Density Ratio Estimation and Neyman Pearson Classification with Missing Data

Josh Givens
University of Bristol

Song Liu
University of Bristol

Henry W J Reeve
University of Bristol

Abstract

Density Ratio Estimation (DRE) is an important machine learning technique with many downstream applications. We consider the challenge of DRE with missing not at random (MNAR) data. In this setting, we show that using standard DRE methods leads to biased results while our proposal (M-KLIEP), an adaptation of the popular DRE procedure KLIEP, restores consistency. Moreover, we provide finite sample estimation error bounds for M-KLIEP, which demonstrate minimax optimality with respect to both sample size and worst-case missingness. We then adapt an important downstream application of DRE, Neyman-Pearson (NP) classification, to this MNAR setting. Our procedure both controls Type I error and achieves high power, with high probability. Finally, we demonstrate promising empirical performance both synthetic data and real-world data with simulated missingness.

1 INTRODUCTION

Density Ratio estimation (DRE) is the problem of estimating the ratio between two probability density functions (PDFs). DRE's ability to characterise the relationship between two PDFs naturally lends itself to many applications such as outlier detection (Azmandian et al., 2012), Generative Adversarial Networks (GANs) (Nowozin et al., 2016), and general binary classification. One issue with DRE is that it is especially sensitive to missing data due to the large impact a "few key points" can have on the procedure (Liu et al., 2017). While there are a vast number of DRE approaches (Sugiyama et al., 2008, 2010; Kanamori et al., 2009; Choi et al., 2021), none of them explicitly account for the case of missing data. Some work has been done regarding the impact of missing data on DRE (Josse et al., 2019), and while this work does explore a wide variety of missing data methods it does so while assuming the data to be missing at random (MAR). There are many applications

in which such an assumption is unrealistic and the probability of an observation being missing depends in some way on the value of the observation itself leading to missing not at random data (MNAR). For example, many measuring instruments are more likely to err when attempting to measure more extreme values, while in questionnaires, participants are less likely to answer a question if they deem their answer to be embarrassing or unfavourable. In such a case, naively applying any classic DRE procedure leads to inconsistent estimations.

Classification is an area of statistic with strong ties to density ratio estimation and is also an area where missing covariates are a common issue. In this paper, we focus on the links of DRE to a particular type of classification problem called Neyman Pearson (NP) classification. NP classification allows the user to construct a classifier with a fixed Type I error rate with high probability (Tong, 2013; Tong et al., 2018). This type of classification has many natural applications as there are often cases where missclassification in one direction is far more damaging than in the other (for example, disease diagnosis or fault detection). We would like to choose our classifier to control the error in said direction. Taking missing data into consideration during NP classification is vital. Without it, our classifier may no longer satisfy our Type I error constraint with high probability. Classic NP classification works in two steps. First, we estimate the class density ratio between the two classes (hence motivating the use of DRE), then we define the classification boundary of our classifier to be a level set of our density ratio estimate, which leads to the satisfaction of our Type I error constraint with high probability (Tong, 2013). To our knowledge, no previous work has been done to make this second stage of NP classification robust to MNAR data.

There is a large body of research into the phenomenon of MNAR data in more general cases (Sportisse et al., 2020; Liu and De, 2015; Li et al., 2013; Seaman and White, 2011), however none of this work has explored its use within DRE specifically. Work exists on MNAR data in the covariates of logistic regression (Ward et al., 2020), a classification method with close ties to DRE. However, Ward et al.'s approach requires more direct knowledge of the missing data distribution rather than just the probability

of an observation being missing conditional on its value.

Our main contributions are:

- We adapt KLIEP (Sugiyama et al., 2008), a classic DRE procedure, to the MNAR setting using an inverse probability weighting approach before expanding this technique to any DRE whose objectives comprise expectations of two different data sets. We give finite sample convergence results for our method and show minimax rate optimality of this convergence when the conditional probability of being missing is known.

- We provide an adaptation to the NP classification procedure, allowing it to control Type I error even with MNAR data and show finite sample power convergence results for our adaptation when the conditional probability of being missing is known.

We also provide extensions to our DRE approaches, which allow us to deal with both partial missingness across multiple dimensions and learning the missing pattern by querying a subset of the missing samples. Finally, we assess and compare the efficacy of our proposed methods by testing them on various simulated and real-world data with synthetic missingness.

2 MISSING DATA & DENSITY RATIO ESTIMATION

2.1 Motivation

Density Ratio Estimation (DRE) is a versatile tool with many downstream applications such as binary classification (including NP classification, which we will discuss in Section 3.4), GANs (Nowozin et al., 2016; Ding et al., 2020), and covariate shift (Sugiyama et al., 2008; Tsuboi et al., 2009). In DRE problems, we want to estimate $r^* := p_1/p_0$, with p_1, p_0 the PDFs for two distributions. DRE approaches estimate r^* using IID samples from these two distributions. Existing DRE approaches are designed to handle fully observed data sets. However, when observations are missing, these approaches do not apply straightforwardly.

DRE requires two sets of observations. When two samples are collected at different times under different contexts, they are especially susceptible to different missing patterns. An example is binary classification for diagnosing an illness. In this setting, it is natural to expect that samples collected from healthy individuals and patients will have different MNAR patterns due to privacy concerns.

Within this setting, using both data sets by simply discarding all missing values will lead to a biased classification and degraded performance.

We aim to make DRE robust to Missing Not at Random (MNAR) datasets. In this setting, an observation's probability of being missing depends upon that observation's

true value and the MNAR pattern is different between two samples. We formally introduce our problem set-up below.

2.2 Missing Data Framework

Let Z^1, Z^0 be two RVs taking values in measurable space $(\mathcal{Z}, \mathcal{B}_\mathcal{Z})$ with densities (Radon Nikodym derivatives) p_1 and p_0 respectively w.r.t. (with respect to) a measure μ on $(\mathcal{Z}, \mathcal{B}_\mathcal{Z})$ and assume that p_0 is strictly positive. For $\omega \in \{0, 1\}$, let X^ω be a RV taking values in the measurable space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ with $\mathcal{X} := \mathcal{Z} \cup \{\varnothing\}$, $\mathcal{B}_\mathcal{X} := \sigma(\mathcal{B}_\mathcal{Z} \cup \{\{\varnothing\}\})$ whose distribution is uniquely defined as follows:

$$\mathbb{P}(X^\omega = \varnothing | Z^\omega) := \varphi^\omega(Z)$$
$$\mathbb{P}(X^\omega = x | Z^\omega) := \mathbb{1}\{Z^\omega = x\}(1 - \varphi^\omega(Z^\omega)),$$

where $\varphi^\omega : \mathcal{Z} \to [0, 1)$ is a measurable function. Here we take $X = \varnothing$ to represent the observation being missing and thus $\varphi(z)$ represents the probability of an observation being missing given its "true" value is z. Additionally, let p'_ω be the density of $X^\omega | X^\omega \neq \varnothing$ w.r.t μ. Throughout, unless stated otherwise, we assume φ to be known.

For $\omega \in \{0, 1\}$ and $n_\omega \in \mathbb{N}$ we define $\{Z_i^\omega\}_{i=1}^{n_\omega}$ and $\{X_i^\omega\}_{i=1}^{n_\omega}$ to be IID copies of Z^ω, X^ω respectively and abbreviate them by $D_\omega := \{Z_i^\omega\}_{i=1}^{n_\omega}$, $D'_\omega := \{X_i^\omega\}_{i=1}^{n_\omega}$. Finally we define, $D := (D_1, D_0)$, $D' := (D'_1, D'_0)$.

2.3 Notations

Throughout we shall adopt the following two conventions for the purpose of brevity. Firstly, if X, Z, or any other class-specific element is given without a sub/superscript specifying the class, the associated statement is assumed to hold for both classes where all the elements within the statement have the same class. Secondly, given a function $h : \mathcal{Z} \to \mathbb{R}$ we implicitly extend h to a function $h : \mathcal{X} \to \mathbb{R}$ by taking $h(\varnothing) := 0$, unless stated otherwise.

We now introduce some additional notation. For $n \in \mathbb{N}$ we let $[n] := \{1, \ldots, n\}$. For a function $h : \mathcal{Z} \to \mathbb{R}$ we take $\|h\|_\infty := \sup_{z \in \mathcal{Z}} |h(z)|$. Finally, for a RV W in a product space $\mathcal{W}^d$ and any $j \in [d]$ define $W^{(j)}$ to be the jth coordinate of W and $W^{(-j)} = \{W^{(j')}\}_{j' \neq j}$. The Kullback-Leibler (KL) between two density functions with p_1 absolutely continuous w.r.t. p_0 is defined as

$$\mathrm{KL}(p_1 | p_0) := \int_\mathcal{Z} p_1(z) \log \frac{p_1(z)}{p_0(z)} \mu(\mathrm{d}z).$$

2.4 DRE

The aim of DRE is to estimate $r^* : \mathcal{Z} \to [0, \infty)$ defined by $r^* := p_1/p_0$. We will let $\mathcal{H}$ be the set of non-negative measurable functions from $\mathcal{Z}$ to $[0, \infty)$ and let $\mathcal{G}$ be some subset of $\mathcal{H}$ which we intend to use in our approximation of r^*. Note that while $r^* \in \mathcal{H}$, $\mathcal{G}$ may or may not contain the "true" density ratio r^*. We say that $\mathcal{G}$ is correctly specified if $r^* \in \mathcal{G}$ and incorrectly specified otherwise. We now introduce a classic approach for estimating r^* using D.

2.4.1 Kullback-Leibler Importance Estimation Procedure

First proposed in (Sugiyama et al., 2008), Kullback-Leibler Importance Estimation Procedure (KLIEP) is a popular DRE procedure that minimises the KL divergence between p_1 and $r \cdot p_0$. Specifically, it aims to solve the following optimisation problem:

$$\tilde{r} := \arg\min_{r \in \mathcal{G}} \mathrm{KL}(p_1 | r \cdot p_0) \tag{1}$$

$$\text{subject to} : \int_{\mathcal{Z}} r(z) p_0(z) \mu(\mathrm{d}z) = 1.$$

As KL-divergence is strictly for use with probability densities, we need to include the additional constraint that $r \cdot p_0$ integrates out to 1 over $\mathcal{Z}$.

Provided $\mathcal{G}$ is closed under positive scalar multiplication, we can re-write $\tilde{r}$ in terms of an unconstrained optimisation problem. Specifically, we have $\tilde{r} := N^{-1} \cdot r_0$ where

$$r_0 := \underset{r \in \bar{\mathcal{G}}}{\arg\max} \, \mathbb{E}[\log r(Z^1)] - \log \mathbb{E}[r(Z^0)] \tag{2}$$

$$N := \mathbb{E}[r_0(Z^0)].$$

for some select $\bar{\mathcal{G}} \subseteq \mathcal{G} = \{a \cdot g : a \in (0, \infty), \, g \in \bar{\mathcal{G}}\}$. More details on this alongside the lemma that makes this possible can be found in the Appendix A.1. Throughout we will be mostly interested in estimating r_0 and as such will refer to 2 as the population KLIEP objective. As we do not know the true distributions of Z^1, Z^0, we must approximate these expectations with samples from corresponding distributions giving us the sample KLIEP objective.

In the case of multi-dimensional real-valued data, $\mathcal{Z} = \mathbb{R}^p$, we can take $\bar{\mathcal{G}} = \{r_\theta : \mathbb{R}^p \to [0, \infty) | \theta \in \mathbb{R}^d\}$ where

$$r_\theta(z) := \exp(\theta^\top f(z)) \tag{3}$$

for some $f : \mathbb{R}^p \to \mathbb{R}^d$ (Tsuboi et al., 2009; Kanamori et al., 2010; Liu et al., 2017). We refer to this form as the *log-linear* form. Under this form, our solution becomes $\hat{r} = \hat{N}^{-1} r_{\hat{\theta}}$ where

$$\hat{\theta} := \underset{\theta \in \mathbb{R}^d}{\arg\max} \, \frac{1}{n_1} \sum_{i=1}^{n_1} \theta^\top f(Z_i^1) - \log \frac{1}{n_0} \sum_{i=1}^{n_0} \exp(\theta^\top f(Z_i^0)) \tag{4}$$

$$\hat{N} := \frac{1}{n_0} \sum_{i=1}^{n_0} \exp(\theta^\top f(Z_i^0))$$

which gives us a convex optimisation problem.

We now define $\tilde{\theta}$ as the minimiser of the population KLIEP objective under our log-linear model (i.e. the population analogue of (4)) allowing us to view $\hat{\theta}$ as an estimator for $\tilde{\theta}$. Indeed, with some mild restrictions on f and for $n_{\min} := \min\{n_0, n_1\} \geq C_0 \log(1/\delta)$, we have that w.p. $1 - \delta$

$$\|\hat{\theta} - \tilde{\theta}\| \leq \sqrt{\frac{C_0 \log(1/\delta)}{n_{\min}}} \tag{5}$$

where C_0 is a constant depending on $f(Z_0), f(Z_1), \tilde{\theta}, d$. Details of this result are found in Appendix A.2.

An additional bound for the accuracy of $\hat{N}$ is given in the Appendix section A.2.1. However, as we will see in applications such as NP classification, we only require the knowledge of r^* up to strictly increasing transform and so it is often unnecessary to estimate N.

Within this paper, we mostly focus on KLIEP however for completeness we also present the general class of f-Divergence based density ratio estimators and describe some notable cases in Appendix B.1. We now go on to adapt this procedure for use with MNAR data.

3 PROPOSED METHODS

In our setting we do not observe samples from Z and so cannot use (4) to approximate the population KLIEP objective. Instead, we only have samples from X. The following result relates expectations of Z with expectations of X.

Lemma 1. *We have that $p' = C \cdot (1 - \varphi) \cdot p$, for some $C \in \mathbb{R}$. Hence, for any measurable function $g : \mathcal{Z} \to \mathbb{R}$*

$$\mathbb{E}[g(Z)] = \mathbb{E}\left[\frac{\mathbb{1}\{X \neq \varnothing\}}{1 - \varphi(X)} g(X)\right]. \tag{6}$$

For proof of this Lemma see Appendix A.3). This importance weighting technique is an approach that is already used to tackle MNAR data contexts outside of DRE (Li et al., 2013; Seaman and White, 2011). From this Lemma, we get that for some $C' \in \mathbb{R}$,

$$r' := \frac{p_1'}{p_0'} = C' \cdot \frac{1 - \varphi^1}{1 - \varphi^0} \cdot r^*$$

Consequently, r' is not proportional to r^* when $\varphi^1 \neq \varphi^0$ with either non-constant. As a result, using KLIEP with only the observed values of our MNAR data will be biased due to the fact it estimates r' rather than r^*. We now use this Lemma to inform our adapted estimation procedures.

3.1 KLIEP with Missing Data

We can now approximate the population KLIEP objective by replacing expectations in (2) with the important weighted expectations in Lemma 1. This replacement modifies the sample KLIEP objective to work with D' as follows:

$$\hat{r}_0 := \underset{r \in \bar{\mathcal{G}}}{\arg\max} \, \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{\mathbb{1}\{X_i^1 \neq \varnothing\}}{1 - \varphi^1(X_i^1)} \log r(X_i^1)$$

$$- \log \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{\mathbb{1}\{X_i^0 \neq \varnothing\}}{1 - \varphi^0(X_i^0)} r(X_i^0) \tag{7}$$

$$\hat{N} := \frac{1}{n_0} \sum_{i=1}^{n} \frac{\mathbb{1}\{X_i^0 \neq \varnothing\}}{1 - \varphi^0(X_i^0)} \hat{r}_0(X_i^0).$$

We refer to this approach as *M-KLIEP* and to (7) as the *sample M-KLIEP objective*. We now take, $\hat{\theta}'$ to be the estimated parameter when r takes the log-linear form (3). To show the efficacy of M-KLIEP, we prove that $\hat{\theta}'$, converges to $\tilde{\theta}$ under mild conditions. Let $m_{\text{eff}} \equiv m_{\text{eff}}(n_0, n_1, \varphi^0, \varphi^1)$ denote the *effective sample size* defined by

$$m_{\text{eff}} := \min\{n_0 \cdot (1 - \|\varphi^0\|_\infty), n_1 \cdot (1 - \|\varphi^1\|_\infty)\}.$$

Theorem 2. *Let $\hat{\theta}'$ be defined as above. Assume that $\|f\|_\infty \leq \infty$ and $\sigma_{\min} := \text{Var}(f(Z^0)) > 0$. Then, there exists some constant $C_0' \geq 1$, depending only on $\|f\|_\infty$, $\sigma_{\min}$, $\|\tilde{\theta}\|$, and d, such that for any $\delta \in (0, 1/2]$ and $m_{\text{eff}}(n_0, n_1, \varphi^0, \varphi^1) > C_0' \cdot \log(1/\delta)$ we have*

$$\mathbb{P}\left(\|\hat{\theta} - \tilde{\theta}\| > \sqrt{\frac{C_0' \log(1/\delta)}{m_{\text{eff}}(n_0, n_1, \varphi^0, \varphi^1)}}\right) \leq \delta.$$

Proof is given in Appendix A.4. This result shows that M-KLIEP, unlike applying KLIEP on only observed values, does not suffer from the inconsistent ratio estimation problem we mentioned earlier.

The following result shows that this bound is minimax optimal w.r.t m_{eff}.

Theorem 3. *Given any estimator $\hat{\theta} \equiv \hat{\theta}(D')$ of $\tilde{\theta}$, and any $(w_0, w_1) \in [0, 1]^2$, $(n_0, n_1) \in \mathbb{N}^2$ and $\delta \in (0, 1/4]$ there exists distributions P_0, P_1 on $\mathcal{X}$ which satisfies the conditions of Theorem 2, with $\|\varphi^\omega\|_\infty \leq w_\omega$ such that if $D'_\omega \sim P_\omega^{n_\omega}$ for $\omega \in \{0, 1\}$ then*

$$\mathbb{P}\left(\|\hat{\theta} - \theta\| > \frac{1}{2} \wedge \sqrt{\frac{10 \log(1/(4\delta))}{m_{\text{eff}}(n_0, n_1, \varphi^0, \varphi^1)}}\right) \geq \delta.$$

3.2 General Extension

While we have described the adaptation for KLIEP, this approach can be applied to any expectation-based DRE procedure with the following form:

$$\underset{r \in \mathcal{G}}{\arg\min}\, h(\mathbb{E}[g_1(Z^1)], \mathbb{E}[g_0(Z^0)])$$

where $h : \mathbb{R}^{d_0} \times \mathbb{R}^{d_1} \to \mathbb{R}$, $g_\omega : \mathcal{Z} \to \mathbb{R}^{d_\omega}$. We can then approximate these expectations using Lemma 1 as we did for KLIEP. An example of this with f-Divergence based DRE is given in the Appendix section B.1.1.

3.3 Extensions to Partial Missingness across Multiple Dimensions

We now extend our approaches to the case of partial missingness in multi-dimensional settings. Throughout this section let $d \in \mathbb{N}$ and replace our original space $\mathcal{Z}$ for $\mathcal{Z}^d$ (and

similarly $\mathcal{X}$ with $\mathcal{X}^d$). We start off with the assumption on our data missingness structure that for all $j \in [d]$,

$$\mathbb{P}(X^{(j)} = \varnothing | Z, X^{(-j)}) = \varphi_j(Z^{(j)})$$
$$\mathbb{P}(X^{(j)} = x^{(j)} | Z, X^{(-j)}) = \mathbb{1}\{Z^{(j)} = x^{(j)}\}(1 - \varphi_j(Z^{(j)}))$$

As such, the components are missing independently from one another with probabilities only depending on their own true value. To further simplify proceedings, we assume a naive Bayes style condition. Namely we include the restriction that for all $j' \neq j$, $Z^{(j')} \perp Z^{(j)}$ so that $(Z^{(j)}, X^{(j)}) \perp (Z^{(j')}, X^{(j')})$. This implies that

$$r^* := \prod_{j=1}^d r_j^* \qquad \text{where} \qquad r_j^* := \frac{p_1^{(j)}}{p_0^{(j)}}$$

with $p^{(j)}$ be the p.d.f. of $Z^{(j)}$.

This allows us to separately estimate the density ratio over each dimension and then take the product to get the joint density ratio (DR). Within each dimension, only data from that dimension is relevant (due to the aforementioned independence) and so we can estimate the density ratio on each dimension using exclusively the data from that dimension with our current methods.

3.4 NP Classification with Missing data

We now apply our adapted procedures to the problem of NP classification. First, we introduce NP classification.

