, and privilegedly observed variable Z, we say that Y is *Privilegedly Missing at Random* (PMAR) if the missingness mechanism is independent of the missing variable given all other fully and privilegedly observed variables, i.e. $Y \perp\!\!\!\perp S \mid X, Z$.

When the data used for regression is PMAR, we refer to the missingness mechanism as *privilegedly ignorable*.

As a motivation of the PMAR setting, we revisit the example from the introduction on the follow-up study of HIV-positive individuals [Hernán and Robins, 2021]. Recall that we have a certain antiretroviral drug dosage (X), the 3-year risk of death (Y), and the severity of side-effects (Z). For the treated individuals, high levels of immunosuppression (U) induce a greater risk of death, but also increase the severity of their symptoms (Z). When symptoms become too severe, patients may drop out of the study, censoring the measurement of Y. Additional to the described causal relations, any other dependencies might be present between U, X, Y and Z, in which case we still remain in the PMAR setting. A compatible causal graph is depicted in Figure 2a, suppressing any bidirected edges to reduce clutter. Practically, when gathering data for training a prediction model of outcome Y given a dosage X, one should measure the symptoms Z to correct for any bias that is induced via U. In this example one can only measure symptoms *after* the drug has been administered, which makes the information Z privileged.

Another setting where PMAR is applicable, is when individuals or items are actively selected or rejected using predictions from an existing model $\hat{Y} = \hat{\mathbb{E}}[Y|X]$, e.g. in the selection of patients for costly CT scans, dismissal of patients from the intensive care, but also in industrial applications as targeted sales, fraud detection etc. When available data that is affected by such a model is used to retrain the model, and selection is not *only* determined by the old model (but for example also by an experts opinion, variable Z in Figure 2b), one has to correct for the bias that is induced by this additional selection mechanism. Similarly, we might want to replace a current selection model $\hat{\mathbb{E}}[Y|Z]$ with a new model $\hat{\mathbb{E}}[Y|X]$ that uses a different feature set (Figure 2c), or we might have a combination of both. In either case, one should correct for any bias induced via Z when learning a model for $\mathbb{E}[Y|X]$ by jointly measuring X and Z while gathering training data, and then appropriately correcting for any bias, as demonstrated in the following section.



(a) (b) (c)

Figure 2: Causal graphs relating to examples of PMAR.

3 IMPUTATION, WEIGHTING, AND DOUBLE ROBUSTNESS UNDER PMAR

In this section, we propose three estimation procedures for training a regression model $\mathbb{E}[Y|X]$ in the PMAR setting: a repeated regression method, a weighted regression method, and a doubly robust method. Throughout, we consider the following example.

Example 1. Consider the data generating process

$$\begin{aligned} X &= \varepsilon_X \\ Z &= 3\sin(X) + \varepsilon_Z \\ Y &= \frac{1}{2}X + Z + \varepsilon_Y \\ S &\sim \text{Bernoulli}(p_S(X, Z)) \end{aligned} \tag{3}$$

with independent Gaussian random variables $\varepsilon_X, \varepsilon_Z$ and ε_Y. The selection probability is defined as $p_S(x, z) := \sigma(x)\ell(\sigma(z))$ with sigmoid $\sigma(x) = 1/(1 + e^x)$ and $\ell(x) := \frac{19}{20}x + \frac{1}{20}$, i.e. a linear scaling of $\sigma(z)$ between 0.05 and 1 to ensure positivity in the z direction.

Figure 3 displays the graphical model related to this data generating process. Note that we can indeed read off that $Y \perp\!\!\!\perp S \mid X, Z$, so if Z is only available during training, the missingness

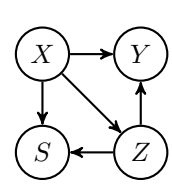


Figure 3

mechanism is PMAR. Figure 4 depicts $n = 400$ draws from the generating model (3). In this example, we have $\#\{S = 1\} = 123$. The black dots indicate (X, Y) pairs when Y is observed ($S = 1$), and the grey dots indicate (X, Y) pairs where Y is unobserved ($S = 0$). The green line shows the true regression line $\mathbb{E}[Y|X = x] = \frac{1}{2}x + 3\sin(x)$, and the black line shows the regression line of the naive, biased estimate $\hat{\mathbb{E}}[Y|X, S = 1]$. In this example and throughout this paper we use thin plate splines regression [Duchon, 1977, Wood, 2003] as implemented in the `mgcv` package [Wood, 2015].

