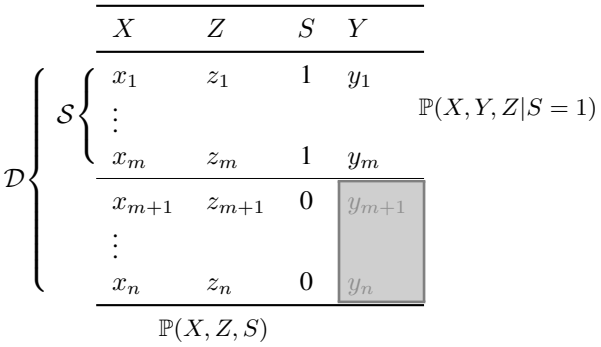
Correcting for Selection Bias and Missing Response in Regression using Privileged Information (Supplementary Material)

Philip Boeken[1,2] **Noud de Kroon**[1] **Mathijs de Jong**[2] **Joris M. Mooij**[1] **Onno Zoeter**[2]

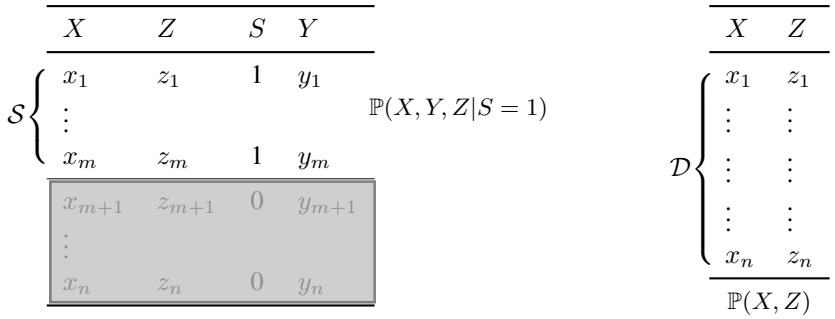
[1]Korteweg-de Vries Institute for Mathematics, University of Amsterdam, The Netherlands
[2]Booking.com, The Netherlands

1 SELECTION BIAS AND MISSINGNESS

A schematic display of the available data under missingness and under selection bias with external data is provided in Figure 1. Note that under missingness we can estimate $\mathbb{P}(S = 1|X, Z)$ from the dataset $\mathcal{D}$; under selection bias this is not possible.



(a) Missing response



(b) Selection bias with external data

Figure 1: Available data under missingness and under selection bias with external data. Grayed-out areas indicate unobserved data.

In the PMAR setting we have $Y \perp\!\!\!\perp S \mid X, Z$ and we are only given values of X at test time; our target is to estimate the function $\mathbb{E}[Y|X]$.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

2 IMPORTANCE WEIGHTING

For estimating the parameter β^* in the regression model $\mathbb{E}[Y|X] = g(X;\beta^*)$ we often specify a loss function ℓ and perform empirical risk minimisation (1) as an approximation of the optimal parameter in terms of the true risk (2).

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \ell(g(X_i;\beta), Y_i) \tag{1}$$

$$\beta^* = \arg\min_{\beta} \mathbb{E}[\ell(g(X;\beta), Y)] \tag{2}$$

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

$$\begin{aligned}
\mathbb{E}[f(X,Y)] &= \int f(x,y)p(x,y,z)\mathrm{d}(x,y,z) \\
&= \int f(x,y)p(x,y,z)\frac{p(x,y,z|S=1)}{p(x,y,z|S=1)}\mathrm{d}(x,y,z) \\
&= \int f(x,y)\frac{p(x,y,z)\mathbb{P}(S=1)}{p(x,y,z,S=1)}p(x,y,z|S=1)\mathrm{d}(x,y,z) \\
&= \int f(x,y)\frac{\mathbb{P}(S=1)}{\mathbb{P}(S=1|x,y,z)}p(x,y,z|S=1)\mathrm{d}(x,y,z) \\
&= \int f(x,y)\frac{\mathbb{P}(S=1)}{\mathbb{P}(S=1|x,z)}p(x,y,z|S=1)\mathrm{d}(x,y,z) \\
&= \int f(x,y)w(x,z)p(x,y,z|S=1)\mathrm{d}(x,y,z) \\
&= \mathbb{E}[w(X,Z)f(X,Y)|S=1],
\end{aligned} \tag{3}$$