NP classification constructs a classifier that strictly controls the miss-classification of one class while minimising miss-classification in the other (Cannon et al., 2002; Scott and Nowak, 2005; Tong, 2013; Tong et al., 2018). Specifically, in NP classification we aim to learn a classifier $\phi : \mathcal{Z} \to \{0, 1\}$ which solves the following constrained optimisation problem

$$\min_\phi \mathbb{P}(\phi(Z^1) = 0)$$
$$\text{subject to: } \mathbb{P}(\phi(Z^0) = 1) \leq \alpha. \qquad (8)$$

We refer to the classifier which solves the above problem as the *NP oracle classifier* at level α and refer to class 0 as the *error controlled class*. From the Neyman-Pearson lemma, it can be shown that the classification boundary of the oracle classifier is a level set of r^* (Tong, 2013). This motivates the use of DRE to approximate r^* and in turn the oracle classifier. In contrast to hypothesis testing, we do not know the distribution of Z^0. As such, there is no means of constructing a classifier that is guaranteed to satisfy (8). Instead, we create a classifier $\hat{\phi}_D$ from our data D. This classifier has a small pre-specified probability of violating our Type I error constraint. In other words for a given small $\delta > 0$, and classification procedure $\hat{\phi}_D$,

$$\mathbb{P}\left(\mathbb{P}\left(\hat{\phi}_D(Z^0) = 1 \middle| D\right) \leq \alpha\right) \geq 1 - \delta. \qquad (9)$$

A procedure satisfying (9) is described in Tong et al. (2018). This procedure uses any density ratio estimate alongside additional data from the error-controlled class to choose a classification threshold. The estimated density ratio alongside the threshold is used for classification. As a result, different DRE approaches lead to different NP classification procedures. More information on the details of this classification procedure can be found in Appendix B.2.

As we know the optimal classifier comes from the true density ratio, we can assess the accuracy of various DRE procedures by the efficacy of the associated NP classification procedure. We asses the efficacy of an NP classification procedure $\hat{\phi}_D$ by the expected power which is defined as

$$\mathbb{E}\left[\mathbb{P}\left(\hat{\phi}_D(Z^1) = 1 \middle| D\right)\right] = \mathbb{P}(\hat{\phi}_D(Z^1) = 1).$$

We can also assess the efficacy of a classifier by it's Type II error which we define to be $R_1(\phi) := \mathbb{P}(\phi(Z^0) = 1)$.

3.4.1 Adapting NP Classification to Missing Data

Current NP classification approaches require direct samples from Z^0 in order to select the appropriate classification threshold. Therefore, we will need to adapt it for the MNAR setting.

We can again use Lemma 1 to approximate $\mathbb{P}(\hat{r}(Z^0) > C)$ using D'_0 and choose a threshold based on this. Specifically, if for any $h : \mathcal{Z} \to \mathbb{R}$ we define

$$W_i^{(C,h)} := \frac{\mathbb{1}\{X_i^0 \neq \varnothing\}}{1 - \varphi^0(X_i^0)} \mathbb{1}\{h(X_i^0) > C\}$$

then $\mathbb{E}[W_i^{(C,h)}] = \mathbb{P}(h(Z^0) > C)$. We also introduce the effective sample size for class 0 defined by $m_{\text{eff}}^0 := n_0(1 - \|\varphi^0\|_\infty)$. We now use this result to prove the key lemma that informs our adaptation to the threshold selection within our NP classification algorithm.

Lemma 4. *For a given measurable $h : \mathcal{Z} \to \mathbb{R}$ let's choose*

$$\hat{C}_{\alpha,\delta,h} \equiv \hat{C}_{\alpha,\delta,h}(D'_0)$$

$$:= \inf\left\{ C \in \mathbb{R} : \frac{1}{n}\sum_{i\in[n]} W_i^{(C,h)} \leq \alpha - \Delta_{m_{\text{eff}}^0,\delta} \right\}$$

where $\Delta_{m_{\text{eff}}^0,\delta} := \sqrt{\dfrac{16\log(1/\delta)}{m_{\text{eff}}^0}}.$

Then, for any $\delta \in (0, 1/2]$,

$$\mathbb{P}\left[\mathbb{P}(h(Z^0) > \hat{C}_{\alpha,\delta,h} \mid D') > \alpha\right] \leq \delta.$$

Proof. Fix $C_{\alpha,h}^* := \inf\left\{C \in \mathbb{R} : \mathbb{P}(h(Z^0) > C) \leq \alpha\right\}$. As $\mathbb{P}(h(Z^0) > C)$ is a decreasing right continuous func-

tion of C we have that $\mathbb{P}(h(Z^0) > C_{\alpha,h}^*) \leq \alpha$. Hence,

$$\mathbb{P}\left[\mathbb{P}(h(Z^0) > \hat{C}_{\alpha,\delta,h} \mid D'_0) > \alpha\right]$$

$$= \mathbb{P}(\hat{C}_{\alpha,\delta,h} < C_{\alpha,h}^*)$$

$$\leq \mathbb{P}\left(\frac{1}{n}\sum_{i\in[n]} W_i^{(C_{\alpha,h}^*)} \leq \alpha - \Delta_{m_{\text{eff}}^0,\delta}\right) \leq \delta,$$

where we used Lemma 11 from Appendix A.4 in the final step. $\square$

Therefore this gives rise to the following NP classification Algorithm 1. We will assume that we have access to an additional n_0 IID copies of X_0 which we will label as $X_{n_0+1}^0, \ldots, X_{2n_0}^0$ however this is just for notational convenience and the algorithm and associated theory adapts to any number of samples from X_0. We now let $g : \mathbb{R} \to \mathbb{R}$ be any strictly increasing function.

Algorithm 1 Missing NP Classification Procedure

1: Use $\{X_i^1\}_{i=1}^{n_1}, \{X_i^0\}_{i=1}^{n_0}$ to estimate $g \circ r^*$ with $g \circ \hat{r}$ by any DRE procedure.

2: For $i \in \{1, \ldots, n_0\}$ compute $\hat{r}_i := g \circ \hat{r}(X_{i+n_0}^0)$ and $w_i := \mathbb{1}\{X_{i+n_0}^0 \in \mathcal{Z}\}(1 - \varphi^0(X_{i+n_0}^0))^{-1}$ with $g \circ \hat{r}(\varnothing) := -\infty$.

3: Sort $\hat{r}_1, \ldots, \hat{r}_{n_0}$ in increasing order to get $\hat{r}_{(1)}, \ldots, \hat{r}_{(n_0)}$ with $\hat{r}_{(i)} \leq \hat{r}_{(i+1)}$ and associated $w_{(1)}, \ldots, w_{(n_0)}$.

4: Set $i^* = \min\{i \in \{1, \ldots, n_0\} | \frac{1}{n_0}\sum_{j=i}^{n_0} w_{(j)} \leq \alpha - \Delta_{m_{\text{eff}}^0,\delta}\}$.

5: Let $\hat{C}_{\alpha,\delta,g\circ\hat{r}}(D') := \hat{r}_{(i^*)}$ and define $\hat{\phi}_{D'}$ by

$$\hat{\phi}_{D'}(z) := \mathbb{1}\{g \circ \hat{r}(z) > \hat{C}_{\alpha,\delta,g\circ\hat{r}}\} \quad \text{for all } z \in \mathcal{Z}.$$

We can freely introduce g within Algorithm 1 as the classifier produced is invariant to strictly increasing transformations. Crucially we can do this without knowing g specifically and are only required to know $g \circ \hat{r}$. This added flexibility has many advantages, for example, we are now only required to learn r^* up to a multiplicative constant.

We now analyse the performance of Algorithm 1 in conjunction with M-KLIEP. To this end, define $\hat{\phi}_{D'}$ as in Algorithm 1 with $\hat{r} = r_{\hat{\theta}'}$ as defined in Theorem 2. We now compare this with its population analogue $\tilde{\phi}$ defined by

$$\tilde{\phi}(z) := \mathbb{1}\{g \circ r_{\tilde{\theta}}(z) > C_{\alpha,g\circ r_{\tilde{\theta}}}^*\}$$

for all $z \in \mathcal{Z}$, where $C_{\alpha,h}^*$ is defined as in the proof of Lemma 4.

Theorem 5. *Assume the conditions of Theorem 2 hold and let $\hat{\phi}_{D'}, \tilde{\phi}$ be defined as above. Further assume that for all*

$\theta \in \mathbb{R}^d, z \in \mathcal{Z}, \|\nabla_\theta g \circ r_\theta(z)\| \leq L$. *Finally, suppose there exists positive constants* $B_0, B_1, \gamma_0, \gamma_1, a$ *s.t.* h *satisfies the following conditions*

$$\mathbb{P}\left(h(Z^0) \in [C_{\alpha,h}^*, C_{\alpha,h}^* + \zeta]\right) \geq B_0^{-1}\zeta^{\gamma_0}$$

$$\mathbb{P}\left(h(Z^0) \in [C_{\alpha,h}^* - \zeta, C_{\alpha,h}^* + \zeta]\right) \leq B_1\zeta^{\gamma_1},$$

for all $\zeta \in (0, \ B_0 a^{1/\gamma_0}]$. *Then there exists a constant* $C_2 > 0$ *depending only on* $C_0', B_0, B_1, \gamma_0, \gamma_1, a$, *and* L *such that for all* $\delta \in (0, \frac{1}{2}]$, *with* $m_{\text{eff}} \geq C_2 \log(1/\delta)$, *we have*

$$\mathbb{P}(|R_1(\hat{\phi}_{D'}) - R_1(\tilde{\phi})| > \varepsilon') \leq 2\delta.$$

where

$$\varepsilon' := C_2 \left[\left(\frac{\log(1/\delta)}{m_{\text{eff}}^0}\right)^{\frac{\gamma_1+1}{2\gamma_0} \wedge \frac{1}{2}} + \left(\frac{\log(1/\delta)}{m_{\text{eff}}}\right)^{\frac{\gamma_1+1}{2}} \right].$$

Remark 1. *If* $\mathcal{G}$ *is correctly specified and* Z^1, Z^0 *are continuous then the NP lemma gives us that* $\tilde{\phi}$ *is the oracle classifier (Tong, 2013). Hence we have shown convergence in Type II error to the oracle classifier in this case.*

Remark 2. *Taking* $g = \log$ *gives* $L = b$ *under the constraints imposed in Theorem 2.*

We have now fully adapted NP classification to the MNAR setting. Namely, we have proposed an algorithm to construct an NP classifier using MNAR data and shown this classifier simultaneously control Type I error with high probability while also converging to the oracle classifier when $\mathcal{G}$ is correctly specified. We test our adaptations on some synthetic examples.

4 SYNTHETIC DATA EXAMPLES

Here we empirically evaluate our proposed methods alongside CC-KLIEP (naive **C**omplete **C**ase **KLIEP**) which simply discards any missing value in the dataset and estimating the ratio using only observed data. Both M-KLIEP and CC-KLIEP are tested on simulated datasets. Details on the data generating processes can be found in Appendix C.

4.1 5-dimensional Correctly Specified Example

In this example we take both classes to have multivariate Gaussian distributions with the same variance. We then induced MNAR missingness in class 0 and no missingness in class 1. We use the log-linear form for r_θ with $f(z) = z$ so that $\mathcal{G}$ is correctly specified. 100 simulations of the above data generation procedure were run for $n_0 = n_1$ with n_0 ranging from 100 to 1,500 and $\tilde{\theta}$ estimated by M-KLIEP and CC-KLIEP. These simulations were then used to estimate the mean square (Euclidean) distance (MSD) between our estimate, $\hat{\theta}'$, and $\tilde{\theta}$ for both procedures. These estimates alongside 99% C.I.s (confidence intervals) were calculated
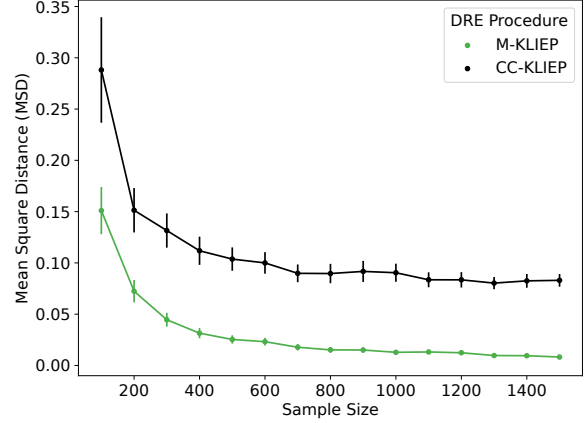

Figure 1: MSD between $\hat{\theta}'$ and $\tilde{\theta}$ for varying n with 99% C.I.s.

which are presented in Figure 1. In the plot we can see clear evidence of the complete case approach being asymptotically biased as the error plateaus around 0.08 while the error under M-KLIEP converges to 0 as n_0, n_1 increase. We now go on to illustrate the affect of our DRE procedure on NP classification.

4.2 Neyman Pearson Classification

For this example we take both classes to be 2-dimensional Gaussian mixtures. We then induce MNAR missingness in class 0 and no missingness in class 1. M-KLIEP and CC-KLIEP are then used to estimate r^* using the log-linear form with $f(x) = x$ again. NP classifiers were then fit using these estimators alongside the true r^*.
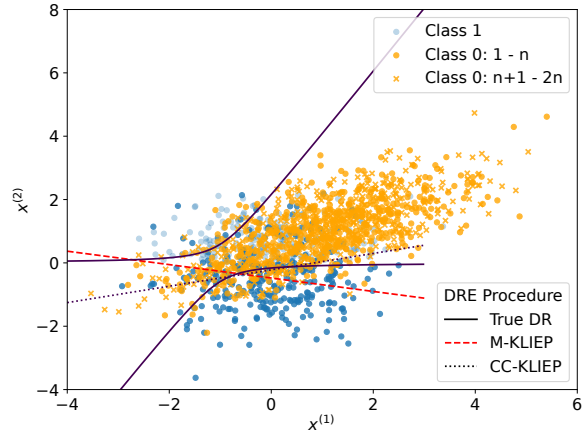

Figure 2: Scatter plot of D alongside classification boundaries produced from corresponding D' via various procedures. If $X_i = \varnothing$ then Z_i is faded out.

Figure 2 shows one run of this experiment with $n_0 = n_1 = 500$. We see that M-KLIEP provides a good approximation to the classifier that uses r^* despite the model being incorrectly specified while CC-KLIEP is biased due to the way
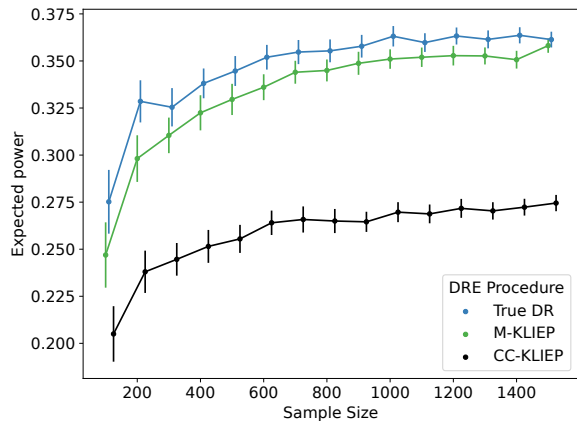
Figure 3: Expected Power for varying n with 99% C.I.s

it discards corrupted points.

Repeated Simulations

We now run 100 simulations of the above experiment for various $n_1 = n_0$ ranging from 100 to 1500. We then use these to estimate the expected power of the procedures alongside 99% C.I.s . Figure 3 shows the results of this experiment.

We see similar results to our correctly specified DRE experiment with the complete case approach performing poorly. As we can see, despite our model not being able to encompass the true density ratio, our estimate density ratio via M-KLIEP still gives a reasonable approximation to r^* and therefore gives a good NP classification procedure. Due to $\mathcal{G}$ being incorrectly specified, M-KLIEP never converges to the expected power using the true DR. Additional synthetic experiments are given in Appendix D

5 REAL-WORLD DATASETS WITH SIMULATED MISSINGNESS

We now go on to perform NP classification experiments with real-world data. To make this scenario more realistic we will assume the φ to be unknown and aim to estimate them. We briefly tackle how to do this in the section below.

5.1 Learning the Missingness Function

Thus far we have assumed φ is known. We now consider the challenge of learning φ by querying the latent values Z_i for a subset of our missing data.

Here we are motivated by settings in which we can send a small sub-sample of observations which erred (i.e. a subset of $\{i \in [n] : X_i = \varnothing\}$) off for further investigation to obtain their respective true values Z_i. Now we can learn the missingness function φ by fitting a logistic regression with samples of $(Z_i, \mathbb{1}\{X_i = \varnothing\})$ and treating $\mathbb{1}\{X_i = \varnothing\}$ as the response variable.

As we have not observed Z_i for every $i \in [n]$, we cannot use a standard logistic regression to learn φ. However, we can use an adaption of logistic regression designed to deal with miss-representative class proportions presented in King and Zeng (2001). More details on this can be found in the Appendix in section B.3. We now try our methods on real-world data to assess their efficacy.

5.2 Set-Up

We choose datasets and classification problems where the NP classification is properly motivated: One would want to strictly control miss-classification for one class in these datasets. We have chosen 3 datasets which we refer to as "Fire", "CTG", and "weather";

- **CTG dataset**: This data set contains 11 different summaries of Cardiotocography (CTG) data for 2126 foetuses where we take each foetus to be an observation. Alongside this is a classification of each foetus as "Healthy", "Suspect", or "Pathologic". We aim to predict whether the foetus is classified as "Healthy" or not. We take the error controlled class to be "Suspect" or "Pathologic".

- **Fire dataset**: This data set contains 62,630 readings of 12 atmospheric measures such as temperature, humidity, and CO2 content. We aim to predict whether a fire is present. We take the presence of fire as the error controlled class.

- **Weather data**: This data set contains 142,193 observations of 62 dimensions giving various weather readings for a given day in a given location in Australia and whether it rained the following day. We aim to predict whether it rains the following day. We take the occurrence of rain as the error controlled class.

In all cases, we artificially induce missing observations separately across each dimension and only in the non error-controlled class. We construct each dimension-wise missing function $\varphi_j, j \in [d]$ as $\varphi_j(z) = (1 + \exp\{\tau_j(a_{j,0} + a_{j,1}z)\})$ where $a_{j,0}, a_{j,1} \in [0, \infty)$ and $\tau_j \in \{-1, 1\}$.

We constructed NP classifiers using M-KLIEP and CC-KLIEP under our naive Bayes framework introduced in Section 3.3. For M-KLIEP we test both cases where φ_j is known and φ_j is learned. Estimating φ_j is done using the method described in Section 5.1 and each φ_j is estimated separately using only data from that dimension. We also estimate the density ratio using the fully observed data via standard KLIEP as a benchmark.

The power of the NP classifiers produced with these DRE procedures was then calculated on fully observed testing data. We repeat this process multiple times with new random test/train splits and τ_j at each iteration to estimate the pseudo expected power and corresponding 95% C.I.s. We apply this technique at different target Type I errors and
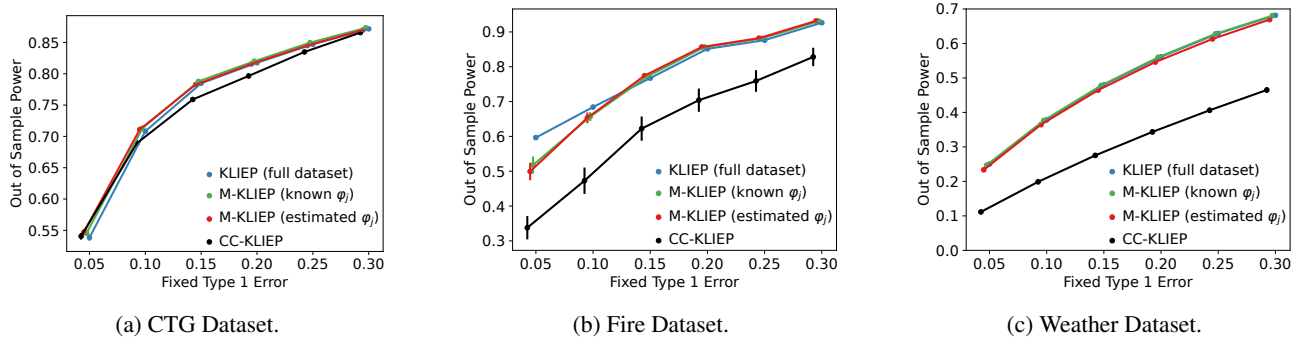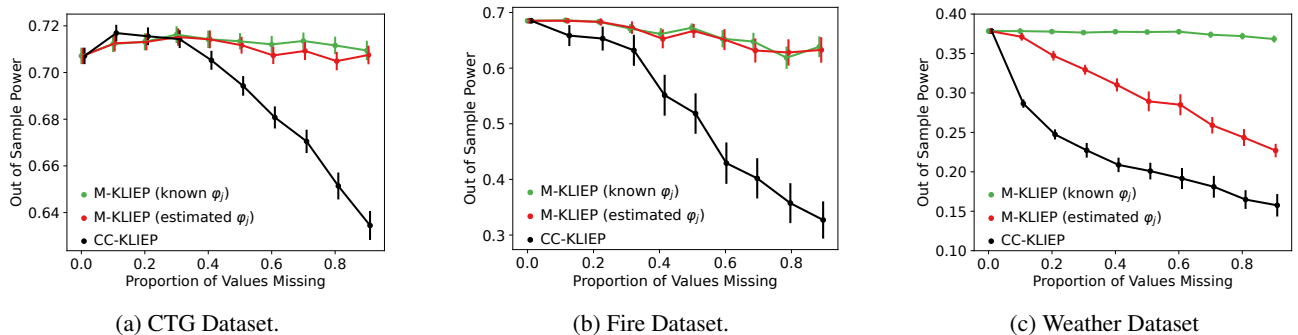
(a) CTG Dataset. (b) Fire Dataset. (c) Weather Dataset.

