Domain Adaptation under Missingness Shift

Helen Zhou, Sivaraman Balakrishnan, Zachary C. Lipton

Carnegie Mellon University
{ hlzhou, sbalakri, zlipton}@andrew.cmu.edu

Abstract

Rates of missing data often depend on record-keeping policies and thus may change across times and locations, even when the underlying features are comparatively stable. In this paper, we introduce the problem of Domain Adaptation under Missingness Shift (DAMS). Here, (labeled) source data and (unlabeled) target data would be exchangeable but for different missing data mechanisms. We show that if missing data indicators are available, DAMS reduces to covariate shift. Addressing cases where such indicators are absent, we establish the following theoretical results for underreporting completely at random: (i) covariate shift is violated (adaptation is required); (ii) the optimal linear source predictor can perform arbitrarily worse on the target domain than always predicting the mean; (iii) the optimal target predictor can be identified, even when the missingness rates themselves are not; and (iv) for linear models, a simple analytic adjustment yields consistent estimates of the optimal target parameters. In experiments on synthetic and semi-synthetic data, we demonstrate the promise of our methods when assumptions hold. Finally, we discuss a rich family of future extensions.

1 Introduction

As of October 2021, following extensive awareness campaigns and mass distribution efforts promoting COVID-19 vaccines, approximately 79.2% of the U.S. population over age 18 had received at least one dose (CDC, 2022). And yet, when collaborating with a regional healthcare provider, we found only 40.5% of 121,329 adults tested for COVID-19 were tagged indicating positive vaccination status in the electronic medical record (EMR). This was not a regional anomaly—cross referencing with vaccination data from the CDC, between 75.7% and 90.3% of the adult population in the region had actually received at least one dose. A more plausible explanation is that many patients were vaccinated outside of the hospital system (e.g., at a pharmacy or football stadium) but that this information was never reported to the hospital system and thus never captured in the EMR.

Now suppose that our collaborator decided to update their intake form to include a question about vaccination status. Overnight, the rate of patients being tagged in the EMR as vaccinated would increase dramatically. Absent any shift in the actual health status of patients, the distribution of observed data would still shift, owing to this sudden change in clerical practices. In real-world healthcare settings, such changes in missingness rates are common. Furthermore, as in our vaccination example, indicators disambiguating which features are genuinely negative (vs. missing) cannot be taken for granted. Faced with data from different time periods or locations, each characterized by different patterns of missing data, how should machine learning (ML) practitioners leverage the available data to get the best possible predictor on a target domain? While missing data and formal models of distribution shift are both salient concerns of the ML community, no work to date provides guidance on how to adjust a predictor under such shocks.

In this work, we introduce *missingness shift*, where distributional shocks arise due to changes in the pattern of missingness (Figure 1). In this setup, all domains share a fixed underlying distribution $P(X, Y)$, and observed covariates $\widetilde{X}$ are produced by stochastically zeroing out a subset of the underlying *clean* covariates, i.e., each $\widetilde{X} = X \odot \xi$ for some $\xi \in \{0, 1\}^d$. We propose the **Domain Adaptation under Missingness Shift** problem, where the learner aspires to recover the optimal target predictor given *labeled* data from the source distribution $P^s(\widetilde{X}, Y)$, and unlabeled data from the target (deployment) distribution $P^t(\widetilde{X})$.

We focus primarily on a special DAMS setting where the components of ξ's (one per feature) are drawn from independent Bernoullis with unknown constant probabilities. First, we show that when missingness indicators

$(1-\xi)$ are available, missingness shift is an instance of covariate shift. However, absent indicators, missingness shift constitutes neither covariate shift nor label shift. Thus, adaptation is required. We demonstrate that under DAMS, the optimal source predictor may even exhibit arbitrarily higher MSE than just guessing the label mean $\mathbb{E}[Y]$. One natural strategy might be to relate the source and target distributions to the underlying clean distribution, which we show is identified when missingness rates are known. However, we show that the missingness rates are not, in general, identifiable. Fortunately, as we prove, the target distribution (and thus optimal target predictor) is nevertheless identifiable, requiring only that we estimate the (observable) relative proportions of nonzero values for each covariate across domains. Using these relative proportions, we derive a simple adjustment formula that yields the optimal linear predictor on the target domain. Additionally, we provide a non-parametric, model-agnostic procedure which attempts to transform source data into labeled data i.i.d. to the target distribution. Finally, we confirm the validity of our approach and demonstrate empirical gains in settings where our assumptions hold through synthetic and semi-synthetic experiments.

2 Related Work

There is a rich history of learning under various missing data mechanisms when missing data indicators are available (Rubin, 1976; Robins et al., 1994; Little and Rubin, 2019; Gelman et al., 2020). Common practices for handling missing data include discarding all samples with missingness (complete-case analysis) (Little and Rubin, 2019), imputing with mean or last value carried forward, combining inferences from multiple imputations (Rubin, 1996; Van Buuren and Groothuis-Oudshoorn, 2011), matching-based algorithms, iterative regression imputation (Stekhoven and Bühlmann, 2012; Le Morvan et al., 2021), building missingness indicators into model architecture (Le Morvan et al., 2020a), and including missingness indicators as features (Groenwold et al., 2012; Lipton et al., 2016; Little and Rubin, 2019). However, these techniques require indicators for whether each covariate is missing in the first place.

In single cell RNA sequencing, missing data indicators are often absent in count data due to dropout, where a tiny proportion of the transcripts in each cell are sequenced, so expressed transcripts can go undetected and are instead assigned a zero value. This is often dealt with by leveraging domain-specific knowledge to inform probabilistic models, such as assuming a zero-inflated negative binomial distribution of counts (Risso et al., 2018), using a mixture model to identify likely missing values before imputing with nonnegative least squares regression (Li and Li, 2018), adopting a Bayesian approach to estimate a posterior distribution of gene expressions (Huang et al., 2018), or graph-based methods on a lower dimensional manifold derived from principal component analysis (Van Dijk et al., 2018).

In survey data, underreporting (i.e. missingness without indicators) arises in binary data when respondents give false negative responses to questions. As noted in Sechidis et al. (2017), this can be viewed as a form of misclassification bias. In its simplest form, an underreported variable has *specificity* $p(\widetilde{x}=0|x=0)=1$ and *sensitivity* $p(\widetilde{x}=1|x=1)<1$ (one minus the rate of missingness). If sensitivity is independent of outcome Y, this is referred to as *non-differential* misclassification, which often, but not always biases measures of association towards zero (Dosemeci et al., 1990; Brenner and Loomis, 1994). Given knowledge of the specificities and sensitivities, prior work has derived adjusted estimators for the log-odds ratio (Chu et al., 2006) and relative-risk (Rahardja and Young, 2021) under non-differential exposure misclassification. Recent work has also provided conditions under which the joint distribution $p(y,\widetilde{a}|x)$ (outcome y, single binary underreported exposure $\widetilde{a}$, and fully observed covariates x) is identifiable (Adams et al., 2019).

In our setting, for binary covariates, estimating the missingness rates takes the form of learning from positive and unlabeled data (Elkan and Noto, 2008; Bekker and Davis, 2020). Here, identification of the missingness rates hinges on the existence of a separable positive subdomain (Garg et al., 2021), which may not hold in problems of interest. Many canonical distribution shift problems address adaptation under different forms of structure, including covariate shift (Shimodaira, 2000; Zadrozny, 2004; Huang et al., 2006; Sugiyama et al., 2007; Gretton et al., 2009), label shift (Saerens et al., 2002; Storkey, 2009; Zhang et al., 2013; Lipton et al., 2018; Garg et al., 2020), and concept drift (Tsymbal, 2004; Gama et al., 2014). We show that missingness shift with missing data indicators can often be reinterpreted as a form of covariate shift, but to our knowledge, missingness shift without indicators does not fit neatly into any previous setting.

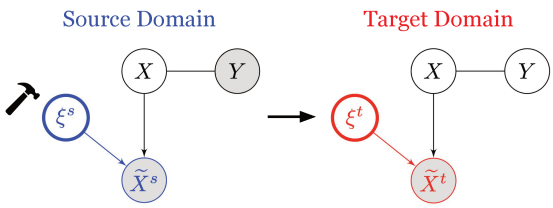


Figure 1: DAMS with UCAR. The source and target data are drawn from the same $P(X, Y)$, but differ in how ξ (and hence $\widetilde{X}$) takes its value. Shaded nodes are observed. Observed covariates are generated as $\widetilde{X} = X \odot \xi$. The undirected edge between X and Y indicates that they can have an arbitrary bidirectional relationship.

3 DAMS Problem Setup

First, we define notation for (1) missing data; (2) missingness shift; and (3) the DAMS problem. Then, motivated by the medical setting, we focus on a specific form of DAMS (Figure 1) for the remainder of the paper.

Let us denote *clean* covariates $X \in \mathbb{R}^d$ and labels $Y \in \mathbb{R}$. Let X_j denote the jth covariate, for $j \in \{1, 2, ..., d\}$.

Missing Data In every environment e with missing data, we do not directly observe X, but instead observe *corrupted* covariates:

$$\widetilde{X} = X \odot \xi,$$

where $\xi \in \{0, 1\}^d$ and $(X, Y, \xi) \sim P^e$ for distribution P^e. Note that mask ξ is the complement of missing data indicators $(1 - \xi)$. In this paper, we assume no missingness in Y in labeled data. An important assumption of missing data problems is how ξ takes its value, e.g. independent of other covariates, dependent on other covariates, etc. Furthermore, ξ may or may not be observed.

Definition 1 (Missingness Shift)**.** *Consider a source domain s and target domain t in which X and Y are drawn from the same underlying distribution, i.e. $P(X, Y) = P^s(X, Y) = P^t(X, Y)$.* Missingness shift *occurs when the missing data mechanism differs between s and t, i.e. $P^s(\xi|\cdot) \neq P^t(\xi|\cdot)$.*

Domain Adaptation under Missingness Shift Suppose missingness shift occurs between source domain s and target domain t. Given observations of corrupted labeled source data $\{(\widetilde{X}^{s,i}, Y^{s,i})\}_{i=1}^{n_s}$ where $(\widetilde{X}^{s,i}, Y^{s,i}) \sim P^s(\widetilde{X}, Y)$, as well as corrupted unlabeled target data $\{\widetilde{X}^{t,i}\}_{i=1}^{n_t}$ where $\widetilde{X}^{t,i} \sim P^t(\widetilde{X})$, the goal of DAMS is to learn an optimal predictor on the corrupted target domain data. In this paper, we focus on regression-type tasks, where optimality is measured by the squared error on the corrupted target domain data, and we seek the optimal predictor $\mathbb{E}_{(\widetilde{X}^t, Y) \sim P^t}[Y|\widetilde{X}^t]$.

As we will show (in Section 4), DAMS is particularly challenging when missing data indicators are *not available*. This setting without observing ξ is trickiest when there are a substantial number of true 0 values that now become indistinguishable from missing values. Without knowledge of which data are missing versus true 0s, conventional techniques for imputing missing entries do not apply. To make this difficult setting tractable, we define the DAMS with underreporting completely at random (UCAR) setting, which we focus on in this paper.

DAMS with UCAR Assume that ξ (unobserved) is drawn independently of other variables, and parameterized by *constant (but unknown) missingness rates* $m^s \in [0, 1]^d$ in source and $m^t \in [0, 1]^d$ in target. That is, $\forall j \in \{1, 2, ..., d\}$, we have independently drawn $\xi_j^s \sim \text{Bernoulli}(1 - m_j^s)$ and $\xi_j^t \sim \text{Bernoulli}(1 - m_j^t)$, abbreviated as:

$$\xi^s \sim \text{Bernoulli}(1 - m^s)$$
$$\xi^t \sim \text{Bernoulli}(1 - m^t).$$

For binary data, this setting without missingness indicators is known as *underreporting*. We thus refer to this setting as underreporting completely at random, but note our results are not limited to binary data.

4 Cost of Non-Adaptivity

Here, we provide intuition on the cost of not adapting the source predictor to the target domain in DAMS with UCAR. Let us start with a simple motivating example. Define the risk of an estimator $\widehat{h}$ to be $r(\widehat{h}) = \mathbb{E}[(Y - \widehat{h}(X))^2]$.

Example 1 (Redundant Features). *Let $m^s = [1-\epsilon, \epsilon]$ and $m^t = [\epsilon, 1-\epsilon]$. Consider the data generating process:*

$$\begin{aligned} Z &= u_Z & \\ X_1 &= Z & u_Z &\sim \mathcal{N}(0, \sigma_z^2) \\ X_2 &= Z & u_Y &\sim \mathcal{N}(0, \sigma_y^2) \\ Y &= Z + u_Y & \end{aligned}$$

where σ_z is a positive constant, Z is a latent variable, X_1 and X_2 are observed, and Y is the outcome of interest.

Remark 1. *In Example 1, as $\epsilon \to 0$, the optimal linear source and target predictors have coefficients $\beta_*^s \to [0, 1]$ and $\beta_*^t \to [1, 0]$. The risk on target data $r^t(\beta_*^s) \to Var(Y)$.*

That is, failing to adapt to the target levels of missingness results in performance no better than simply guessing the label mean (proof in Appendix A). Now, let us consider a slightly more complex example.

Example 2 (Confounded Features). *Now, suppose that $m^s = [0, 0]$ and $m^t = [1, 0]$. For some constants a, b, c consider the following data generating process:*

$$\begin{aligned} X_1 &= \nu_1 & \nu_1 &\sim \mathcal{N}(0, 1) \\ X_2 &= aX_1 + \nu_2 & \nu_2 &\sim \mathcal{N}(0, 1) \\ Y &= bX_1 + cX_2 + \nu_Y & \nu_Y &\sim \mathcal{N}(0, 1) \end{aligned}$$

Remark 2. *In Example 2, the optimal source and target predictors are $\beta_*^s = [b, c]$ and $\beta_*^t = [0, \frac{ab}{a^2+1} + c]$. By setting $a = -\frac{b}{c}$, we can show that for any $\tau > 1$, there exists values of a, b, c such that $r^t(\beta_*^s) > \tau Var(Y)$.*

Here, failing to adapt to target levels of missingness can result in performance arbitrarily *worse* than predicting the constant label mean (proof in Appendix A).

Observing ξ, Reduction to Covariate Shift In DAMS with UCAR, missing data indicators are absent. By contrast, *suppose we observed missingness indicators* $(1 - \xi)$ (and hence ξ). Then, we show that missingness shift is an instance of covariate shift, where the optimal predictor does not change across domains. This result holds not only when ξ is drawn independently of other covariates, but also when it is dependent on other completely observed covariates (proof in Appendix B). Here, when ξ is "drawn independently of other covariates," as described in the DAMS with UCAR setup (Section 3), we have that $\xi \sim \text{Bernoulli}(1 - m)$ for some constant vector of missingness rates $m \in [0, 1]^d$. When ξ is drawn depending only on other completely observed covariates, we have that some subset of covariates $X_c \subseteq X$ is completely observed (i.e. no missingness), and the missingness of the other covariates $X_m = X \setminus X_c$ depends on X_c. That is, $\xi \sim \text{Bernoulli}(f(X_c))$ for some function $f : \mathbb{R}^{|X_c|} \to [0, 1]^{|X_m|}$. Mohan and Pearl (2021) classifies these missingness mechanisms as MCAR (missing completely at random) and v-MAR (variant of the missingness at random described by Rubin (1976)), respectively.

Proposition 1 (Reduction to Covariate Shift). *Assume we observe ξ. Consider augmented covariates $\tilde{x}' = (\tilde{x}, \xi)$. When ξ is drawn independently of other covariates or depending only on other completely observed covariates, missingness shift satisfies the covariate shift assumption, i.e, $P^s(Y|\widetilde{X}' = \tilde{x}') = P^t(Y|\widetilde{X}' = \tilde{x}')$.*

Covariate shift problems are well-studied (Shimodaira, 2000; Zadrozny, 2004; Huang et al., 2006; Sugiyama et al., 2007; Gretton et al., 2009). When source and target distributions have shared support, covariate shift only requires adaptation under model misspecification (Shimodaira, 2000), where the most common approach is to re-weight examples according to $p^t(x)/p^s(x)$, rendering the (re-weighted) training and target data exchangeable. However, even given missingness indicators, DAMS may still require some care. For example, in the augmented covariate space (with missing data indicators), one might need more complex models than in the original covariate space. When re-weighting is necessary, the structure of the DAMS problem might be leveraged to estimate importance

weights more efficiently, or to identify the optimal target predictor in certain cases where missingness introduces non-overlapping support. However, because our work is primarily motivated by underreporting in the medical setting, we focus our attention on the case where missingness indicators are absent.

