

$$X_i^e = g(X_{S_i}^e) + \epsilon^e, \text{ for some } i \in \{1, \dots, m\}, \quad (1)$$

$$\hat{Y}_i^{\text{test}} =: \mathbb{E}_{\mathcal{P}_e}[X_i|X_+,] - \mathbb{E}_{\mathcal{P}_e}[X_i|X_+, V = 0]$$

The noise variables ϵ_Y and ϵ_3 are i.i.d. $\mathcal{N}(0, \sigma^2)$. Suppose particular $e \in \mathcal{E}$ becomes difficult as β_1^e and μ_2^e vary with

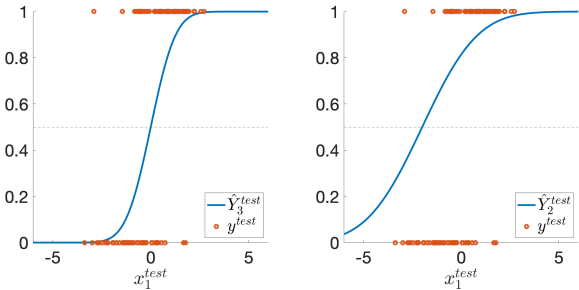


Fig. 1: Comparisons of \hat{Y}_3^{test} (left) and \hat{Y}_2^{test} (right), where

$$\mathbb{E}_{\mathcal{P}_e}[X_{\alpha}|X_+=x_+]-\mathbb{E}_{\mathcal{P}_e}[X_{\alpha}|X_+=x_+,V=0]$$

Definition 1. For $k \in \{1, \dots, m\}$, $S \subseteq \{1, \dots, m\} \backslash k$, and $h(X_S, Y) := \mathbb{E}_{\mathcal{P}_e}[X_k|X_S, Y]$, the pair (k, S) satisfies the

$$\mu(\wedge_S, \top) = \mu(\wedge_S, \bot) \quad , \tag{8}$$

- 1) $X_k^e = g(X_R^e, Y^e) + \epsilon^e$ as in (1) ,
- 2) $X_Q^e \perp\!\!\!\perp X_k^e \mid (X_R^e, Y^e)$.

$$\mathbb{E}_{\mathcal{P}_e}[Y|\phi_e(X) = (x_Q, x_R, z)]$$

(9)

□

$$\left\{ \begin{array}{l} X_1^e := f_1^e(X_{PA(X_1^e)}^e, \, \epsilon_1^e), \end{array} \right. \tag{10}$$

$$\text{all } i \in \{0,\ldots,m\}.$$

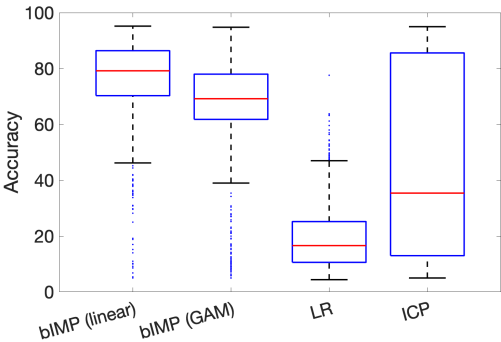
$$\text{some } i \in \{0,\ldots,m\}, \text{ let } f_i^e(X_{PA(X^e)}^e, \epsilon_i^e) = g(X_{PA(X^e)}^e) +$$

}

Input: Y^e , for each $e \in \mathcal{E}_{\text{train}}$, and X^e , for each $e \in \mathcal{E}_{\text{obs}}$

$$\frac{1}{|\mathcal{I}_{\text{inv}}|} \sum_{(k,S) \in \mathcal{T}_{\text{inv}}} \hat{Y}_{k,S}^{\text{test}}$$

for each $e \in \mathcal{E}_{\text{train}}$ **and** $i \in \{0, 1\}$ **do**
 $\mathbf{R}_i^e = \mathbf{X}_{k,Y=i}^e - \hat{g}_i(\mathbf{X}_{S,Y=i}^e)$
 \mathbf{R}
 $\text{pval}_i^e = t\text{-test}(\mathbf{R}_i^e, \mathbf{R}_i^{-e})$