Figure 4: Out of sample power with pseudo 95% C.I.s for various different target Type I errors.

(a) CTG Dataset. (b) Fire Dataset. (c) Weather Dataset

Figure 5: Out of sample power with pseudo 95% C.I.s for various φ with varying missing proportions.

varying $a_{j,0}$ to construct datasets with different missing proportions. Additional detail is given in Appendix E.2.

5.3 Results

In Figure 4 we see that in both the Fire and Weather data, M-KLIEP significantly outperforms CC-KLIEP for all values of α. For the CTG data we see significant outperformance for $\alpha \in [0.1, 0.25]$ and comparable performance in the extremes. Surprisingly, we see that M-KLIEP performs equally well when φ is either learned or known. Further, M-KLIEP perform comparably with KLIEP run on fully observed datasets. In Figure 5 we see that M-KLIEP outperforms CC-KLIEP for larger missingness proportions. As we would expect, the more missing points, the worse CC-KLIEP performs. For the Fire and CTG data we see no loss in performance when learning φ. For the Weather data, an increase in the missing proportion leads to a decrease in performance of M-KLIEP using a learned φ. Remarkably, there is no performance loss when running M-KLIEP using the true φ in all datasets.

Results from additional experiments comparing with an iterative imputation approach are given in Appendix section E.3.1.

6 DISCUSSION & CONCLUSIONS

DRE is a widely used machine learning technique with a diverse range of applications. In this paper we have shown that, when data is MNAR, naively performing DRE by discarding all missing observations can lead to inaccurate estimates of the density ratio functions. We have proposed a novel procedure M-KLIEP as well as adaptations to a broad family of DRE procedures to account for the MNAR phenomenon when the missingness structure is known. For M-KLIEP we have presented finite sample bounds under a commonly used parametric form showing convergence at rate $\sqrt{m_{\text{eff}}}$ where m_{eff} is the effective sample size. We have then extended these approaches in multiple ways to partial missingness across multiple dimensions under the naive Bayes framework. Finally, we have presented a technique to estimate missing patterns from data by querying the true values of a few missing observations.

We then studied a downstream application of DRE in NP classification. We have adapted NP classification to MNAR data settings and shown that our adaptation ensures satisfaction of our Type I error constraints with high probability. We have also provided finite sample convergence results for the expected power of our NP classification procedure and hence shown that it converges to the optimal power under certain conditions.

We have shown all these adaptations to work well in practice on simulated data and applied them to real world data with known and unknown synthetic missing patterns. In particular, our adapted NP classifier using M-KLIEP has been shown to negate losses in accuracy/power incurred by MNAR phenomenon in most cases. We briefly explore the societal impact of our work in Appendix F.

A natural extension of this work would be relaxing the naive Bayes assumption required for the handling partial missingness (Section 3.3) to a more realistic set-up where dependencies are properly modelled and accounted for. This extension would enable applications in areas where the naive Bayes framework is unrealistic such as image classification. Another possible direction would be exploring the set-up where the missing pattern is known to only belong to some class as discussed in Sportisse et al. (2020). This approach could be further expanded using a Bayesian framework by assigning of some prior belief on the missing pattern. Finally, our adaptation to NP classification could be extended further by adapting it to the case of partial missingness in both classes.

Acknowledgments

References

Azmandian, F., Dy, J. G., Aslam, J. A., and Kaeli, D. R. (2012). Local kernel density ratio-based feature selection for outlier detection. *Journal of Machine Learning Research*, 25:49–64.

Cannon, A., Howse, J., Hush, D., and Scovel, C. (2002). Learning with the neyman-pearson and min-max criteria. *Los Alamos National Laboratory, Tech. Rep. LA-UR*, pages 02–2951.

Choi, K., Liao, M., and Ermon, S. (2021). Featurized density ratio estimation. In de Campos, C. and Maathuis, M. H., editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 172–182. PMLR.

Ding, X., Wang, Z. J., and Welch, W. J. (2020). Subsampling generative adversarial networks: Density ratio estimation in feature space with softplus loss. *IEEE Transactions on Signal Processing*, 68:1910–1922.

Josse, J., Prost, N., Scornet, E., and Varoquaux, G. (2019). On the consistency of supervised learning with missing values. *arXiv preprint arXiv:1902.06931*, pages 1–43.

Kanamori, T., Hido, S., and Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(48):1391–1445.

Kanamori, T., Suzuki, T., and Sugiyama, M. (2010). Theoretical analysis of density ratio estimation. *IEICE Transactions*, 93-A:787–798.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2):137–163.

Li, L., Shen, C., Li, X., and Robins, J. M. (2013). On weighting approaches for missing data. *Statistical Methods in Medical Research*, 22(1):14–30. PMID: 21705435.

Liu, S., Takeda, A., Suzuki, T., and Fukumizu, K. (2017). Trimmed density ratio estimation. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Liu, Y. and De, A. (2015). Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study. *International Journal of Statistics in Medical Research*, 4:287–295.

Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-GAN: Training generative neural samplers using variational divergence minimization. *Advances in Neural Information Processing Systems*, pages 271–279.

Scott, C. and Nowak, R. (2005). A neyman-pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819.

Seaman, S. R. and White, I. R. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295.

Sportisse, A., Boyer, C., and Josse, J. (2020). Imputation and low-rank estimation with missing not at random data. *Statistics and Computing*, 30(6):1629–1643.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2010). Density ratio estimation: A comprehensive review. *RIMS Kokyuroku*, pages 10–31.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., Von Bünau, P., and Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746.

Tong, X. (2013). A plug-in approach to neyman-pearson classification. *Journal of Machine Learning Research*, 14(92):3011–3040.

Tong, X., Feng, Y., and Li, J. J. (2018). Neyman-pearson classification algorithms and np receiver operating characteristics. *Science Advances*, 4(2):eaao1659.

Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., and Sugiyama, M. (2009). Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155.

Ward, R. C., Axon, R. N., and Gebregziabher, M. (2020). Approaches for missing covariate data in logistic regres-

sion with MNAR sensitivity analyses. *Biometrical Journal*, 62(4):1025–1037.

Density Ratio Estimation and Neyman Pearson Classification with Missing Data: Supplementary Material

A ADDITIONAL PROOFS

Here we prove key results from the paper however before we do we need to first introduce some additional notation. We let p''_ω to refer to the density of X^ω with w.r.t. the measure ν where ν is the unique measure on $\mathcal{Z} \cup \{\varnothing\}$ s.t. (such that) $\nu(A) = \mu(A)$ for any $A \in \mathcal{B}_\mathcal{Z}$ and $\nu(\{\varnothing\}) = 1$. For square matrices $A, B \in \mathbb{R}^{d \times d}$ we take $A \succeq B$ to mean that $A - B$ is positive semi-definite and define $\lambda_{\min}(A)$ to be the smallest eigenvalue of A.

A.1 KLIEP Objective Simplification

Lemma 6. *Let $\mathcal{G}$ be closed under positive scalar scalar multiplication so that for all $r \in \mathcal{G}$, $a > 0$, $ar \in \mathcal{G}$. Define $\tilde{r}$ to be the solution of the constrained optimisation problem*

$$\min_{r \in \mathcal{G}} \mathrm{KL}(p_1 | rp_0)$$
$$subject\ to : \int_\mathcal{Z} r(z)p_0(z)\mu(\mathrm{d}z) = 1.$$

Then we can re-write the solution to the above optimisation problem as $\tilde{r} := N^{-1} \cdot r_0$ where

$$r_0 := \underset{r \in \bar{\mathcal{G}}}{argmax}\ \mathbb{E}[\log r(Z^1)] - \log \mathbb{E}[r(Z^0)]$$
$$N := \mathbb{E}[r_0(Z^0)]$$

where $\bar{\mathcal{G}}$ is $\bar{\mathcal{G}} \subseteq \mathcal{G}$ s.t. for any $r \in \bar{\mathcal{G}}$, there exists $r_0 \in \bar{\mathcal{G}}, a > 0$ s.t. $a \cdot r_0 = r$.

Proof. We have that

$$\underset{r \in \mathcal{G}}{argmin}\ \mathrm{KL}(p_1 | rp_0) = \int_\mathcal{Z} p_1(z) \log \frac{p_1}{rp_0(z)}$$
$$= \underset{r \in \mathcal{G}}{argmin}\ \mathbb{E}[\log \frac{p_1(Z^1)}{r(Z^1)p_0(Z^1)}]$$
$$= \underset{r \in \mathcal{G}}{argmin}\ \mathbb{E}[\log p_1(Z^1)] - \mathbb{E}[\log p_0(Z^1)] - \mathbb{E}[\log r(Z^1)]$$
$$= \underset{r \in \mathcal{G}}{argmax}\ \mathbb{E}[\log r(Z^1)].$$

As such we can re-write the optimisation problem as

$$\min_{r \in \mathcal{G}} \mathbb{E}[\log r(Z^1)]$$
$$subject\ to : \mathbb{E}[r(Z^0)] = 1.$$

To simplify this further we now show that

$$\{r \in \mathcal{G} | \mathbb{E}[r(Z^0)] = 1\} = \{\mathbb{E}[r(Z^0)]^{-1}r | r \in \mathcal{G}\}.$$

To this end let $r \in \mathcal{G}$, then as G is closed under positive scalar multiplication, $\mathbb{E}[r(Z^0)]^{-1}r \in \mathcal{G}$. Hence, as $\mathbb{E}\left[\mathbb{E}[r(Z^0)]^{-1}r(Z^0)\right] = 1$, we have that $\mathbb{E}[r(Z^0)]^{-1}r \in \{r \in \mathcal{G} | \mathbb{E}[r(Z^0)] = 1\}$. As the other inclusion is trivial, we have shown equality between the two sets of functions.

From this we can now re-write the optimisation as

$$\underset{\mathbb{E}[r(Z^0)]^{-1}r\in\{\mathbb{E}[r(Z^0)]^{-1}r|r\in\mathcal{G}\}}{\operatorname{argmax}}\mathbb{E}[\log(\mathbb{E}[r(Z^0)]^{-1}r(Z^1))]$$

$$=\underset{\mathbb{E}[r(Z^0)]^{-1}r\in\{\mathbb{E}[r(Z^0)]^{-1}r|r\in\mathcal{G}\}}{\operatorname{argmax}}\mathbb{E}[\log r(Z^1)]-\log\mathbb{E}[r(Z^0)]$$

which gives the desired result. □

Remark 3. *We can take $\bar{\mathcal{G}}$ s.t. for any $r \in \bar{\mathcal{G}}$, $a' > 0$, if $r, a'r \in \mathcal{G}$ then $a' = 1$. Hence for certain choices of $\mathcal{G}$ we can choose $\bar{\mathcal{G}}$ so that r_0 is unique*

A.2 KLIEP Finite Sample Proofs

Before we can prove Theorem 9 there are a few simple results we need first. The first is a generalisation of Bernstein bounds to random vectors presented in Kohler and Lucchi (2017).

Lemma 7 (Kohler and Lucchi (2017)). *Let $W_1, \ldots, W_n$ be IID copies of a non-negative random vector W over $\mathbb{R}^d$ with $\mathbb{E}[W] = \mu$. Now let $b, \sigma > 0$ be s.t. $\|W\| \le b$ a.s. and $\mathbb{E}[\|W\|^2] \le \sigma^2$. Then for any $\varepsilon < \frac{\sigma^2}{b}$*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}W_i-\mu\right\|\le\varepsilon\right)\ge 1-e^{1/4}\exp\left\{-\frac{n\varepsilon^2}{8\sigma^2}\right\}$$

We also need a Lemma bounding true and sample covariance in terms of the eigen-space

Lemma 8. *Let W be a RV over $\mathbb{R}^d$ with $\|W\| \le b$ a.s. and let $W_1 \ldots, W_n$ to be IID copies of W. Define the sample covariance of $\{W_i\}_{i\in[n]}$ by*

$$\widehat{Var}(W;\{W_i\}_{i\in[n]}):=\frac{1}{n}\sum_{i=1}^{n}W_iW_i^\top-\left(\frac{1}{n}\sum_{i=1}^{n}W_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}W_i\right)^\top$$

then provided $\sigma_{\min} \le 4b^2$

$$\mathbb{P}\left(\lambda_{\min}(\widehat{Var}(W;\{W_i\}_{i\in[n]})\ge\frac{1}{2}\lambda_{\min}(Var(W))\right)\ge 1-(d+e^{1/4})\exp\left\{\frac{\sigma_{\min}n}{32b^2}\right\}$$

where $\sigma_{\min}:=\lambda_{\min}(Var(W))$.

Proof. Define the centred RVs $Y_1, \ldots, Y_n$ by $Y_i := W_i - \mathbb{E}[W_i]$ and the random matrices R_n, S_n by

$$S_n:=\frac{1}{n}\sum_{i=1}^{n}Y_iY_i^\top$$

$$R_n:=\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)^\top$$

so that $R_n - S_n = \widehat{Var}(Y_1, \ldots, Y_n) = \widehat{Var}(W_1, \ldots, W_n)$.

Firstly we use the matrix Chernoff inequalities (Tropp, 2015) to get that

$$\mathbb{P}\left(\lambda_{\min}(S_n)\ge\frac{3}{4}\sigma_{\min}\right)\ge 1-d\exp\left\{-\frac{a'\sigma_{\min}n}{b^2}\right\}$$

where $a' := 3/4 \log(3/4) + 1/4$.

Now note that for any $x \in \mathbb{R}^d$ $\lambda_{\max}(xx^\top) = \|x\|^2$ so that $\lambda_{\max}(R_n) = \|\frac{1}{n}\sum_{i=1}^n Y_i\|^2$. Additionally, by Lemma 7, provided $\sigma_{\min} \le 4b^2$,

$$\mathbb{P}\left(\lambda_{\max}(R_n) \le \frac{1}{4}\sigma_{\min}\right) = \mathbb{P}\left(\|\frac{1}{n}\sum_{i=1}^n Y_i\| < \frac{1}{2}\sqrt{\sigma_{\min}}\right) \ge 1 - e^{1/4}\exp\left\{-\frac{\sigma_{\min}n}{32b^2}\right\}$$

Taking the intersection of these two events we get that w.p. at least $1 - (d + e^{1/4})\exp\{\frac{-\sigma_{\min}n}{32b^2}\}$

$$\begin{aligned}
\lambda_{\min}(\widehat{\mathrm{Var}}(Y_1,\ldots,Y_n)) &= \lambda_{\min}(S_n - R_n) \\
&\ge \lambda_{\min}(S_n) - \lambda_{\max}(R_n) \\
&\ge \frac{3}{4}\sigma_{\min} - \frac{1}{4}\sigma_{\min} \\
&= \frac{1}{2}\lambda_{\min}(\mathrm{Var}(W))
\end{aligned}$$

$\square$

We can now use this to state and prove our theorem giving finite sample bounds for the estimation error of KLIEP.

Theorem 9. *Define r_θ as in (3) and the population and empirical losses L, $\hat{L}$ by*

$$L(\theta) := -\mathbb{E}[\log(r_\theta(Z^1)] + \log\mathbb{E}[r_\theta(Z^0)]$$

$$\hat{L}(\theta, D) := -\frac{1}{n_1}\sum_{i=1}^{n_1}\log(r_\theta(Z_i^1)) + \log\frac{1}{n_0}\sum_{i=1}^{n_0}r_\theta(Z_i^0).$$

Let $\tilde{\theta}, \hat{\theta}$ be the minimisers of these respective losses. Now assume that $\|f\|_\infty < \infty$ and $\lambda_{\min}(\mathrm{Var}(f(Z^0))) := \sigma_{\min} > 0$. Then we have that for any $\delta < \frac{1}{2}$ and $n_{\min} \ge C_0\log(1/\delta)$,

$$\mathbb{P}\left(\|\hat{\theta}' - \tilde{\theta}\| \le \sqrt{\frac{C_0\log(1/\delta)}{n_{\min}}}\right) \ge 1 - \delta$$

with C_0 a constant depending upon $\|f\|_\infty, \sigma_{\min}, d, \tilde{\theta}$.

Proof. We firstly show that $\hat{L}$ is a convex function for θ for any sample $\mathcal{D} \in \mathcal{Z}^{n_0, n_1}$.

The first term is linear in θ so clearly convex. For the second term, note that it can be written as $g \circ h_{\mathcal{D}^0}(\theta)$ with $g: \mathbb{R}^{n_0} \to \mathbb{R}$, $h: \mathbb{R}^d \to \mathbb{R}^{n_0}$ given by

$$h_{\mathcal{D}^0}(\theta) = \begin{pmatrix} \theta^\top f(x_1^0) \\ \vdots \\ \theta^\top f(x_{n_0}^0) \end{pmatrix}$$

$$g(w) = \log\sum_{i=1}^{n_0}\exp\{w_i\}$$

Now again for any $\mathcal{D}^0$, h is linear in θ and therefore convex additionally g is convex and non-decreasing therefore $g \circ h$ is convex.

Now we state some important bounds which we will use in our proof. If we let $b := \|f\|_\infty$ $M := \exp\{\|\tilde{\theta}\|b\}$ then for any $z \in \mathcal{Z}$,

- $1/M \le r_{\tilde{\theta}}(z) \le M$
- $\|\nabla r_{\tilde{\theta}}(z)\| \le bM$
- $\left\|\frac{\nabla r_{\tilde{\theta}}(z)}{r_{\tilde{\theta}}(x)}\right\| \le b$

Finally we need to bound the minimum eigenvalue of the Hessian of the empirical loss function. To this end note that for any $\bar{\theta} \in B_\eta(\tilde{\theta}) := \{\theta \in \mathbb{R}^d \mid \|\tilde{\theta} - \theta\| \leq \eta\}$,

$$
\begin{aligned}
\nabla_\theta^2 \hat{L}(\bar{\theta}; D) &:= \frac{\frac{1}{n_0^2} \sum_{i,j} r_{\bar{\theta}}(Z_i^0) r_{\bar{\theta}}(Z_i^0) [f(Z_i^0) - f(Z_j^0)][f(Z_i^0) - f(Z_j^0)]^\top}{\frac{1}{n_0^2} \sum_{i,j} r_{\bar{\theta}}(Z_i^0) r_{\bar{\theta}}(Z_i^0)\}} \\
&\succeq \frac{\frac{1}{n_0^2} \sum_{i,j} M^{-2} 2 e^{-2\eta b}[f(Z_i^0) - f(Z_j^0)][f(Z_i^0) - f(Z_j^0)]^\top}{\frac{1}{n_0^2} \sum_{i,j} e^{2\eta b} r_{\tilde{\theta}}(Z_i^0) r_{\tilde{\theta}}(Z_i^0)} \\
&\succeq \frac{1}{M^2 e^{4\eta b}} \frac{\frac{1}{n_0^2} \sum_{i,j} [f(Z_i^0) - f(Z_j^0)][f(Z_i^0) - f(Z_j^0)]^\top}{\frac{1}{n_0^2} \sum_{i,j} r_{\tilde{\theta}}(Z_i^0) r_{\tilde{\theta}}(Z_j^0)} \\
&=: \frac{2}{M^2 e^{4\eta b}} \frac{\widehat{\mathrm{Var}}(f(Z^0); D_0)}{\widehat{\mathbb{E}}[r_{\tilde{\theta}}(Z^0); D_0]^2}
\end{aligned}
$$

where, for square matrices A, B we take $A \succeq B$, to mean $A - B$ is positives semi-definite.