UCAR as Regularization While the optimal predictor does not change across domains when ξ are observed (as the covariate shift assumption holds), it is less obvious *how missingness without indicators impacts the optimal predictor.* To build intuition on the effect of underreporting completely at random, we note that applying mask ξ, which zeros out covariates with some probability, resembles the mechanism of dropout in neural networks. Using similar theoretical arguments as in how dropout acts as a form of regularization (Wager et al., 2013), we show that for linear models, the phenomenon of UCAR in data with constant missingness rate m translates into a form of regularization on the resulting predictor (proof in Appendix C). First, we show that for generalized linear models, UCAR results in a regularization effect. Here, generalized linear models are defined as $p_\beta(y|x) = h(y)\exp\{yx \cdot \beta - A(x \cdot \beta)\}$, where $h(y)$ is a quantity independent of x and β, and $A(\cdot)$ is the log partition function, and the negative log likelihood objective is $l_{x,y}(\beta) = -\log p_\beta(y|x)$. Then, considering linear regression, we show that the regularization penalty can be viewed as a form of L2 regularization.

Theorem 4.1. *Under UCAR with missingness rates $m \in [0,1)^d$, the minimizer $\widehat{\beta}$ of the negative log likelihood of the corrupted training data $\widetilde{X}$ scaled by $\frac{1}{1-m}$ is given by:*

$$\begin{aligned}\widehat{\beta} &= \arg\min_{\beta\in\mathbb{R}^d} \sum_{i=1}^{n} \mathbb{E}_\xi[l_{\widetilde{x}^{(i)},y^{(i)}}(\beta)] \\ &= \arg\min_{\beta\in\mathbb{R}^d} \sum_{i=1}^{n} l_{x^{(i)},y^{(i)}}(\beta) + R(\beta),\end{aligned}$$

where $l_{\widetilde{x}^{(i)},y^{(i)}}(\beta)$ and $l_{x^{(i)},y^{(i)}}(\beta)$ are the negative log likelihoods of a corrupted sample and the corresponding clean sample (respectively). For linear regression, the regularization term $R(\beta)$ is given by:

$$R(\beta) = \frac{1}{2}\left(\beta\widetilde{\Delta}_{diag}\right)^\top\left(\beta\widetilde{\Delta}_{diag}\right),$$

where we define $\widetilde{\Delta}_{diag} = diag\left(\sqrt{\frac{m}{1-m}}\right) diag(I)^{1/2}$, where $diag\left(\sqrt{\frac{m}{1-m}}\right)$ refers to a diagonal matrix with $\sqrt{\frac{m_j}{1-m_j}}$ on the diagonal, and $diag(I)^{1/2}$ refers to the square root of the diagonal of the Fisher information matrix.

Thus, for linear regression, applying missingness rates m to data scaled by $\frac{1}{1-m}$ can be viewed as a form of L2 regularization of β scaled by $\widetilde{\Delta}_{\text{diag}}$.

5 Identification Results

This section shows that in DAMS with UCAR, the clean joint distribution p is identifiable from the corrupted joint distribution $\widetilde{p}$ with missingness rates $m \in [0,1)^d$ when m is known (Lemma 5.1). However, m is not in general identifiable directly from the observed corrupted data (Remark 4). Instead, we identify *relative* rates of non-missingness from the corrupted data across domains (Remark 5), which can in turn be used to identify the labeled target distribution $\widetilde{p}^t$ from the labeled source distribution $\widetilde{p}^s$ (Theorem 5.2).

First, we define some notation useful for our identification results. Consider vectors $a \in \mathbb{R}^d$ and $b \in \mathbb{R}^d$. Let $a \prec b$ denote that $\forall j \in \{1,2,...,d\}$, we have $a_j < b_j$. Similarly, let $a \succeq b$ denote that $\forall j \in \{1,2,...,d\}$, $a_j \geq b_j$.

To help clarify the relationship between corrupted and clean distributions, we define the notion of m-reachability.

Definition 2 (m-reachable)**.** *We say b is m-reachable from a (denoted $a \rightsquigarrow b$) if $\exists \xi \in \{0,1\}^d$ such that $b = a \odot \xi$.*

Remark 3 (Characteristics of m-reachability)**.** *If $a \rightsquigarrow b$, then the dimensions of a that are 0 must be a subset of the ones that are 0 in b. Additionally, any dimensions that are nonzero in both a and b must match in value.*

For example, if we observe a data point $b = [1,1,1]$, the only data point a for which $a \rightsquigarrow b$ is $a = [1,1,1]$. If $b = [1,1,0]$, then possible values of a are $a = [1,1,c]$ for any value of $c \in \mathbb{R}$. In binary data, $a \rightsquigarrow b \iff a \succeq b$.

Let $p_{x,y} = P(X = x, Y = y)$ denote the probability of some set of covariates $x \in \mathbb{R}^d$ and label $y \in \mathbb{R}$ in the clean distribution, and let $\widetilde{p}_{x,y} = P(\widetilde{X} = x, Y = y)$ denote the same in the corrupted distribution. Throughout the paper we use notation for discrete X, but note that it is straightforward to extend the results to continuous X (e.g. by replacing summations with integrals, etc.). Summing over all possible values of $z \in \mathbb{R}^d$ from which x is m-reachable, $\widetilde{p}$ can be expressed in terms of p and m:

$$\widetilde{p}_{x,y} = \sum_{z:z \rightsquigarrow x} p_{z,y} \cdot \prod_{j=1}^{d} (1 - m_j)^{[x_j]_{\neq 0}} m_j^{[z_j]_{\neq 0} - [x_j]_{\neq 0}} \tag{1}$$

where $[x]_{\neq 0} \triangleq \mathbb{1}[x \neq 0]$ is an indicator function for nonzero values. While it is obvious that one can obtain $\widetilde{p}$ from p, we show, surprisingly, that the above system is in fact invertible.

Lemma 5.1. *Given m, where $m \prec 1$, the clean distribution p is identifiable from the corrupted distribution $\widetilde{p}$.*

Roughly, the proof of Lemma 5.1 (in Appendix D) rearranges equation (1) and uses Remark 3 to observe that any entry $p_{(x,y)}$ can be expressed in terms of $\widetilde{p}$, m, and entries of p with fewer zeros. Using proof by induction on the number of zeros (0 to d), one can show that p is identifiable from $\widetilde{p}$.

Returning to the DAMS problem, *given m^s and m^t*, one could in theory identify p from $\widetilde{p}^s$ thru Lemma 5.1, and then use equation (1) to derive $\widetilde{p}^t$. Unfortunately, however, *missingness rates are not in general identifiable* from the observed corrupted data.

Remark 4. *Missingness rates are not in general identifiable directly from corrupted data. To see this, consider the following simple counterexample. Consider two distinct possible source distributions $A \sim$ Bernoulli(0.5) and $B \sim$ Bernoulli(0.25). Application of missingness with rates $m_A = 0.5$ to A and $m_B = 0$ to B yields identical corrupted distributions $\widetilde{A} \sim$ Bernoulli(0.25) and $\widetilde{B} \sim$ Bernoulli(0.25). Thus, the rates are not identifiable.*

While missingness rates are not in general identified given corrupted data from a single domain, one might hope to nevertheless *relate* the missingness rates between source and target domains. For this, we leverage nonzero values. Whereas observed zeros are a mixture of zeroed-out values and true zeros, all observed nonzeros were nonzero in the clean data. Thus, the relative proportions of nonzeros are informative of relative *non-missingness rates* $1 - m$. For a covariate X_j, where $j \in \{1, ..., d\}$, denote the true proportion of nonzeros in the underlying data as $q_j = P(X_j \neq 0)$. Then, the proportion of observed nonzeros in the corrupted data is $P(\widetilde{X}_j \neq 0) = (1 - m_j)q_j$. Vectorized, $P(\widetilde{X} \neq 0) = (1 - m) \odot q$.

Remark 5. *The ratio between non-missingness rates $1 - m^t$ and $1 - m^s$ is given by:*

$$\frac{1 - m^t}{1 - m^s} = \frac{(1 - m^t) \odot q}{(1 - m^s) \odot q} = \frac{P^t(\widetilde{X} \neq 0)}{P^s(\widetilde{X} \neq 0)} \triangleq 1 - r^{s \to t}, \tag{2}$$

where the divisions are element-wise. Note that the second-to-last expression is estimable from observed data.

We refer to $r^{s \to t} = 1 - \frac{1-m^t}{1-m^s} = \frac{m^t - m^s}{1-m^s}$ as the *relative missingness rates* between s and t. Interestingly, while identification of the *clean* distribution from a corrupted distribution (Lemma 5.1) may be difficult due to unidentifiability of m^s and m^t in general (Remark 4), we leverage identifiability of $r^{s \to t}$ to show that *adapting* from one corrupted distribution to another corrupted distribution does not require identification of the clean distribution.

Theorem 5.2. *For source and target distributions $\widetilde{p}^s$ and $\widetilde{p}^t$ with unknown missingness rates m^s and m^t (respectively), where $m^s \prec 1$, $\widetilde{p}^t$ is identifiable from $\widetilde{p}^s$ given $r^{s \to t}$:*

$$\widetilde{p}^t_{x,y} = \sum_{z:z \rightsquigarrow x} \widetilde{p}^s_{z,y} \cdot \prod_{j=1}^{d} (1 - r_j^{s \to t})^{[x_j]_{\neq 0}} (r_j^{s \to t})^{[z_j]_{\neq 0} - [x_j]_{\neq 0}}. \tag{3}$$

That is, while the precise missingness rates m^s and m^t may be unidentifiable in general from corrupted data, one can identify relative missingness rates $r^{s \to t}$ (Remark 5) and use them to directly identify $\widetilde{p}^t$ from $\widetilde{p}^s$ (proof in Appendix E), rather than explicitly using the clean distribution as an intermediate step. Note that the form of (3) matches that of (1), with missingness rates set to $m = r^{s \to t}$.

6 Estimation Results

We discuss estimation of optimal target predictors from labeled source data $\{(\widetilde{X}^{s,i}, Y^{s,i})\}_{i=1}^{n_s}$, drawn from $P^s(\widetilde{X}, Y)$ and unlabeled target data $\{\widetilde{X}^{t,i}\}_{i=1}^{n_t}$, drawn from $P^t(\widetilde{X})$.

Non-parametric adjustment procedure for nonnegative relative missingness The parallels between equations (3) and (1) suggest an intuitive non-parametric procedure when $m^s \preceq m^t$, so that $r^{s\to t} \succeq 0$ (Algorithm 1). To obtain data distributed identically to $\widetilde{X}^t$, one can sample masks $\xi^{s\to t}$ with missingness rates $r^{s\to t}$ and apply them to $\widetilde{X}^s$. Let us define a *missingness filter* applied to each datapoint $x \in \mathbb{R}^d$ as $\nu_{s\to t}(x) = x \odot \xi^{s\to t}$, where $\xi^{s\to t} \sim \text{Bernoulli}(1 - r^{s\to t})$. When a missingness filter is applied to a dataset, $\xi^{s\to t}$ is independently drawn for every data point. A proof showing that labeled data $\{(\nu_{s\to t}(\widetilde{X}^{s,i}), Y^{s,i})\}_{i=1}^{n_s}$ are drawn i.i.d. to $P^t(\widetilde{X}, Y)$ is in Appendix G. For any desired model class, we can now train a predictor on this labeled data. When $m^s \preceq m^t$, we call this adjustment a *proper* adjustment as it yields a predictor trained on data i.i.d. to labeled target data.

When $m^s \not\preceq m^t$, i.e. $r^{s\to t} \not\succeq 0$, it is less obvious what the proper non-parametric adjustment procedure implied by Theorem 5.2 might be. As a stopgap measure, we experiment with using a missingness filter of rate $\max\{r^{s\to t}, 0\}$ (Algorithm 1), but call this an improper adjustment as it does not create data i.i.d to the target distribution.

Algorithm 1 Non-parametric adjustment procedure
(proper adjustment when $m^s \preceq m^t$)

1: Compute $\widehat{q}_j^t = \frac{\text{count}(\widetilde{x}_j^t \neq 0)}{n_t}$, $\widehat{q}_j^s = \frac{\text{count}(\widetilde{x}_j^s \neq 0)}{n_s}$, and $\widehat{r}^{s\to t} = 1 - \frac{\widehat{q}^t}{\widehat{q}^s}$.
2: Compute $\widetilde{r}^{s\to t} = \max\{\widehat{r}^{s\to t}, 0\}$ (element-wise max). Note that if $\widehat{r}^{s\to t} \succeq 0$, then $\widehat{r}^{s\to t} = \widetilde{r}^{s\to t}$.
3: Apply a missingness filter with rate $\widetilde{r}^{s\to t}$ to source data to get $\{(\widetilde{\nu}_{s\to t}(\widetilde{X}^{s,i}), Y^{s,i})\}_{i=1}^{n_s}$.
4: Fit a predictor on data $\{(\widetilde{\nu}_{s\to t}(\widetilde{X}^{s,i}), Y^{s,i})\}_{i=1}^{n_s}$.

Step 1 of Algorithm 1 estimates the relative missingness $r^{s\to t}$ from data. Using Hoeffding's inequality, we show that with high probability, the estimated $\widehat{r}^{s\to t}$ is close to $r^{s\to t}$ (proof in Appendix F).

Theorem 6.1. *With probability at least* $1 - \delta$,

$$\left|\widehat{r}^{s\to t} - r^{s\to t}\right| \leq \frac{1}{\widehat{q}^s}\left(\sqrt{\frac{\log(4/\delta)}{2n_t}} + (1 - r^{s\to t})\sqrt{\frac{\log(4/\delta)}{2n_s}}\right).$$

A proper non-parametric adjustment requires $r^{s\to t} \succeq 0$. Next, we derive a closed-form expression for the optimal linear target predictor for any given relative missingness.

Closed-Form Solution for Optimal Linear Predictor Define the optimal predictor as the one that minimizes mean squared error. Given observations of source covariates $\widetilde{X}^s$ and their corresponding labels Y^s, as well unlabeled target covariates $\widetilde{X}^t$, we seek the optimal linear predictor $f_*^t(x^t) = \beta_*^t x^t$ for the target domain. Indeed, β_*^t can be expressed in terms of quantities estimable from data (proof in Appendix H.1).

Proposition 2. *The optimal linear target predictor is given by:*

$$\beta_*^t = \mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t]^{-1}\left(r^{s\to t} \odot \mathbb{E}[\widetilde{X}^{s\top}Y^s]\right). \tag{4}$$

Thus, *without knowing the levels of missingness*, as long as $m^s \prec 1$, the optimal linear predictor for the target domain is nevertheless estimable, using target unlabeled data to derive the covariance $\mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t]$. As we show in Appendix H, it is also possible to compute the entries of $\mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t]$ using only source data and relative missingness.

Proposition 3. *For $i \neq j$, where $i \in \{1, 2, .., d\}$, $j \in \{1, 2, .., d\}$, we have*

$$\mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t]_{ij} = (1 - r_i^{s\to t})(1 - r_j^{s\to t})\mathbb{E}[\widetilde{X}^{s\top}\widetilde{X}^s]_{ij} \tag{5}$$

$$\mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t]_{ii} = (1 - r_i^{s\to t})\mathbb{E}[\widetilde{X}^{s\top}\widetilde{X}^s]_{ii}. \tag{6}$$

Although $\mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t]$ could be estimated from either source or target covariates, in practice with finite samples it might be beneficial to utilize both. For example, to adjust for sample size of the source and target datasets, one could take a weighted average of the estimates of $\mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t]$, where the weights of the source-derived and target-derived estimates are $\alpha_s = \frac{n_s}{n_s+n_t}$ and $\alpha_t = \frac{n_t}{n_s+n_t}$, respectively. This attempts to adjust for the variance of estimation error due to the different sample sizes, however it does not account for estimation error in the relative missingness rate. We leave further exploration of these weightings to future work. Algorithm 2 describes the estimation procedure for linear models adjusted for the target domain.

Algorithm 2 Adjusted linear model learning procedure

1: Compute $\widehat{q}_j^t = \frac{\text{count}(\widetilde{x}_j^t \neq 0)}{n_t}$, $\widehat{q}_j^s = \frac{\text{count}(\widetilde{x}_j^s \neq 0)}{n_s}$, and $\widehat{r}^{s\to t} = 1 - \frac{\widehat{q}^t}{\widehat{q}^s}$ for all $j \in \{1, 2, .., d\}$.

2: Estimate target-based $\widehat{M}^t = \widehat{\mathbb{E}}[\widetilde{X}^{t\top}\widetilde{X}^t]$ from unlabeled target samples.