Note that, additional to what is shown in Figure 4, we have all (X, Z) pairs available. Our goal is to fit a regression model that is close to the green line, $\mathbb{E}[Y|X]$. △

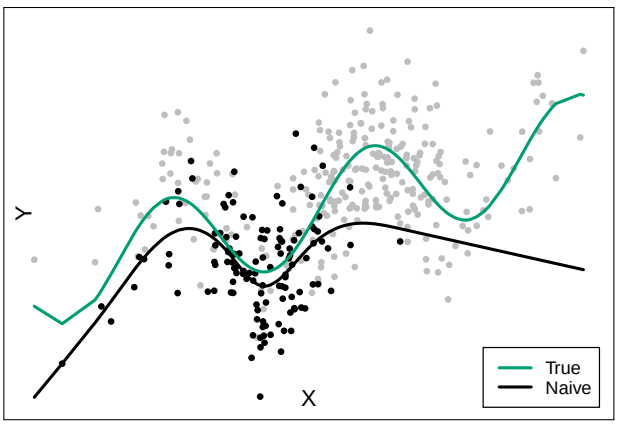


Figure 4: Dataset from Example 1, with observed (black) and unobserved data (grey); the true regression line $\mathbb{E}[Y|X]$ (green) and the naively estimated regression line $\hat{\mathbb{E}}[Y|X, S = 1]$ (black).

Extrapolation and positivity Many missingness methods assume *positivity* of the selection mechanism, i.e. $\mathbb{P}(S = 1|X, Z) > 0$ a.s., as is for example required for the identification equation (1). This assumption ensures that asymptotically, we do not run into the problem of extrapolation. For finite samples, we might run into extrapolation issues when $\mathbb{P}(S = 1|X, Z)$ is close to zero. In the example of Figure 4 we have no positivity for large values of X; we will see that extrapolation can still be made possible by using privileged data Z that is highly predictive of Y.

3.1 REPEATED REGRESSION

By the law of total expectation and the PMAR conditional independence $Y \perp\!\!\!\perp S \mid X, Z$ we can write the conditional expectation $\mathbb{E}[Y|X]$ as follows

$$\begin{aligned}\mathbb{E}[Y|X] &= \mathbb{E}[\mathbb{E}[Y|X, Z]|X] \\ &= \mathbb{E}[\mathbb{E}[Y|X, Z, S = 1]|X],\end{aligned} \tag{4}$$

similar to equation (1). We formulate an estimation procedure based on this expression by estimating each conditional expectation with a regression model. That is, we first regress Y on X and Z using the data $\mathcal{S}$, which results in the regression model $\hat{\mathbb{E}}[Y|X, Z, S = 1]$. Using the unbiased data (x_i, z_i) for $i \in \mathcal{D}$ we construct *pseudo-labels* $\tilde{Y}_i = \hat{\mathbb{E}}[Y|X = x_i, Z_i = z_i, S = 1]$. Now, we regress $\tilde{Y}$ on X using the data $(X_1, \tilde{Y}_1), ..., (X_{|\mathcal{D}|}, \tilde{Y}_{|\mathcal{D}|})$, which yields the final model:

$$\hat{\mu}_{RR}(x) = \hat{\mathbb{E}}[\tilde{Y}|X = x]. \tag{5}$$

Note that this method requires datasets $\mathcal{S}$ and $\mathcal{D}$, and assumes that $Y \perp\!\!\!\perp S \mid X, Z$. So, if this conditional independence is satisfied, it can directly be applied in both the missingness setting, and in the selection bias setting where we have 'external data' $\mathbb{P}(X, Z)$.