Thus if we define $0 < \sigma_{\min} := \lambda_{\min}(\mathrm{Var}[f(Z^0)])$ then Lemma 8 tells us that, provided $\frac{1}{4}\sigma_{\min} < 1$,

$$
\mathbb{P}\left(\lambda_{\min}(\widehat{\mathrm{Var}}(f(Z^0); D_0)) \geq \frac{1}{2}\sigma_{\min}\right) \geq 1 - (d + e^{1/4}) \exp\left\{-\frac{1}{8}\left(\frac{\sigma_{\min}}{4b^2} \wedge 1\right) n\right\}
$$

Additionally if we let $\mu_2 := \mathbb{E}[r_{\tilde{\theta}}(Z^0)]$ we can use Lemma 7 to get that,

$$
\mathbb{P}\left(\widehat{\mathbb{E}}[r_{\tilde{\theta}}(Z^0); D_0] \leq \frac{3}{2}\mu_2\right) \geq 1 - e^{1/4} \exp\left\{-\frac{1}{8}\left(\frac{\mu_2^2}{4M^2} \wedge 1\right) n_0\right\}
$$

Thus for any given $\bar{\theta} \in B_\eta(\tilde{\theta})$ w.p.

$$
\mathbb{P}\left(\underbrace{\lambda_{\min}(\nabla_\theta^2 \hat{L}(\bar{\theta}; D)) \geq \frac{4\sigma_{\min}}{9M^2 e^{4\eta b}\mu_2^2}}_{:=A}\right) \geq 1 - (d + 2e^{1/4}) \exp\left\{-\frac{1}{8}\left(\frac{\mu_2^2}{4b^2} \wedge \frac{\sigma_{\min}}{4b^2} \wedge 1\right) n_0\right\} \tag{10}
$$

Now we have these bounds we can continue with the main body of the proof. For some $\eta > 0$, define $\hat{\theta}'$ to be the solution to the following constrained optimisation problem

$$
\min_{\theta \in B_\eta(\tilde{\theta})} \hat{L}(\theta; D)
$$

While, $\hat{\theta}'$ is a purely theoretical quantity, we know that when $\|\hat{\theta}' - \tilde{\theta}\| \leq 1$, $\hat{\theta} = \hat{\theta}'$ and so we can use it to get finite sample bounds for $\hat{\theta}$.

By the necessary KKT conditions, we know that there exists $\mu^* > 0$ s.t.

$$
\nabla_\theta \hat{L}(\hat{\theta}') + \mu^*(\hat{\theta}' - \tilde{\theta}) = 0
$$

in turn giving us that

$$
\begin{aligned}
0 &= \langle \hat{\theta}' - \tilde{\theta}, \nabla_\theta \hat{L}(\hat{\theta}') + \mu^*(\hat{\theta}' - \tilde{\theta}) \rangle \\
&= \langle \hat{\theta}' - \tilde{\theta}, \nabla_\theta \hat{L}(\hat{\theta}') \rangle + \mu^* \|(\hat{\theta}' - \tilde{\theta})\|^2 \\
&= \langle \hat{\theta}' - \tilde{\theta}, \nabla_\theta \hat{L}(\tilde{\theta}) + \nabla_\theta^2 \hat{L}(\bar{\theta})(\hat{\theta}' - \tilde{\theta}) \rangle + \mu^* \|(\hat{\theta}' - \tilde{\theta})\|^2.
\end{aligned}
$$

Using multidimensional MVT over each coordinate with $\bar{\theta}$ some point in the hypercube with opposite corners given by $\tilde{\theta}, \hat{\theta}'$ (which is a subset of $B_\eta(\tilde{\theta})$). We then get that

$$\|\hat{\theta}' - \tilde{\theta}\|^2 \lambda_{\min}(\nabla_\theta^2 \hat{L}(\bar{\theta})) \leq (\hat{\theta}' - \tilde{\theta})^\top \nabla_\theta^2 \hat{L}(\bar{\theta})(\hat{\theta}' - \tilde{\theta}) + \mu^* \|(\hat{\theta}' - \tilde{\theta})\|^2$$
$$= \langle \tilde{\theta} - \hat{\theta}', \nabla_\theta \hat{L}(\tilde{\theta}) \rangle$$
$$\leq \|\hat{\theta}' - \tilde{\theta}\| \|\nabla_\theta \hat{L}(\tilde{\theta})\|$$

.

Therefore we have that

$$\|\hat{\theta}' - \tilde{\theta}\| \leq \frac{1}{\lambda_{\min}(\nabla_\theta^2 \hat{L}(\bar{\theta}))} \|\nabla_\theta \hat{L}(\tilde{\theta})\|.$$

Hence if A from (10) holds with our given $\bar{\theta} \in B_\eta(\tilde{\theta})$ we have that

$$\|\hat{\theta}' - \tilde{\theta}\| \leq \frac{9M^2 e^{4\eta b} \mu_2^2}{4\sigma_{\min}} \|\nabla_\theta \hat{L}(\tilde{\theta})\| \quad \text{a.s. .}$$

From now on we will work on the event A so that bounding $\|\hat{\theta} - \tilde{\theta}\|$ simply requires us to bound $\|\nabla_\theta \hat{L}(\tilde{\theta})\|$.

As $\nabla_\theta L(\tilde{\theta}) = 0$ we have

$$\|\nabla_\theta \hat{L}(\tilde{\theta})\| = \|\nabla_\theta \hat{L}(\tilde{\theta}) - \nabla_\theta L(\tilde{\theta})\|$$
$$= \left\| \mathbb{E}\left[\frac{\nabla_\theta r_{\tilde{\theta}}(Z^1)}{r_{\tilde{\theta}}(Z^1)}\right] - \frac{\mathbb{E}[\nabla_\theta r_{\tilde{\theta}}(Z^0)]}{\mu_2} - \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\nabla_\theta r_{\tilde{\theta}}(Z_i^1)}{r_{\tilde{\theta}}(Z_i^1)} + \right.$$
$$\left. \frac{\frac{1}{n_0}\sum_{i=1}^{n_0}\nabla_\theta r_{\tilde{\theta}}(Z_i^0)}{\frac{1}{n_0}\sum_{i=1}^{n_0} r_{\tilde{\theta}}(Z_i^0)} \right\|$$
$$\leq \underbrace{\left\| \mathbb{E}\left[\frac{\nabla_\theta r_{\tilde{\theta}}(Z^1)}{r_{\tilde{\theta}}(Z^1)}\right] - \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\nabla_\theta r_{\tilde{\theta}}(Z_i^1)}{r_{\tilde{\theta}}(Z_i^1)} \right\|}_{:=I_1} +$$
$$\underbrace{\left\| \frac{\mathbb{E}[\nabla_\theta r_{\tilde{\theta}}(Z^0)]}{\mathbb{E}[r_{\tilde{\theta}(Z)}^0]} - \frac{\frac{1}{n_0}\sum_{i=1}^{n_0}\nabla_\theta r_{\tilde{\theta}}(Z_i^0)}{\frac{1}{n_0}\sum_{i=1}^{n_0} r_{\tilde{\theta}}(Z_i^0)} \right\|}_{:=I_2}$$

We now go on to bound I_1, I_2, using using generalised Hoeffding bounds.

We first introduce some additional notation to allow us to proceed. Define RVs W_1, W_2 and constants μ_1, μ_2 as follows

$$\mu_1 := \mathbb{E}[\nabla_\theta r_\theta(Z^0)] \qquad\qquad W_1 := \frac{1}{n_0}\sum_{i=1}^{n_0}\nabla_\theta r_{\tilde{\theta}}(Z_i^0)$$

$$\mu_2 := \mathbb{E}[r_{\tilde{\theta}}(Z^0)] \qquad\qquad W_2 := \frac{1}{n_0}\sum_{i=1}^{n_0} r_{\tilde{\theta}}(Z_i^0)$$

Then we have

$$I_2 := \left\| \frac{\mu_1}{\mu_2} - \frac{W_1}{W_2} \right\|$$
$$= \frac{\|\mu_1 W_2 - \mu_2 W_1\|}{\mu_2 W_2}$$
$$\leq \frac{\|\mu_1 - W_1\|}{\mu_2} + \frac{|\mu_2 - W_2|\|\mu_1\|}{\mu_2 W_2}$$

Thus to have

$$I_1 + I_2 < \frac{4\varepsilon\sigma_{\min}}{9M^2 e^{4\eta b}\mu_2^2},$$

it is sufficient to have

$$I_1 < \frac{2\varepsilon\sigma_{\min}}{9M^2 e^{4\eta b}\mu_2^2},$$

$$\|\mu_1 - W_1\| < \frac{\varepsilon\sigma_{\min}}{9M^2 e^{4\eta b}\mu_2},$$

$$|\mu_2 - W_2| < \frac{\varepsilon\sigma_{\min}}{18M^2 e^{4\eta b}\|\mu_1\|}$$

provided $\varepsilon < \frac{9M^2 e^{4\eta b}\mu_2\|mu_1\|}{\sigma_{\min}}$. Furthermore we can get probabilities on each of these events using Lemma 7 once again. These probabilities are

$$\mathbb{P}\left(I_1 < \frac{2\varepsilon\sigma_{\min}}{9M^2 e^{4\eta b}\mu_2^2}\right) \geq 1 - e^{1/4}\exp\left\{-\frac{1}{8}\left(\frac{4\varepsilon^2\sigma_{\min}^2}{81b^2 M^4 e^{8\eta b}\mu_2^4} \wedge 1\right)n_1\right\}$$

$$\mathbb{P}\left(\|\mu_1 - W_1\| \leq \frac{\varepsilon\sigma_{\min}}{9M^2 e^{4\eta b}\mu_2}\right) \geq 1 - e^{1/4}\exp\left\{-\frac{1}{8}\left(\frac{\varepsilon^2\sigma_{\min}^2}{81b^2 M^6 e^{8\eta b}\mu_2^2} \wedge 1\right)n_0\right\}$$

$$\mathbb{P}\left(|\mu_2 - W_2| \leq \frac{\varepsilon\sigma_{\min}}{18M^2 e^{4\eta b}\|\mu_1\|}\right) \geq 1 - e^{1/4}\exp\left\{-\frac{1}{8}\left(\frac{\varepsilon^2\sigma_{\min}^2}{324M^6 e^{8\eta b}\|\mu_1\|^2} \wedge 1\right)n_0\right\}$$

Therefore using unions bounds and taking $\eta = \|\tilde{\theta}\|$ we have that

$$\mathbb{P}(\|\hat{\theta}' - \tilde{\theta}\| \leq \varepsilon) \geq \mathbb{P}\left(\|\nabla_\theta \hat{L}(\tilde{\theta})\| \leq \frac{\varepsilon\sigma_{\min}}{M^4 e^{4\|\tilde{\theta}\|b}}, A\right)$$
$$\geq 1 - \alpha\exp\left\{-C(\varepsilon^2 \wedge \gamma)n_{\min}\right\}$$

where

$$\alpha := d + 5e^{1/4}$$

$$C := \frac{1}{8}\min\left\{\frac{4\sigma_{\min}^2}{81b^2 M^4 e^{8\|\tilde{\theta}\|b}\mu_2^4}, \frac{\varepsilon^2\sigma_{\min}^2}{81b^2 M^6 e^{8\|\tilde{\theta}\|b}\mu_2^2}, \frac{\varepsilon^2\sigma_{\min}^2}{324M^6 e^{8\|\tilde{\theta}\|b}\|\mu_1\|^2}\right\}$$

$$\gamma := \frac{1}{8C}\min\left\{\frac{\mu_2^2}{4b^2}, \frac{\sigma_{\min}}{4b^2}, 1\right\}$$

As $\hat{L}$ is convex for any sample we know that if $\hat{\theta}'$ is in the interior of $B_1(\tilde{\theta})$ then $\hat{\theta}' = \hat{\theta}$. Therefore, for any $0 < \varepsilon < \eta$ the same result holds replacing $\hat{\theta}'$ with $\hat{\theta}$.

hence now gives us that for $\frac{\log(\alpha/\delta)}{Cn_{\min}} \leq \gamma \wedge \|\tilde{\theta}\|$ w.p. $1 - \delta$

$$\|\hat{\theta} - \tilde{\theta}\| \leq \sqrt{\frac{\log(\alpha/\delta)}{Cn_{\min}}}$$

If we assume $\delta \leq \frac{1}{2}$ then $\log(\alpha/\delta) \leq (\log(\alpha) + 1)\log(1/\delta)$. Therefore we have that provided $n_{\min} \geq \frac{(\log(\alpha)+1)\log(1/\delta)}{\|\tilde{\theta}\|\wedge\gamma}$ w.p. $1 - \delta$,

$$\|\hat{\theta} - \tilde{\theta}\| \leq \sqrt{\frac{(\log(\alpha) + 1)\log(1/\delta)}{Cn_{\min}}}.$$

Finally, taking $C_0 := \frac{\log(\alpha)+1}{\min\{C,\|\tilde{\theta}\|,\gamma\}}$ gives our desired result. $\qquad\square$

A.2.1 Bound on KLIEP Normalising Constant

We now give a finite sample bound for the normalising constant calculated in KLIEP

Corollary 10. *Define* $N^*, \hat{N}$ *to be*

$$N^* := \mathbb{E}[r_{\tilde{\theta}}(Z^0)]$$

$$\hat{N} := \frac{1}{m_0} \sum_{i=1}^{m_0} r_{\hat{\theta}}(Z_i^0)$$

with $\hat{\theta}, \tilde{\theta}, r_{\tilde{\theta}}$ *defined as before. Then we have that for* $\delta \in (0, 1/2]$, *provided* $n_{\min} > C_N \log(1/\delta)$,

$$\mathbb{P}\left(\|N^* - \hat{N}\| \leq \sqrt{\frac{C_N \log(1/\delta)}{n_{\min}}} \right) \geq 1 - \delta$$

with C_N *a constant depending upon* $C_0, b, \|\tilde{\theta}\|$.

Proof. We condition upon the same events we condition upon in Theorem 9. Specifically for $\delta \in (0, 1/2]$, provided $n_{\min} > 2C_0 \log(1/\delta)$, w.p. at least $1 - \delta$ the following two conditions hold:

$$\|\hat{\theta} - \tilde{\theta}\| \leq \sqrt{\frac{C_0 \log(1/\delta)}{n_{\min}}} \qquad\qquad |\mu_2 - W_2| < \frac{\varepsilon \sigma_{\min}}{18 M^2 e^{4\eta b} \|\mu_1\|}$$

with $\varepsilon = \sqrt{C_0 \log(\alpha/\delta)}$. The first conditions implies

$$|r_{\tilde{\theta}}(z) - r_{\hat{\theta}}(z)| \leq e^{\|\tilde{\theta}\| b} \sqrt{\frac{2 C_0 \log(1/\delta)}{n_{\min}}}$$

Therefore we have that

$$|N^* - \hat{N}| \leq \left| \mathbb{E}[r_{\tilde{\theta}}(Z^0)] - \frac{1}{n_0} \sum_{i=1}^{n} r_{\tilde{\theta}}(Z_i^0) \right| + \left| \frac{1}{n_0} \sum_{i=1}^{n_0} r_{\hat{\theta}}(Z_i^0) - \frac{1}{n_0} \sum_{i=1}^{n_0} r_{\tilde{\theta}}(Z_i^0) \right|$$

$$\leq \sqrt{\frac{C_1 \log(1/\delta)}{n_{\min}}} + e^{\|\tilde{\theta}\| b} \sqrt{\frac{C_0 \log(1/\delta)}{n_{\min}}}$$

with

$$C_1 := \frac{\sigma_{\min} \sqrt{C_0''}(\log(\alpha) + 1)}{18 M^2 e^{4\|\tilde{\theta}\| b} \|\mu_1\|}$$

Taking $C_N = (\sqrt{C_0} + e^{\|\tilde{\theta}\| b} \sqrt{C_1})^2$ gives our desired result. $\qquad\square$

A.3 Proof of Lemma 1

Proof. Throughout, unless stated otherwise, all integrals are taken w.r.t. μ (note that for any integral over some subset A of $\mathcal{Z}$ this is equivalent to taking the integral w.r.t ν.)

First we prove that

$$p''(z) = (1 - \varphi(x))p(z)$$

for μ almost every $z \in \mathcal{Z}$. by showing that for all $A \in \mathcal{B}_{\mathcal{Z}}$

$$\mathbb{P}(X \in A) = \int_A (1 - \varphi(x))p(x)\nu(\mathrm{d}x)$$

Indeed

$$\int_A (1 - \varphi(x))p(x)\mathrm{d}x = \mathbb{E}[\mathbb{1}\{Z \in A\}(1 - \varphi(Z))]$$

$$= \mathbb{E}[\mathbb{1}\{Z \in A\}] - \mathbb{E}[\mathbb{1}\{Z \in A\}\mathbb{1}\{X = \varnothing\}] \quad \text{by definition of } \varphi(Z)$$

$$= \mathbb{E}[\mathbb{1}\{Z \in A\}\mathbb{1}\{X \neq \varnothing\}] \quad \text{by linearity of expectation}$$

$$= \mathbb{P}(Z \in A, X \neq \varnothing)$$

$$= \mathbb{P}(Z \in A, X = Z)$$

$$= \mathbb{P}(X \in A)$$

We now have that for any $A \in \mathcal{B}_{\mathcal{Z}}$,

$$\mathbb{P}(X \in A | X \neq \varnothing) = \mathbb{P}(X \neq \varnothing)^{-1}\mathbb{P}(X \in A)$$

$$= \mathbb{P}(X \neq \varnothing)^{-1} \int_A p''(x)\nu(\mathrm{d}x)$$

$$= \mathbb{P}(X \neq \varnothing)^{-1} \int_A (1 - \varphi(x)p(x)\mu(\mathrm{d}x)$$

and so $p' = \mathbb{P}(X \neq \varnothing)^{-1} \cdot (1 - \varphi) \cdot p$.

We also have that p can be extended to a density over $\mathcal{X}$ by taking $p(\varnothing) = 0$. This combined with the above result gives us that

$$\frac{p(x)}{p''(x)} := \frac{\mathbb{1}\{x \neq \varnothing\}}{1 - \varphi(x)}$$

for ν almost every $x \in \mathcal{X}$.

Hence we can use importance weighting to get our desired result. $\qquad\square$

A.4 Proof M-KLIEP Finite Sample Bounds

Lemma 11. *Let W, V be two RVs over $\mathbb{R}^d$ with p.d.f.s p_W p_V w.r.t. some measure ν. Assume that $g := \frac{p_W}{p_V}$ is well defined and $g(w) \leq \alpha$ for all $w \in \mathbb{R}^d$. Finally suppose that $\|W\|, \|V\| \leq b$ a.s. Then we have that for any $0 < \varepsilon < b$*

$$\mathbb{P}\left(\left|\mathbb{E}[W] - \frac{1}{n}g(V)V\right| \leq \varepsilon\right) \geq 1 - e^{1/4}\exp\left\{-\frac{\varepsilon^2 n}{8\alpha b^2}\right\}$$

Proof. This proof is a direct corollary of Lemma 7 First we clearly have that $\|g(V)V\| \leq \alpha b$ a.s.. Second we have that

$$\mathbb{E}(\|g(V)V\|^2) = \mathbb{E}[g(V)^2\|V\|^2]$$

$$= \mathbb{E}[g(W)\|W^2\|]$$

$$\leq \alpha\mathbb{E}[\|W\|^2]$$

$$\leq ab^2.$$

Hence as $\mathbb{E}[g(V)V - \mathbb{E}[W]] = 0$, we can use the Bernstein inequality to get

$$\mathbb{P}\left(\left|\mathbb{E}[W] - \frac{1}{n}g(V)V\right| \leq \varepsilon\right) \geq 1 - e^{1/4}\exp\left\{-\frac{\varepsilon^2 n}{8\alpha b}\right\}$$

$$\square$$

Remark 4. *If we want to remove the requirement that $\varepsilon < b$ we can re-write this as*

$$\mathbb{P}\left(\left|\mathbb{E}[W] - \frac{1}{n}g(V)V\right| \leq \varepsilon\right) \geq 1 - e^{1/4}\exp\left\{-\left(\frac{\varepsilon^2}{b^2} \wedge 1\right)\frac{n}{8\alpha}\right\}$$

Lemma 12. *Let W, V be two RVs over $\mathbb{R}^d$, with densities p_W, p_V respectively s.t. $\|W\|, \|V\| < b$ a.s.. Assume that $g := \frac{p_W}{p_V} : \mathbb{R}^d \to R^p$ is well defined with $g(w) < \alpha$ for all $w \in \mathbb{R}^d$. . Let $V_1, \ldots, V_n$ be IID copies of V. Now define the importance weighted sample covariance by of W estimated by $\{V_i\}_{i \in [n]}$*

$$\widehat{Var}(W; \{V_i\}_{i \in [n]}) := \left(\frac{1}{n} \sum_{i=1}^n g(V_i) V_i V_i^\top \right) \left(\frac{1}{n} \sum_{i \in [n]} g(V_i) \right) - \left(\frac{1}{n} \sum_{i=1}^n g(V_i) V_i \right) \left(\frac{1}{n} \sum_{i=1}^n g(V_i) V_i \right)^\top$$

then provided $\sigma_{\min} \leq 4b^2$

$$\mathbb{P}\left(\lambda_{\min}(\widehat{Var}(W; \{V_i\}_{i \in [n]})) \geq \frac{1}{2}\sigma_{\min} \right) \geq 1 - (d + 2e^{1/4}) \exp\{ \frac{a'\sigma_{\min} n}{\alpha b^2} \}$$

where $a' := \frac{(2-\sqrt{3})^2}{32}$ and $\sigma_{\min} := \lambda_{\min}(Var(W))$.