3: Estimate source-based $\widehat{M}^s = \widehat{\mathbb{E}}[\widetilde{X}^{t\top}\widetilde{X}^t]$ by computing for all $i \neq j$, where $i \in \{1, 2, .., d\}$, $j \in \{1, 2, .., d\}$:

$$\widehat{M}^s_{ij} = (1 - \widehat{r}_i^{s\to t})(1 - \widehat{r}_j^{s\to t})\widehat{\mathbb{E}}[\widetilde{X}^{s\top}\widetilde{X}^s]_{ij}$$
$$\widehat{M}^s_{ii} = (1 - \widehat{r}_i^{s\to t})\widehat{\mathbb{E}}[\widetilde{X}^{s\top}\widetilde{X}^s]_{ii}$$

4: Construct a combined weighted estimate of $\widehat{\mathbb{E}}[\widetilde{X}^{t\top}\widetilde{X}^t]$: $\widehat{M} = \alpha_s\widehat{M}^s + \alpha_t\widehat{M}^t$

5: Estimate $\widehat{\mathbb{E}}[\widetilde{X}^{s\top}Y^s]$ from source samples, and compute

$$\widehat{\beta}^t = \widehat{M}^{-1}\left(\widehat{r}^{s\to t} \odot \widehat{\mathbb{E}}[\widetilde{X}^{s\top}Y^s]\right).$$

7 Experiments

We apply Algorithms 1 and 2 to synthetic, semi-synthetic, and real data settings. We compare the performance of four variations of predictors: (1) the oracle predictor (Oracle), trained with target labeled data and tested on a held-out target test set; (2) the source predictor (Source), trained on source labeled data without any adjustments; (3) the closed-form adjustment (Closed-form Adj.) for linear predictors, given by Algorithm 2; and (4) the non-parametric adjustment (Non-param. Adj.), given by Algorithm 1. We also do MissForest imputation of both source and target data, treating all zeros as missing values, and train a source predictor to evaluate on target (Imputed).

In synthetic and semi-synthetic experiments, the data is split 4:1:4:1 to create source training, source test, target training, and target test sets. Different levels of missingness are applied completely at random to source and target datasets. Code is provided at https://github.com/acmi-lab/Missingness-Shift.

Synthetic data experiments We draw 10,000 samples from two simple data-generating processes:

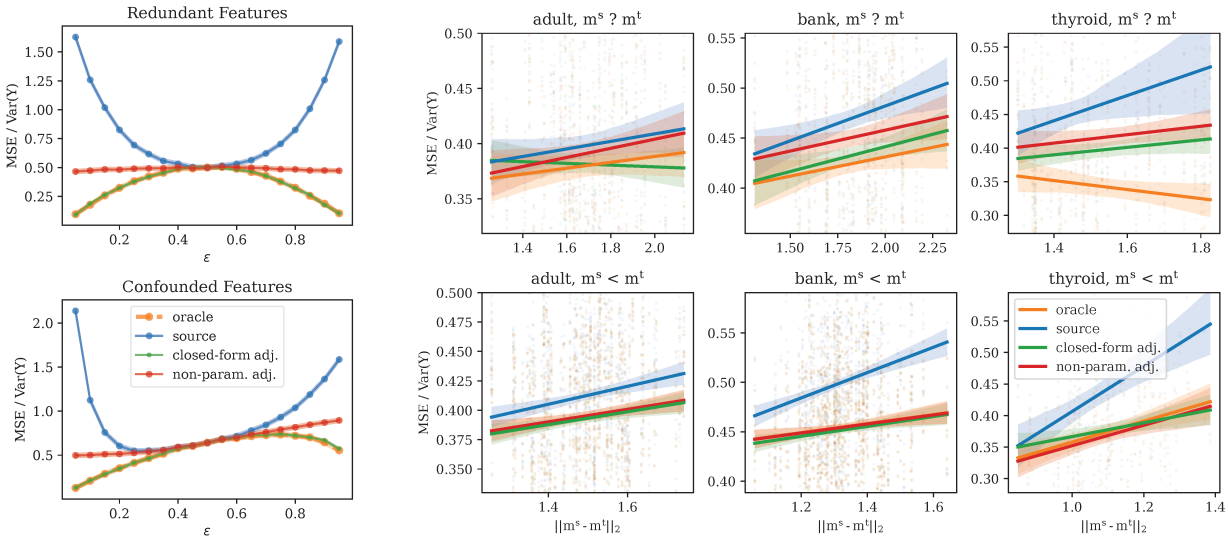
Scenario 1: "Redundant Features"

$$\begin{aligned} u_y &\sim \mathcal{N}(0,1) \\ Z &\sim \text{Bernoulli}(0.5) \\ X_1 &= Z \\ X_2 &= Z \\ Y &= Z + u_y \end{aligned}$$

Scenario 2: "Confounded Features"

$$\begin{aligned} u_{x_2} &\sim \mathcal{N}(0,1) \\ u_y &\sim \mathcal{N}(0,1) \\ X_1 &\sim \text{Bernoulli}(0.5) \\ X_2 &= \text{expit}(2X_1 + u_{x_2}) \\ Y &= X_1 - X_2 + u_y \end{aligned}$$

In both, we apply missingness with rates $m^s = [1-\epsilon, \epsilon]$ and $m^t = [\epsilon, 1-\epsilon]$ for varying ϵ between 0.05 and 0.95 in increments of 0.05, with 20 runs for each ϵ, and evaluate the performance of linear predictors (Figure 2a). At $\epsilon = 0.5$, the source and target domains are identically distributed, so Oracle, Source, Closed-form Adj., and Non-param. Adj. all attain the same mean squared error scaled by variance of the label (MSE/Var(Y)). As ϵ approaches 0 or 1, however, the error in the Source predictor grows rapidly whereas the Oracle and Closed-form Adj. errors decrease. Since $m^s \not\approx m^t$, as expected, Non-param. Adj. cannot fully match the target distribution, and has intermediate performance.



(a) Target domain error of linear models vs. ϵ. Oracle & closed-form overlap.

(b) Target domain error of linear models as the L2-norm between m^s and m^t varies. Best-fit line with 95% confidence intervals from bootstrapping.

Figure 2: MSE/Var(Y) of linear models on (a) synthetic and (b) semisynthetic data across varying m^s and m^t.

For $\epsilon = 0.1$, we compare linear regression, XGBoost, and MLP (Table 1). In both Scenario 1 and 2, the linear closed-form adjustment dramatically outperforms the source linear predictor. However, in Scenario 1, source XGBoost and MLP almost match the performance of their respective oracles, and source XGBoost outperforms the linear oracle. On the other hand, in Scenario 2, the linear closed-form adjustment outperforms source XGBoost and MLP.

Semi-synthetic data experiments Using the adult ($n = 48842$), bank ($n = 48188$), and thyroid binding protein ($n = 2800$) UCI datasets (Dua and Graff, 2017), which contain a mixture of categorical and numerical variables, we construct semi-synthetic datasets by borrowing the covariates, but replacing the labels with synthetically generated labels that are linear functions of the clean covariates. That is, we train using new labels $y_{new} = \beta X$, for randomly sampled $\beta_j \sim \text{Uniform}(0, 10), \forall j \in \{1, 2, ..., d\}$, and original covariates X. Source and target missingness rates are sampled under two regimes: (1) To test the proper non-parametric adjustment, where $m^s \preceq m^t$, we sample $m^s_j \sim \text{Uniform}(0, 0.5)$ and $m^t_j \sim m^s_j + (1 - m^s_j)\epsilon$, where $\epsilon \sim \text{Uniform}(0, 0.5)$. (2) To simulate a more general form of missingness shift, we sample $m^s_j, m^t_j \sim \text{Uniform}(0, 0.9)$, abbreviated as m^s ? m^t. For additional experiment and data preprocessing details, see Appendix I.

Overall, where adjusted models are applicable/proper, they perform at least as well as (and often better than) source unadjusted models when compared within each model class (Table 1). Among linear models, the closed-form and non-parametric adjustments consistently outperform the source predictors. In nonlinear models, only the non-parametric adjustment applies, and this adjustment is only proper if $m^s \preceq m^t$. Among nonlinear models, if $m^s \preceq m^t$, either Non-param. and Source tie, or Non-param. performs best. When m^s ? m^t, Non-param. (improper adjustment) often has the second-best or best performance (especially when no other adjustments apply). Ignoring model class, the best-performing model for each semi-synthetic dataset is an adjusted model. Plotting the line of best fit for MSE/Var(Y) of the linear models versus the L2 distance between m^s and m^t, we note that the Source predictor tends to have the stronger positive slope than the Oracle, Closed-form Adj., or Non-parametric Adj. models (Figure 2b).

Real data experiments To explore the applicability of our methods to naturally-occurring missingness shifts, we use the FIDDLE data pre-processing pipeline (Tang et al., 2020) on the eICU Collaborative Research Database (Pollard et al., 2018), which contains data from critical care units across several hospitals. FIDDLE extracts binary feature vectors capturing several patient characteristics, including demographics, physiological measurements, labs, medications, etc. We extract the binary 48-hour mortality outcome for patients in two of the hospitals with the most data ($n_1 = 3006$, $n_2 = 2663$), and verify that the prevalences of the covariates are different across these two hospitals. Additional data and experiment details are provided in Appendix I.

Table 1: Target domain MSE/Var(Y), averaged across various missingness levels on synthetic and semi-synthetic data. Confidence intervals are provided in Appendix I. The first two columns are synthetic datasets (Redundant Features and Confounded Features), and the last three columns are semi-synthetic UCI datasets.

	Rednd.	**Confnd.**	**Adult**		**Bank**		**Thyroid**	
	$m^s \; ? \; m^t$	$m^s \; ? \; m^t$	$m^s \preceq m^t$	$m^s \; ? \; m^t$	$m^s \preceq m^t$	$m^s \; ? \; m^t$	$m^s \preceq m^t$	$m^s \; ? \; m^t$
Linear Regression Models								
Oracle	0.178	0.206	0.420	0.362	0.338	0.433	0.298	0.251
Source	1.259	1.103	0.437	0.380	0.371	0.480	0.350	0.320
Imputed	1.002	0.918	0.490	0.483	0.501	0.592	0.306	0.358
Closed-form	**0.186**	**0.209**	0.422	**0.363**	0.339	**0.442**	0.316	**0.291**
Non-param.	0.473	0.492	**0.420**	0.373	**0.338**	0.459	**0.293**	**0.291**
XGBoost Models								
Oracle	0.166	0.200	0.398	0.354	0.287	0.453	0.316	0.274
Source	**0.166**	0.475	**0.399**	**0.379**	0.305	**0.500**	**0.310**	**0.352**
Imputed	1.002	1.157	0.512	0.521	0.492	0.708	0.355	0.441
Non-param.	0.425	**0.473**	**0.399**	0.392	**0.287**	0.503	**0.310**	0.381
MLP Models								
Oracle	0.166	0.201	0.389	0.343	0.295	0.458	0.279	0.230
Source	**0.184**	**0.321**	0.399	0.357	0.322	0.499	0.320	0.303
Imputed	1.003	0.924	0.480	0.468	0.484	0.668	0.304	0.345
Non-param.	0.436	0.470	**0.389**	**0.355**	**0.294**	**0.487**	**0.278**	**0.272**

Table 2: Target domain performance of linear models on eICU 48-hour mortality prediction, where source s and target t can be Hospital 1 (H1) or Hospital 2 (H2). Here, underreporting occurs naturally in the data. Since all features are binary, imputation of all zeros behaves poorly, leading to baseline performance. AUPRC refers to average precision.

Model Class	s	t	MSE	AUROC	AUPRC
Oracle	H1	H1	0.103 (0.088 – 0.117)	0.713 (0.652 – 0.775)	0.236 (0.156 – 0.317)
Source	H2	H1	0.143 (0.135 – 0.151)	**0.593 (0.563 – 0.623)**	**0.146 (0.122 – 0.170)**
Imputed	H2	H1	**0.089 (0.081 – 0.097)**	0.500 (0.500 – 0.500)	0.097 (0.088 – 0.106))
Closed-form Adj.	H2	H1	0.439 (0.223 – 0.655)	0.540 (0.509 – 0.571)	0.123 (0.103 – 0.143)
Non-param. Adj.	H2	H1	0.142 (0.133 – 0.150)	0.555 (0.537 – 0.573)	0.126 (0.108 – 0.144)
Oracle	H2	H2	0.121 (0.100 – 0.142)	0.601 (0.528 – 0.675)	0.167 (0.103 – 0.230)
Source	H1	H2	0.122 (0.113 – 0.131)	**0.576 (0.545 – 0.608)**	**0.144 (0.120 – 0.169)**
Imputed	H1	H2	**0.090 (0.082 – 0.098)**	0.500 (0.500 – 0.500)	0.099 (0.089 – 0.109)
Closed-form Adj.	H1	H2	0.373 (0.327 – 0.420)	0.556 (0.523 – 0.588)	0.122 (0.104 – 0.141)
Non-param. Adj.	H1	H2	0.196 (0.182 – 0.210)	0.511 (0.503 – 0.520)	0.109 (0.095 – 0.123)

We train linear models to predict mortality, and evaluate MSE, AUROC, and AUPRC. Since the preprocessed data only contains binary features, MissForest imputation of all zeros results in a dataset consisting entirely of ones, and the linear model learns to simply predict the label mean and only achieves baseline performance. Estimated relative missingness indicates that $m^s \not\preceq m^t$ (Appendix I), so the non-parametric estimation procedure is not expected to produce labeled data i.i.d. to the target distribution. The source predictor achieves highest AUROC and AUPRC.

Note, however, that beyond missingness levels, there are also several other aspects of the data distribution that likely differ between these two hospitals. Different hospitals likely have different underlying $P(X, Y)$, and in practice, missingness could be dependent on other covariates (e.g. a doctor may choose not to perform a test based on patient state). Thus, fundamental assumptions of our adaptation methods are likely violated in this dataset.

8 Discussion

This work introduces the domain adaptation under missingness shift (DAMS) problem, and explores DAMS under the underreporting completely at random (UCAR) assumption. Our synthetic and semi-synthetic experiments demonstrate that when assumptions hold, the proposed methods (when applicable/proper), tend to outperform or perform at least as well as unadjusted source predictors in the same model class (Table 1). In linear models, our proposed adjustments (linear closed-form and non-param. adj.) consistently outperform the source predictors, and sometimes, the benefits of adaptation can even outweigh the bias incurred by restricting to linear models. For example, in the Confounded Features, Bank $m^s \ ? \ m^t$, and Thyroid datasets, linear adjusted models outperform all Source models, regardless of model class. Note that even if the underlying relationship between clean unobserved covariates X and label Y is linear, after X is corrupted by missingness to create observed corrupted covariates $\widetilde{X}$, the new relationship between $\widetilde{X}$ and Y is often nonlinear (a phenomenon which has also been noted by Le Morvan et al. (2020b)). Correspondingly, the best MLP and XGBoost models tend to outperform the best linear models (Table 1).

The best-performing model(s) in each of the synthetic and semi-synthetic datasets, except for the synthetic Redundant Features dataset, use a proposed adjustment (Table 1). Although the adjustments perform best in the synthetic Redundant Features dataset when restricted to the linear model class, the best-performing model in this dataset overall is a source XGBoost model, which matches the performance of the oracle. In addition to the flexibility of the XGBoost model, which improves the oracle XGBoost over the oracle linear model, a likely reason for improvement of Source XGBoost over Non-param. Adj. can be found in the particular setup of this scenario. Here, $X_1 = X_2 = Z$, and $Y = Z + u_y$, where $u_y \sim \mathcal{N}(0, 1)$, and so given knowledge of either X_1 or X_2, prediction of Y is straightforward. The only applicable adjustment, Non-param. Adj. (improper, since $m^s \not\preceq m^t$), would zero out much of the data to bring the missingness rate in X_1 from 0.9 to 0.1, thus making prediction harder. There are also multiple settings in which Source XGBoost performs similarly to Non-param. Adj. XGBoost (Confounded Features, Adult $m^s \preceq m^t$, Bank $m^s \ ? \ m^t$, and Thyroid $m^s \ ? \ m^t$). On the other hand, for the MLP model class, the non-parametric adjustment outperforms all source predictors in the semi-synthetic datasets. Thus, depending on the model class, non-parametric adjustment may not always have a consistent effect on performance.

The generally worse performance of imputation in synthetic and semi-synthetic experiments (Table 1) helps highlight the difficulty of not having missing data indicators. Learning without missing data indicators is fundamentally more difficult than learning with them, and methods which might make sense when missing data indicators are present (e.g. imputation) can be ill-defined when the indicators are absent. In the eICU dataset, for example, all covariates were binary, and so imputing all 0's only left 1's to train on. As a result, MissForest learned to predict 1 for everything, rendering these binary features useless. Nevertheless, we included a comparison with imputation of all zeros in the other datasets, as it could still be useful for continuous variables.

The experiments with real eICU data also help demonstrate that it is important to clarify assumptions on whether one is truly in a DAMS with UCAR setting, as failure to do so could result in predictors that perform worse than if no adaptation had been done in the first place (Table 2). Ideally, in real-world data, DAMS with UCAR might be useful around a sudden change in clerical practices where the underlying $P(X, Y)$ is similar before and after the change, and underreporting is completely at random (e.g. determined based a blanket policy independent of covariates). In the absence of such data, however, we instead included synthetic and semisynthetic data where the missingness shift with UCAR assumptions hold, and also included a real critical care (eICU) dataset containing multiple hospitals for thoroughness. While our proposed techniques for DAMS with UCAR do not work particularly well on real eICU data, we also note that we have no particular reason to believe that missingness shift is especially prominent between the hospitals compared to factors such as selection bias (very different cohort), label shift, or changes in prevalences of disease, among others. Finding appropriate real world empirical testbeds and analyzing sensitivity to assumption violations are important directions for future work.