Example 1 (*continued*). Recall that the data for Y is generated by $Y = \frac{1}{2}X + Z + \varepsilon_Y$, so for fitting a model $\hat{\mathbb{E}}[Y|X, Z, S = 1]$, although we can only fit this model on a small part of the data (123 out of 400 observations) it is a relatively easy (linear) model to fit. We compute the pseudo-labels $\tilde{Y}_i = \hat{\mathbb{E}}[Y|X = x_i, Z = z_i, S = 1]$, as depicted with orange crosses in Figure 5. Regressing the imputed values on X yields the final model $\hat{\mu}_{RR}$. △

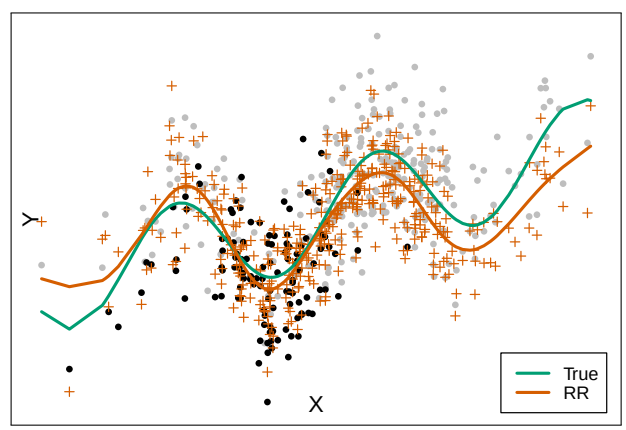


Figure 5: Repeated regression. Orange crosses indicate imputed values, and the orange line is the regression line resulting from repeated regression.

Regression imputation is explicitly used for estimating the mean under missingness via $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X, S = 1]]$ by Bang and Robins [2005]. Similarly, the *average causal effect* can be estimated by $\mathbb{E}[Y|\,\mathrm{do}(X)] = \mathbb{E}[\mathbb{E}[Y|X, Z]]$ for a valid *adjustment set* Z [Hernán and Robins, 2021]. For these estimators, the inner expectation is estimated with a regression model, and the outer expectation is calculated by taking the mean of the pseudo-labels. The novelty in the estimator (5) lies in the fact that the outer expectation is estimated by regression instead of the empirical mean, and the observation that this estimator is applicable to the PMAR setting. Note that in general, automatically generated confidence intervals resulting from the outer regression

$\hat{\mathbb{E}}[\tilde{Y}|X]$ are not valid. To this end, *multiple imputation* methods are often used. As this requires the modelling of the full distribution $\mathbb{P}(Y|X, Z)$, we do not consider this. For an overview of multiple imputation methods in GEE regression, see Ditlhong et al. [2018].

3.2 IMPORTANCE WEIGHTING

For estimating the parameter β^* in the regression model $\mathbb{E}[Y|X] = g(X; \beta^*)$ we often specify a loss function ℓ and do empirical risk minimisation (6) as an approximation of the parameter that is optimal in terms of the true risk (7).

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \ell(g(X_i; \beta), Y_i) \tag{6}$$

$$\beta^* = \arg\min_{\beta} \mathbb{E}[\ell(g(X; \beta), Y)] \tag{7}$$

Writing $f(x, y) := \ell(g(x; \beta), y)$, we can express the risk in terms of the distribution conditional on $S = 1$ using importance weighting:

$$\mathbb{E}[f(X, Y)] = \mathbb{E}[w(X, Z) f(X, Y)|S = 1], \tag{8}$$

with *importance weights* $w(x, z) := \mathbb{P}(S = 1)/\mathbb{P}(S = 1|X = x, Z = z)$. A derivation of these importance weights can be found in the supplementary material. Since $\beta^* = \arg\min_\beta \mathbb{E}[w(X, Z)\ell(g(X; \beta), Y)|S = 1]$, when we have observations $(X_i, Z_i, Y_i) \sim \mathbb{P}(X, Z, Y|S = 1)$ for $i \in \mathcal{S}$, we can directly do empirical risk minimization on this dataset using the weighted loss:

$$\hat{\beta}_w := \arg\min_{\beta} \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} w(X_i, Z_i)\ell(g(X_i; \beta), Y_i), \tag{9}$$

and have $\hat{\beta}_w \approx \beta^*$. Many implementations of regression algorithms allow the user to specify such sample weights. When the weights are not known in the missingness setting, they can be estimated [Cortes et al., 2008]. Practically, when the selection probability $\mathbb{P}(S = 1|x_i, z_i)$ is nearly zero for particular drawn values x_i, z_i, these probabilities are 'clipped' (i.e. transformed to remain lower bounded by some pre-specified strictly positive value) to reduce variance. In the experiments, we consider a linear map of the array of selection probabilities to $[1/20, 1]$, as this yields the best performance among different clipping strategies.