Proof. Define the centred RVs $Y_1, \ldots, Y_n$ by $Y_i := V_i - \mathbb{E}[W_i]$ and the random matrices R_n, S_n by

$$S_n := \frac{1}{n} \sum_{i=1}^n g(V_i) Y_i Y_i^\top$$

$$R_n := \left(\frac{1}{n} \sum_{i=1}^n g(V_i) Y_i \right) \left(\frac{1}{n} \sum_{i=1}^n g(V_i) Y_i \right)^\top$$

$$a_n := \frac{1}{n} \sum_{i \in [n]} g(V_i)$$

so that $a_n R_n - S_n = \widehat{Var}(W; \{Y_i\}_{i \in [n]})$. Simple algebraic manipulation gives us that $a_n R_n - S_n = \widehat{Var}(W; \{V_i\}_{i \in [n]})$.

Firstly we can use the matrix Chernoff inequalities (Tropp, 2015), alongside the fact that $\mathbb{E}[S_n] = Var(W)$ and $\lambda_{\max}(\frac{1}{n} g(V_i) Y_i Y_i^\top) \leq \frac{\alpha b^2}{n}$ to get

$$\mathbb{P}\left(\lambda_{\min}(S_n) \geq \frac{\sqrt{3}}{2}\sigma_{\min} \right) \geq 1 - d\exp\left\{ -\frac{a\sigma_{\min} n}{\alpha b^2} \right\}$$

Now note that for any $x \in \mathbb{R}^d$ $\lambda_{\max}(xx^\top) = \|x\|^2$ so that $\lambda_{\max}(R_n) = \|\frac{1}{n} Y_i\|^2$. Hence, as $\mathbb{E}[g(V_i) Y_i] = 0$, we can use vector Bernstein bounds to get that provided $\sigma_{\min} \leq 4b^2$

$$\mathbb{P}\left(\lambda_{\max}(R_n) \leq \frac{1}{4}\sigma_{\min} \right) = \mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^n g(V_i) Y_i \right\| < \frac{1}{2}\sqrt{\sigma_{\min}} \right)$$

$$\geq 1 - e^{1/4} \exp\left\{ -\frac{\sigma_{\min} n}{32\alpha b^2} \right\}.$$

Finally by Lemma 11,

$$\mathbb{P}\left(a_n \geq \frac{\sqrt{3}}{2} \right) \geq 1 - e^{1/4} \exp\left\{ \frac{(2-\sqrt{3})^2 n}{32\alpha} \right\}$$

Taking the intersection of these three events we get that w.p. at least $1 - (d + 2e^{1/4}) \exp\left\{ -\left(\frac{a'\sigma_{\min}}{b^2} \wedge 1 \right) \frac{n}{8\alpha} \right\}$

$$\lambda_{\min}(\widehat{Var}(W; \{(V_i, g(V_i))\}_{i \in [n]})) = \lambda_{\min}(a_n S_n - R_n)$$

$$\geq \lambda_{\min}(a_n S_n) - \lambda_{\max}(R_n)$$

$$= a_n \lambda_{\min}(S_n) - \lambda_{\max}(R_n)$$

$$\geq \frac{\sqrt{3}}{2} \frac{\sqrt{3}}{2} \sigma_{\min} - \frac{1}{4}\sigma_{\min}$$

$$= \frac{1}{2}\lambda_{\min}(Var(W))$$

where $a' := \frac{(2-\sqrt{3})^2}{4}$ $\qquad\square$

We can now state the Lemma which gives rise to Theorem 9.

Lemma 13. *Let r_θ take the log-linear form and for for $\theta \in \mathbb{R}^d$ define and $L, \hat{L}'$ by*

$$L(\theta) := -\mathbb{E}[r_\theta(Z^1)] + \log \mathbb{E}[r_\theta(Z^0)] \hat{L}'(\theta; D') := -\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{\mathbb{1}\{X_i^1 \neq \varnothing\}}{1 - \varphi^1(X_i^1)} \log r_\theta(X_i^1)$$

$$+ \log \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{\mathbb{1}\{X_i^0 \neq \varnothing\}}{1 - \varphi^0(X_i^0)} r_\theta(X_i^0).$$

Finally, define the constant $\tilde{\theta}$ and the RV $\hat{\theta}$, by

$$\tilde{\theta} := \underset{\theta \in \mathbb{R}^d}{argmax}\, L(\theta)$$

$$\hat{\theta} := \underset{\theta \in \mathbb{R}^d}{argmax}\, \hat{L}'(\theta, D').$$

Now suppose that $\|\varphi^0\|_\infty < 1$, $\|\varphi^1\|_\infty < 1$, $\|f\|_\infty < \infty$ and additionally let $\sigma_{\min} := \lambda_{\min}(Var(f(Z^0))) > 0$

Then we have that for any $\delta < \frac{1}{2}$ and $m_{\text{eff}} \min \geq C_0' \log(1/\delta)$,

$$\mathbb{P}\left(\|\hat{\theta}' - \tilde{\theta}\| \leq \sqrt{\frac{C_0' \log(1/\delta)}{m_{\text{eff}}}} \right) \geq 1 - \delta$$

where C_0' a constant depending upon $\|f\|_\infty, \sigma_{\min}, d, \|\tilde{\theta}\|$.

Proof. Firstly we define the population version of $\hat{L}'$ by L' as follows

$$L'(\theta) = \mathbb{E}\left[\frac{\mathbb{1}\{X^1 \neq \varnothing\}}{1 - \varphi^1(X^1)} \log r(X^1) \right] - \log \mathbb{E}\left[\frac{\mathbb{1}\{x^0 \neq \varnothing\}}{1 - \varphi^0(X^0)} \right].$$

Then from Lemma 1 we have that $L(\theta) = L'(\theta)$. As a result we can re-define $\tilde{\theta}$ to be

$$\tilde{\theta} := \underset{\theta \in \mathbb{R}^d}{argmin}\, L'(\theta)$$

For proof of convexity we note that again the term involving $\mathcal{D}'^+$ (our sample from $\{X_i^1\}_{i=1}^{n_1}$) is linear in θ and the second term is convex for the same reason as $\hat{L}$ is with the caveat that we replace $g : \mathbb{R}^{n_0} \to \mathbb{R}$ by $g'_{\mathcal{D}^0} : \mathbb{R}^{n_0} \to \mathbb{R}$ defined by

$$g(x) = \log \sum_{i=1}^n w_i \exp\{x_i\} \quad \text{where}$$

$$w_i^0 := \frac{\mathbb{1}\{x_i^0 \neq \varnothing\}}{1 - \varphi^0(x_i^0)}.$$

As each $w_i > 0$, this modified log-sum-exp is also convex by the same argument which makes the unmodified log-sum-exp function convex. That is,

$$\nabla_x^2 g'_{\mathcal{D}^0}(z) = \frac{\text{diag}(u)(1^\top u) - uu^\top}{(1^\top u)^2}$$

where $u_i = w_i^0 \exp(z_i)$. If we then let $v \in \mathbb{R}^{n_0}$, we get

$$v^\top \nabla_x^2 g'_{\mathcal{D}^0}(z) = \frac{v^\top \text{diag}(u)v(1^\top u) - v^\top uu^\top v}{(1^\top u)^2}$$

$$= \frac{(\sum_{i=1}^{n_0} v_i^2 u_i)(\sum_{i=1}^{n_1} u_i) - (\sum_{i=1}^{n_0} v_i u_i)}{(1^\top u)} > 0$$

where the final inequality is given by the Cauchy-Schwartz inequality.

For ease of notation we define $b := \|f\|_\infty$ and $M := \exp\|\tilde{\theta}\|b$. We now aim to bound $\lambda_{\min}(\nabla^2_\theta \hat{L}'(\bar{\theta}))$ from below. To this end we have

$$
\nabla^2_\theta \hat{L}(\theta; D) := \frac{\frac{1}{n_0^2}\sum_{i,j} w_i^0 w_j^0 r_{\bar{\theta}}(X_i^0) r_{\bar{\theta}}(X_j^0)[f(X_i^0) - f(X_j^0)][f(X_i^0) - f(X_j^0)]^\top}{\frac{1}{n^2}\sum_{i,j} w_i^0 w_j^0 r_{\bar{\theta}}(X_i^0) r_{\bar{\theta}}(X_j^0)}
$$

$$
\succeq \frac{\frac{1}{n_0^2}\sum_{i,j} M^{-2}e^{-2\eta b} w_i^0 w_j^0[f(X_i^0) - f(X_j^0)][f(X_i^0) - f(X_j^0)]^\top}{\frac{1}{n^2}\sum_{i,j} e^{2\eta b} w_i^0 w_j^0 r_{\bar{\theta}}(X_i^0) r_{\bar{\theta}}(X_j^0)}
$$

$$
\succeq \frac{1}{M^2 e^{4\eta b}} \frac{\frac{1}{n_0^2}\sum_{i,j} w_i^0 w_j^0[f(X_i^0) - f(X_j^0)][f(X_i^0) - f(X_j^0)]^\top}{\frac{1}{n_0}\sum_{i,j} w_i^0 w_j^0 r_{\tilde{\theta}}(X_i^0) r_{\tilde{\theta}}(X_j^0)}
$$

$$
=: \frac{2}{M^4 e^{4\eta b}} \frac{\widehat{\mathrm{Var}}_0\left[f(Z^0); D_1', \{w_i^0\}_{i=1}^{n_0}\right]}{\left(\widehat{\mathbb{E}}'\left[r_{\tilde{\theta}}(Z^0); D_1'\right]\right)^2}
$$

where, for square matrices A, B we take $A \succeq B$, to mean $A - B$ is positives semi-definite. We can then bound $\widehat{\mathrm{Var}}_0$ and $\widehat{\mathbb{E}}'$ similarly to before to get that

$$
\mathbb{P}\left(\lambda_{\min}(\widehat{\mathrm{Var}}(f(Z), D_1')) \geq \frac{1}{2}\sigma_{\min}\right) \geq 1 - (d + 2e^{1/4})\exp\{-\left(\frac{a'\sigma_{\min}}{b^2} \wedge 1\right)\frac{n_0(1 - \varphi_{\max}^0)}{8}\}
$$

$$
\mathbb{P}\left(\mathbb{E}'\left[r_{\tilde{\theta}}(Z^0); D_1'\right] \leq \frac{3}{2}\mathbb{E}[r_{\tilde{\theta}}(X^0)]\right) \geq 1 - e^{1/4}\exp\left\{-\left(\frac{\mu_2}{4M^2} \wedge 1\right)\frac{n_0(1 - \varphi_{\max}^0)}{8}\right\}
$$

where again $\sigma_{\min} := \lambda_{\min}(\mathrm{Var}(f(Z^0)))$, $\mu_2 := \mathbb{E}[r_{\tilde{\theta}}(Z^0)]$.

Therefore we have that

$$
\mathbb{P}\left(\underbrace{\lambda_{\min}(\nabla^2_\theta \hat{L}'; (\theta; D')) \geq \frac{4\sigma_{\min}}{9M^2 e^{4\eta b}\mu_2^2}}_{:=A'}\right) \geq 1 - (d + 3e^{1/4})\exp\left\{-\left(\frac{\mu_2}{4M^2} \wedge \frac{a'\sigma_{\min}}{b^2} \wedge 1\right)\frac{n_0(1 - \varphi_{\max}^0)}{8}\right\} \quad (11)
$$

We now have all the bounds required to continue with the main body of the proof. Let $\hat{\theta}'$ be defined as the solution to the following constrained optimisation problem.

$$
\min_{\theta \in B_\eta(\tilde{\theta})} \hat{L}(\theta, D')
$$

By and identical argument to Theorem 9 we then get that

$$
\|\hat{\theta}' - \tilde{\theta}\| \leq \frac{1}{\lambda_{\min}(\nabla^2_\theta \hat{L}'(\bar{\theta}))}\|\nabla_\theta \hat{L}'(\tilde{\theta})\|.
$$

for some $\bar{\theta} \in B_\eta(\tilde{\theta})$.

Hence if A' from (11) holds with our given $\bar{\theta} \in B_\eta(\tilde{\theta})$ we have that

$$
\|\hat{\theta}' - \tilde{\theta}\| \leq \frac{9M^2 e^{4\eta b}\mu_2^2}{4\sigma_{\min}}\|\nabla_\theta \hat{L}(\tilde{\theta})\| \quad \text{a.s. .}
$$

From now on we will work on the event A so that bounding $\|\hat{\theta} - \tilde{\theta}\|$ simply requires us to bound $\|\nabla_\theta \hat{L}(\tilde{\theta})\|$.

We further get that $\|\nabla_\theta \hat{L}'(\tilde{\theta})\| \leq I_1' + I_2'$ where

$$I_1' = \left\| \mathbb{E}\left[\frac{\nabla_\theta r_{\tilde{\theta}}(Z^1)}{r_{\tilde{\theta}}(Z^1)} \right] - \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{\mathbb{1}\{X_i^1 \neq \varnothing\}}{1 - \varphi^1(X_i^1)} \frac{\nabla_\theta r_{\tilde{\theta}}(X_i^1)}{r_{\tilde{\theta}}(X_i^1)} \right\|$$

$$I_2' = \left\| \frac{\mathbb{E}[\nabla_\theta r_{\tilde{\theta}}(X^0)]}{\mathbb{E}[r_{\tilde{\theta}}(X^0)]} - \frac{\frac{1}{n_0} \sum_{i=1}^{n_0} \frac{\mathbb{1}\{X_i^0 \neq \varnothing\}}{1 - \varphi^0(X_i^0)} \nabla_\theta r_{\tilde{\theta}}(X_i^0)}{\frac{1}{n_0} \sum_{i=1}^{n_0} \frac{\mathbb{1}\{X_i^0 \neq \varnothing\}}{1 - \varphi^0(X_i^0)} r_{\tilde{\theta}}(X_i^0)} \right\|.$$

Define RVs W_1', W_2'

$$W_1' := \frac{1}{n_0} \sum_{i \in [n_0]} \frac{\mathbb{1}\{X_i^0 \neq \varnothing\}}{1 - \varphi^0(X_i^0)} \nabla_\theta r_{\tilde{\theta}}(X_i^0)$$

$$W_2' := \frac{1}{n_0} \sum_{i \in [n_0]} \frac{\mathbb{1}\{X_i^0 \neq \varnothing\}}{1 - \varphi^0(X_i^0)} r_{\tilde{\theta}}(X_i^0)$$

and $\mu_1 := \mathbb{E}[\nabla_\theta r_{\tilde{\theta}}(Z^0)]$. We can then bound I_2' to get

$$I_2' := \leq \frac{\|\mu_1 - W_1'\|}{\mu_2} + \frac{|\mu_2 - W_2'|\|\mu_1\|}{\mu_2 W_2'}$$

Thus to have

$$I_1' + I_2' < \frac{4\sigma_{\min}}{\varepsilon 9 M^2 e^{4\eta b} \mu_2^2},$$

it is sufficient to have

$$I_1' < \frac{2\varepsilon \sigma_{\min}}{9 M^2 e^{4\eta b} \mu_2^2},$$

$$\|\mu_1 - W_1'\| < \frac{\varepsilon \sigma_{\min}}{9 M^2 e^{4\eta b} \mu_2},$$

$$|\mu_2 - W_2'| < \frac{\varepsilon \sigma_{\min}}{18 M^2 e^{4\eta b} \|\mu_1\|}.$$

provided $\varepsilon < \frac{9 M^2 e^{4\eta b} \mu_2 \|mu_1\|}{\sigma_{\min}}$.

The probabilities for these events are

$$\mathbb{P}\left(I_1' < \frac{2\varepsilon \sigma_{\min}}{9 M^2 e^{4\eta b} \mu_2^2} \right) \geq 1 - e^{1/4} \exp\left\{ -\frac{1}{8} \left(\frac{4\varepsilon^2 \sigma_{\min}^2}{81 b^2 M^4 e^{8\eta b} \mu_2^4} \wedge 1 \right) n_1 (1 - \varphi_{\max}^1) \right\}$$

$$\mathbb{P}\left(\|\mu_1 - W_1'\| \leq \frac{\varepsilon \sigma_{\min}}{9 M^2 e^{4\eta b} \mu_2} \right) \geq 1 - e^{1/4} \exp\left\{ -\frac{1}{8} \left(\frac{\varepsilon^2 \sigma_{\min}^2}{81 b^2 M^6 e^{8\eta b} \mu_2^2} \wedge 1 \right) n_0 (1 - \varphi_{\max}^0) \right\}$$

$$\mathbb{P}\left(|\mu_2 - W_2'| \leq \frac{\varepsilon \sigma_{\min}}{18 M^2 e^{4\eta b} \|\mu_1\|} \right) \geq 1 - e^{1/4} \exp\left\{ -\frac{1}{8} \left(\frac{\varepsilon^2 \sigma_{\min}^2}{324 M^6 e^{8\eta b} \|\mu_1\|^2} \wedge 1 \right) n_0 (1 - \varphi_{\max}^0) \right\}$$

Therefore by the same argument as Theorem 9 we get

$$\mathbb{P}(\|\hat{\theta} - \tilde{\theta}\| \leq \varepsilon) \geq \mathbb{P}\left(\|\nabla_\theta \hat{L}(\tilde{\theta})\| \leq \frac{\varepsilon \sigma_{\min}}{M^4 e^{4\eta b}}, A' \right)$$

$$\geq 1 - \alpha' \exp\left\{ -C' \left(\varepsilon^2 \wedge \gamma \right) m_{\text{eff}} \right\}.$$

where

$$\alpha' := d + 6e^{1/4}$$

$$C' := \frac{1}{8} \min \left\{ \frac{4\sigma_{\min}^2}{81b^2 M^4 e^{8\|\tilde{\theta}\|b} \mu_2^4}, \frac{\varepsilon^2 \sigma_{\min}^2}{81b^2 M^6 e^{8\|\tilde{\theta}\|b} \mu_2^2}, \frac{\varepsilon^2 \sigma_{\min}^2}{324 M^6 e^{8\|\tilde{\theta}\|b} \|\mu_1\|^2} \right\}$$

$$\gamma' := \frac{1}{8C} \min \left\{ \frac{a'\mu_2^2}{M^2}, \frac{\sigma_{\min}}{4b^2}, 1 \right\}$$

By an identical argument to the proof of Theorem 9 in Section A.2 this gives us that for us that for $\delta \leq \frac{1}{2}$, provided $m_{\text{eff}} \geq \frac{\log(\alpha')\log(1/\delta)}{\gamma' \wedge \|\tilde{\theta}\|}$ then

$$\mathbb{P}\left(\|\tilde{\theta} - \hat{\theta}\| \leq \sqrt{\frac{\log(\alpha')\log(1/\delta)}{C' m_{\text{eff}}}} \right) \geq 1 - \delta.$$

Now taking $C_0' := \frac{\log(\alpha')}{\min\{C', \|\tilde{\theta}\|, \gamma'\}}$ gives our desired result. $\square$

A.4.1 Bound on Normalisation term in M-KLIEP

Corollary 14. *Define $N^*, \hat{N}'$ to be*

$$N^* := \mathbb{E}[r_{\tilde{\theta}}(Z^0)]$$

$$\hat{N} := \frac{1}{m_0} \sum_{i=1}^{m_0} \frac{\mathbb{1}\{X_i^0 \neq \varnothing\}}{1 - \varphi^0(x_i^0)} r_{\hat{\theta}'}(X_i^0)$$

with $\hat{\theta}', \tilde{\theta}, r_{\tilde{\theta}}$ defined as before. Then we have that for $\delta \in (0, 1/2]$, provided $n_{\min} > C_N' \log(1/\delta)$,

$$\mathbb{P}\left(\|N^* - \hat{N}\| \leq \frac{1}{1 - \|\varphi^0\|_\infty} \sqrt{\frac{C_N' \log(1/\delta)}{m_{\text{eff}}}} \right) \geq 1 - \delta$$

Proof. We work assuming the same events hold as in Theorem 2. Crucially we have that for $\delta \in (0, 1/2]$, provided $m_{\text{eff}} \geq C_0' \log(1/\delta)$, w.p. at least $1 - \delta$ the following two events hold

$$\|\hat{\theta} - \tilde{\theta}\| \leq \sqrt{\frac{2C_0' \log(1/\delta)}{n_{\min}}} \qquad\qquad |\mu_2 - W_2'| < \frac{\varepsilon \sigma_{\min}}{18 M^2 e^{4\eta b} \|\mu_1\|}$$

We will refer to these inequalities as A and B respectively. A then implies,

$$\left| \frac{\mathbb{1}\{x \neq \varnothing\}}{1 - \varphi^0(x)} r_{\tilde{\theta}}(x) - \frac{\mathbb{1}\{x \neq \varnothing\}}{1 - \varphi^0(x)} r_{\hat{\theta}}(z) \right| \leq \frac{1}{1 - \|\varphi^0\|_\infty} e^{\|\tilde{\theta}\|b} \sqrt{\frac{C_0' \log(1/\delta)}{n_{\min}}}$$

Therefore A and B together give

$$|N^* - \hat{N}'| \leq \left| \mathbb{E}[r_{\tilde{\theta}}(Z^0)] - \frac{1}{n_0} \sum_{i=1}^{n} r_{\tilde{\theta}}(Z_i^0) \right| + \left| \frac{1}{n_0} \sum_{i=1}^{n_0} r_{\hat{\theta}'}(Z_i^0) - \frac{1}{n_0} \sum_{i=1}^{n_0} r_{\tilde{\theta}}(Z_i^0) \right|$$

$$\leq \sqrt{C_1' \frac{\log(1/\delta)}{m_{\text{eff}}}} + \frac{1}{1 - \|\varphi^0\|_\infty} e^{\|\tilde{\theta}\|b} \sqrt{\frac{C_0' \log(1/\delta)}{m_{\text{eff}}}}$$

with

$$C_1 := \frac{\sigma_{\min} \sqrt{C_0'}(\log(\alpha) + 1)}{18 M^2 e^{4\|\tilde{\theta}\|b} \|\mu_1\|}.$$

Therefore, taking $C_N' = (\sqrt{C_0'} + \exp\{\|\tilde{\theta}\|b\} \sqrt{C_1'})^2$ gives our desired result. $\square$

A.5 M-KLIEP Lower Bound

For each $\zeta \in (0, \infty)$ let's define a density f_ζ of a measure supported on $[0, 1]$ by

$$f_\zeta(y) = \frac{\zeta e^{\zeta y}}{e^\zeta - 1},$$

for $y \in [0, 1]$, and $f_\zeta(y) = 0$ for $y \notin [0, 1]$. In addition, let f_0 denote the uniform density on $[0, 1]$.