Beyond the UCAR setting, there are several open avenues for further research in domain adaptation under missingness shift. Allowing underreporting to depend on other covariates would significantly broaden the set of applicable real-world cases, as doctors often take certain measurements as needed in their diagnostic process. Moreover, future works could explore other variations of graphical model structures (Figure 1) for expressing models of missingness shift.

Bibliography

Adams, R., Ji, Y., Wang, X., and Saria, S. (2019). Learning models from data with measurement error: Tackling underreporting. In *International Conference on Machine Learning (ICML)*, pages 61–70. PMLR.

Bekker, J. and Davis, J. (2020). Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760.

Brenner, H. and Loomis, D. (1994). Varied forms of bias due to nondifferential error in measuring exposure. *Epidemiology*, pages 510–517.

CDC (2022). Covid-19 vaccinations in the united states,jurisdiction.

Chu, H., Wang, Z., Cole, S. R., and Greenland, S. (2006). Sensitivity analysis of misclassification: a graphical and a bayesian approach. *Annals of Epidemiology*, 16(11):834–841.

Dosemeci, M., Wacholder, S., and Lubin, J. H. (1990). Does nondifferential misclassification of exposure always bias a true effect toward the null value? *American Journal of Epidemiology*, 132(4):746–748.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Elkan, C. and Noto, K. (2008). Learning classifiers from only positive and unlabeled data. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 213–220.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37.

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. (2020). A unified view of label shift estimation. *Advances in Neural Information Processing Systems (NeurIPS)*.

Garg, S., Wu, Y., Smola, A. J., Balakrishnan, S., and Lipton, Z. (2021). Mixture proportion estimation and pu learning: A modern approach. *Advances in Neural Information Processing Systems (NeurIPS)*, 34.

Gelman, A., Hill, J., and Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2009). Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning*, 3(4):5.

Groenwold, R. H., White, I. R., Donders, A. R. T., Carpenter, J. R., Altman, D. G., and Moons, K. G. (2012). Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *Canadian Medical Association Journal*, 184(11):1265–1269.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. (2006). Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems (NeurIPS)*, 19.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J. I., Raj, A., Li, M., and Zhang, N. R. (2018). Saver: gene expression recovery for single-cell rna sequencing. *Nature methods*, 15(7):539–542.

Le Morvan, M., Josse, J., Moreau, T., Scornet, E., and Varoquaux, G. (2020a). Neumiss networks: differentiable programming for supervised learning with missing values. *Advances in Neural Information Processing Systems*, 33:5980–5990.

Le Morvan, M., Josse, J., Scornet, E., and Varoquaux, G. (2021). What's a good imputation to predict with missing values? In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11530–11540. Curran Associates, Inc.

Le Morvan, M., Prost, N., Josse, J., Scornet, E., and Varoquaux, G. (2020b). Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *International Conference on Artificial Intelligence and Statistics*, pages 3165–3174. PMLR.

Li, W. V. and Li, J. J. (2018). An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications*, 9(1):1–9.

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 (ICML)*. PMLR.

Lipton, Z. C., Kale, D. C., Wetzel, R., et al. (2016). Modeling missing data in clinical time series with rnns. *Machine Learning for Healthcare*, 56.

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

Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037.

Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. (2018). The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13.

Rahardja, D. and Young, D. M. (2021). Confidence intervals for the risk ratio using double sampling with misclassified binomial data. *Journal of Data Science*, 9(4):529–548.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell rna-seq data. *Nature Communications*, 9(1):1–17.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489.

Saerens, M., Latinne, P., and Decaestecker, C. (2002). Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. *Neural Computation*.

Sechidis, K., Sperrin, M., Petherick, E. S., Luján, M., and Brown, G. (2017). Dealing with under-reported variables: An information theoretic solution. *International Journal of Approximate Reasoning*.

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

Stekhoven, D. J. and Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Storkey, A. (2009). When training and test sets are different: characterizing learning transfer. *Dataset Shift in Machine Learning*, 30:3–28.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P., and Kawanabe, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. *Advances in Neural Information Processing Systems (NeurIPS)*, 20.

Tang, S., Davarmanesh, P., Song, Y., Koutra, D., Sjoding, M. W., and Wiens, J. (2020). Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 27(12):1921–1934.

Tsymbal, A. (2004). The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 106(2):58.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45:1–67.

Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729.

Wager, S., Wang, S., and Liang, P. S. (2013). Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. In *International Conference on Machine Learning (ICML)*, page 114.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. (2013). Domain adaptation under target and conditional shift. In *International Conference on Machine Learning (ICML)*. PMLR.

A Motivating Examples

Example 1 (Redundant Features)

Let $m_s = [1-\epsilon, \epsilon]$ and $m_t = [\epsilon, 1-\epsilon]$. Consider the following data generating process:

$$\begin{aligned} Z &= u_Z & \\ X_1 &= Z & u_Z \sim \mathcal{N}(0, \sigma_Z^2) \\ X_2 &= Z & u_Y \sim \mathcal{N}(0, \sigma_Y^2) \\ Y &= Z + u_Y & \end{aligned}$$

where Z is a latent variable, X_1 and X_2 are observed covariates, and Y is the label we wish to predict.

We start by summarizing the findings, and then provide the full algebraic justification. The optimal (risk-minimizing) linear predictor on the source data is given by:

$$\beta_*^s = \left[\frac{\epsilon}{1-\epsilon+\epsilon^2}, \frac{1-\epsilon}{1-\epsilon+\epsilon^2}\right]$$

And for the target data:

$$\beta_*^t = \left[\frac{1-\epsilon}{1-\epsilon+\epsilon^2}, \frac{\epsilon}{1-\epsilon+\epsilon^2}\right]$$

The excess risk of the source predictor on the target data is given by:

$$\begin{aligned} r^t(\beta_*^s) - r^t(\beta_*^t) &= (\beta_*^s - \beta_*^t)^\top \mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t](\beta_*^s - \beta_*^t) \\ &= \sigma_Z^2 \cdot \frac{(1-2\epsilon)^2(1-2\epsilon+2\epsilon^2)}{(1-\epsilon+\epsilon^2)^2} \end{aligned}$$

As $\epsilon \to 0$, we have:

$$\begin{aligned} \beta_*^s &\to [0, 1] \\ \beta_*^t &\to [1, 0] \\ r^t(\beta_*^s) - r^t(\beta_*^t) &\to \sigma_Z^2 \\ r^t([0,0]) - r^t(\beta_*^t) &\to \sigma_Z^2 \end{aligned}$$

that is, the source classifier performs no better than simply predicting 0 (the mean of Y). Thus, $r^t(\beta_*^s) \to \sigma_Z^2 + \sigma_Y^2 = \mathrm{Var}(Y)$

Proof. In the example, we have:

$$\begin{aligned} \mathbb{E}[X^T X] &= \begin{bmatrix} \sigma_Z^2 & \sigma_Z^2 \\ \sigma_Z^2 & \sigma_Z^2 \end{bmatrix} \\ \mathbb{E}[X^T Y] &= \begin{bmatrix} \sigma_Z^2 \\ \sigma_Z^2 \end{bmatrix} \end{aligned}$$

We apply the expressions for $\mathbb{E}[\widetilde{X}^T\widetilde{X}]$ and $\mathbb{E}[\widetilde{X}^\top Y]$ derived in Appendix H:

$$\begin{aligned} \mathbb{E}[\widetilde{X}^\top\widetilde{X}] &= (1-m)(1-m)^\top \odot \mathbb{E}\left[X^\top X\right] + \mathrm{diag}\left(m(1-m^\top)\right)\mathrm{diag}\left(\mathbb{E}\left[X^\top X\right]\right) \\ &= \begin{bmatrix} 1-m_1 & (1-m_1)(1-m_2) \\ (1-m_1)(1-m_2) & 1-m_2 \end{bmatrix} \odot \mathbb{E}\left[X^\top X\right] \\ \mathbb{E}[\widetilde{X}^\top Y] &= (1-m) \odot \mathbb{E}[X^\top Y] \end{aligned}$$

to get:

$$\begin{aligned}
\mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t] &= \begin{bmatrix} 1-\epsilon & \epsilon(1-\epsilon) \\ \epsilon(1-\epsilon) & \epsilon \end{bmatrix} \cdot \sigma_Z^2 \\
\mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t]^{-1} &= \frac{1}{\sigma_Z^2\epsilon(1-\epsilon)(1-\epsilon+\epsilon^2)} \begin{bmatrix} \epsilon & -\epsilon(1-\epsilon) \\ -\epsilon(1-\epsilon) & 1-\epsilon \end{bmatrix} \\
\mathbb{E}[\widetilde{X}^{t\top}Y] &= \sigma_Z^2 \begin{bmatrix} 1-\epsilon \\ \epsilon \end{bmatrix} \\
\beta_*^t &= \mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t]^{-1}\mathbb{E}[\widetilde{X}^{t\top}Y] \\
&= \frac{1}{\epsilon(1-\epsilon)(1-\epsilon+\epsilon^2)} \begin{bmatrix} \epsilon(1-\epsilon)+-\epsilon^2(1-\epsilon) \\ -\epsilon(1-\epsilon)^2+\epsilon(1-\epsilon) \end{bmatrix} \\
&= \frac{1}{\epsilon(1-\epsilon)(1-\epsilon+\epsilon^2)} \begin{bmatrix} \epsilon(1-\epsilon)(1-\epsilon) \\ \epsilon(1-\epsilon)(-(1-\epsilon)+1) \end{bmatrix} \\
&= \frac{1}{1-\epsilon+\epsilon^2} \begin{bmatrix} 1-\epsilon \\ \epsilon \end{bmatrix}.
\end{aligned}$$

Similarly,

$$\beta_*^s = \frac{1}{1-\epsilon+\epsilon^2} \begin{bmatrix} \epsilon \\ 1-\epsilon \end{bmatrix},$$

so we can compute

$$\begin{aligned}
\beta_*^s - \beta_*^t &= \frac{1}{1-\epsilon+\epsilon^2} \begin{bmatrix} 2\epsilon-1 \\ -2\epsilon+1 \end{bmatrix} \\
&= \frac{1-2\epsilon}{1-\epsilon+\epsilon^2} \begin{bmatrix} -1 \\ 1 \end{bmatrix}
\end{aligned}$$

Now, excess risk is computed as follows:

$$\begin{aligned}
r^t(\beta_*^s) - r^t(\beta_*^t) &= (\beta_*^s - \beta_*^t)^\top \mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t](\beta_*^s - \beta_*^t) \\
&= \frac{(1-2\epsilon)^2}{(1-\epsilon+\epsilon^2)^2} \begin{bmatrix} -1 \\ 1 \end{bmatrix}^\top \begin{bmatrix} 1-\epsilon & \epsilon(1-\epsilon) \\ \epsilon(1-\epsilon) & \epsilon \end{bmatrix} \cdot \sigma_Z^2 \cdot \begin{bmatrix} -1 \\ 1 \end{bmatrix} \\
&= \frac{\sigma_Z^2(1-2\epsilon)^2(1-2\epsilon+2\epsilon^2)}{(1-\epsilon+\epsilon^2)^2}
\end{aligned}$$

As $\epsilon \to 0$, we can see that $r^t(\beta_*^s) - r^t(\beta_*^t) \to \sigma_Z^2$.

Additionally, we can compute the excess risk of the constant zero classifier:

$$\begin{aligned}
r^t([0,0]) - r^t(\beta_*^t) &= \beta_*^{t\top}\mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t]\beta_*^t \\
&= \frac{1}{(1-\epsilon+\epsilon^2)^2} \begin{bmatrix} 1-\epsilon \\ \epsilon \end{bmatrix}^\top \begin{bmatrix} 1-\epsilon & \epsilon(1-\epsilon) \\ \epsilon(1-\epsilon) & \epsilon \end{bmatrix} \cdot \sigma_Z^2 \cdot \begin{bmatrix} 1-\epsilon \\ \epsilon \end{bmatrix} \\
&= \frac{\sigma_Z^2}{(1-\epsilon+\epsilon^2)^2} \begin{bmatrix} (1-\epsilon)^2+\epsilon^2(1-\epsilon) \\ \epsilon(1-\epsilon)^2+\epsilon^2 \end{bmatrix}^\top \begin{bmatrix} 1-\epsilon \\ \epsilon \end{bmatrix} \\
&= \frac{\sigma_Z^2(1-\epsilon+\epsilon^2)}{(1-\epsilon+\epsilon^2)^2} \begin{bmatrix} (1-\epsilon) \\ \epsilon \end{bmatrix}^\top \begin{bmatrix} 1-\epsilon \\ \epsilon \end{bmatrix} \\
&= \frac{\sigma_Z^2(1-\epsilon+\epsilon^2)}{(1-\epsilon+\epsilon^2)^2} \left[(1-\epsilon)^2+\epsilon^2\right] \\
&= \frac{\sigma_Z^2(1-2\epsilon+2\epsilon^2)}{1-\epsilon+\epsilon^2}
\end{aligned}$$

As $\epsilon \to 0$, we can see that $r^t([0,0]) - r^t(\beta_*^t) \to \sigma_Z^2$. □

Example 2 (Confounded Features)

Now, suppose that $m_s = [0, 0]$ and $m_t = [1, 0]$. For some constants a, b, c consider the following data generating process:

$$\begin{aligned} X_1 &= \nu_1 & \nu_1 &\sim \mathcal{N}(0, 1) \\ X_2 &= aX_1 + \nu_2 & \nu_2 &\sim \mathcal{N}(0, 1) \\ Y &= bX_1 + cX_2 + \nu_Y & \nu_Y &\sim \mathcal{N}(0, 1). \end{aligned}$$

We will show that the optimal source and target predictors are $\beta^s_* = [b, c]$ and $\beta^t_* = [0, \frac{ab}{a^2+1} + c]$. By setting $a = -\frac{b}{c}$, we will show that for any $\tau > 1$, there exists values of a, b, c such that $r^t(\beta^s_*) > \tau \text{Var}(Y)$.

Proof. First, we compute β^s_* (where $m_s = [0, 0]$):

$$\begin{aligned} \mathbb{E}[\widetilde{X}^{s\top}\widetilde{X}^s] &= \mathbb{E}\left[X^\top X\right] \\ &= \begin{bmatrix} 1 & a \\ a & a^2+1 \end{bmatrix} \\ \mathbb{E}[\widetilde{X}^{s\top}\widetilde{X}^s]^{-1} &= \begin{bmatrix} a^2+1 & -a \\ -a & 1 \end{bmatrix} \\ \mathbb{E}[\widetilde{X}^{s\top}Y] &= \mathbb{E}[X^\top Y] \\ &= \begin{bmatrix} b+ac \\ ab+a^2c+c \end{bmatrix} \\ \beta^s_* &= \mathbb{E}[\widetilde{X}^{s\top}\widetilde{X}^s]^{-1}\mathbb{E}[\widetilde{X}^{s\top}Y] \\ &= \begin{bmatrix} a^2+1 & -a \\ -a & 1 \end{bmatrix} \cdot \begin{bmatrix} b+ac \\ ab+a^2c+c \end{bmatrix} \\ &= \begin{bmatrix} b \\ c \end{bmatrix}. \end{aligned}$$

Thus, $\beta^s_* = [b, c]$.

Now, let us compute β^t_* (where $m_t = [1, 0]$). Since X_1 is entirely missing and X_2 is completely observed, we only regress on X_2:

$$\begin{aligned} \mathbb{E}[\widetilde{X}_2^{t\top}\widetilde{X}_2^t] &= \mathbb{E}[X_2^\top X_2] \\ &= (a^2+1) \\ \mathbb{E}[\widetilde{X}_2^{t\top}\widetilde{X}_2^t]^{-1} &= \frac{1}{a^2+1} \\ \mathbb{E}[\widetilde{X}_2^{t\top}Y] &= ab + a^2c + c \\ \mathbb{E}[\widetilde{X}_2^{t\top}\widetilde{X}_2^t]^{-1}\mathbb{E}[\widetilde{X}_2^{t\top}Y] &= \frac{ab+a^2c+c}{a^2+1} \\ &= \frac{ab}{a^2+1} + c. \end{aligned}$$

Thus, $\beta^t_* = \left[0, \frac{ab}{a^2+1} + c\right]$.

Now, let us compute $\text{Var}(Y)$. Note that $\mathbb{E}[Y] = 0$, so $\text{Var}(Y) = \mathbb{E}[Y^2]$. Also, note that ν_1, ν_2, ν_Y are independent:

$$\begin{aligned} \text{Var}(Y) &= \text{Var}(bX_1 + cX_2 + \nu_Y) \\ &= \mathbb{E}[(b\nu_1 + c(a\nu_1 + \nu_2) + \nu_Y)^2] \\ &= (b+ac)^2 + c^2 + 1 \\ &= b^2 + 2abc + a^2c^2 + c^2 + 1. \end{aligned}$$

Thus, $\text{Var}(Y) = b^2 + 2abc + a^2c^2 + c^2 + 1$.