Example 1 (*continued*). The fitted IW regression model is depicted in Figure 6. Here, we used the true weights $w(x_i, z_i)$. The size of the observed points indicate the associated weight. Note that the IW regression only uses the black points (i.e. the (x_i, y_i) pairs for which $S_i = 1$). Comparing to repeated regression (Figure 5), we see that IW extrapolates relatively poorly, possibly due to the effect of importance weighting on regularization [Xu et al., 2021]. On the other hand, IW interpolates better than the naive model. △

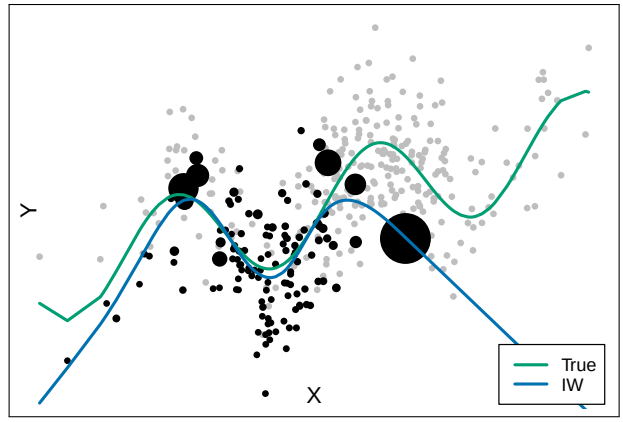


Figure 6: IW regression with true weights.

The idea of weighting observations stems from the Horvitz-Thompson estimator for the population mean [Horvitz and Thompson, 1952]. The use of importance weights in GEE regression in the MAR setting is analysed by [Robins et al., 1994, Robins and Rotnitzky, 1995], and its MNAR counterpart is described by Scharfstein et al. [1999]. For weighted SVM- and kernel regression under covariate shift and target shift, see Zhang et al. [2013].

3.3 DOUBLY ROBUST REGRESSION

We now follow a relatively standard procedure to combine the two previous models into a doubly robust model [Bang and Robins, 2005, Kang and Schafer, 2007, Dudík et al., 2014]. First, we consider the repeated regression method $\hat{\mu}_{RR}$ from Section 3.1. We calculate the residuals of this method on the available Y values: we set $R_i := Y_i - \hat{\mu}_{RR}(X_i)$ for all $i \in \mathcal{S}$. We model the conditional expectation of the residuals given X using the IW regression method from Section 3.2, as

$$\hat{r}_{IW}(x) := g(X; \hat{\beta}_w) \approx \mathbb{E}[R|X] \tag{10}$$

where $\hat{\beta}_w = \arg\min_\beta |\mathcal{S}|^{-1} \sum_{i \in \mathcal{S}} w(X_i, Z_i)\ell(g(X_i; \beta), R_i)$ and where we use the same importance weights w as in Section 3.2. Then we define the *doubly robust* model

$$\hat{\mu}_{DR}(x) := \hat{\mu}_{RR}(x) + \hat{r}_{IW}(x). \tag{11}$$

Here, double robustness refers to the fact that for this model to be consistent, only one of $\hat{\mu}_{RR}$ and $\hat{r}_{IW}$ has to be consistent.

Example 1 (*continued*). For sake of exposition, we apply the doubly robust method with a misspecified model $\hat{\mu}_{RR}(x)$. Recall that $\hat{\mu}_{RR}(x) = \hat{\mathbb{E}}[\tilde{Y}|X = x]$, where $\tilde{Y}$ are the values that are imputed with the model $\hat{\mathbb{E}}[Y|X, Z, S = 1]$. To make the model deliberately misspecified, we use the same $\tilde{Y}$ as in Section 3.1 (more specifically, as in Figure 5), but for the outer regression $\hat{\mathbb{E}}[\tilde{Y}|X]$ we employ polynomial regression

with degree 5. The resulting models are depicted in Figure 7. We see that $\hat{\mu}_{DR}$ extrapolates poorly, but it interpolates better than the misspecified model $\hat{\mu}_{RR}$. △

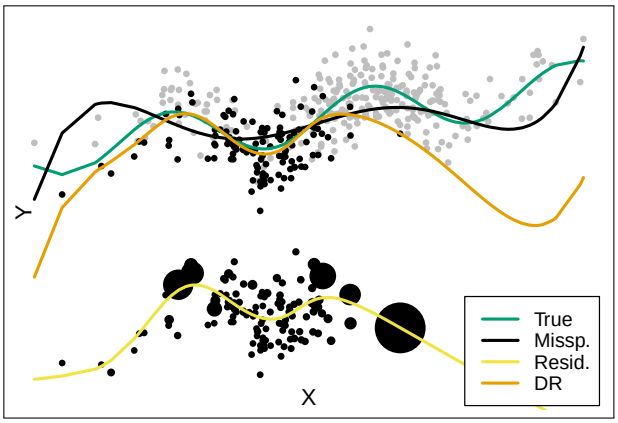


Figure 7: Doubly robust regression. At the top the (X, Y) pairs are plotted, with the true model (green) and misspecified RR model (black). The residuals of the misspecified model are plotted at the bottom. The IW regression model for the residuals (yellow) is then added to the misspecified model, which yields the doubly robust model (orange).

For overview papers on doubly robust estimation of the mean, see Bang and Robins [2005] and Kang and Schafer [2007]. For doubly robust regression in the MNAR setting, see Rotnitzky et al. [1998] and Scharfstein et al. [1999]. Doubly robust estimation of causal effects is treated by Bhattacharya et al. [2022], among others.

4 EVALUATION

When all values of Y are known, performance is typically assessed using MSE $:= |\mathcal{D}|^{-1} \sum_{i \in \mathcal{D}} (\hat{y}_i - y_i)^2$. Under missingness, this quantity cannot be calculated on a train, test or validation set as it requires the unobserved values y_i for $i \notin \mathcal{S}$. Quantities that *can* be calculated on the training set are as follows:

$$\text{MSE-n} := \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} (\hat{y}_i - y_i)^2 \tag{12}$$

$$\text{MSE-}w := \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} w(x_i, z_i)(\hat{y}_i - y_i)^2 \tag{13}$$

$$\text{MSE-}\tilde{y} := \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} (\hat{y}_i - \tilde{y}_i)^2, \tag{14}$$

which are respectively the naively calculated MSE, the reweighted MSE using the true weights, and the MSE calculated using the imputed values $\tilde{y}$. We define MSE-$\hat{w}$ similarly, in terms of the estimated weights.

Example 1 (*continued*). The IW and DR models are learned both with true and estimated weights (denoted with -t and -e respectively), and all weights are clipped. We calculate the different mean squared error metrics for the estimated models on 500 independently drawn test sets of $n = 400$ samples. The results are provided in Table 1. We emphasize

	MSE	MSE-n	MSE-$\tilde{y}$	MSE-w	MSE-$\hat{w}$
Naive	24.07	**7.56**	19.54	16.19	19.79
RR	**8.81**	8.13	**4.44**	**8.50**	**9.34**
IW-t	40.42	8.40	35.58	22.22	24.67
IW-e	38.44	8.28	33.60	21.49	23.08
DR-t	25.91	8.13	21.27	14.82	16.65
DR-e	24.09	8.00	19.43	13.96	15.52
True	8.03	7.78	4.63	7.87	8.87

Table 1: Mean squared errors of different regression methods applied to Example 1, averaged over 500 test sets.

that MSE is not calculable in practice, and note that MSE-n, MSE-w and MSE-$\hat{w}$ yield qualitatively different results than the oracle MSE. MSE-$\tilde{y}$ provides the same ordering as MSE, but in practice the reliance on the imputation model can be undesirable. We observe RR appropriately corrects for bias, as its MSE lies close to the MSE of the true function $\mathbb{E}[Y|X]$. Additional to the oracle MSE, we calculate the MSEs specifically for data points where the regression model is interpolating (between the minimum and maximum values of X for which $S = 1$) and extrapolating (the complement of these points):

	MSE	MSE-interp.	MSE-extrap.
RR	**8.81**	**8.43**	**10.17**
IW-t	40.42	12.00	138.22
IW-e	38.44	11.77	129.41

We observe that IW extrapolates rather poorly, especially compared to RR. △

When the true weights are not known, one can only evaluate the proposed bias correction methods on biased data by relying on an auxiliary model, i.e. either the imputation model $\hat{\mathbb{E}}[Y|X, Z, S = 1]$ or $\hat{\mathbb{P}}(S = 1|X, Z)$. In general, if the practitioner is able to evaluate on an unbiased random sample, this would be much more reliable than evaluation on a biased sample.