Lemma 15. *For all $\zeta \in [0, 2]$ we have*

$$\mathrm{KL}(f_0, f_\zeta) \leq \frac{\zeta^2}{20}.$$

Proof. Given any $\zeta \in (0, 2]$ we have

$$\mathrm{KL}(f_0, f_\zeta) = \int_0^1 \log\left(\frac{e^\zeta - 1}{\zeta e^{\zeta y}}\right) dy = \log\left\{\frac{\sinh(\zeta/2)}{(\zeta/2)}\right\} \leq \log\left(1 + \frac{\zeta^2}{20}\right) \leq \frac{\zeta^2}{20},$$

where the penultimate inequality uses the bound (Klén et al., 2010, Lemma 3.3 (i)). The case where $\zeta = 0$ is immediate. $\square$

Let P_0 be the uniform distribution on $A_d := [0, 1/\sqrt{2}] \times [0, 1/\sqrt{2(d-1)}]^{d-1}$, take $X^0 = Z^0 \sim P_0$ and $C_d := \sqrt{2^d \cdot (d-1)^{d-1}}$. Moreover, given $\zeta \in [0, 2]$, $w_1 \in [0, 1]$ and $\tau \in \{0, 1\}$ we write $P_1(\zeta, w_1, \tau)$ for the distribution on $X^1 = (X_j^1)_{j \in [d]}$ constructed by choosing Z^1 with density $(z_j)_{j \in [d]} \mapsto C_d \cdot f_{\tau \cdot \zeta}(\sqrt{2} z_1)$ on A_d, and choosing φ^1 so that $\varphi^1(z) = w_1$ for all $z \in [0, 1]^d$. We write $\overline{P}_{n_0, n_1}(\zeta, w_1, \tau)$ for the joint distribution on $D' := (D_0', D_1')$ where $D_0' \sim P_0^{n_0}$ and $D_1' \sim P_1(\zeta, w_1, \tau)^{n_1}$.

Lemma 16. *Given $\zeta \in [0, 2]$, $n_1 \in \mathbb{N}$ and $w_1 \in [0, 1]$ we have*

$$\mathrm{KL}\left\{\overline{P}_{n_0, n_1}(\zeta, w_1, 0), \overline{P}_{n_0, n_1}(\zeta, w_1, 0)\right\} \leq \frac{n_1(1 - w_1)\zeta^2}{20}.$$

Proof. We have

$$\mathrm{KL}\left\{P_1(\zeta, w_1, 0), P_1(\zeta, w_1, 1)\right\} = \int_{\mathcal{X}} \log\left(\frac{dP_1(\zeta, w_1, 0)}{dP_1(\zeta, w_1, 1)}\right) dP_1(\zeta, w_1, 0)$$

$$= (1 - w_1) \cdot \int_{A_d} \log\left(\frac{C_d \cdot f_0(\sqrt{2} \cdot z_1)}{C_d \cdot f_\zeta(\sqrt{2} \cdot z_1)}\right) \left(C_d \cdot f_0(\sqrt{2} \cdot z_1)\right) dz_1$$

$$= (1 - w_1) \cdot \int_{[0,1]} \log\left(\frac{f_0(y)}{f_\zeta(y)}\right) f_0(y) dy$$

$$= (1 - w_1) \cdot \mathrm{KL}(f_0, f_\zeta) \leq \frac{(1 - w_1) \cdot \zeta^2}{20},$$

where we used Lemma 15 in the final step. Hence, by the product rule for Kullback-Leibler divergence (Tsybakov, 2009, Chapter 2) we have

$$\mathrm{KL}\left\{\overline{P}_{n_0, n_1}(\zeta, w_1, 0), \overline{P}_{n_0, n_1}(\zeta, w_1, 0)\right\}$$

$$= n_0 \cdot \mathrm{KL}(P_0, P_0) + n_1 \cdot \mathrm{KL}\left\{P_1(\zeta, w_1, 0), P_1(\zeta, w_1, 1)\right\} \leq \frac{n_1(1 - w_1)\zeta^2}{20}.$$

$\square$

We are now ready to complete the proof of Theorem 3.

Proof of Theorem 3. We assume, without loss of generality, that $n_1 w_1 \leq n_0 w_0$. Suppose $Z^0 \sim P_0$ and Z^1 has density $(z_j)_{j \in [d]} \mapsto \sqrt{2} \cdot f_{\tau \cdot \zeta}(\sqrt{2} \cdot z_1)$ for some $\zeta \in [0, 1/\sqrt{2}]$, $\tau \in \{0, 1\}$ then the density ratio between the densities of Z^1 and Z^0 is proportional to $e^{\theta(\zeta, \tau)^\top z}$ on $[0, 1]^d$ where $\theta(\zeta, \tau) = (\sqrt{2} \cdot \tau \cdot \zeta, 0, \ldots, 0)^\top \in \mathbb{R}^d$, and hence $\|\theta(\zeta, \tau)\| \leq 1$. Next, we convert $\hat{\theta} = (\hat{\theta}_j)_{j \in [d]}$ into an estimator $\hat{\tau}$ by $\hat{\tau} = \mathbb{1}\{\hat{\theta}_j > \zeta/\sqrt{2}\}$. Consequently, for $\tau \in \{0, 1\}$ we have

$$\|\hat{\theta} - \theta(\zeta, \tau)\|_2 \geq |\hat{\theta}_1 - \zeta\tau| \geq \frac{\zeta}{\sqrt{2}} \cdot \mathbb{1}\{\hat{\tau} \neq \tau\}.$$

Hence, by (Tsybakov, 2009, Theorem 2.2(iii)) for at least one $\tau \in \{0, 1\}$ we have

$$\min_{\tau \in \{0,1\}} \mathbb{P}_{D' \sim \overline{P}_{n_0, n_1}(\zeta, w_1, \tau)} \left\{ \|\hat{\theta} - \theta(\zeta, \tau)\|_2 \geq \frac{\zeta}{\sqrt{2}} \right\}$$

$$\geq \min_{\tau \in \{0,1\}} \mathbb{P}_{D' \sim \overline{P}_{n_0, n_1}(\zeta, w_1, \tau)} (\hat{\tau} \neq \tau) \geq \frac{1}{4} \cdot \exp\left(-\frac{n_1(1 - w_1)\zeta^2}{20}\right).$$

To complete the proof we take

$$\zeta := \min\left\{ \sqrt{\frac{20 \log(1/(4\delta))}{\min\{n_0(1 - w_0), n_1(1 - w_1)\}}}, \frac{1}{\sqrt{2}} \right\}.$$

$\square$

A.6 Proof of Missing NP Classifier Finite Sample Bounds

Before we can proof this result we need to state a (relatively trivial) result on conditional probabilities.

Lemma 17. *Let X, Y be RVs on measurable space $(\mathcal{Z}, \mathcal{B}_\mathcal{Z})$ and define events $F \in \sigma(X)$, and $E_0, E_1 \in \mathcal{B}_\mathcal{Z}$. Suppose that $F \cap E_0 \subseteq E_1$ then*

$$\mathbb{P}(F \cap \{\mathbb{E}[\mathbb{1}_{E_0} Y | X] > \mathbb{E}[\mathbb{1}_{E_1} Y | X]\}) = 0.$$

Proof. Define the event $C := \{\mathbb{E}[\mathbb{1}_{E_0} Y | X] > \mathbb{E}[\mathbb{1}_{E_1} Y | X]\}$ then by construction, as $F \cap C \in \sigma(X)$, this gives

$$\mathbb{E}[\mathbb{1}_{E_0 \cap F \cap C} Y] = \mathbb{E}[\mathbb{E}[\mathbb{1}_{E_0} Y | X] \mathbb{1}_{F \cap C}]$$
$$\geq \mathbb{E}[\mathbb{E}[\mathbb{1}_{E_1} Y | X] \mathbb{1}_{F \cap C}]$$
$$= \mathbb{E}[\mathbb{1}_{E_1 \cap F \cap C} Y]$$

where equality holds throughout if and only if $\mathbb{P}(F \cap C) = 0$. Alternatively, as $F \cap E_0 \subseteq E_1$ we have that $F \cap E_0 \cap C \subseteq F \cap E_1 \cap C$ so that

$$\mathbb{E}[\mathbb{1}_{F \cap C \cap E_0} Y] \leq \mathbb{E}[\mathbb{1}_{F \cap C \cap E_1} Y].$$

Therefore equality holds and $\mathbb{P}(F \cap C) = 0$. $\square$

We can now go on to prove Theorem 5

Proof of Theorem 5. This proof is adapted from Tong (2013).

For notational simplicity we define functions $\hat{h}, \tilde{h} : \mathcal{Z} \to \mathbb{R}$ by

$$\hat{h}(z) := g \circ r_{\hat{\theta}}$$
$$\tilde{h}(z) := g \circ r_{\tilde{\theta}}$$

so that $R_1(\hat{\phi})$, $R_1(\tilde{\phi})$ can be re-written as

$$\hat{\phi} = \mathbb{P}(\hat{h}(Z^1) \leq \hat{C}_{\alpha, \delta, \hat{h}})$$
$$\tilde{\phi} = \mathbb{P}(\tilde{h}(Z^1) \leq C^*_{\alpha, \tilde{h}})$$

Define events G_0, G_1 by

$$G_0 := \{\tilde{h}(Z^0) > C^*_{\alpha,\tilde{h}}, \ \hat{h}(Z^0) \leq \hat{C}_{\alpha,\delta,\hat{h}}\}$$
$$G_1 := \{\tilde{h}(Z^0) \leq C^*_{\alpha,\tilde{h}}, \ \hat{h}(Z^0) > \hat{C}_{\alpha,\delta,\hat{h}}\}$$

We now have that

$$R_1(\hat{\phi}) - R_1(\tilde{\phi}) = \mathbb{E}\left[\mathbb{1}G_0 \cup G_1 \left| h(\tilde{Z}^0) - C^*_{\alpha,\tilde{h}}\right|\right] + 2C^*_{\alpha,\tilde{h}}(\mathbb{P}(\tilde{h}(Z^0) \geq C^*_{\alpha,\tilde{h}}) - \mathbb{P}(\hat{h}(Z^0) \geq \hat{C}_{\alpha,\delta,\hat{h}}))$$

For ease of notation we will define $\Delta_{m^0_{\text{eff}},\delta}$ as in Lemma 4, that is,

$$\Delta_{m^0_{\text{eff}},\delta} := \sqrt{\frac{16\log(1/\delta)}{m^0_{\text{eff}}}}$$

Now define 2 events $E_0, \ E_1$ as follows

$$E_0 := \{\|\hat{\theta} - \tilde{\theta}\| \leq \varepsilon''\}$$
$$E_1 := \{\mathbb{P}(\hat{h}(X^0)) \geq \hat{C}_{\alpha,\delta,\hat{h}}|D') \geq \alpha - 2\Delta_{m^0_{\text{eff}},\delta}\}$$

where

$$\varepsilon'' := \sqrt{\frac{C'_0\log(1/\delta)}{m_{\text{eff}}}}.$$

Theorem 2 and Lemma 4 give us that both these events occur w.p. $1 - \delta$ so that $\mathbb{P}(E_0 \cap E_1) \geq 1 - 2\delta$. We now aim to show that,

$$E_0 \cap E_1 \subseteq \{R_1(\hat{\phi}) - R_1(\tilde{\phi}) < \varepsilon'\}.$$

We immediately have that

$$E_1 \subseteq \{\mathbb{P}(\tilde{h}(Z^0) \geq C^*_{\alpha,\tilde{h}}) - \mathbb{P}(\hat{h}(Z^0) \geq \hat{C}_{\alpha,\delta,\hat{h}}|D') \leq 2\Delta_{\delta,m_{\text{eff}}}\}$$

Additionally, we note that as $\|\nabla_\theta g \circ r_\theta(z)\| \leq L$ for all z, θ, so that

$$E_0 \subseteq \{\sup_{z \in \mathcal{Z}}\|\tilde{h}(z) - \hat{h}(z)\|^2 \leq \varepsilon''L\}.$$

We now aim to bound $\hat{C}_{\alpha,\delta,g}$ above. Taking the intersection of these two events we get that

$$E_0 \cap E_1 \subseteq \{\mathbb{P}(\tilde{h}(Z^0) \geq \hat{C}_{\alpha,\delta,\hat{h}} - \varepsilon''L|\hat{C}_{\alpha,\delta,\hat{h}}) \geq \alpha - 2\Delta_{m^0_{\text{eff}},\delta}\}$$

On the other hand, the lower bound on our condition gives that $2\Delta_{m^0_{\text{eff}},\delta} > a$,

$$2\Delta_{m^0_{\text{eff}},\delta} \leq \mathbb{P}[C^*_{\alpha,\tilde{h}} < \tilde{h}(Z^0) \leq C^*_{\alpha,\tilde{h}} + (2B_0\Delta_{m^0_{\text{eff}},\delta})^{1/\gamma_0}]$$
$$= \mathbb{P}[\tilde{h}(X^0) \leq C^*_{\alpha,\tilde{h}} + (2B_0\Delta_{m^0_{\text{eff}},\delta})^{1/\gamma_0}] - \mathbb{P}[\tilde{h}(X^0) \leq C^*_{\alpha,\tilde{h}}]$$
$$= \mathbb{P}[\tilde{h}(X^0) \leq C^*_{\alpha,\tilde{h}} + (2B_0\Delta_{m^0_{\text{eff}},\delta})^{1/\gamma_0}] - (1 - \alpha).$$

Combining these two results we get that

$$E_0 \cap E_1 \subseteq \{\mathbb{P}[\tilde{h}(X^0) \geq C^*_{\alpha,\tilde{h}} + (2B_0\Delta_{m^0_{\text{eff}},\delta})^{1/\gamma_0}] \leq \alpha - 2\Delta_{m^0_{\text{eff}},\delta} \leq \mathbb{P}[\tilde{h}(X^0) \geq \hat{C}_{\alpha,\delta,\tilde{h}} - \varepsilon'L|\hat{C}_{\alpha,\delta,\hat{h}}]\},$$

and hence, $E_0 \cap E_1 \subseteq \{\hat{C}_{\alpha,\delta,g} < C^*_{\alpha,\tilde{h}} + \varepsilon'' L + (2B_0\Delta_{m^0_{\text{eff}},\delta})^{1/\gamma_0}\}$ which implies that

$$E_0 \cap E_1 \subseteq \{C^*_{\alpha,\tilde{h}} \leq \tilde{h}(X^0) \leq C^*_{\alpha,\tilde{h}} + 2\varepsilon'' L + (2B_0\Delta_{m^0_{\text{eff}},\delta})^{1/\gamma_0}\} =: G'_0.$$

We can now use this to bound the expectation. Indeed using Lemma 17 we have that

$$E_0 \cap E_1 \subseteq \{\mathbb{E}[\mathbb{1}\{G_0\}|\tilde{h}(Z^0) - C^*_{\alpha,\tilde{h}}| \,|D'] \leq \mathbb{E}[\mathbb{1}\{G'_0\}|\tilde{h}(Z^0) - C^*_{\alpha,\tilde{h}}||D']\}$$

We we also have

$$\begin{aligned}
\mathbb{E}[\mathbb{1}\{G'_0\}|\tilde{h}(Z^0) - C^*_{\alpha,\tilde{h}}||D'] &\leq \left(2\varepsilon'' L + (2B_0\Delta_{m^0_{\text{eff}},\delta})^{1/\gamma_0}\right)\mathbb{P}(G'_0) \\
&\leq B_2(2\varepsilon'' L + (2B_0\Delta_{m^0_{\text{eff}},\delta})^{1/\gamma_0})^{\gamma_1+1}
\end{aligned}$$

with

$$B_2 := \min\left\{\frac{1}{B_0 a^{\gamma_1/\gamma_0}}, B_1\right\}$$

where the final step uses our probability lower bound.

By an identical argument again

$$E_0 \cap E_1 \subseteq \{\mathbb{E}[\mathbb{1}_{G_1}|\tilde{h}(Z^0) - C^*_{\alpha,\tilde{h}}| \,|D'] \leq B_2(2\varepsilon'' L + (2B_0\Delta_{m^0_{\text{eff}},\delta})^{1/\gamma_0})^{\gamma_1+1}\}$$

and so we get

$$E_0 \cap E_1 \subseteq \{R_1(\hat{\phi}) - R_1(\tilde{\phi}) \leq 2B_2(2\varepsilon'' L + (2B_0\Delta_{m^0_{\text{eff}},\delta})^{1/\gamma_0})^{\gamma_1+1} + 2C^*_{\alpha,\tilde{h}}\Delta_{\delta,m^0_{\text{eff}}}\}.$$

Taking $C_2 := 4B_2(L \vee B_0)(\gamma_1 + 1) + 2C^*_{\alpha,\tilde{h}}$ therefore gives $E_0 \cap E_1 \subseteq \{R_1(\hat{\phi}) - R_1(\tilde{\phi}) \leq \varepsilon'\}$ and hence,

$$\begin{aligned}
\mathbb{P}(R_1(\hat{\phi}) - R_1(\tilde{\phi}) \leq \varepsilon) &\leq \mathbb{P}(R_1(\hat{\phi}) - R_1(\tilde{\phi}) \leq \varepsilon, \, E_0 \cap E_1) \\
&= \mathbb{P}(E_0 \cap E_1) \\
&= 1 - 2\delta.
\end{aligned}$$

$\square$

B   SUPPLEMENTARY METHODS

B.1 f-Divergence based DRE

We now present another example DRE which relates to the f-Divergence. First we state the following theorem on which these approaches are built

Theorem 18. *Let $f : (0,\infty) \to \mathbb{R}$ be convex and lower-semicontinuous function and define $f' : \mathbb{R}_+ \to \mathbb{R}$ to be the derivative of f. Firstly, there exists unique $f^* : A \to \mathbb{R}$ with $A \subseteq \mathbb{R}$ by*

$$f^*(t) := \sup_{u \in (0,\infty)} ut - f(u)$$

Secondly,

$$\underset{T \in \mathcal{T}}{\text{argmin}} \; \mathbb{E}[T(Z^1)] - \mathbb{E}[f^*(T(Z^0))] = T^*$$

where, $T^* := f' \circ r^*$ and $\mathcal{T}$ is some set of non-negative functions containing T^*.