where in the fifth equation we use the conditional independence $Y \perp\!\!\!\perp S \mid X, Z$, and where we define the *importance weights*

$$w(x,z) := \frac{\mathbb{P}(S=1)}{\mathbb{P}(S=1|x,z)}. \tag{4}$$

Since $\beta^* = \arg\min_\beta \mathbb{E}[w(X,Z)\ell(g(X;\beta),Y)|S=1]$, when we have observations $(X_1,Z_1,Y_1),...,(X_n,Z_n,Y_n) \sim \mathbb{P}(X,Z,Y|S=1)$, we can directly perform empirical risk minimization on this dataset using the weighted loss:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} w(x_i,z_i)\ell(g(x_i;\beta), y_i). \tag{5}$$

3 SIMULATIONS

We performed a brute-force search of all *Acyclic Directed Mixed Graphs* (ADMGs) that satisfy $X \not\perp Y$ and the PMAR pattern of d-separations $Y \not\perp S$, $Y \not\perp S \mid X$, and $Y \perp S \mid X, Z$, using the `pcalg` package [Kalisch et al., 2012], resulting in 550 ADMGs. For reference, all 55 DAGs are depicted in figures 2 and 3, where the graphs are categorised by whether S has any children or not. The remaining 495 ADMGs are not depicted here.

For each of the 550 ADMGs, we simulate 50 datasets according to the following procedure. Throughout, let V, S, X etc. denote vertices in the graph, and let x_V, x_S, x_X denote vectors of dimension $n = 2000$ for the simulated values.

- First, we replace any bidirected edge pointing to variables $\boldsymbol{V}$ by a variable $U_{\boldsymbol{V}}$, and let $\boldsymbol{V}$ be the children of $U_{\boldsymbol{V}}$. This turns the ADMG into a DAG.
- Then, we calculate a topological order of the DAG.
- For every variable V in the topological order, we simulate 2000 observations as follows:
  - If V has no parents, then
    - if $V = S$, draw $x_S \sim$ Bernoulli$(1/3)$;

* otherwise, $V \neq S$ and we draw $x_V \sim \mathcal{RD}$, as defined below.

– Otherwise, denote the parents of V with $\boldsymbol{Pa}$ and their value $\boldsymbol{x_{Pa}}$, and then

* if $V = S$, draw $(x_S)_i \sim \text{Bernoulli}(p((x_{\boldsymbol{Pa}})_i))$ where

$$p((x_{\boldsymbol{Pa}})_i) := \prod_{v \in \boldsymbol{Pa}} \sigma((x_v)_i) \tag{6}$$

and $\sigma(x) := (1 + e^{20x})^{-1}$;

* otherwise, $V \neq S$ and
  · draw a random function f_V from a Gaussian process on $\mathbb{R}^{|\boldsymbol{Pa}\setminus\{S\}|}$ as $f_V \sim \mathcal{GP}(0, K_M)$;
  · draw noise $\varepsilon_V \sim \mathcal{RD}$ and set

$$x_V := f_V(\boldsymbol{x}_{\boldsymbol{Pa}\setminus\{S\}}) + \frac{1}{2}\varepsilon_V; \tag{7}$$

  · if $S \in \boldsymbol{Pa}$, calculate the empirical standard deviation $c := \text{sd}(x_V)$ and set $(x_V)_i := (x_V)_i - 3c$ for all i where $(x_S)_i = 1$;
  · then, standardize x_V.