5 EXPERIMENTS

The example that is used throughout this paper is quite extreme in the sense that the model $\mathbb{E}[Y|X, Z]$ is relatively simple and $\mathbb{E}[Y|X]$ is rather complex, which gives RR a clear advantage over IW. To investigate the proposed methods in a more general setting, we assess performance on an extensive simulation setup and on the Boston Housing dataset.

5.1 SIMULATIONS

To assess the performance of the proposed methods empirically, we first identify which graphs with variables X, Y, Z and S satisfy the pattern of d-separations $X \not\perp Y$ (such that regression $\mathbb{E}[Y|X]$ makes sense), $Y \not\perp S|X$ and $Y \perp S \mid X, Z$ (PMAR). There are 126 *Acyclic Directed Mixed Graphs* (ADMGs) that fit this pattern. An overview of the valid ADMGs is provided in the supplements. For each ADMG we simulate 60 datasets, and for each dataset we draw $n = 1000$ samples according to an additive noise structural equation model, where all equations for X, Y, Z are random functions, independently drawn from a Gaussian process, and the additive noise is the pushforward of a Unif$[0, 1]$ distribution through a random function that is also drawn from a Gaussian process [Mooij et al., 2016]. We let the selection probability be $p(x, z) = \sigma(x)\sigma(z)$, when X and Z are parents of S in the ADMG. For a complete description of the simulation setup, we refer to the supplements. On each dataset, we fit a 'naive' model $\hat{\mathbb{E}}[Y|X, S = 1]$ and the models $\hat{\mu}_{RR}, \hat{\mu}_{IW}$ and $\hat{\mu}_{DR}$ from Section 3. We also fit a 'true' model $\hat{\mathbb{E}}[Y|X]$ for comparison. In Table 2 we report the different mean squared error metrics

	MSE	MSE-$\tilde{y}$	MSE-w	MSE-$\hat{w}$
Naive	2.75 (8.5)	2.47 (8.3)	0.80 (0.8)	0.81 (0.8)
RR	**1.80** (1.8)	**0.56** (0.8)	1.09 (0.9)	1.05 (0.9)
IW-t	3.90 (12.1)	3.57 (11.9)	0.82 (0.9)	0.82 (0.8)
IW-e	3.62 (11.3)	3.26 (10.8)	0.82 (0.9)	0.81 (0.8)
DR-t	4.16 (14.0)	4.18 (15.5)	**0.69** (0.4)	0.72 (0.4)
DR-e	3.70 (11.9)	3.68 (13.2)	0.71 (0.4)	**0.71** (0.4)
True	0.88 (0.4)	1.49 (2.2)	0.94 (0.6)	0.92 (0.5)

Table 2: MSE results over 7560 simulated datasets.

of the methods, averaged over all 7560 tests sets (standard deviations are shown in parentheses). We see that the MSE of RR is less than half the MSE of the best IW method. The median MSEs follow the relation RR $<$ Naive $<$ IW-t, confirmed with respective p-values $1.5 \cdot 10^{-40}$ and $3.9 \cdot 10^{-25}$ of Wilcoxon's signed rank test. In ADMGs where X is a parent of S there is a clear region of the support of X with no positivity of S, hence a clear region where the regression model is extrapolating, similar to Example 1. When singling out these ADMGs we calculate the average MSE, and additionally the MSEs on the parts where the regression model is inter- and extrapolating, as shown in the table below. These results indicate that IW does not extrapolate well.