Now, let us compute $r^t(\beta_*^s)$. Let $[\beta_*^s]_2$ denote the second dimension of β_*^s. We have:

$$\begin{aligned}
r^t(\beta_*^s) &= \mathbb{E}[(Y - \widetilde{X}_2^t[\beta_*^s]_2)^2] \\
&= \mathbb{E}[Y^2] - 2\mathbb{E}[\widetilde{X}_2^t[\beta_*^s]_2 Y] + \mathbb{E}[(\widetilde{X}_2^t[\beta_*^s]_2)^2] \\
&= \text{Var}[Y^2] - 2\mathbb{E}[X_2[\beta_*^s]_2 Y] + \mathbb{E}[(X_2[\beta_*^s]_2)^2] \\
&= \text{Var}[Y^2] - 2\mathbb{E}[(a\nu_1 + \nu_2)c(b\nu_1 + c(a\nu_1 + \nu_2) + \nu_Y)] + \mathbb{E}[((a\nu_1 + \nu_2)c)^2] \\
&= b^2 + 2abc + a^2c^2 + c^2 + 1 - 2[ac(b + ac) + c^2] + [a^2c^2 + c^2] \\
&= b^2 + 1.
\end{aligned}$$

Thus, we have $\frac{r^t(\beta_*^s)}{\text{Var}(Y)} = \frac{b^2+1}{b^2+2abc+a^2c^2+c^2+1}$. If we set $a = -\frac{b}{c}$, then we have:

$$\begin{aligned}
\frac{r^t(\beta_*^s)}{\text{Var}(Y)} &= \frac{b^2 + 1}{b^2 + 2abc + a^2c^2 + c^2 + 1} \\
&= \frac{b^2 + 1}{b^2 - 2b^2 + b^2 + c^2 + 1} \\
&= \frac{b^2 + 1}{c^2 + 1}.
\end{aligned}$$

Now suppose that for some $\tau > 1$, we would like $r^t(\beta_*^s) > \tau\text{Var}(Y)$. Then, it is easy to see that we can simply choose b large enough, c small enough, and $a = -\frac{b}{c}$, such that $\frac{b^2+1}{c^2+1} > \tau$. □

B DAMS with Indicators as an Instance of Covariate Shift

This section contains a proof of Proposition 1: Assume we observe ξ. Let us consider an augmented set of covariates $\tilde{x}' = (\tilde{x}, \xi)$. When ξ is drawn independently of other covariates or depending only on other completely observed covariates, we will show that missingness shift satisfies the covariate shift assumption, i.e, $P^s(Y|\widetilde{X}' = \tilde{x}') = P^t(Y|\widetilde{X}' = \tilde{x}')$.

First, let us formalize what it means for ξ to be drawn independently of other covariates or depending only on other completely observed covariates:

(a) **Independent of other covariates** When ξ is drawn independently of other covariates, as described in the DAMS with UCAR setup (Section 3), we have that $\xi \sim \text{Bernoulli}(1-m)$ for some constant vector of missingness rates $m \in [0,1]^d$.

(b) **Depending only on other completely observed covariates** Now, suppose that some subset of covariates $X_c \subseteq X$ is completely observed (i.e. no missingness), and the missingness of the other covariates $X_m = X \setminus X_c$ depends on X_c. That is, $\xi \sim \text{Bernoulli}(f(X_c))$ for some function $f : \mathbb{R}^{|X_c|} \to [0,1]^{|X_m|}$.

Since (b) is more general than (a), we adopt notation from (b) throughout our proof, and then argue why it also holds for (a).

Proof. Consider some augmented set of covariates taking values $\tilde{x}' = (\tilde{x}_m, \xi, x_c)$. To prove that the covariate shift assumption holds, let us start by considering the left-hand side of the equation. Applying Bayes' Rule, we have:

$$P^s(Y|\widetilde{X}' = \widetilde{x}') = P^s(Y|\widetilde{X}^s_m = \widetilde{x}_m, \xi^s = \xi, X_c = x_c) = \frac{P^s(Y, \widetilde{X}^s_m = \widetilde{x}_m, \xi^s = \xi, X_c = x_c)}{\sum_y P^s(Y = y, \widetilde{X}^s_m = \widetilde{x}_m, \xi^s = \xi, X_c = x_c)}$$

We can rewrite the numerator as follows:

$$\begin{aligned}
P^s(Y, \widetilde{X}^s_m = \widetilde{x}_m, \xi^s = \xi, X_c = x_c) &= \sum_{x_m : x_m \odot \xi = \widetilde{x}_m} P(Y, \widetilde{X}^s_m = \widetilde{x}_m, \xi^s = \xi, X_m = x_m, X_c = x_c) \\
&= \sum_{x_m : x_m \odot \xi = \widetilde{x}_m} P(Y, \xi^s = \xi, X_m = x_m, X_c = x_c) \\
&= \sum_{x_m : x_m \odot \xi = \widetilde{x}_m} P(\xi^s = \xi | Y, X_m = x_m, X_c = x_c) \cdot P(Y, X_m = x_m, X_c = x_c) \\
&= \sum_{x_m : x_m \odot \xi = \widetilde{x}_m} P(\xi^s = \xi | X_c = x_c) \cdot P(Y, X_m = x_m, X_c = x_c) \\
&= P(\xi^s = \xi | X_c = x_c) \sum_{x_m : x_m \odot \xi = \widetilde{x}_m} P(Y, X_m = x_m, X_c = x_c),
\end{aligned}$$

where the first line follows from marginalizing over all possible values of X_m, the second line comes from the fact that $\widetilde{x}_m$ is determined given x_m and ξ, the third line comes from Bayes' Rule, the fourth line comes the fact that ξ only depends on X_c, and the last line comes from pulling the first term out of the summation.

Plugging back into the expression for $P^s(Y|\widetilde{X}' = \widetilde{x}')$, we have:

$$\begin{aligned}
P^s(Y|\widetilde{X}' = \widetilde{x}') &= \frac{P^s(Y, \widetilde{X}^s_m = \widetilde{x}_m, \xi^s = \xi, X_c = x_c)}{\sum_y P^s(Y = y, \widetilde{X}^s_m = \widetilde{x}_m, \xi^s = \xi, X_c = x_c)} \\
&= \frac{P(\xi^s = \xi | X_c = x_c) \sum_{x_m : x_m \odot \xi = \widetilde{x}_m} P(Y, X_m = x_m, X_c = x_c)}{\sum_y P(\xi^s = \xi | X_c = x_c) \sum_{x_m : x_m \odot \xi = \widetilde{x}_m} P(Y = y, X_m = x_m, X_c = x_c)} \\
&= \frac{\sum_{x_m : x_m \odot \xi = \widetilde{x}_m} P(Y, X_m = x_m, X_c = x_c)}{\sum_{x_m : x_m \odot \xi = \widetilde{x}_m} P(Y = y, X_m = x_m, X_c = x_c)},
\end{aligned}$$

which does not contain source-specific quantities (everything is in terms of the underlying distribution). By the same logic,

$$P^t(Y|\widetilde{X}' = \widetilde{x}') = \frac{\sum_{x_m : x_m \odot \xi = \widetilde{x}_m} P(Y, X_m = x_m, X_c = x_c)}{\sum_{x_m : x_m \odot \xi = \widetilde{x}_m} P(Y = y, X_m = x_m, X_c = x_c)}.$$

Thus, $P^s(Y|\widetilde{X}' = \widetilde{x}') = P^t(Y|\widetilde{X}' = \widetilde{x}')$ as desired. When ξ is instead drawn independently of other covariates, as in (a) above, we note that all of the steps of the proof follow through simply by removing X_c. Additionally, while all of the above expressions apply to discrete X, extension to continuous X is straightforward (e.g. replace summations with integrals, and constants with sets or intervals). □

C Constant Missingness as L2 Regularization

This section contains a proof of Theorem 4.1. This proof is based off of that presented in Wager et al. (2013)'s work showing dropout to be a form of adaptive regularization. Instead of assuming a single constant dropout rate across all covariates, however, our proof extends to varying rates of missingness (i.e. different constant dropout rates) for different covariates.

Proof. Assume we know the constant missingness rates m. For mathematical convenience, we preprocess $\widetilde{x}$ by multiplying each dimension by the corresponding $\frac{1}{1-m_j}$. For the remainder of this derivation, this preprocessed data is referred to as $\widetilde{x}$.

Similar to Wager et al. (2013), we start with an analysis of generalized linear models and then consider the case of linear regression. Minimizing the expected negative log likelihood $l_{\widetilde{x}^{(i)}, y^{(i)}}(\beta)$ of a generalized linear model $p_\beta(y|x) = h(y)\exp\{yx \cdot \beta - A(x \cdot \beta)\}$, we have:

$$\begin{aligned}
\widehat{\beta} &= \arg\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} \mathbb{E}_\xi[l_{\widetilde{x}^{(i)}, y^{(i)}}(\beta)] \\
\sum_{i=1}^{n} \mathbb{E}_\xi[l_{\widetilde{x}^{(i)}, y^{(i)}}(\beta)] &= \sum_{i=1}^{n} \mathbb{E}_\xi[-\log p_\beta(y^{(i)}|\widetilde{x}^{(i)})] \\
&= \sum_{i=1}^{n} \mathbb{E}_\xi[-(\log h(y^{(i)}) + y^{(i)}\widetilde{x}^{(i)}\beta - A(\widetilde{x}^{(i)} \cdot \beta))] \\
&= \sum_{i=1}^{n} -\log h(y^{(i)}) - y^{(i)}\mathbb{E}_\xi[\widetilde{x}^{(i)}]\beta + \mathbb{E}_\xi[A(\widetilde{x}^{(i)} \cdot \beta)] \\
&= \sum_{i=1}^{n} -\log h(y^{(i)}) - y^{(i)}\left(x^{(i)} \odot \frac{1-m}{1-m}\right)\beta + \mathbb{E}_\xi[A(\widetilde{x}^{(i)} \cdot \beta)] \\
&= \sum_{i=1}^{n} -(\log h(y^{(i)}) + y^{(i)}x^{(i)}\beta - A(x^{(i)}\beta)) - A(x^{(i)}\beta) + \mathbb{E}_\xi[A(\widetilde{x}^{(i)} \cdot \beta)] \\
&= \sum_{i=1}^{n} l_{x^{(i)}, y^{(i)}}(\beta) + \mathbb{E}_\xi[A(\widetilde{x}^{(i)} \cdot \beta)] - A(x^{(i)}\beta) \\
&= \sum_{i=1}^{n} l_{x^{(i)}, y^{(i)}}(\beta) + R(\beta)
\end{aligned}$$

where $R(\beta) \triangleq \sum_{i=1}^{n} \mathbb{E}_\xi[A(\widetilde{x}^{(i)} \cdot \beta)] - A(x^{(i)}\beta)$. How do we interpret $R(\beta)$?

First, we do a second order Taylor expansion of A around $x\beta$. Note that linear regression has a second order log partition function. Thus, for linear regression this expansion is exact:

$$A(y) \approx A(x\beta) + A'(x\beta)(y - x\beta) + \frac{1}{2}A''(x\beta)(y - x\beta)^2$$

$$
\begin{aligned}
A(\widetilde{x}\beta) &\approx A(x\beta) + A'(x\beta)(\widetilde{x}\beta - x\beta) + \frac{1}{2}A''(x\beta)(\widetilde{x}\beta - x\beta)^2 \\
&= A(x\beta) + A'(x\beta)(\widetilde{x} - x)\beta + \frac{1}{2}A''(x\beta)(\widetilde{x}\beta - x\beta)^2
\end{aligned}
$$

Now, we can compute the first term of $R(\beta)$:

$$
\begin{aligned}
\mathbb{E}_\xi[A(\widetilde{x}\cdot\beta)] &\approx \mathbb{E}_\xi[A(x\beta)] + \mathbb{E}_\xi[A'(x\beta)(\widetilde{x} - x)\beta] + \mathbb{E}_\xi[\frac{1}{2}A''(x\beta)(\widetilde{x}\beta - x\beta)^2] \\
&= A(x\beta) + 0 + \frac{1}{2}A''(x\beta)\mathbb{E}_\xi[(\widetilde{x}\beta - x\beta)^2] \\
&= A(x\beta) + \frac{1}{2}A''(x\beta)\text{Var}_\xi(\widetilde{x}\beta)
\end{aligned}
$$

where the second step follows because $\mathbb{E}_\xi[\widetilde{x}] = x$. Thus, $R(\beta)$ is given by:

$$
\begin{aligned}
R(\beta) &= \sum_{i=1}^{n} \mathbb{E}_\xi[A(\widetilde{x}^{(i)}\cdot\beta)] - A(x^{(i)}\beta) \\
&\approx \sum_{i=1}^{n} A(x^{(i)}\beta) + \frac{1}{2}A''(x^{(i)}\beta)\text{Var}_\xi(\widetilde{x}^{(i)}\beta) - A(x^{(i)}\beta) \\
&= \sum_{i=1}^{n} \frac{1}{2}A''(x^{(i)}\beta)\text{Var}_\xi(\widetilde{x}^{(i)}\beta) \\
&\triangleq R^q(\beta).
\end{aligned}
$$

Note that the first term corresponds to variance of $y^{(i)}$, and the second term corresponds to the variance of the estimated GLM parameter due to noising, or in the linear case, $\text{Var}(y^{(i)})$. Additionally, note that for linear regression $R(\beta) = R^q(\beta)$ since the approximate equality comes from the Taylor series approximation.

Analyzing $\text{Var}_\xi(\widetilde{x}^{(i)}\beta)$,

$$
\begin{aligned}
\text{Var}_\xi(\widetilde{x}^{(i)}\beta) &= \sum_{j=1}^{d} \text{Var}_\xi(\widetilde{x}_j^{(i)}\beta_j) \\
&= \sum_{j=1}^{d} \text{Var}_\xi\left(\frac{x_j^{(i)}}{1-m_j}\cdot b_j\cdot\beta_j\right) \\
&= \sum_{j=1}^{d} \left(\frac{x_j^{(i)}}{1-m_j}\right)^2 \beta_j^2(1-m_j)(m_j) \\
&= \sum_{j=1}^{d} \frac{m_j}{1-m_j}\left(x_j^{(i)}\right)^2\beta_j^2
\end{aligned}
$$

where $b_j \sim \text{Bernoulli}(1-m_j)$. Thus, $R^q(\beta)$ is given by:

$$
R^q(\beta) = \frac{1}{2}\sum_{i=1}^{n} A''(x^{(i)}\beta)\sum_{j=1}^{d}\frac{m_j}{1-m_j}\left(x_j^{(i)}\right)^2\beta_j^2.
$$

Let $V(\beta) \in \mathbb{R}^{n\times n}$ be diagonal with entries $A''(x^{(i)}\beta)$, and $X \in \mathbb{R}^{n\times d}$ be the design matrix with rows $x^{(i)}$. For linear regression, $V(\beta)$ is given by the identity matrix. Then, we can rewrite $R^q(\beta)$ as:

$$
R^q(\beta) = \frac{1}{2}\left(\beta\odot\sqrt{\frac{m}{1-m}}\right)^\top \text{diag}(X^\top V(\beta)X)\left(\beta\odot\sqrt{\frac{m}{1-m}}\right)
$$

$$
\begin{aligned}
R^q(\beta) &= \frac{1}{2}\left(\beta \odot \frac{m}{1-m}\right)^\top \text{diag}(I)\left(\beta \odot \frac{m}{1-m}\right) \\
&= \frac{1}{2}\left(\text{diag}(I)^{1/2}\beta \odot \frac{m}{1-m}\right)^\top \left(\text{diag}(I)^{1/2}\beta \odot \frac{m}{1-m}\right) \\
&= \frac{1}{2}\left(\beta\widetilde{\Delta}_{\text{diag}}\right)^\top \left(\beta\widetilde{\Delta}_{\text{diag}}\right)
\end{aligned}
$$

where $\widetilde{\Delta}_{\text{diag}} = \text{diag}\left(\sqrt{\frac{m}{1-m}}\right)\text{diag}(I)^{1/2}$, where $\text{diag}\left(\sqrt{\frac{m}{1-m}}\right)$ refers to a diagonal matrix with the vector quantities on the diagonal, and $\text{diag}(I)^{1/2}$ refers to the square root of the diagonal of the Fisher information matrix. Thus, for linear regression, applying missingness rates $m \in [0,1]^d$ to data scaled by $\frac{1}{1-m}$ can be viewed as an attempt to apply L2 regularization of β scaled by $\widetilde{\Delta}_{\text{diag}}$. □

D Identification of Clean Distribution from Corrupted Distribution

This section proves Lemma 5.1, which states that the clean distribution p is identified from the corrupted distribution $\widetilde{p}$ given missingness rates m, and $m \prec 1$.

Proof. Let $\mathcal{A}^k$ denote the set of possible values of x where at most k of the dimensions of x are 0. We would like to show that $\forall k \in \{0,1,...,d\}$, $\forall a \in \mathcal{A}^k$, the clean distribution $p_{a,y}$ is identifiable (and hence $p_{x,y}$ is identifiable) for all values of x and y. We proceed with a proof by induction on k.