Drawing from a Gaussian process The kernels used for calculating the covariance matrix for drawing from a Gaussian process are the Matérn kernel and the squared exponential kernel, being

$$K_M(x, y) := (1 + \sqrt{5}\|x - y\| + \frac{5}{3}\|x - y\|^2)e^{-\sqrt{5}} \tag{8}$$

$$K_{SE}(x, y) := \frac{1}{4}e^{\frac{2}{9}\|x-y\|^2} \tag{9}$$

respectively, where $\|\cdot\|$ denotes the Euclidean norm. Then, given input $x_i \in \mathbb{R}^d$ for $i = 1, ..., n$ and kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, denote $\boldsymbol{x} := (x_i)_{i=1}^n$ and draw $f(\boldsymbol{x}) \sim \mathcal{N}(0, (K(x_i, x_j))_{i,j})$, where i and j run over $\{1, ..., n\}$ in the kernel matrix $(K(x_i, x_j))_{i,j}$. This draw is denoted with $f \sim \mathcal{GP}(0, K)$.

Drawing from a random distribution Drawing noise from a random distribution, denoted with $\varepsilon \sim \mathcal{RD}$, is done as follows:

- First, draw 2000 i.i.d. samples $U \sim \text{Unif}[0, 1]$.
- Then, draw a random function $f_\varepsilon \sim \mathcal{GP}(0, K_{SE})$
- Set $\varepsilon := f_\varepsilon(U)$, and standardize.

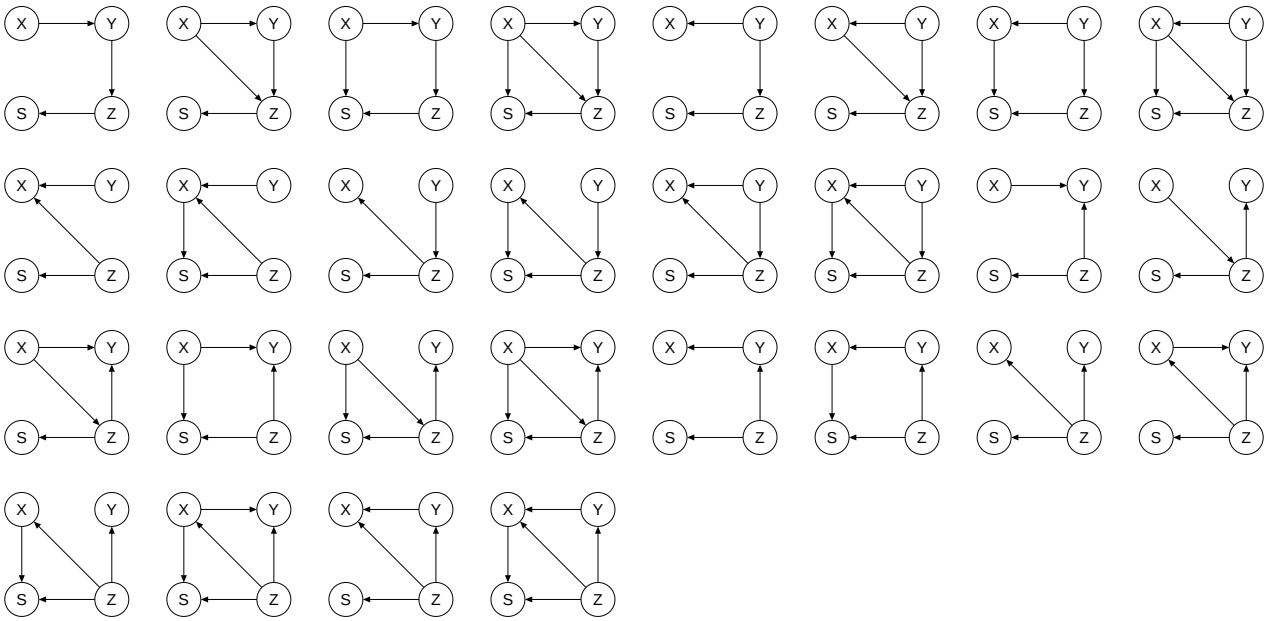


Figure 2: PMAR DAGs where S is a sink node.