As a direct result of this, for any set of positive real values functions $\mathcal{G}$ containing r^* we have that

$$r^* := \underset{r \in \mathcal{G}}{\text{argmax}} \; \mathbb{E}[f'(r(Z^1))] - \mathbb{E}[f^*(f'(r(Z^0))]$$

We can then approximate this using D to get

$$\frac{1}{n_1} \sum_{i=1}^{n_1} f'(r(Z_i^1)) - \frac{1}{n_0} \sum_{i=1}^{n_0} f^*(f'(r(Z_i^0)))$$

Note that while we have theoretical guarantees of consistency in the case of correctly specified $\mathcal{G}$, we have much less justification for this approach in the case of incorrectly specified $\mathcal{G}$. This is simply because from a heuristic perspective, we have no idea if

$$\tilde{r} := \underset{r \in \mathcal{G}}{\text{argmax}} \; \mathbb{E}[f'(r(Z^1))] - \mathbb{E}[f^*(f'(r(Z^0)))]$$

is sensible approximation of r^* when $\mathcal{G}$ is miss-specified. As a result, one could argue that KLIEP is a more principled approach as in the objective we are approximating we choose some notion of the "closest" $r \in \mathcal{G}$ to r^*.

We now go on to expand this process in the case of 2 popular choice of f

JS-Divergence

If we take $f(u) := u \log(u) - (u+1) \log\left(\frac{1+u}{2}\right)$ then the associated f-divergence is the JS-Divergence. The corresponding $f' : (0, \infty) \to \mathbb{R}$, $f^* : (\log(2), \infty) \to \mathbb{R}$ are then given by

$$f'(t) := \log\left(\frac{2t}{1+t}\right) \qquad\qquad f^*(t) := -\log(2 - \exp(t))$$

giving us that

$$r^* := \underset{r \in \mathcal{G}}{\text{argmax}} \; \mathbb{E}[f'(r(Z^1))] - \mathbb{E}[f^*(f'(r(Z^0)))]$$

$$= \underset{r \in \mathcal{G}}{\text{argmax}} \; \mathbb{E}\left[\log\left(\frac{2r(Z^1)}{1+r(Z^1)}\right)\right] + \mathbb{E}\left[\log\left(\frac{2}{1+r(Z^0)}\right)\right]$$

$$= \underset{r \in \mathcal{G}}{\text{argmax}} \; \mathbb{E}\left[\log\left(\frac{r(Z^1)}{1+r(Z^1)}\right)\right] + \mathbb{E}\left[\log\left(\frac{1}{1+r(Z^0)}\right)\right]$$

for correctly specified $\mathcal{G}$. The form that our approximation for this take is highly similar to the form of the logistic regression approach and is in fact identical when $n_1 = n_0$. The only difference being that this approach adjusts for the class imbalance directly in the estimator while the logistic regression approach adjusts for it after the fact.

KL-Divergence

If we take $f(t) := t \log(t)$ then the corresponding f-divergence is the KL-divergence. This choice of f has corresponding $f^* : \mathbb{R} \to \mathbb{R}$, $f' : (0, \infty) \to R$ defined by

$$f^*(t) := \exp\{t - 1\} \qquad\qquad f'(t) := 1 + \log(t).$$

Thus we get that

$$r^* := \underset{r \in \mathcal{G}}{\text{argmax}} \; \mathbb{E}[f'(r(X^1))] - \mathbb{E}[f^*(f'(r(X^0)))]$$

$$= \underset{r \in \mathcal{G}}{\text{argmax}} \; \mathbb{E}[\log(r(X^1))] - \mathbb{E}[r(X^0)].$$

Note the objective of this estimator differs from KLIEP as $\mathbb{E}[r(X^0)]$ is not logged.

B.1.1 Adaptations to MNAR Data

Adapting this to work with D' simply equates to modifying the objective to be

$$\frac{1}{n_1}\sum_{i=1}^{n_0}\frac{\mathbb{1}\{X_i^1 \neq \varnothing\}}{1-\varphi^1(X_i^1)}f'(X_i^1) - \frac{1}{n_0}\sum_{i=1}^{n_0}\frac{\mathbb{1}\{X_i^0 \neq \varnothing\}}{1-\varphi^0(X_i^0)}f^*(f'(X_i^0)).$$

B.2 Original Neyman-Pearson Classification

We now describe the Neyman-Pearson classification procedure laid out in (Tong et al., 2018). This procedure is describe in Algorithm 2 and constructs a classifier from our data. We assume we have n_0 additional copies of Z^0 which we label $Z_{n_0+1}^0, \ldots, Z_{2n_0}^0$ however the algorithm adapts to any number of samples from either distribution. We also let $g : \mathbb{R} \to \mathbb{R}$ be any strictly increasing function. Note that the classifier produced by the algorithm does not depend on g.

Algorithm 2 Neyman-Pearson Classification Procedure

1: Use $\{Z_i^1\}_{i=1}^{n_1}, \{Z_i^0\}_{i=1}^{n_0}$ to estimate r^* with $\hat{r}$ by any DRE procedure.

2: Set $i^* = \min\{i \in \{1, \ldots, n_0\}|\mathbb{P}(W \geq i) \leq \delta\}$ where $W \sim \text{Binomial}(n_0, 1-\alpha)$

3: For $i \in \{1, \ldots, n_0\}$ compute $\hat{r}_i := g \circ \hat{r}(Z_{n_0+i}^0)$

4: Sort $\hat{r}_1, \ldots, \hat{r}_{n_0}$ in increasing order to get $\hat{r}_{(1)}, \ldots, \hat{r}_{(n_0)}$ with $\hat{r}_{(i)} \leq \hat{r}_{(i+1)}$

5: Define $\hat{C}_{\alpha,\delta,g\circ\hat{r}} := \hat{r}_{(i^*)}$

6: Define the classifier $\hat{\phi}_D$ by

$$\hat{\phi}_D(z) = \mathbb{1}\{g \circ \hat{r}(z) > C_{\alpha,\delta,g\circ\hat{r}}\} \quad \text{for all } z \in \mathcal{Z}.$$

Proposition 19. *Algorithm 2 constructs a Neyman-Pearson classifier satisfying Equation (9).*

Proof. As the algorithm does not depend on g we take g as the identity in the proof. We adapt the proof given in (Tong et al., 2018). Let $\hat{r} \in \mathcal{H}$ and define A to be the event

$$A := \{\mathbb{P}(\hat{\phi}_D(Z^0) = +|D) > \alpha\}.$$

We then aim to show that $\mathbb{P}(A|\hat{r}) \leq \delta$. To this end let $C_{\alpha,\hat{r}}^* := \inf\{C \in \mathbb{R}|\mathbb{P}(\hat{r}(Z^0) > C|\hat{r}) \leq \alpha\}$.

Now as $\mathbb{P}(\hat{r}(Z^0) > C|\hat{r})$ is a right continuous decreasing function we have that $\mathbb{P}(\hat{r}(Z^0) > C_{\alpha,\hat{r}}|\hat{r}) \leq \alpha$.

Thus we can re-write A

$$\begin{aligned} A &= \{\hat{C} < C^*\} \\ &= \{\hat{r}_{(i^*)} < C^*\} \\ &= \{i^* \text{ or more of } \hat{r}_1, \ldots, \hat{r}_{n_0} \text{ are less than } C^*\} \\ &= \{\sum_{i=1}^{n_0} B_i \geq i^*\} \quad \text{where } B_i := \mathbb{1}\{\hat{r}_i < C^*\}. \end{aligned}$$

Hence we have

$$\mathbb{P}(A|\hat{r}) = \mathbb{P}(\sum_{i=1}^{n_0} B_i \geq i^*|\hat{r}).$$

As $Z_{n_0+1}^0, \ldots, Z_{2n_0}^0$ are independent, we have that $B_1, \ldots, B_{n_0}$ are independent given $\hat{r}$. Furthermore, we have that $q := \mathbb{P}(B_i = 1) \leq 1 - \alpha$. If we now define $W' \sim \text{Binomial}(n_0, q)$, $W \sim \text{Binomial}(n_0, 1-\alpha)$ we get that.

$$\begin{aligned} \mathbb{P}(A|\hat{r}) &= \mathbb{P}(W' \geq i^*) \\ &\leq \mathbb{P}(W \geq i^*) \quad \text{as } q \leq 1 - \alpha \\ &\leq \delta \quad \text{by definition of } i^*. \end{aligned}$$

Finally, as $\mathbb{P}(A|\hat{r}) \leq \delta$ for all non-negative measurable $\hat{r}$ we have that $\mathbb{P}(A) \leq \delta$. $\qquad\square$

B.3 Logistic Regression Re-adjustment

Here we describe how to use the adjusted logistic regression described in King and Zeng (2001) to learn the φ under the scenario presented in Section 5.1.

First define $I = \{i \in [n] | z_i \text{ is observed}\}$ so that for $i \in I$ either $X_i \neq \varnothing$ or we have queried Z_i. We then refer to $\{(W_i, Z_i)\}_{i \in I}$ as our set of fully observed samples. These fully observed samples will have disproportionally few $W_i = 1$ however $\{Z_i | W_i = 1, i \in I\}$, $\{Z_i | W_i = 0, i \in I\}$ still give exact samples from $Z | W = 1$, $Z | W = 0$ respectively. King and Zeng (2001) propose a re-adjustment to logistic regression specifically for this case in which the true proportion each response are different to the proportion in your given data.

Let $\hat{\beta}_0, \hat{\beta}_1$ be the logistic regression estimators for the coefficients of the intercept and slope respectively from our fully observed samples. Now let $\hat{\tau}$ be some consistent estimate of $\mathbb{P}(W = 1)$ and $\hat{\tau}_I = \frac{\sum_{i \in I} \mathbb{1}\{W_i = 1\}}{|I|}$. Now define

$$\hat{\beta}_0' = \hat{\beta}_0 - \log \left\{ \frac{(1 - \hat{\tau})\hat{\tau}_I}{\hat{\tau}(1 - \hat{\tau}_I)} \right\}$$

Then $\hat{\beta}_0'$ and $\hat{\beta}_1$ are consistent estimates of the logistic regression parameters. In our case we will simply take $\hat{\tau} = \frac{\sum_{i \in [n]} w_i = 1}{n}$. This allows us to simplify and get

$$\hat{\beta}_0' = \hat{\beta}_0 - \log \{m/n_1\}$$

where $n_1 = \sum_{i \in [n]} \mathbb{1}\{w_i = 1\}$ and $m = \sum_{i \in I} \mathbb{1}\{w_i = 1\}$ that is, the number of observations we have queried.

C SYNTHETIC EXPERIMENT DETAILS

C.1 5-dimensional Correctly Specified Case

For setting we take Z_1, Z_0 to have PDFs defined by

$$p_1(z) = N(z; \mu_1, I)$$
$$p_0(z) = N(z; \mu_0, I)$$

with $\mu_0 = 0$, $\mu_1 = (0.1, 0.1, 0.1, 0.1, 0.1)^\top$ and where $N(z; \mu, \Sigma)$ is the PDF of a multivariate normal distribution with mean μ and variance Σ evaluated at z. We take $\varphi^1(x) = \frac{1}{2}\mathbb{1}\{x^\top a > 0\}$, $\varphi^0 = 0$ where $a = (1, 1, 1, 1, 1)^\top$.

To estimate the density ratio, we use $r_\theta(x) := \exp\{\theta^\top x\}$ making the model correctly specified with "true" parameter $\tilde{\theta} := -\mu_1$. for given $n \in \mathbb{N}$, n IID samples are drawn from both X^0 and X^1. M-KLIEP and CC-KLIEP are then fit using gradient simple gradient descent to obtain parameter estimates. This process is repeated 100 times for each n and with $n \in \{100, 200, 300, \dots, 1500\}$.

C.2 Mixed Gaussian Neyman-Pearson Case

For this experiment we take Z^1 and Z^0 to be 2-dimensional Gaussian mixtures with the following densities

$$p_1(z) = \frac{1}{2} N \left(z; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, I \right) + \frac{1}{2} N \left(z; \begin{pmatrix} -1 \\ 4 \end{pmatrix}, I \right)$$

$$p_0(z) = \frac{1}{2} N \left(z; \begin{pmatrix} 1 \\ 0 \end{pmatrix}, I \right) + \frac{1}{2} N \left(z; \begin{pmatrix} 0 \\ 4 \end{pmatrix}, I \right)$$

Now let $\varphi^1 = 0.9\mathbb{1}_{Z^\top a > 2}(Z)$ where $a = (0, 1)^\top$, $\varphi^0 = 0$ and define X^1, X^0 on $\mathbb{R}^2 \cup \{\varnothing\}$ as before. For given $n \in \mathbb{N}$ we generate n samples from X^0, X^1 and use these to estimate $\tilde{r}$ by both M-KLIEP and CC-KLIEP. An additional $Z^0 = X^0$ and these used to produce a classifier via Algorithm 2 with $\alpha = \delta = 0.1$. These samples are also used to produce a classifier via Algorithm 2 with $\hat{r}$ replaced by r^*, the true density ratio. One instance of this procedure with $n = 500$ is presented in Figure 2

To estimate the power of the produced classifiers, 1,000,000 samples from Z^1 are produced and the proportion classified as 1 recorded. This process is then repeated for 100 times for each $n \in \{100, 200, 300, \dots, 1500\}$. The estimated powers of the classifiers produced from these iterations is then used to estimate the expected power of the procedures alongside 99% confidence intervals
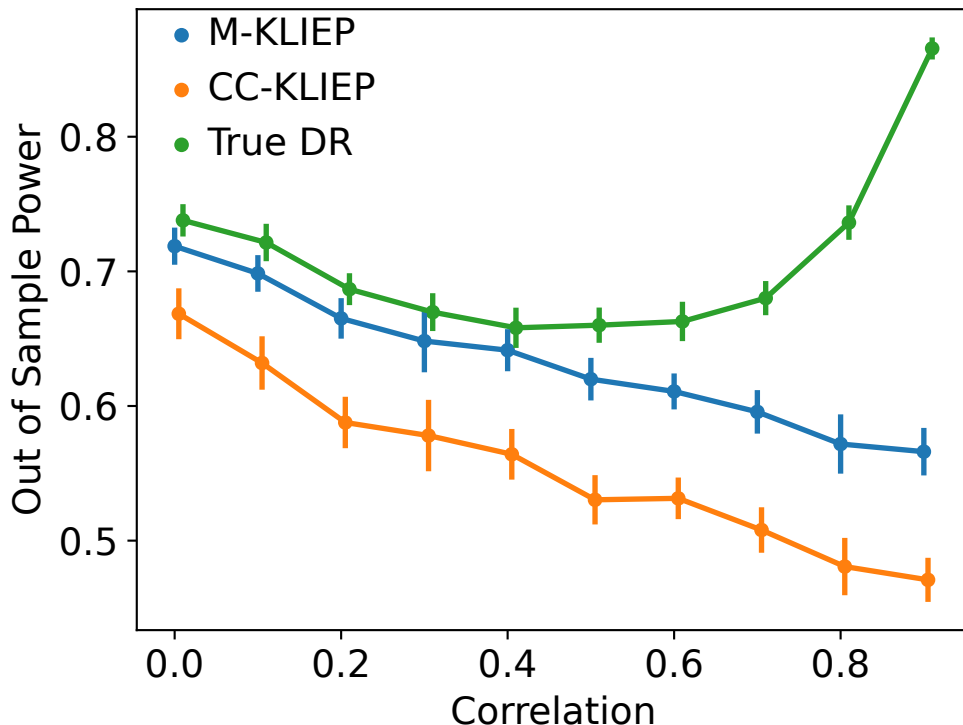
Figure 6: The expected power of the Naive Bayes NP classifier for various levels of correlation between features.

D   ADDITIONAL SYNTHETIC EXPERIMENTS

D.1   Naive Bayes Assumption Test

We wanted to test the effect of our Naive Bayes assumption on our ability to estimate the density ratio. For this experiment we take $Z_0, \ Z_1$ to be distributed as follows:

$$Z_1 \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$
$$Z_0 \sim N \left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

for varying $\rho \in [0, 1]$. To test our Naive Bayes assumption alongside our M-KLIEP approach we also induced non-uniform missingness separately in the features of the class 0 data. The missingness functions used were $\varphi_1^1(x) = 0.8 \cdot \mathbb{1} \{x > 0\}$, $\varphi_2^1(x) = 0.8 \cdot \mathbb{1} \{x < 0\}$ and then no missingness for class 0 i.e. $\varphi_1^0 \equiv \varphi_2^0 \equiv 0$.

100 samples from X_0, X_1 were then generated, a density ratio fit through the Naive Bayes assumption, and then 100 more samples from X_0 (equivalent to Z_0) were used to produce an NP classifier. Figure 6 shows you the average power of the classifier alongside 95% C.I.s from 100 montecarlo simulations.

As we can see the Naive Bayes approaches only start seriously deviating from the performance of the true density at around $\rho = 0.5$ We also reassuringly see that M-KLIEP consistently outperforms CC-KLIEP.
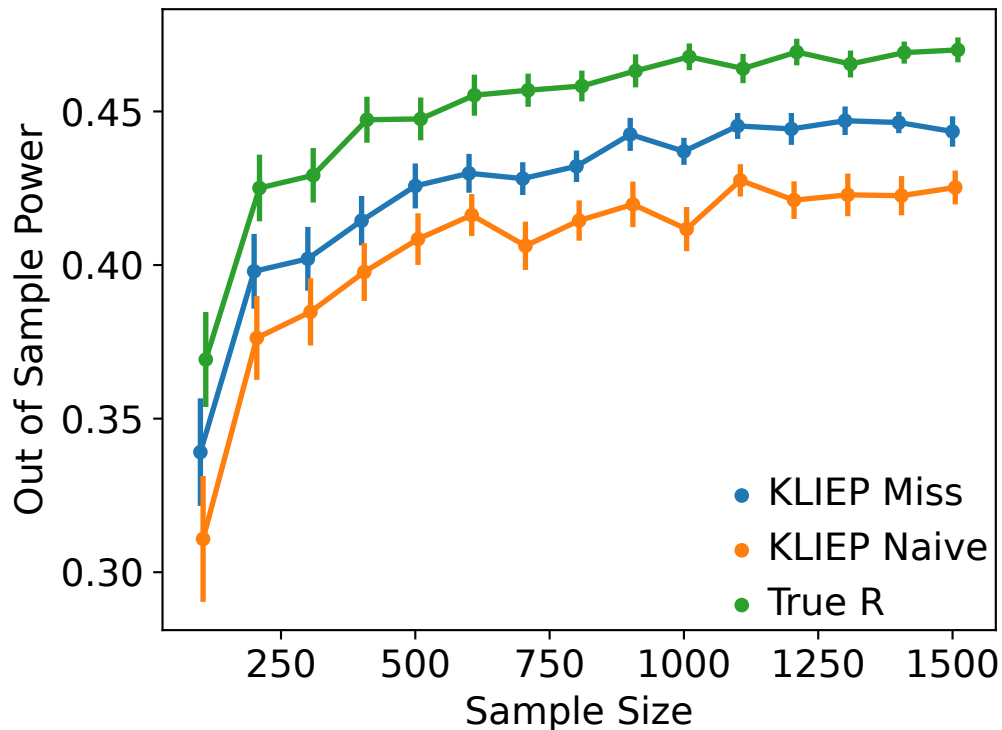
Figure 7: Expected power of NP classifier when used in conjunction with multiple DRE approaches for varying sample size. In this case our model is misspecified.

D.2 Differing Variance Misspecified Test

We have tried an additional misspecified case set up as follows

$$Z_1 \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

$$Z_0 \sim N\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\right)$$

We then introduced complete missingness in class 1 with $\varphi^1(x) = 0.8 \cdot \mathbb{1}\{x_1 > 0\}$ and no missingness in class 0 ($\varphi^0 \equiv .$) We then used the log-linear form for our density ratio estimate with $f(x) = x$ which leads to an incorrectly specified model ($f(x) = (x^\top, x^{2^\top})^\top$ would lead to a correctly specified model.)