- *Base case ($k = 0$)*:

  Consider $\mathcal{A}^0$, the set of possible values of x where none of the dimensions of x are 0. For any subset $a \subseteq \mathcal{A}^0$, we can write:

  $$\widetilde{p}_{a,y} = \prod_{j=1}^{d}(1-m_j)p_{a,y}$$

  which can be rearranged to recover p_a from $\widetilde{p}_a$ and m, which are both known:

  $$p_{a,y} = \prod_{j=1}^{d}\frac{1}{1-m_j}\widetilde{p}_{a,y}.$$

  Thus $p_{a,y}$ is identified for $a \subseteq \mathcal{A}^0$.

- *Inductive Step*: Assume $p_{a,y}$ is identified for $a \subseteq \mathcal{A}^k$. Consider some $a' \subseteq \mathcal{A}^{k+1}$. Using equation (1), we have:

  $$
  \begin{aligned}
  \widetilde{p}_{a',y} &= \sum_{b:b\rightsquigarrow a'} p_{b,y}\cdot\prod_{j=1}^{d}(1-m_j)^{[a'_j]\neq 0}m_j^{[b_j]\neq 0-[a'_j]\neq 0} \\
  &= p_{a',y}\cdot\prod_{j=1}^{d}(1-m_j)^{[a'_j]\neq 0} + \sum_{\substack{b:b\rightsquigarrow a',\\ b\neq a'}} p_{b,y}\cdot\prod_{j=1}^{d}(1-m_j)^{[a'_j]\neq 0}m_j^{[b_j]\neq 0-[a'_j]\neq 0}
  \end{aligned}
  $$

  Recall from Remark 3 that if $b \rightsquigarrow a'$, then the dimensions of b that are 0 must be a subset of the ones that are 0 in a'. Additionally, any dimensions that are nonzero in both b and a' must match in value. This implies that if there are the same number of zeros in b and a', then $b = a'$. The remaining b where $b \rightsquigarrow a'$ have at least one less zero than a'. Thus, the set of $\{b : b \rightsquigarrow a', b \neq a'\} \in \mathcal{A}^k$, and by our inductive hypothesis, $p_{b,y}$ are identified when $b \in \mathcal{A}^k$. As a result, we can identify the second term in the equation above (the summation over b's), and rearranging the equation, we can identify $p_{a',y}$ as $\widetilde{p}$ and m are known.

Thus, by the principle of mathematical induction, p_a is identified for $a \in \mathcal{A}^k$, $\forall k \in \{0,1,...,d\}$. Therefore, given m, we have identified the clean distribution from the corrupted distribution. Additionally, while all of the above expressions apply to discrete X, extension to continuous X is straightforward (e.g. replace summations with integrals, and constants with sets or intervals). □

E Identification of Labeled Target Distribution from the Labeled Source Distribution

Here we prove Theorem 5.2, which states that:

$$\widetilde{p}^t_{x,y} = \sum_{z:z\rightsquigarrow x} \widetilde{p}^s_{z,y} \cdot \prod_{j=1}^{d} (1 - r_j^{s\to t})^{[x_j]\neq 0} (r_j^{s\to t})^{[z_j]\neq 0 - [x_j]\neq 0}$$

Proof. Applying equation (1), the corrupted source and target distributions can be written as:

$$\widetilde{p}^s_{a,y} = \sum_{b:b\rightsquigarrow a} p_{b,y} \cdot \prod_{j=1}^{d} (1 - m_{sj})^{[a_j]\neq 0} m_{sj}^{[b_j]\neq 0 - [a_j]\neq 0}$$

$$\widetilde{p}^t_{a,y} = \sum_{c:c\rightsquigarrow a} p_{c,y} \cdot \prod_{j=1}^{d} (1 - m_{tj})^{[a_j]\neq 0} m_{tj}^{[c_j]\neq 0 - [a_j]\neq 0}$$

We apply relative missingness $r = r^{s\to t} = \frac{m_t - m_s}{1 - m_s}$ to source distribution $\widetilde{p}^s$, denoting this new distribution as $\widetilde{p}^{s\to t}$:

$$\begin{aligned}
\widetilde{p}^{s\to t}_{a,y} &= \sum_{b:b\rightsquigarrow a} \widetilde{p}^s_{b,y} \cdot \prod_{j=1}^{d} (1 - r_j)^{[a_j]\neq 0} r_j^{[b_j]\neq 0 - [a_j]\neq 0} \\
&= \sum_{b:b\rightsquigarrow a} \sum_{c:c\rightsquigarrow b} p_{c,y} \cdot \prod_{j=1}^{d} (1 - m_{sj})^{[b_j]\neq 0} m_{sj}^{[c_j]\neq 0 - [b_j]\neq 0} \cdot \prod_{j=1}^{d} (1 - r_j)^{[a_j]\neq 0} r_j^{[b_j]\neq 0 - [a_j]\neq 0} \\
&= \sum_{c:c\rightsquigarrow b} p_{c,y} \sum_{b:b\rightsquigarrow a} \cdot \prod_{j=1}^{d} (1 - m_{sj})^{[b_j]\neq 0} m_{sj}^{[c_j]\neq 0 - [b_j]\neq 0} \cdot \prod_{j=1}^{d} (1 - r_j)^{[a_j]\neq 0} r_j^{[b_j]\neq 0 - [a_j]\neq 0} \\
&= \sum_{c:c\rightsquigarrow b} p_{c,y} \sum_{b:b\rightsquigarrow a} \cdot \prod_{j=1}^{d} (1 - m_{sj})^{[b_j]\neq 0} m_{sj}^{[c_j]\neq 0 - [b_j]\neq 0} \cdot \prod_{j=1}^{d} \left(\frac{1 - m_{tj}}{1 - m_{sj}}\right)^{[a_j]\neq 0} \left(\frac{m_{tj} - m_{sj}}{1 - m_{sj}}\right)^{[b_j]\neq 0 - [a_j]\neq 0} \\
&= \sum_{c:c\rightsquigarrow b} p_{c,y} \sum_{b:b\rightsquigarrow a} \prod_{j=1}^{d} (1 - m_{sj})^{\mathbb{1}\{[c_j]\neq 0 = [b_j]\neq 0 = 1, [a_j]\neq 0 = 0\} + \mathbb{1}\{[c_j]\neq 0 = [b_j]\neq 0 = [a_j]\neq 0 = 1\}} \\
&\qquad \cdot m_{sj}^{\mathbb{1}\{[c_j]\neq 0 = 1, [b_j]\neq 0 = [a_j]\neq 0 = 0\}} \\
&\qquad \cdot \left(\frac{1 - m_{tj}}{1 - m_{sj}}\right)^{\mathbb{1}\{[c_j]\neq 0 = [b_j]\neq 0 = [a_j]\neq 0 = 1\}} \\
&\qquad \cdot \left(\frac{m_{tj} - m_{sj}}{1 - m_{sj}}\right)^{\mathbb{1}\{[c_j]\neq 0 = [b_j]\neq 0 = 1, [a_j]\neq 0 = 0\}} \\
&= \sum_{c:c\rightsquigarrow b} p_{c,y} \sum_{b:b\rightsquigarrow a} \prod_{j=1}^{d} m_{sj}^{\mathbb{1}\{[c_j]\neq 0 = 1, [b_j]\neq 0 = [a_j]\neq 0 = 0\}} \\
&\qquad \cdot (1 - m_{tj})^{\mathbb{1}\{[c_j]\neq 0 = [b_j]\neq 0 = [a_j]\neq 0 = 1\}} \\
&\qquad \cdot (m_{tj} - m_{sj})^{\mathbb{1}\{[c_j]\neq 0 = [b_j]\neq 0 = 1, [a_j]\neq 0 = 0\}} \\
&= \sum_{c:c\rightsquigarrow a} p_{c,y} \cdot \left(\prod_{j:[c_j]\neq 0 = [a_j]\neq 0 = 1} 1 - m_{tj}\right) \cdot \left(\prod_{j:[c_j]\neq 0 = [a_j]\neq 0 = 0} 1\right) \\
&\qquad \cdot \sum_{b:b\rightsquigarrow a} \left(\prod_{j:[c_j]\neq 0 = 1, [a_j]\neq 0 = 0} m_{sj}^{1 - [b_j]\neq 0} (m_{tj} - m_{sj})^{[b_j]\neq 0}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{c:c\rightsquigarrow a} p_{c,y} \cdot \left(\prod_{j:[c_j]_{\neq 0}=[a_j]_{\neq 0}=1} 1 - m_{tj} \right) \cdot \left(\prod_{j:[c_j]_{\neq 0}=[a_j]_{\neq 0}=0} 1 \right) \\
&\qquad \cdot \sum_{[b]_{\neq 0}\in\{0,1\}^d} \left(\prod_{j:[c_j]_{\neq 0}=1,[a_j]_{\neq 0}=0} m_{sj}^{1-[b_j]_{\neq 0}} (m_{tj} - m_{sj})^{[b_j]_{\neq 0}} \right) \\
&= \sum_{c:c\rightsquigarrow a} p_{c,y} \cdot \left(\prod_{j:[c_j]_{\neq 0}=[a_j]_{\neq 0}=1} 1 - m_{tj} \right) \cdot \left(\prod_{j:[c_j]_{\neq 0}=[a_j]_{\neq 0}=0} 1 \right) \cdot \left(\prod_{j:[c_j]_{\neq 0}=1,[a_j]_{\neq 0}=0} m_{t_j} \right) \\
&= \sum_{c:c\rightsquigarrow a} p_{c,y} \prod_{j=1}^{d} (1 - m_{tj})^{[a_j]_{\neq 0}} m_{tj}^{[c_j]_{\neq 0} - [a_j]_{\neq 0}} \\
&= \widetilde{p}^t_{a,y}
\end{aligned}
$$

as desired. The steps are explained in words below:

- Plug in equation for corrupted source distribution.
- Switch summation order and factor out $p_{c,y}$.
- Plug in for r.
- Note that $[c_j]_{\neq 0} - [b_j]_{\neq 0} = 1$ only if $[c_j]_{\neq 0} = 1$ and $[b_j]_{\neq 0} = 0$. Use similar reasoning for the remaining, keeping in mind that $[c]_{\neq 0} \succeq [b]_{\neq 0} \succeq [a]_{\neq 0}$. Simplify.
- Since all elements of the sum have $\mathbb{1}\{[c]_{\neq 0} \succeq [b]_{\neq 0} \succeq [a]_{\neq 0}\}$, it is also true that $\mathbb{1}\{[c]_{\neq 0} \succeq [a]_{\neq 0}\}$.
- If $[a_i]_{\neq 0} = [c_i]_{\neq 0} = 1$, then $[b_i]_{\neq 0} = 1$ necessarily.
- Note that if $c \rightsquigarrow b \rightsquigarrow a$ and $[c_i]_{\neq 0} = 1, [a_i]_{\neq 0} = 0$, then $\forall i, b_i \in \{0, c_i\}$. We can then perform a change of variables in the summation, now summing over $[b]_{\neq 0} \in \{0,1\}^d$ instead.
- We use the following identity for arbitrary d-dimensional vectors a and b:

$$
\sum_{u\in\{0,1\}^d} \prod_j a_j^{u_j} b_j^{1-u_j} = \prod_j (a_j + b_j)
$$

To gain intuition for why this is the case, let's start with $d = 2$:

$$
\begin{aligned}
LHS &= \sum_{u\in\{0,1\}^d} \prod_j a_j^{u_j} b_j^{1-u_j} \\
&= \sum_{u\in\{0,1\}^2} a_1^{u_1} b_1^{1-u_1} a_2^{u_2} b_2^{(1-u_2)} \\
&= \sum_{u\in[(1,1),(1,0),(0,1),(0,0)]} a_1^{u_1} b_1^{1-u_1} a_2^{u_2} b_2^{(1-u_2)} \\
&= a_1a_2 + a_1b_2 + b_1a_2 + b_1b_2 \\
RHS &= \prod_j (a_j + b_j) \\
&= (a_1 + b_1)(a_2 + b_2) \\
&= a_1a_2 + a_1b_2 + b_1a_2 + b_1b_2
\end{aligned}
$$

Notice that the right-hand side is a product of sums $(a_j + b_j)$, of which there are d terms. When expanding this product of sums into a sum of products, each term in the sum of products will include either a_j or b_j for all $j \in 1, 2, ..., d$. Summing over all possible choices of either a_j or b_j for all j is then equivalent to summing over all possible values of a binary d-dimensional vector u. Thus, we get the left-hand side of the identity.

- The remaining steps are straightforward simplifications to get a form matching equation (3).
- Note that while all of the above expressions apply to discrete X, extension to continuous X is straightforward (e.g. replace summations with integrals, and constants with sets or intervals).

□

F Error Bound for Estimating Non-Missing Proportions

This is a proof of Theorem 6.1. To estimate the non-missingness proportion $q = P(\widetilde{X} = 1)$ within ϵ of the true non-missingness proportion with probability at least $1 - \delta$, we use Hoeffding's bound to show:

$$\begin{aligned} P(|\widehat{q} - q| \geq \epsilon) &\leq 2\exp(-2ne^2) = \delta \\ \implies -2n\epsilon^2 &= \log(\delta/2) \\ \implies n &= \frac{\log(2/\delta)}{2\epsilon^2} \\ \implies |\widehat{q} - q| &= \sqrt{\frac{\log(2/\delta)}{2n}}. \end{aligned}$$

Now, we show that with high probability, the estimate for $1 - r^{s \to t} = \frac{q_t}{q_s}$ is close to the true value. This part of the derivation is similar to that used in Garg et al. (2021). Using triangle inequality,

$$\begin{aligned} \left| \frac{\widehat{q}_t}{\widehat{q}_s} - \frac{q_t}{q_s} \right| &= \left| \frac{q_s \widehat{q}_t - \widehat{q}_s q_t}{\widehat{q}_s q_s} \right| \\ &= \frac{1}{\widehat{q}_s q_s} \left| q_s \widehat{q}_t - q_s q_t + q_s q_t - \widehat{q}_s q_t \right| \\ &\leq \frac{1}{\widehat{q}_s q_s} \left| q_s \widehat{q}_t - q_s q_t \right| + \frac{1}{\widehat{q}_s q_s} \left| q_s q_t - \widehat{q}_s q_t \right| \\ &\leq \frac{1}{\widehat{q}_s} \left| \widehat{q}_t - q_t \right| + \frac{q_t}{\widehat{q}_s q_s} \left| q_s - \widehat{q}_s \right|. \end{aligned}$$

On the right hand side, we use the union bound and plug in $\delta/2$ for δ in Hoeffding's bound. Plugging in, we then have that with probability at least $1 - \delta$,

$$\begin{aligned} \left| \frac{\widehat{q}_t}{\widehat{q}_s} - \frac{q_t}{q_s} \right| &\leq \frac{1}{\widehat{q}_s} \left(\sqrt{\frac{\log(4/\delta)}{2n_t}} + \frac{q_t}{q_s} \sqrt{\frac{\log(4/\delta)}{2n_s}} \right) \\ \implies \left| \widehat{r}^{s \to t} - r^{s \to t} \right| &\leq \frac{1}{\widehat{P}^s(\widetilde{x} = 1)} \left(\sqrt{\frac{\log(4/\delta)}{2n_t}} + (1 - r^{s \to t}) \sqrt{\frac{\log(4/\delta)}{2n_s}} \right). \end{aligned}$$

G Justification for the Non-parametric Procedure with Non-Negative Relative Missingness

Simple Justification Since (3) matches the form of (1) except with $m = r^{s\to t}$, applying missingness with rate $r^{s\to t}$ to the source distribution will yield samples independent and identically distributed to the target distribution. That is, plugging in $\widetilde{p}^s$ for p and $r^{s\to t}$ for m, we have:

$$\begin{aligned}\widetilde{p}_{x,y} &= \sum_{z:z\rightsquigarrow x} p_{z,y}\cdot\prod_{j=1}^{d}(1-m_j)^{[x_j]_{\neq 0}}m_j^{[z_j]_{\neq 0}-[x_j]_{\neq 0}}\\ &= \sum_{z:z\rightsquigarrow x} \widetilde{p}^s_{z,y}\cdot\prod_{j=1}^{d}(1-r_j^{s\to t})^{[x_j]_{\neq 0}}(r_j^{s\to t})^{[z_j]_{\neq 0}-[x_j]_{\neq 0}}\\ &= \widetilde{p}^t_{x,y}\end{aligned}$$

where the first line is (1) and the third line follows from (3).