	MSE	MSE-interp.	MSE-extrap.
RR	**2.02** (2.2)	1.17 (1.1)	**3.06** (4.7)
IW-t	7.11 (16.9)	1.08 (0.6)	13.82 (32.6)
IW-e	6.27 (15.5)	**1.08** (0.6)	11.85 (28.8)

5.2 BOSTON HOUSING DATA

In absence of a real-world dataset where there is missingness *and* the missing value is known, we consider the Boston Housing dataset [Harrison Jr and Rubinfeld, 1978] and simulate a selection mechanism ourselves. We consider the problem of predicting the value of owner-occupied homes in US Dollars (Y) from the number of rooms per dwelling (X). We let selection depend on X, and on the 'percentage of people of lower status of the population in the town where the house is situated' (Z), which strongly correlates with Y. Selection is simulated by setting $\mathbb{P}(S = 1|X, Z) := \sigma(f_1(X))\sigma(f_2(Z))$, where f_1, f_2 are random functions drawn from a Gaussian process. For each dataset, the 506 available observations are randomly split into evenly sized train and test sets. The average MSEs are provided in Table 3, and show the same qualitative results as the simulations. When calculating the MSE of the inter- and extrapolation regions, we observe that RR performs better than IW at both tasks (table shown in supplements).

	MSE	MSE-$\tilde{y}$	MSE-w	MSE-$\hat{w}$
Naive	1.23 (2.5)	0.84 (2.3)	0.48 (0.7)	1.79 (5.9)
RR	**0.71** (0.3)	**0.26** (0.3)	0.44 (0.4)	1.63 (4.8)
IW-t	2.18 (4.9)	1.77 (4.9)	0.56 (1.5)	2.40 (9.2)
IW-e	1.75 (4.4)	1.40 (4.7)	0.50 (0.9)	1.98 (7.4)
DR-t	1.92 (3.7)	1.65 (3.4)	**0.23** (0.2)	0.39 (1.2)
DR-e	2.43 (5.4)	2.20 (5.5)	0.25 (0.2)	**0.21** (0.3)
True	0.45 (0.1)	0.54 (0.4)	0.36 (0.3)	0.85 (2.2)

Table 3: Results of 500 biased instantiations of the Boston Housing dataset.

6 DISCUSSION AND CONCLUSION

In this work, we have motivated the use of *privileged information* when selection or missingness is nonignorable for estimating a regression model, and introduced the *Privilegedly Missing at Random* (PMAR) setting. We formulated the repeated regression method, the IW regression method, and the doubly robust combination of the two. We note that evaluation of regression methods on biased data is not straightforward and relies on auxiliary models. Experiments show that repeated regression can appropriately correct for bias, and with considerable advantage over IW regression. In particular, in recursive regression, extrapolation is facilitated by privileged data Z that is predictive of Y. IW does not have such a property, and extrapolates worse than RR.

Further research can be done on the statistical properties of the proposed methods, e.g. on bounding the regret of DR in terms of the regret of RR and IW (possibly under misspecification), or on the interplay of importance weighting, regularization, and extrapolation.

References

H. Bang and J. M. Robins. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61(4):962–972, 2005. ISSN 0006-341X. URL https://www.jstor.org/stable/3695907.

E. Bareinboim, J. Tian, and J. Pearl. Recovering from Selection Bias in Causal and Statistical Inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, page 12, 2014.

R. Bhattacharya, R. Nabi, and I. Shpitser. Semiparametric Inference For Causal Effects In Graphical Models With Hidden Variables. *Journal of Machine Learning Research*, 23(295):1–76, 2022. ISSN 1533-7928. URL http://jmlr.org/papers/v23/20-296.html.

C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample Selection Bias Correction Theory. In Y. Freund, L. Györfi, G. Turán, and T. Zeugmann, editors, *Algorithmic Learning Theory*, volume 5254, pages 38–53. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-87986-2 978-3-540-87987-9. doi: 10.1007/978-3-540-87987-9_8. URL http://link.springer.com/10.1007/978-3-540-87987-9_8.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. ISSN 0035-9246. URL https://www.jstor.org/stable/2984875.

K. E. Ditlhong, O. O. Ngesa, and A. Y. Kombo. A Comparative Analysis of Generalized Estimating Equations Methods for Incomplete Longitudinal Ordinal Data with Ignorable Dropouts. *Open Journal of Statistics*, 8(5):770–792, Sept. 2018. doi: 10.4236/ojs.2018.85051. URL http://www.scirp.org/Journal/Paperabs.aspx?paperid=87364.

J. Duchon. Splines minimizing rotation-invariant seminorms in Sobolev spaces. In W. Schempp and K. Zeller, editors, *Constructive Theory of Functions of Several Variables*, Lecture Notes in Mathematics, pages 85–100, Berlin, Heidelberg, 1977. Springer. ISBN 978-3-540-37496-1. doi: 10.1007/BFb0086566.