For various values of n, we then generate n samples from X_0, X_1 and a density ratio was then fit using M-KLIEP and CC-KLIEP. An additional n samples from X_0 (equivalent to Z_0) were used to produce an NP classifier.

Figure 7 shows one simulation in this case alongside the classification boundaries produced. We can see that with this misspecified parametric boundary our model will always approximate the true boundary relatively crudely however we can clearly see that the M-KLIEP boundary is a more sensible approximation than the CC-KLIEP boundary

Figure 8 shows you the estimated expected power of the classifiers produced alongside 95% C.I.s from 100 montecarlo simulations (where 1,000,000 additional samples from Z_1 were used to estimate the power of the classifier in each simulation.) This definitively shows M-KLIEP has better performance than CC-KLIEP. While M-KLIEP does perform worse than the true classification boundary this decrease is not drastic.
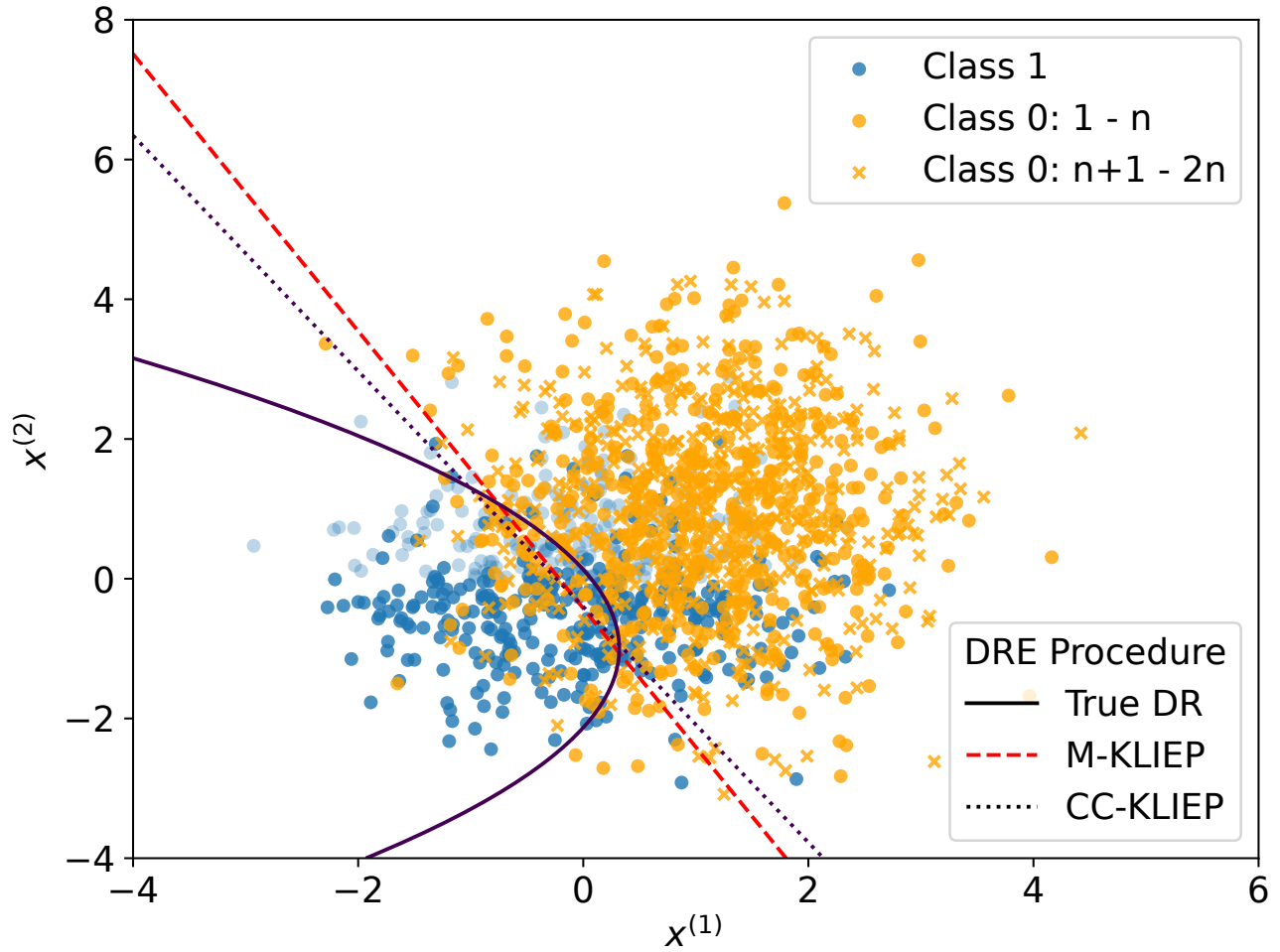
Figure 8: Scatter plot of D alongside classification boundaries produced from corresponding D' via various procedures. If $X_i = \varnothing$ then Z_i is faded out.
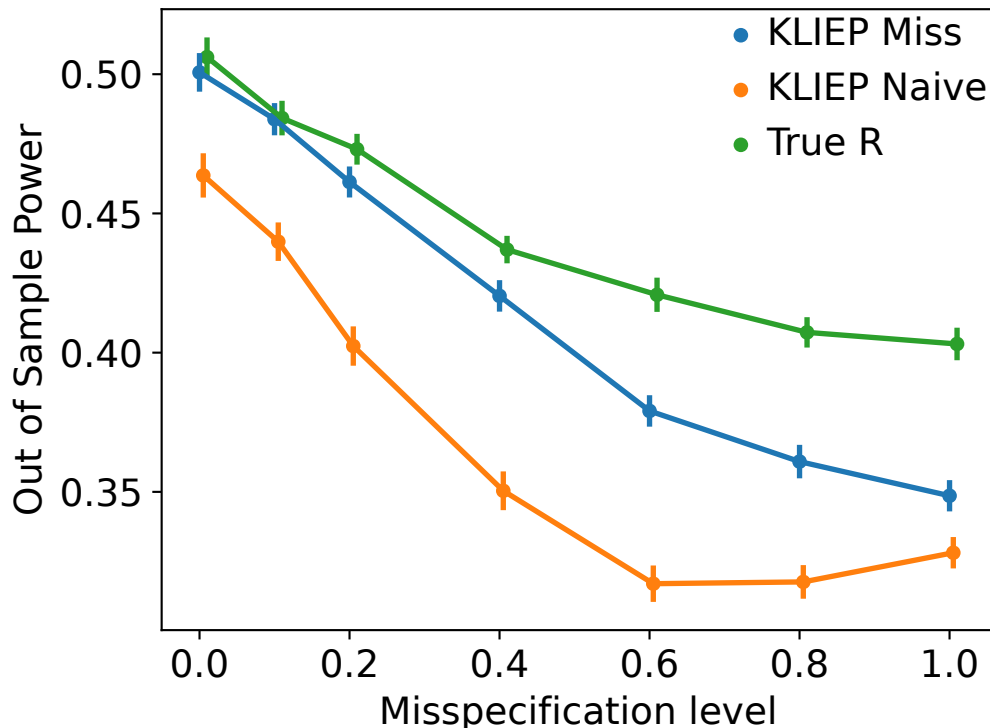
Figure 9: Expected power of NP classifier produced via various DRE techniques for varying levels of misspecification. We take misspecification to be $2 * \rho$.

D.3 Varying Misspecificaiton Level

In this experiment we wanted to examine the effect of the level of misspecification on our model. To do this we let Z_1, Z_0 have pdfs defined as follows

$$p_1(z) = (1 - \rho)N\left(z; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, I\right) + \rho N\left(z; \begin{pmatrix} 2 \\ 0 \end{pmatrix}, I\right)$$

$$p_0(z) = \frac{1}{2}N\left(z; \begin{pmatrix} 1 \\ 0 \end{pmatrix}, I\right)$$

for various $\rho \in [0, 0.5]$. We then introduced complete missingness in class 1 with $\varphi^1(x) = 0.8 \cdot \mathbb{1}\{x_1 > 0\}$ and no missingness in class 0 ($\varphi^0 \equiv$.) We then used the log-linear form for our density ratio estimate with $f(x) = x$ which leads to an incorrectly specified model for $\rho > 0$ with larger values of ρ leading to greater levels of misspecification.

We then generate 100 samples from X_0, X_1 and a density ratio was fit using M-KLIEP and CC-KLIEP. An additional 100 samples from X_0 (equivalent to Z_0) were used to produce an NP classifier. Figure 9 shows you the average power of the classifier alongside 95% C.I.s from 100 montecarlo simulations (where 1,000,000 additional samples from Z_1 were used to estimate the power of the classifier in each simulation.)

E   REAL WORLD EXPERIMENT DETAILS

Here we give additional information on the experiments performed on the real world data. We start off by describing each of the three datasets.

E.1   Datasets

E.1.1   CTG

Foetal Cardiotocograms (CTGs) measure the babies heart function during labour. These are then used by doctors and nurses and then used as a diagnostic tool to assess the foetus' health. In our data-set we have numerical summaries of foetal CTGs which alongside the diagnoses of one of "normal", "suspect", or "pathologic" that doctors ascribed to associated

foetuses. We aim to create an automated diagnosis procedure which will determine from the summaries whether the foetus is "normal" (class 1) and therefore needs no additional attention/ intervention, or one of "suspect" or "pathologic" (Class 0) to and hence needing further follow up. It is clear that NP classification as a well suited classification procedure to this paradigm as we would like to strictly control the probability of miss-classifying a "suspect" or "pathologic" baby as "normal".

The data is taken from: https://archive.ics.uci.edu/ml/datasets/cardiotocography and was first presented in de Campos et al. (2000).

This data contains observations of 2129 foetuses, 1655 from class 1 and 474 from class 0. From the data we select 10 features which we describe below

We split the data as follows, 237 & 1555 samples from Class 0 & 1 to fit $\hat{r}$, 237 samples from Class 0 are used to fit the threshold of the classifier, and 100 samples from Class 1 are used to estimate the power of the classifier. When φ are learned we query 10 missing samples from each feature.

E.1.2 Fire

For multiple different fires and non-fires, various different environmental readings were taken such as Temperature, Humidity and CO2 concentration. The aim of this is to be able to classify whether or not a fire is present to create a sort of IOT (internet of things) smoke detector to detect the presence of fire. Again NP classification is clearly a good fit as falsely detecting a fire is far less damaging than missing a fire. As such, we take the presence of fire as our error controlled class (Class 0) as we want to strictly control the probability of not detecting a fire which is present. This data is taken from: https://www.kaggle.com/datasets/deepcontractor/smoke-detection-dataset.

This dataset consists of 62,630 observations of 12 features. The observations consist of 17,873 observations from Class 1 and 44,757 from Class 0. Before carrying out any of our procedures we perform minimal feature manipulation by trimming extreme values for some of the features as described below:

- **TVOC(ppb), eCO2(ppm)**: Values are trimmed to not exceed 1000.

- **PM1.0, PM2.5**: Values are trimmed to not exceed 1.

- **NC0.5, NC2.5**: Values are trimmed to not exceed 5.

E.1.3 Weather

A dataset containing various weather reading for different days in Australia collected with the aim of using the previous days weather to predict whether there is a chance on the following day. We want to be able to choose what we mean by a "chance" in terms of the probability of our classifier correctly predicting rain when it is present therefore motivating the use of NP classification. As such we take the event of rain the following day to be our error controlled class (Class 0.) This data is taken from: https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package.

This dataset consists of 142,193 observations of 20 features. The observations consist of 110,316 observations from Class 1 and 31,877 from Class 0.

Before carrying out any of our procedures we perform minimally feature manipulation by trimming extreme values for some of the features. Factor variables are split into indicator variables for each of the possible outcomes, with one outcome having no indicator to avoid redundancy. Continuous features are trimmed to be within 5 IQR (inter-quartile range) of the median. This data manipulation lead to a total of 62 dimensions in our final data.

E.2 Experimental Designs

For each of the datasets, the same overall experimental design is used. Firstly data is randomly split into train (Class 0 & 1)/calibrate (Class 0)/test (Class 1) datasets of sizes given below. Any missing values originally in the data set are then imputed using simple mean imputation. Next the data is normalised to reduce the risk of numerical precision issues within our algorithms. The Class 0 training data is then corrupted along each dimension using φ_j of the form $\varphi_j(z) := (1 + \exp\{\tau_j(a_{0,j} + a_{1,j}z)\})^{-1}$. This φ_j is then estimated by querying points as described in Section 5.1. The density ratio is then fit by M-KLIEP, CC-KLIEP on the training data. Alongside this, we also fit the density ratio using M-KLIEP with the true φ_j and KLIEP with the original non-corrupted training data as benchmarks. A classifier is then constructed from these estimated density ratios using the calibration data by the procedure described in Algorithm 2 with given α, δ. The testing data is then used to estimate the power of these classifiers.

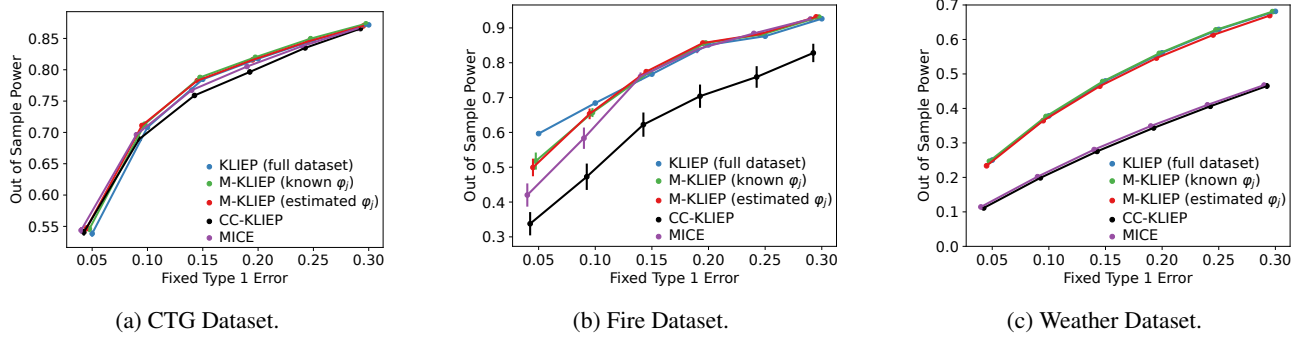(a) CTG Dataset.　　　　(b) Fire Dataset.　　　　(c) Weather Dataset.

Figure 10: Out of sample power with pseudo 95% C.I.s for various different target Type I errors with iterative imputation approach included.

Multiple iterations of this process are performed with random train/calibrate/test splits (of fixed size) and random $\tau_j \in \{-1, 1\}$ for each iteration. The estimated powers from these iterations are then used to calculate the pseudo expected power and 95% C.I.s

For experiment 1, we vary $\alpha \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ with $\delta = 0.05$, $a_{j,0} = -\mu_j/\sigma_j$, $a_{j,1} = 1/\sigma_j$ where μ_j, and σ_j are the sample mean and variance of feature j.

For experiment 2 this process is repeated with varying $a_{j,1}$ with $\alpha = 0.1, \delta = 0.05, a_{j,1} = \sigma_j$. The $a_{j,0}$ are chosen to produce the following proportions of missing values for each feature: $\{0.1, 0.2, 0.3, \ldots, 0.9\}$.

Dataset specific experimental design information can be found below

E.2.1　Dataset specific experimental design

CTG　The size of the data splits are as follows, Class 0 train = 237, Class 1 train = 1555, Class 0 calibrate = 237, and Class 1 test 100. We query 10 missing samples from each feature when learning φ_j. For both experiments 1,000 iterations are run.

Fire　The sample sizes of the data splits are as follows, Class 0 train = 20,000, Class 1 train = 12,873, Class 0 calibrate = 24,757, and Class 1 test = 5,000. We query 50 missing samples from each feature when learning φ_j. For both experiments, 100 iterations are run.

Weather　The size of the data splits are as follows, Class 0 train = 15,000, Class 1 train = 100,316, Class 0 calibrate = 16,877, and Class 1 test = 10,000. We query 50 missing samples from each feature when learning φ_j. For experiment 1, 100 iterations are run; for experiment 2, 30 iterations are run.

E.3　Additional Missing Data Approaches

We also compared our approached to an iterative imputation approach implemented using the `sklearn.impute.IterativeImputer` module in Python using the default estimator `BayesianRidge()`. For fair comparison, this imputed data was then used to perform Naive Bayes DRE to make it comparable to our approach. Results of these additional experiments are given in figures 10 & 11 and described below.

E.3.1　Additional Results

Here we present the same results as the paper but with the iterative imputation method included. Figure 10 shows the out of sample power of various NP classifiers for varying Type I error we can see that the additional iterative imputation method seems to perform comparably or worse than our method for each data set

Figure 11 shows the out of sample power of various NP classifiers for varying levels of missingness. Here we see that as the level of missingness increases, the iterative imputation method degrades significantly leading it to perform significantly worse than our approach. It does however consistently perform better than the complete case approach

F　SOCIETAL IMPACT

We now briefly discuss the societal impact of our work. As we have shown, ignoring MNAR data can lead to degradation in performance. Perhaps worse than this however is that, without taking account of our MNAR structure, use of the original
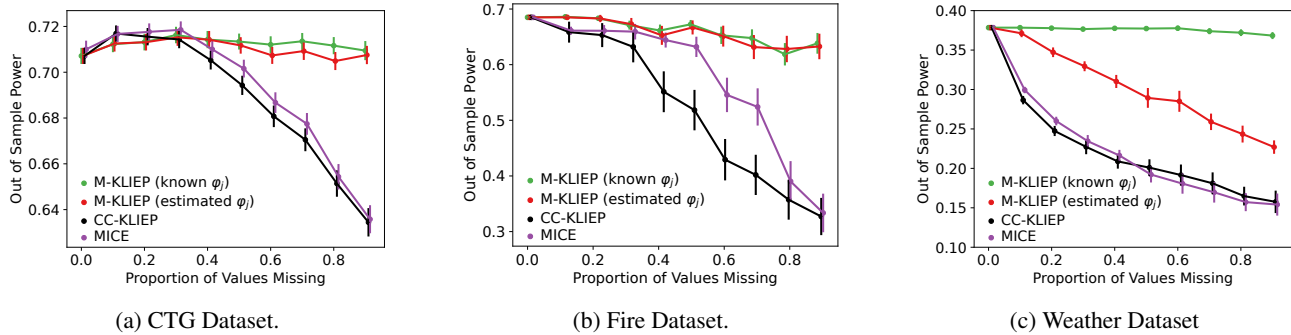
(a) CTG Dataset.   (b) Fire Dataset.   (c) Weather Dataset

Figure 11: Out of sample power with pseudo 95% C.I.s for various φ and varying missing proportions with iterative imputation approach included.

NP classification procedure won't guarantee our desired Type I error with high probability. This could lead to classifiers which perform far worse on our error controlled class than estimated. The impact of this could be serious in the case of say disease detection where we believe our classifier to be detecting a far higher proportion of diseased individuals than it truly is. The impact of MNAR data on analysis has been explored before in more general settings (Rutkowski, 2011; Padgett et al., 2014; Goldberg et al., 2021) with the key takeaway always being that ignoring the structure of the missingness can negatively impact your results and lead to false inference.

References

de Campos, D. A., Bernardes, J., Garrido, A., de sá, J. M., and Pereira-leite, L. (2000). Sisporto 2.0: a program for automated analysis of cardiotocograms. *The Journal of maternal-fetal medicine*, 9 5:311–8.

Goldberg, S. B., Bolt, D. M., and Davidson, R. J. (2021). Data missing not at random in mobile health research: Assessment of the problem and a case for sensitivity analyses. *J Med Internet Res*, 23(6):e26749.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2):137–163.

Klén, R., Visuri, M., and Vuorinen, M. (2010). On jordan type inequalities for hyperbolic functions. *Journal of Inequalities and Applications*, 2010:1–14.

Kohler, J. M. and Lucchi, A. (2017). Sub-sampled cubic regularization for non-convex optimization. In *ICML*.

Padgett, C. R., Skilbeck, C. E., and Summers, M. J. (2014). Missing data: The importance and impact of missing data from clinical research. *Brain Impairment*, 15(1):1–9.

Rutkowski, L. (2011). The impact of missing background data on subpopulation estimation. *Journal of Educational Measurement*, 48(3):293–312.

Tong, X. (2013). A plug-in approach to neyman-pearson classification. *Journal of Machine Learning Research*, 14(92):3011–3040.

Tong, X., Feng, Y., and Li, J. J. (2018). Neyman-pearson classification algorithms and np receiver operating characteristics. *Science Advances*, 4(2):eaao1659.

Tropp, J. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer: Paris.