Alternative Justification Suppose that $m^t \succeq m^s$, where $\succeq$ denotes whether all elements of m^t are greater than or equal to all corresponding elements of m^s, that is, $m_j^t \geq m_j^s$ for $j = 1, 2, ..., d$. Below, we show that the data generating process for the target data is equivalent to applying a missingness filter with relative missingness rate $r^{s\to t}$ applied to the source data. To draw a point from the source, target, and transformed distribution, respectively, one first draws a clean data point $(x, y) \sim P(X, Y)$, where $x \in \mathbb{R}^d, y \in \mathbb{R}$, and then applies the respective missingness filter to the clean covariates:

$$\begin{aligned}\widetilde{x}^s &= \nu_s(x) = x\odot\xi^s\\ \widetilde{x}^t &= \nu_t(x) = x\odot\xi^t\\ \widetilde{x}^{s\to t} &= \nu_{s\to t}(\nu_s(x)) = x\odot\xi^s\odot\xi^{s\to t}\end{aligned}$$

where $\xi^t \sim \text{Bernoulli}(1-m^t)$, $\xi^s \sim \text{Bernoulli}(1-m^s)$, and $\xi^{s\to t} \sim \text{Bernoulli}(1-r^{s\to t})$. Combining Bernoullis, we have:

$$\begin{aligned}\xi^s\odot\xi^{s\to t} &= \begin{cases}1 & \text{w.p. } \left(1-\frac{m^t-m^s}{1-m^s}\right)\cdot(1-m^s)\\ 0 & otherwise\end{cases}\\ &= \begin{cases}1 & \text{w.p. } (1-m^t)\\ 0 & otherwise\end{cases} = \xi^t\end{aligned}$$

Thus, for true relative missing rates $r^{s\to t}$, we have $\nu_t(x) = \nu_{s\to t}(\nu_s(x))$. Since the data generating process after applying $\nu_{s\to t}$ to source data is now identical to the data generating process of the target dataset, we have $\{(\nu_{s\to t}(\widetilde{X}^{s,i}), Y^{s,i})\}_{i=1}^{n_s}$ drawn independent and identically distributed to $P^t(\widetilde{X}, Y)$.

H Optimal Linear Predictors

H.1 Optimal linear target predictor, derived from target covariances

For each dimension j, the covariance between corrupted data $\widetilde{X}_j$ with missingness rate m and its labels Y is $\text{Cov}(\widetilde{X}_j, Y) = \text{Cov}(X_j \cdot \xi_j, Y) = (1 - m_j)\text{Cov}\,(X_j, Y)$. Thus,

$$\begin{aligned}
\text{Cov}(X, Y) &= \frac{1}{1-m} \odot \text{Cov}(\widetilde{X}, Y) \\
\mathbb{E}[X^\top Y] &= \text{Cov}\,(X, Y) + \mathbb{E}[X]^\top \mathbb{E}[Y] \\
&= \frac{1}{1-m} \odot \text{Cov}\left(\widetilde{X}, Y\right) + \frac{1}{1-m} \odot \mathbb{E}[\widetilde{X}]^\top \mathbb{E}[Y] \\
&= \frac{1}{1-m} \odot \mathbb{E}[\widetilde{X}^\top Y].
\end{aligned}$$

Plugging into the ordinary least squares regression solution,

$$\begin{aligned}
\beta_*^t &= \mathbb{E}[\widetilde{X}^{t\top} \widetilde{X}^t]^{-1} \mathbb{E}[\widetilde{X}^{t\top} Y^t] \\
&= \mathbb{E}[\widetilde{X}^{t\top} \widetilde{X}^t]^{-1} \left((1 - m_t) \odot \mathbb{E}[X^\top Y]\right) \\
&= \mathbb{E}[\widetilde{X}^{t\top} \widetilde{X}^t]^{-1} \left(\frac{1 - m_t}{1 - m_s} \odot \mathbb{E}[\widetilde{X}^{s\top} Y^s]\right) \\
&= \mathbb{E}[\widetilde{X}^{t\top} \widetilde{X}^t]^{-1} \left(r^{s \to t} \odot \mathbb{E}[\widetilde{X}^{s\top} Y^s]\right).
\end{aligned}$$

The remainder of this section derives the optimal linear target predictor, where the corrupted target covariance is derived from the corrupted source covariance.

H.2 Means, Variances, and Covariances

This section begins by deriving the relationships between the means, covariances, and variances of the corrupted and clean data. Then, it derives the relationships between corrupted and clean $\mathbb{E}[X^\top X]$. Finally, the derived first and second moments are summarized in Table 3.

Recall that for any covariate x_j, we have:

$$\begin{aligned}
\widetilde{x}_j &= \begin{cases} 0 & \text{w.p. } m_j \\ x_j & \text{w.p. } 1 - m_j \end{cases} \\
&= b_j x_j
\end{aligned}$$

where $b_j \sim \text{Bernoulli}(1 - m_j)$. The **mean** of the corrupted data is given by:

$$\mathbb{E}[\widetilde{X}] = (1 - m) \odot \mathbb{E}\,[X]$$

To derive the covariance matrix of the corrupted data, consider the covariance between two arbitrary distinct covariate dimensions $\widetilde{x}_1$ and $\widetilde{x}_2$. Let $A = b_1$, $B = x_1$, $C = b_2$, and $D = x_2$. Note that A and C are independent of all other variables. Thus,

$$\begin{aligned}
\text{Cov}(\widetilde{x}_1, \widetilde{x}_2) &= \text{Cov}(AB, CD) \\
&= \mathbb{E}[ABCD] - \mathbb{E}[AB]\mathbb{E}[CD] \\
&= \mathbb{E}[ABCD] - \mathbb{E}[A]\mathbb{E}[B]\mathbb{E}[C]\mathbb{E}[D] \\
&= \mathbb{E}[A]\mathbb{E}[C](\mathbb{E}[BD] - \mathbb{E}[B]\mathbb{E}[D]) \\
&= \mathbb{E}[A]\mathbb{E}[C]\text{Cov}(B, D)
\end{aligned}$$

$$
\begin{aligned}
&= (1-m_1)(1-m_2)\text{Cov}\left(x_1, x_2\right) \\
\implies \text{Cov}(x_1, x_2) &= \frac{1}{(1-m_1)(1-m_2)}\text{Cov}(\widetilde{x}_1, \widetilde{x}_2)
\end{aligned}
$$

And similarly,

$$
\begin{aligned}
\text{Cov}(\widetilde{x}_1, y) &= (1-m_1)\text{Cov}\left(x_1, y\right) \\
\implies \text{Cov}\left(x_1, y\right) &= \frac{1}{1-m_1}\text{Cov}(\widetilde{x}_1, y)
\end{aligned}
$$

The variance (entries along the diagonal of the covariance matrix) is given by:

$$
\begin{aligned}
\text{Var}(\widetilde{x}_1) &= \text{Var}\left(b_1 x_1\right) \\
&= \text{Var}(AB) \\
&= (\sigma_A^2 + \mu_A^2)(\sigma_B^2 + \mu_B^2) - \mu_A^2\mu_B^2 \\
&= (m_1(1-m_1) + (1-m_1)^2)\left(\text{Var}(x_1) + \mathbb{E}[x_1]^2\right) - (1-m_1)^2\mathbb{E}[x_1]^2 \\
&= (1-m_1)\left(\text{Var}(x_1) + \mathbb{E}[x_1]^2\right) - (1-m_1)^2\mathbb{E}[x_1]^2 \\
&= (1-m_1)\left(\text{Var}(x_1) + \mathbb{E}[x_1]^2 - (1-m_1)\mathbb{E}[x_1]^2\right) \\
&= (1-m_1)\left(\text{Var}(x_1) + \mathbb{E}[x_1]^2 - \mathbb{E}[x_1]^2 + m_1\mathbb{E}[x_1]^2\right) \\
&= (1-m_1)\left(\text{Var}(x_1) + m_1\mathbb{E}[x_1]^2\right) \\
&= (1-m_1)\text{Var}(x_1) + m_1(1-m_1)\mathbb{E}[x_1]^2 \\
\text{Var}(x_1) &= \frac{\text{Var}(\widetilde{x}_1)}{1-m_1} - m_1\mathbb{E}[x_1]^2 \\
&= \frac{\text{Var}(\widetilde{x}_1)}{1-m_1} - \frac{m_1}{(1-m_1)^2}\mathbb{E}[\widetilde{x}_1]^2
\end{aligned}
$$

Putting this together, the variance-covariance matrix is given by (elementwise division below):

$$
\begin{aligned}
\text{Cov}(\widetilde{X}, \widetilde{X}) &= (1-m)(1-m)^\top \odot \text{Cov}(X, X) \\
&\quad + \text{diag}(((1-m) - (1-m)^2)\text{Var}(X) + m(1-m)\mathbb{E}[x_1]^2) \\
&= (1-m)(1-m)^\top \odot \text{Cov}(X, X) + \text{diag}(m(1-m)(\text{Var}(X) + \mathbb{E}[X]^2)) \\
&= (1-m)(1-m)^\top \odot \text{Cov}(X, X) \\
&\quad + \text{diag}(m(1-m)^\top)\text{diag}(\text{Cov}(X, X) + \mathbb{E}[X]^\top\mathbb{E}[X]) \\
&= (1-m)(1-m)^\top \odot \text{Cov}(X, X) + \text{diag}(m(1-m)^\top)\text{diag}(\mathbb{E}[X^\top X]) \\
\implies \text{Cov}(X, X) &= \left(\frac{1}{1-m}\right)\left(\frac{1}{1-m}\right)^\top \odot \text{Cov}(\widetilde{X}, \widetilde{X}) \\
&\quad + \text{diag}\left(-\frac{\text{Var}(\widetilde{X})}{(1-m)^2} + \frac{\text{Var}(\widetilde{X})}{1-m} - \frac{m\mathbb{E}\left[\widetilde{X}\right]^2}{(1-m)^2}\right) \\
&= \left(\frac{1}{1-m}\right)\left(\frac{1}{1-m}\right)^\top \odot \text{Cov}(\widetilde{X}, \widetilde{X}) - \text{diag}\left(\frac{m}{(1-m)^2}(\text{Var}(\widetilde{X}) + \mathbb{E}[\widetilde{X}]^2)\right)
\end{aligned}
$$

Thus far, we have been working with the covariance matrix. How do the expressions for covariance relate to $\widetilde{X}^\top\widetilde{X}$ and $\widetilde{X}^\top Y$? We have:

$$
\begin{aligned}
\text{Cov}(\widetilde{X}, \widetilde{X}) &= (1-m)(1-m)^\top \odot \text{Cov}(X, X) + \text{diag}\left(m(1-m)^\top\right)\text{diag}(\mathbb{E}[X^\top X]) \\
\mathbb{E}[\widetilde{X}^\top\widetilde{X}] &= \text{Cov}\left(\widetilde{X}, \widetilde{X}\right) + \mathbb{E}[\widetilde{X}]^\top\mathbb{E}[\widetilde{X}]
\end{aligned}
$$

$$= (1-m)(1-m)^\top \odot (\text{Cov}(X,X) + \mathbb{E}[X]^\top \mathbb{E}[X]) \text{diag}\left(m(1-m)^\top\right) \text{diag}\left(\mathbb{E}\left[X^\top X\right]\right)$$
$$= (1-m)(1-m)^\top \odot \mathbb{E}\left[X^\top X\right] + \text{diag}\left(m(1-m^\top)\right) \text{diag}\left(\mathbb{E}\left[X^\top X\right]\right)$$

Additionally,

$$\begin{aligned}
\text{Cov}\left(X,X\right) &= \left(\frac{1}{1-m}\right)\left(\frac{1}{1-m}\right)^\top \odot \text{Cov}(\widetilde{X},\widetilde{X}) + \text{diag}\left(-\frac{m}{(1-m)^2}\right) \text{diag}\left(\text{Var}(\widetilde{X}) + \mathbb{E}[\widetilde{X}]^2\right) \\
\mathbb{E}\left[X^\top X\right] &= \text{Cov}\left(X,X\right) + \mathbb{E}\left[X\right]^\top \mathbb{E}\left[X\right] \\
&= \left(\frac{1}{1-m}\right)\left(\frac{1}{1-m}\right)^\top \odot \left(\text{Cov}(\widetilde{X},\widetilde{X}) + \mathbb{E}[\widetilde{X}]^\top \mathbb{E}[\widetilde{X}]\right) \\
&\quad + \text{diag}\left(-\frac{m}{(1-m)^2}\right) \text{diag}\left(\text{Var}(\widetilde{X}) + \mathbb{E}[\widetilde{X}]^2\right) \\
&= \left(\frac{1}{1-m}\right)\left(\frac{1}{1-m}\right)^\top \odot \mathbb{E}[\widetilde{X}^\top \widetilde{X}] - \text{diag}\left(\frac{m}{(1-m)^2}\right) \text{diag}\left(\mathbb{E}[\widetilde{X}^\top \widetilde{X}]\right)
\end{aligned}$$

Table 3: Summary of 1st and 2nd moments of corrupted data and clean data

Quantity of Interest	Expression
$\mathbb{E}\left[X\right]$	$\frac{1}{1-m} \odot \mathbb{E}\left[\widetilde{X}\right]$
$\mathbb{E}\left[\widetilde{X}\right]$	$(1-m) \odot \mathbb{E}\left[X\right]$
$\mathbb{E}\left[X^\top X\right]$	$\left(\frac{1}{1-m}\right)\left(\frac{1}{1-m}\right)^\top \odot \mathbb{E}\left[\widetilde{X}^\top \widetilde{X}\right] - \text{diag}\left(\frac{m}{(1-m)^2}\right) \text{diag}\left(\mathbb{E}[\widetilde{X}^\top \widetilde{X}]\right)$
$\mathbb{E}\left[\widetilde{X}^\top \widetilde{X}\right]$	$(1-m)(1-m)^\top \odot \mathbb{E}\left[X^\top X\right] + \text{diag}\left(m(1-m)^\top\right) \text{diag}\left(\mathbb{E}\left[X^\top X\right]\right)$

H.3 Closed Form Solution

Using results from previous sections, we can now derive a closed form solution for the optimal linear classifier for a target domain with missing rates m_t, given labeled data from a source domain with missing rates m_s. We break down this problem by going from corrupted data with some missingness rate to clean data with 0 missingness, and then from clean data with 0 missingness to corrupted data with another level of missingness.

Suppose we are going from corrupted data $\widetilde{X}$ with missing rate m to clean data X with 0 missingness:

$$\begin{aligned}
\text{Cov}\left(X,y\right) &= \frac{1}{1-m} \odot \text{Cov}\left(\widetilde{X},y\right) \\
\mathbb{E}\left[X^\top y\right] &= \text{Cov}\left(X,y\right) + \mathbb{E}\left[X\right]^\top \mathbb{E}\left[y\right] \\
&= \frac{1}{1-m} \odot \text{Cov}\left(\widetilde{X},y\right) + \frac{1}{1-m} \odot \mathbb{E}\left[\widetilde{X}\right]^\top \mathbb{E}\left[y\right] \\
&= \frac{1}{1-m} \odot \mathbb{E}\left[\widetilde{X}^\top y\right] \\
\mathbb{E}\left[X^\top X\right] &= \left(\frac{1}{1-m}\right)\left(\frac{1}{1-m}\right)^\top \odot \mathbb{E}\left[\widetilde{X}^\top \widetilde{X}\right] - \text{diag}\left(\frac{m}{(1-m)^2} \odot \mathbb{E}\left[\widetilde{X}^\top \widetilde{X}\right]\right) \\
\implies \beta &= \left\{\left(\frac{1}{1-m}\right)\left(\frac{1}{1-m}\right)^\top \odot \mathbb{E}\left[\widetilde{X}^\top \widetilde{X}\right] - \text{diag}\left(\frac{m}{(1-m)^2} \odot \mathbb{E}\left[\widetilde{X}^\top \widetilde{X}\right]\right)\right\}^{-1} \frac{1}{1-m} \odot \mathbb{E}\left[\widetilde{X}^\top y\right]
\end{aligned}$$

Going from clean to corrupted data, we have:

$$\begin{aligned}
\mathbb{E}[\widetilde{X}^\top y] &= \text{Cov}(\widetilde{X},y) + \mathbb{E}[\widetilde{X}]^\top \mathbb{E}\left[y\right] \\
&= (1-m) \odot \text{Cov}\left(X,y\right) + (1-m) \odot \mathbb{E}\left[\widetilde{X}\right]^\top \mathbb{E}\left[y\right]
\end{aligned}$$

$$\mathbb{E}[\widetilde{X}^\top \widetilde{X}] = (1-m)(1-m)^\top \odot \mathbb{E}\left[X^\top X\right] + \text{diag}\left(m(1-m^\top)\right)\text{diag}\left(\mathbb{E}\left[X^\top X\right]\right)$$
$$\implies \widetilde{\beta} = \left[(1-m)(1-m)^\top \odot \mathbb{E}\left[X^\top X\right] + \text{diag}\left(m(1-m^\top)\right)\text{diag}\left(\mathbb{E}\left[X^\top X\right]\right)\right]^{-1}(1-m)\odot \mathbb{E}\left[X^\top y\right]$$

Now, we put all of these equations together, going from source corrupted data (S), to clean data (C), to target corrupted data (T).