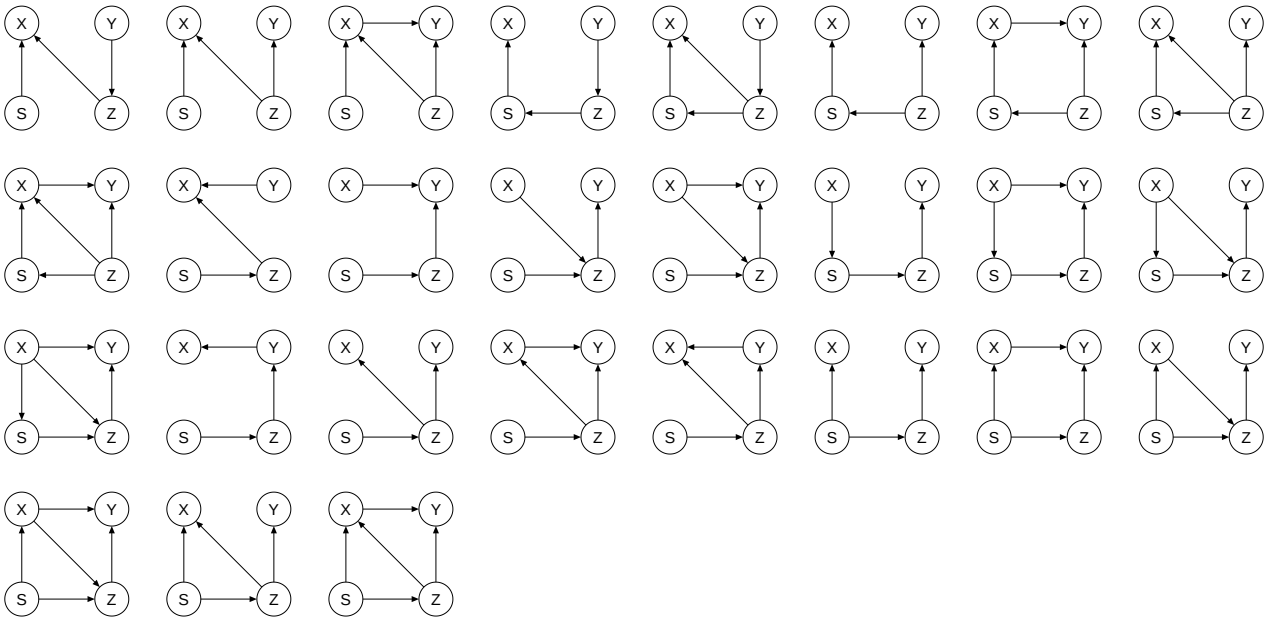


Figure 3: PMAR DAGs where S is not a sink node.

3.1 SIMULATION EXPERIMENTS WITH REGRESSION TREES

In the main paper, we hypothesize that the bad extrapolation performance of weighted regression is caused by a detrimental effect of the weights on the regularization of the regression method. In the main paper we use thin plate spline regression, which doesn't necessarily extrapolate flatly and can diverge away from the true $\mathbb{E}[Y|X]$, yielding large MSE values for IW and DR. Tables 1 and 2 show the simulation results when using regression trees as implemented in the `rpart` package, instead of thin plate regression. The results are numerically less extreme, but qualitatively the same as for thin plate regression, as depicted in the main paper.

	MSE	MSE-$\tilde{y}$	MSE-w	MSE-$\hat{w}$
Naive	1.53 (0.6)	0.62 (0.8)	0.96 (0.6)	0.96 (0.5)
RR	**1.38** (0.6)	**0.27** (0.3)	0.97 (0.6)	0.94 (0.5)
IW-t	1.54 (0.6)	0.63 (0.8)	0.97 (0.6)	0.98 (0.5)
IW-e	1.52 (0.6)	0.61 (0.7)	0.97 (0.6)	0.97 (0.5)
DR-t	1.48 (0.7)	0.58 (0.7)	**0.66** (0.3)	0.73 (0.4)
DR-e	1.46 (0.6)	0.55 (0.7)	0.71 (0.4)	**0.69** (0.3)
True	1.00 (0.2)	0.69 (0.7)	1.05 (0.6)	1.02 (0.5)

Table 1: Results over 27.500 simulated datasets.