M. Dudík, D. Erhan, J. Langford, and L. Li. Doubly Robust Policy Evaluation and Optimization. *Statistical Science*, 29(4), Nov. 2014. ISSN 0883-4237. doi: 10.1214/14-STS500. URL http://arxiv.org/abs/1503.02834.

D. Harrison Jr and D. L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

J. J. Heckman. Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153, Jan. 1979. ISSN 00129682. doi: 10.2307/1912352. URL https://www.jstor.org/stable/1912352?origin=crossref.

M. Hernán and J. M. Robins. *Causal Inference*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, Boca Raton, 2021. ISBN 978-1-4200-7616-5.

D. G. Horvitz and D. J. Thompson. A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. ISSN 0162-1459. doi: 10.2307/2280784. URL https://www.jstor.org/stable/2280784.

J. D. Y. Kang and J. L. Schafer. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4), Nov. 2007. ISSN 0883-4237. doi: 10.1214/07-STS227. URL http://arxiv.org/abs/0804.2958.

R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

K. Mohan and J. Pearl. Graphical Models for Processing Missing Data. *Journal of the American Statistical Association*, 116(534):1023–1037, Apr. 2021. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2021.1874961. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2021.1874961.

J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016. ISSN 1533-7928. URL http://jmlr.org/papers/v17/14-518.html.

J. Pearl. *Causality*. Cambridge university press, 2009.

J. Pearl. A solution to a class of selection-bias problems, 2012.

J. M. Robins and A. Rotnitzky. Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*, 90(429):122–129, 1995. ISSN 0162-1459. doi: 10.2307/2291135. URL https://www.jstor.org/stable/2291135.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. ISSN 0162-1459. doi: 10.2307/2290910. URL https://www.jstor.org/stable/2290910.

P. R. Rosenbaum and D. B. Rubin. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, 79(387):516–524, 1984. ISSN 0162-1459. doi: 10.2307/2288398. URL https://www.jstor.org/stable/2288398.

A. Rotnitzky, J. M. Robins, and D. O. Scharfstein. Semiparametric Regression for Repeated Outcomes with Nonignorable Nonresponse. *Journal of the American Statistical Association*, 93(444):1321–1339, 1998. ISSN 0162-1459. doi: 10.2307/2670049. URL https://www.jstor.org/stable/2670049.

D. B. Rubin. Inference and Missing Data. *Biometrika*, 63(3):581–592, 1976. ISSN 0006-3444. doi: 10.2307/2335739. URL https://www.jstor.org/stable/2335739.

D. O. Scharfstein, A. Rotnitzky, and J. M. Robins. Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999. ISSN 0162-1459. doi: 10.2307/2669923. URL https://www.jstor.org/stable/2669923.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 459–466, Madison, WI, USA, June 2012. Omnipress. ISBN 978-1-4503-1285-1.

M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate Shift Adaptation by Importance Weighted Cross Validation. *The Journal of Machine Learning Research*, 8: 985–1005, Dec. 2007. ISSN 1532-4435.

V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, July 2009. ISSN 0893-6080. doi: 10.1016/j.neunet.2009.06.042. URL https://www.sciencedirect.com/science/article/pii/S0893608009001130.

Wei Fan, I. Davidson, B. Zadrozny, and Philip S. Yu. An Improved Categorization of Classifier's Sensitivity on Sample Selection Bias. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 605–608, Houston, TX, USA, 2005. IEEE. ISBN 978-0-7695-2278-4. doi: 10.1109/ICDM.2005.24. URL http://ieeexplore.ieee.org/document/1565737/.

S. Wood. Package 'mgcv'. *R package version*, 1(29):729, 2015.

S. N. Wood. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114, 2003. ISSN 1467-9868. doi: 10.1111/1467-9868.00374. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00374.

D. Xu, Y. Ye, and C. Ruan. Understanding the role of importance weighting for deep learning, Mar. 2021. URL http://arxiv.org/abs/2103.15209.

B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114, 2004.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain Adaptation under Target and Conditional Shift. In *Proceedings of the 30th International Conference on Machine Learning*, pages 819–827. PMLR, May 2013. URL https://proceedings.mlr.press/v28/zhang13d.html.