(S) $\to$ (C):

$$\mathbb{E}[X^\top X] = \left(\frac{1}{1-m_s}\right)\left(\frac{1}{1-m_s}\right)^\top \odot \mathbb{E}[\widetilde{X}^{s\top}\widetilde{X}^s] - \text{diag}\left(\frac{m_s}{(1-m_s)^2}\odot(\mathbb{E}[\widetilde{X}^{s\top}\widetilde{X}^s])\right)$$
$$\mathbb{E}[X^\top y] = \frac{1}{1-m_s}\odot \text{Cov}\left(\widetilde{X}^s, y\right) + \frac{1}{1-m_s}\odot \mathbb{E}[\widetilde{X}^s]^\top \mathbb{E}[y]$$

(C) $\to$ (T):

$$\begin{aligned}
\mathbb{E}\left[\widetilde{X}^{t\top}\widetilde{X}^t\right] &= (1-m_t)(1-m_t)^\top \odot \mathbb{E}\left[X^\top X\right] + \text{diag}\left(m_t(1-m_t^\top)\right)\text{diag}\left(\mathbb{E}\left[X^\top X\right]\right) \\
&= (1-m_t)(1-m_t)^\top \odot \left[\left(\frac{1}{1-m_s}\right)\left(\frac{1}{1-m_s}\right)^\top \odot \mathbb{E}\left[\widetilde{X}^{s\top}\widetilde{X}^s\right] - \text{diag}\left(\frac{m_s}{(1-m_s)^2}\mathbb{E}\left[\widetilde{X}^{s\top}\widetilde{X}^s\right]\right)\right] \\
&\quad + \text{diag}\left(\frac{m_t}{1-m_t}\right)\odot \text{diag}\left(\mathbb{E}\left[\widetilde{X}^{s\top}\widetilde{X}^s\right] - \text{diag}\left(m_s\right)\mathbb{E}\left[\widetilde{X}^{s\top}\widetilde{X}^s\right]\right) \\
&= (1-m_t)(1-m_t)^\top \odot \left(\frac{1}{1-m_s}\right)\left(\frac{1}{1-m_s}\right)^\top \odot \mathbb{E}\left[\widetilde{X}^{s\top}\widetilde{X}^s\right] \\
&\quad - (1-m_t)(1-m_t)^\top \odot \text{diag}\left(\frac{m_s}{(1-m_s)^2}\odot \mathbb{E}\left[\widetilde{X}^{s\top}\widetilde{X}^s\right]\right) \\
&\quad + \text{diag}\left(\frac{m_t(1-m_t)}{1-m_s}\odot \mathbb{E}\left[\widetilde{X}^{s\top}\widetilde{X}^s\right]\right)
\end{aligned}$$

For $i \neq j$, the off-diagonal entries of the above expression are given by:

$$\mathbb{E}\left[\widetilde{X}^{t\top}\widetilde{X}^t\right]_{ij} = \left(\frac{1-m_{ti}}{1-m_{si}}\right)\left(\frac{1-m_{tj}}{1-m_{sj}}\right)\mathbb{E}\left[\widetilde{X}^{s\top}\widetilde{X}^s\right]_{ij} = (1-r_i^{s\to t})(1-r_j^{s\to t})\mathbb{E}\left[\widetilde{X}^{s\top}\widetilde{X}^s\right]_{ij}$$

The diagonal entries of the above expression are given by:

$$\begin{aligned}
\mathbb{E}\left[\widetilde{X}^{t\top}\widetilde{X}^t\right]_{ii} &= \mathbb{E}\left[\widetilde{X}^{s\top}\widetilde{X}^s\right]_{ii}\left(\left(\frac{1-m_{ti}}{1-m_{si}}\right)^2 - \frac{m_{si}(1-m_{ti})^2}{(1-m_{si})^2} + \frac{m_{ti}(1-m_{ti})}{1-m_{si}}\right) \\
&= \mathbb{E}\left[\widetilde{X}^{s\top}\widetilde{X}^s\right]_{ii}\left((1-r_i^{s\to t})^2 - m_{si}(1-r_i^{s\to t})^2 + m_{ti}(1-r_i^{s\to t})\right) \\
&= \mathbb{E}\left[\widetilde{X}^{s\top}\widetilde{X}^s\right]_{ii}(1-r_i^{s\to t})\left((1-r_i^{s\to t}) - m_{si}(1-r_i^{s\to t}) + m_{ti}\right) \\
&= \mathbb{E}\left[\widetilde{X}^{s\top}\widetilde{X}^s\right]_{ii}(1-r_i^{s\to t})\left(\frac{1-m_{ti}}{1-m_{si}} - \frac{m_{si}-m_{si}m_{ti}}{1-m_{si}} + \frac{m_{ti}-m_{si}m_{ti}}{1-m_{si}}\right) \\
&= \mathbb{E}\left[\widetilde{X}^{s\top}\widetilde{X}^s\right]_{ii}(1-r_i^{s\to t})\left(\frac{1-m_{si}}{1-m_{si}}\right) \\
&= \mathbb{E}\left[\widetilde{X}^{s\top}\widetilde{X}^s\right]_{ii}(1-r_i^{s\to t})
\end{aligned}$$

Additionally,

$$\begin{aligned}
\mathbb{E}\left[\widetilde{X}^{t\top}y\right] &= (1-m_t)\odot \mathbb{E}\left[X^\top y\right] \\
&= (1-m_t)\odot\left(\frac{1}{1-m_s}\odot \text{Cov}\left(\widetilde{X}^s, y\right) + \frac{1}{1-m_s}\odot \mathbb{E}\left[\widetilde{X}^s\right]^\top \mathbb{E}\left[y\right]\right) \\
&= \frac{1-m_t}{1-m_s}\odot \mathbb{E}\left[\widetilde{X}^{s\top}y\right]
\end{aligned}$$

I Experiment Details

Experiments were run on a machine with 28 CPU cores. The linear regression models were implemented from scratch and validated against that of sklearn. The MLPRegressor class from the scikit-learn Python package was used with default hyperparameters, and the XGBoost class from the xgboost Python package was used with default hyperparameters. All experiments (except imputation) are feasible to run within a few hours.

Semi-synthetic experiments on linear models included 10 samples of β, and 50 samples of missingness rates under each regime ($m^s \preceq m^t$ and $m^s \ ? \ m^t$). Semi-synthetic experiments on nonlinear models (XGB, NN) included 5 samples of β and 20 samples of missingness rates under each regime. Across these runs, 95% confidence intervals were computed.

In the imputation experiments, a MissForest imputer from the missingpy Python package was trained on the combination of the source training set and target training set (just on the covariates, without labels). This imputer was then applied to both the source and target test sets. Finally, we train a source classifier on the imputed source labeled data and evaluate its performance on the target unlabeled data. We note that in our experience with the imputation experiments, imputation was somewhat slow (2-3 minutes for each imputation), and so all of our imputed results are reported on 5 samples of β and 20 samples of missingness rates under each regime, across all semi-synthetic datasets.

I.1 Synthetic Data Experiments

Table 4: MSE/Var(Y) on Redundant Features and Confounded Features settings, with 95% confidence intervals computed over varying ϵ between 0.05 to 0.95.

	$m^s \preceq m^t$	$m^s \ ? \ m^t$
Lin. Reg. (oracle)	0.178 (0.172 – 0.185)	0.206 (0.199 – 0.213)
Lin. Reg. (source)	1.259 (1.231 – 1.286)	1.103 (1.076 – 1.129)
Lin. Reg. (imputed)	1.002 (1.002 – 1.002)	0.918 (0.915 – 0.921)
Lin. Reg. (closed-form adj.)	**0.186 (0.180 – 0.193)**	**0.209 (0.205 – 0.213)**
Lin. Reg. (non-param. adj.)	0.473 (0.471 – 0.476)	0.492 (0.489 – 0.495)
XGBoost (oracle)	0.166 (0.160 – 0.172)	0.200 (0.193 – 0.208)
XGBoost (source)	**0.166 (0.160 – 0.172)**	0.475 (0.458 – 0.492)
XGBoost (imputed)	1.002 (1.002 – 1.002)	1.157 (1.102 – 1.211)
XGBoost (non-param. adj.)	0.425 (0.422 – 0.428)	**0.473 (0.468 – 0.478)**
MLP (oracle)	0.166 (0.160 – 0.172)	0.201 (0.195 – 0.208)
MLP (source)	**0.184 (0.165 – 0.202)**	**0.321 (0.300 – 0.342)**
MLP (imputed)	1.003 (1.002 – 1.003)	0.924 (0.918 – 0.930)
MLP (non-param. adj.)	0.436 (0.428 – 0.444)	0.470 (0.465 – 0.474)

I.2 Semi-Synthetic Data Experiments

The UCI datasets Dua and Graff (2017) used in this work are:

- Adult Data Set: The classification task is whether an individual's income exceeds $50K a year based on census data. The dataset contains categorical variables (occupation, education, marital status, etc.), as well as continuous variables (age, hours per week, etc.)
- Bank Marketing Data Set: The classification task is whether a client will subscribe a term deposit. This dataset contains categorical features such as type of job, marital status, education, whether they have a housing loan, etc., as well as continuous variables such as age, number of contacts performed, etc.
- Thyroid Disease Data Set: The classification task is of increased vs. decreased binding protein. This dataset contains binary variables such as whether the patient is pregnant, is male, on thyroxine, has a tumor, etc., as well as continuous variables such as age, TSH, T3, TT4, etc.

For semi-synthetic experiments, we pre-process the UCI data by creating dummy variables from categorical variables, dropping redundant columns, normalizing numerical variables, dropping binary variables with low frequency ($< 5\%$, since we apply additional synthetic missingness in our experiments), and dropping columns with low variance ($< 5\%$). We additionally generate synthetic labels by sampling coefficients $\beta_j \sim \text{Uniform}(0, 10), \forall j \in \{0, 1, 2, ..., d\}$ and computing new synthetic labels $y_{new} = X\beta$. Table 5 contains the MSE/Var(Y) and 95% confidence intervals (from sampling several β and m^s, m^t) of the adult dataset, Table 6 contains the MSE/Var(Y) and 95% confidence intervals of the bank dataset, and Table 7 contains the MSE/Var(Y) and 95% confidence intervals of the thyroid dataset.

Table 5: MSE/Var(Y) on UCI Adult Semi-synthetic Setting, with 95% confidence intervals computed over multiple samples of β and m^s, m^t (described in Section 7).

	$m^s \preceq m^t$	$m^s \ ? \ m^t$
Lin. Reg. (oracle)	0.420 (0.415 – 0.424)	0.362 (0.356 – 0.367)
Lin. Reg. (source)	0.437 (0.433 – 0.442)	0.380 (0.373 – 0.386)
Lin. Reg. (imputed)	0.490 (0.471 – 0.509)	0.483 (0.475 – 0.491)
Lin. Reg. (closed-form adj.)	0.422 (0.417 – 0.426)	**0.363 (0.358 – 0.368)**
Lin. Reg. (non-param. adj.)	**0.420 (0.415 – 0.424)**	0.373 (0.367 – 0.379)
XGBoost (oracle)	0.398 (0.386 – 0.409)	0.354 (0.344 – 0.363)
XGBoost (source)	0.399 (0.387 – 0.410)	0.379 (0.369 – 0.388)
XGBoost (imputed)	0.512 (0.491 – 0.534)	0.521 (0.508 – 0.535)
XGBoost (non-param. adj.)	0.399 (0.387 – 0.410)	0.392 (0.382 – 0.402)
MLP (oracle)	0.389 (0.378 – 0.401)	0.343 (0.334 – 0.352)
MLP (source)	0.399 (0.387 – 0.410)	0.357 (0.348 – 0.367)
MLP (imputed)	0.480 (0.461 – 0.499)	0.468 (0.456 – 0.481)
MLP (non-param. adj.)	**0.389 (0.378 – 0.400)**	**0.355 (0.346 – 0.364)**

Table 6: MSE/Var(Y) on UCI Bank Semi-synthetic Setting, with 95% confidence intervals computed over multiple samples of β and m^s, m^t (described in Section 7).

	$m^s \preceq m^t$	$m^s \ ? \ m^t$
Lin. Reg. (oracle)	0.338 (0.336 – 0.340)	0.433 (0.426 – 0.440)
Lin. Reg. (source)	0.371 (0.369 – 0.373)	0.480 (0.472 – 0.487)
Lin. Reg. (imputed)	0.501 (0.491 – 0.511)	0.592 (0.583 – 0.602)
Lin. Reg. (closed-form adj.)	0.339 (0.337 – 0.340)	**0.442 (0.436 – 0.449)**
Lin. Reg. (non-param. adj.)	**0.338 (0.336 – 0.340)**	0.459 (0.453 – 0.466)
XGBoost (oracle)	0.287 (0.279 – 0.295)	0.453 (0.438 – 0.468)
XGBoost (source)	0.305 (0.297 – 0.313)	**0.500 (0.484 – 0.516)**
XGBoost (imputed)	0.492 (0.482 – 0.503)	0.708 (0.684 – 0.732)
XGBoost (non-param. adj.)	**0.287 (0.279 – 0.295)**	0.503 (0.486 – 0.519)
MLP (oracle)	0.295 (0.287 – 0.303)	0.458 (0.442 – 0.473)
MLP (source)	0.322 (0.314 – 0.330)	0.499 (0.483 – 0.516)
MLP (imputed)	0.484 (0.474 – 0.494)	0.668 (0.645 – 0.690)
MLP (non-param. adj.)	**0.294 (0.286 – 0.302)**	**0.487 (0.471 – 0.503)**

Table 7: MSE/Var(Y) on UCI Thyroid Semi-synthetic Setting, with 95% confidence intervals computed over multiple samples of β and m^s, m^t (described in Section 7).

	$m^s \preceq m^t$	$m^s \ ? \ m^t$
Lin. Reg. (oracle)	0.298 (0.292 – 0.303)	0.251 (0.246 – 0.256)
Lin. Reg. (source)	0.350 (0.342 – 0.357)	0.320 (0.314 – 0.326)
Lin. Reg. (imputed)	0.306 (0.298 – 0.313)	0.358 (0.351 – 0.365)
Lin. Reg. (closed-form adj.)	0.316 (0.310 – 0.322)	**0.291 (0.286 – 0.295)**
Lin. Reg. (non-param. adj.)	**0.293 (0.288 – 0.298)**	**0.291 (0.286 – 0.296)**
XGBoost (oracle)	0.316 (0.304 – 0.328)	0.274 (0.265 – 0.282)
XGBoost (source)	**0.310 (0.298 – 0.322)**	**0.352 (0.341 – 0.362)**
XGBoost (imputed)	0.355 (0.346 – 0.364)	0.441 (0.430 – 0.452)
XGBoost (non-param. adj.)	**0.310 (0.298 – 0.321)**	0.381 (0.370 – 0.392)
MLP (oracle)	0.279 (0.269 – 0.288)	0.230 (0.223 – 0.236)
MLP (source)	0.320 (0.308 – 0.331)	0.303 (0.294 – 0.311)
MLP (imputed)	0.304 (0.296 – 0.311)	0.345 (0.336 – 0.355)
MLP (non-param. adj.)	**0.278 (0.268 – 0.288)**	**0.272 (0.265 – 0.279)**

I.3 Real Data Experiments

The data for these experiments were derived from eICU-CRD (Pollard et al., 2018), a multi-hospital critical care database which uses the PhysioNet Credentialed Health Data License Version 1.5.0. We extract data for predicting 48-hour mortality through the FIDDLE (Tang et al., 2020) preprocessing pipeline with default parameters. FIDDLE extracts both time-varying and fixed features. We collapse the time-varying features by taking the maximum value (note that most features are binary, and none take values less than 0). We extract data from two of the hospitals with the most data, the first of which contains 3,006 data points, and the second of which contains 2,663 data points. The rate of 48-hour mortality in the first hospital is 0.097, and the rate of 48-hour mortality in the second hospital is 0.100. Additionally, we threshold for features that are present that have a prevalence of at least 5% in either of the hospitals and at least 1% in both of the hospitals. Code is provided at https://github.com/acmi-lab/Missingness-Shift. We used target unlabeled data ($\alpha_t = 1, \alpha_s = 0$) to estimate $\mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t]$ for the adjusted linear closed form model because we noticed that the estimation error with limited data made the source estimates less reliable. Due to limited positive samples, in order to evaluate cross-domain performance, a model was trained on all data from one domain and tested on all data from the other. Oracle performance (training and testing on the same domain) was computed from training on a randomly sampled 80% of the data and testing on the remaining 20%. Table 8 contains the estimated relative non-missingness of the top five coefficients for the oracle models from each hospital.

Table 8: The estimated proportion of nonzeros in Hospital 1 (q_1) and Hospital 2 (q_2), estimated relative non-missingness rates $q_2/q_1 = 1 - r^{1\to 2}$, Hospital 1 Oracle coefficient (β_1), and Hospital 2 Oracle coefficient (β_2) for each of the top five features (measure by magnitude of coefficient) from the Oracle linear predictors of Hospital 1 and 2.

	β_1	β_2	q_1	q_2	q_2/q_1
noninvasivemean_max_(78.0, 86.0]	-0.279	-0.364	0.754	0.938	1.244
systemicsystolic_mean_(-94.001, 99.667]	0.271	-0.362	0.333	0.134	0.404
unittype...Neuro ICU	0.055	-0.577	0.194	0.315	1.629
ethnicity...African American	-0.275	0.361	0.141	0.071	0.506
...Intake (ml)...(100.0, 150.0]	0.070	-0.732	0.318	0.045	0.142
...Invasive BP Systolic...(-59.001, 101.0]	-0.571	0.474	0.350	0.130	0.372
cvp_max_(8.0, 12.0]	0.536	-0.476	0.262	0.125	0.477