	MSE	MSE-interp.	MSE-extrap.
Naive	1.58 (0.6)	1.37 (0.6)	1.82 (1.0)
RR	**1.44** (0.6)	**1.22** (0.6)	**1.68** (1.0)
IW-t	1.60 (0.6)	1.38 (0.6)	1.85 (1.1)
IW-e	1.58 (0.6)	1.37 (0.6)	1.82 (1.0)
DR-t	1.55 (0.7)	1.30 (0.6)	1.84 (1.1)
DR-e	1.52 (0.6)	1.28 (0.6)	1.80 (1.0)
True	1.01 (0.2)	1.03 (0.3)	0.98 (0.3)

Table 2: Interpolation and extrapolation results of regression trees for simulated data, on graphs with $X \to S$.

4 BOSTON HOUSING DATA

Considering the Boston Housing Dataset [Harrison Jr and Rubinfeld, 1978], let the variables X, Y and Z be 'the number of rooms per dwelling', 'the value of owner-occupied homes in US Dollars' and 'percentage of people of lower status of the population' respectively. We sample the selection probability

$$p(X, Z) := \sigma(f_1(X))\sigma(f_2(Z)) \tag{10}$$

where $f_1, f_2 \sim \mathcal{GP}(0, K_{SE})$ independent. Then we draw $U_i \sim \text{Unif}[0, 1]$ for $i = 1, ..., 506$, and set

$$S_i := \mathbb{1}\{p(X, Z)_i < U_i\} \tag{11}$$

for all i. The dataset is resampled when $\#\{S = 1\} < 120$. One realisation of such a dataset with an overview of all regression methods is provided in Figure 4.

The MSE and the interpolated and extrapolated variants, as calculated on the Boston Housing dataset, are shown in Table 3. We observe that RR performs better than IW and DR on all three metrics.

References

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

M. Kalisch, M. Mächler, D. Colombo, M. H. Maathuis, and P. Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012. doi: 10.18637/jss.v047.i11.

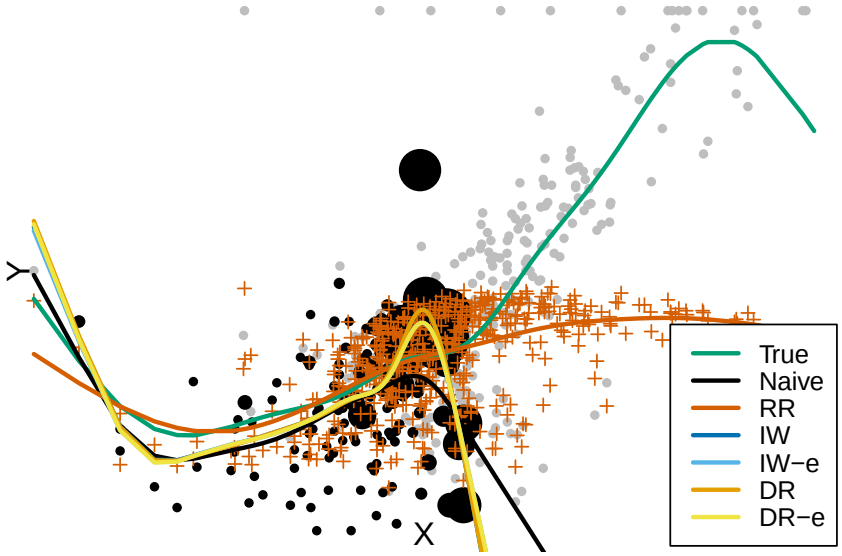


Figure 4: An instantiation of the biased Boston Housing dataset.

	MSE	MSE-interp.	MSE-extrap.
Naive	1.23 (2.5)	0.57 (0.5)	3.88 (9.1)
RR	**0.71** (0.3)	**0.47** (0.2)	**2.05** (4.3)
IW-t	2.18 (4.9)	0.62 (0.8)	7.68 (17.7)
IW-e	1.75 (4.4)	0.58 (0.6)	5.96 (16.1)
DR-t	1.92 (3.7)	0.48 (0.3)	7.37 (16.4)
DR-e	2.43 (5.4)	0.52 (0.4)	9.42 (22.1)

Table 3