$\{3, \dots, 7\}$. For each $i \in \{2, \dots, m\}$ and $e \in \mathcal{E}_{\text{train}}$, $X_i^e \sim$

for $e = e^1$, $[0, 2]$ for $e = e^2$, and $[0, 3]$ for $e = e^{\text{test}}$. Then, where $S_1 = \{2, \dots, m\}$, $Y^e|X_{S_1}^e$ follows a logistic model such that $\mathcal{P}_e(Y = 1|X_{S_1}^e) = 1/(1 + e^{-X_{S_1}^e \beta^e})$ for $e \in \mathcal{E}_{\text{train}}$. For e^{test} , $Y^{\text{test}}|X_{S_1}^{\text{test}}$ follows a probit model such that $Y^{\text{test}} = 1$, if $X_{S_1}^{\text{test}} \beta^{\text{test}} + \epsilon < 0$, where $\epsilon \sim \mathcal{N}(0, 1)$. For all $e \in \mathcal{E}_{\text{obs}}$,

then scaled such that they sum to one. For all $e \in \mathcal{E}_{\text{obs}}$,

Specifically, $g_1(X_{S_1}^e) = X_{S_1}^e \eta_1$ and $g_0(X_{S_1}^e) = X_{S_1}^e \eta_0$. The

Two real-world data. We also include experiments on two real datasets: *census* [18] and *mushroom* [19]. The census

14 societal and demographic variables such as age, education,

whether or not an individual's income exceeded 50k/yr. The

mushroom data below.

meadows	76.0	87.5	46.2
paths	88.1	90.9	11.8

naturally growing mushrooms' size, shape, and color and

or paths. Results in Table II indicate that **bIMP** outperforms

VII. ACKNOWLEDGEMENTS

$e \in \mathcal{E}_{\text{obs}}$. Without loss of generality, let X_i^e be continuous for all $i \in \{1, \dots, m\}$. The pdf of $X_k^e|X_S^e$ for any $e \in \mathcal{E}_{\text{obs}}$ is

$$\begin{aligned}
f_{X_k^e|X_S^e}(x_k|x) &= f_{X_k^e|X_S^e, Y^e}(x_k|x, 1) \cdot p_{Y^e|X_S^e}(1|x) \\
&\quad + f_{X_k^e|X_S^e, Y^e}(x_k|x, 0) \cdot p_{Y^e|X_S^e}(0|x) \\
&= f_{X_k^e|X_S^e, Y^e}(x_k|x, 1) \cdot p_{Y^e|X_S^e}(1|x) \\
&\quad + f_{X_k^e|X_S^e, Y^e}(x_k|x, 0) \cdot [1 - p_{Y^e|X_S^e}(1|x)] \\
&= p_{Y^e|X_S^e}(1|x) [f_{X_k^e|X_S^e, Y^e}(x_k|x, 1) - f_{X_k^e|X_S^e, Y^e}(x_k|x, 0)]
\end{aligned} \tag{12}$$

$$\int_{-\infty}^{\infty} x_k \cdot f_{X_k^e|X_S^e}(x_k|x) dx_k$$

$$- \mathbb{E}_{\mathcal{P}_e}[Y|X_S = x] \cdot \mathbb{E}_{\mathcal{P}_e}[X_k|X_S = x, Y = 0] \tag{13}$$

$$\frac{\mathbb{E}_{\mathcal{P}_e}[X_k|X_S] - \mathbb{E}_{\mathcal{P}_e}[X_k|X_S, Y = 0]}{\mathbb{E}_{\mathcal{P}_e}[X_k|X_S, Y = 1] - \mathbb{E}_{\mathcal{P}_e}[X_k|X_S, Y = 0]}. \tag{14}$$

on e and (II) the denominator of (14) is non-zero. Since $X_S^e = (X_R^e, X_Q^e)$,

$$\begin{aligned}
\mathbb{E}_{\mathcal{P}_e}[X_k|X_S, Y] &= \mathbb{E}_{\mathcal{P}_e}[X_k|X_R, X_Q, Y] \stackrel{(a)}{=} \mathbb{E}_{\mathcal{P}_e}[X_k|X_R, Y] \\
&\stackrel{(b)}{=} \mathbb{E}_{\mathcal{P}_e}[g(X_R, Y) + \epsilon|X_R, Y] = g(X_R^e, Y^e),
\end{aligned} \tag{15}$$

where (a) follows since $X_Q^e \perp\!\!\!\perp X_k^e|X_R^e, Y^e$, (b) follows from

ϵ has zero mean. Thus, the $\mathbb{E}_{\mathcal{P}_e}[X_k|X_S = (x_Q, x_R), Y = y]$ does not depend on e as $\mathbb{E}_{\mathcal{P}_e}[X_k|X_S = (x_Q, x_R), Y = y] =$

□

REFERENCES

- [12] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters, “Invariant
- [13] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters, “Anchor
- [6] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij,