

---

# MissDeepCausal: Causal Inference from Incomplete Data Using Deep Latent Variable Models

---

Imke Mayer<sup>1,2</sup> Julie Josse<sup>2,3</sup> Félix Raimundo<sup>4</sup> Jean-Philippe Vert<sup>4</sup>

## Abstract

Inferring causal effects of a treatment, intervention or policy from observational data is central to many applications. However, state-of-the-art methods for causal inference seldom consider the possibility that covariates have missing values, which is ubiquitous in many real-world analyses. Missing data greatly complicate causal inference procedures as they require an adapted unconfoundedness hypothesis which can be difficult to justify in practice. We circumvent this issue by considering latent confounders whose distribution is learned through variational autoencoders adapted to missing values. They can be used either as a pre-processing step prior to causal inference but we also suggest to embed them in a multiple imputation strategy to take into account the variability due to missing values. Numerical experiments demonstrate the effectiveness of the proposed methodology especially for non-linear models compared to competitors.

## 1. Introduction

Many methods have been developed to estimate the causal effect of an intervention, such as the administration of a treatment, on an outcome such as survival, from observational data, i.e., data that is potentially confounded by selection bias due to the absence of randomization. Classical ones include matching (Iacus et al., 2012), inverse propensity weighting (IPW, Horvitz & Thompson, 1952; Rosenbaum & Rubin, 1983) and doubly robust methods (Robins et al., 1994; Chernozhukov et al., 2018; Wager & Athey, 2018;

Athey et al., 2019). More recent proposals use deep learning methods that ensure balance of the population at the level of representation (Johansson et al., 2016; Shalit et al., 2017), infer the joint distribution of latent and observed confounders, the treatment and the outcome (Louizos et al., 2017) or predict the counterfactuals with GANs (Yoon et al., 2018). For a detailed review of existing literature on treatment effect estimation we refer to Imbens (2004), Lunceford & Davidian (2004) and Guo et al. (2019).

However, state-of-the-art methods still suffer from important shortcomings. In particular, they seldom consider the possibility that covariates have missing values, which is ubiquitous in many real-world situations (Josse & Reiter, 2018) and has been widely discussed in different contexts (Mayer et al., 2019a; van Buuren, 2018; Little & Rubin, 2002). Although this question of missing attributes in the context of treatment effect estimation has been raised early in the development of causal inference (Rosenbaum & Rubin, 1984), there is still a lack of effective and consistent solutions addressing this problem, with a few notable exceptions such as Mattei & Mealli (2009); Seaman & White (2014); Yang et al. (2019); Kallus et al. (2018) which mainly focus on inverse propensity weighting (IPW) methods and Kuroki & Pearl (2014) who discuss identifiability of causal effects under measurement error or unobserved confounders. Recently, Mayer et al. (2019b), in addition to suggesting doubly robust estimators with missing data, classified the existing approaches into two families: the ones that adapt the causal inference assumptions to the missing values setting (D’Agostino Jr & Rubin, 2000; Blake et al., 2019) and the ones (Mattei & Mealli, 2009; Seaman & White, 2014; Kallus et al., 2018) that consider the classical machinery and missingness mechanisms assumptions (Little & Rubin, 2002). While the former are based on the assumption of *unconfoundedness with missing values*, which can be difficult to assess in practice, the latter have been developed under strong parametric assumptions about the outcome, treatment and covariates models, in addition to relying on missing values hypotheses that can also be difficult to meet in practice (Yang et al., 2019).

To avoid relying on the hypothesis of unconfoundedness with missing values or being in the very parametric (and linear) framework of multiple imputation (Mattei & Mealli,

---

\*Equal contribution <sup>1</sup>Centre d’Analyse et de Mathématique Sociales, UMR 8557, EHESS, Paris, France <sup>2</sup>Centre de Mathématiques Appliquées, UMR 7641, École Polytechnique, Palaiseau, France <sup>3</sup>Inria XPOP, Palaiseau, France <sup>4</sup>Brain team, Google Research, Paris, France. Correspondence to: Imke Mayer <imke.mayer@ehess.fr>.

This research was done while JJ was a visiting researcher at Google Brain Paris. IM received funding from the EHESS PhD fellowship program. We thank Pierre-Alexandre Mattei for interesting discussions around deep latent variable models.

2009; Seaman & White, 2014) and matrix factorization (Kallus et al., 2018), we propose a new method for causal inference with missing data, which we call MissDeepCausal. MissDeepCausal is inspired by the work of Kallus et al. (2018) in the sense that we consider a model with latent confounders, and assume that we only have access to covariates with missing values that are noisy proxies of the true latent confounders. However, our approach generalizes and extends the work of Kallus et al. (2018) in different aspects: (i) instead of linear factor analysis models with missing values, we consider non-linear versions using deep latent variable models (Kingma & Welling, 2014; Rezende et al., 2014); (ii) we rely on the missing at random (MAR) (Rubin, 1976) assumption for the missing data mechanisms, and not on the stronger missing completely at random (MCAR) one; (iii) we take into account the posterior distribution of the latent variables given observed data and not only their conditional expectation. This latter point allows us to define a multiple imputation strategy adapted to the latent confounders model, and to couple it with doubly robust treatment effect estimation (Chernozhukov et al., 2018).

In the remainder of this article we first introduce the problem framework and recall existing work for handling missing values in causal inference in Section 2. We then introduce two variants of our MissDeepCausal approach in Section 3. Finally we compare MissDeepCausal empirically with several state-of-the-art methods on simulated data in Section 4.

## 2. Setting, notations and related works

In this section we start by quickly reviewing the problem of causal inference from observational data without missing data. We consider the potential outcomes framework (Rubin, 1974; Imbens & Rubin, 2015) where we have a sample of  $n$  independent and identically distributed (i.i.d.) observations  $(Y_i(0), Y_i(1), W_i, X_i)_{i=1, \dots, n}$  with  $W_i \in \{0, 1\}$  a binary treatment,  $X_i = (X_{i1}, \dots, X_{ip})^\top \in \mathbb{R}^p$  a vector of covariates, and  $(Y_i(0), Y_i(1)) \in \mathbb{R}^2$  the outcomes we would have observed had we assigned control or treatment to the  $i$ -th sample, respectively. The observed outcome for unit  $i$ ,  $Y_i \in \mathbb{R}$  is defined as  $Y_i \triangleq W_i Y_i(1) + (1 - W_i) Y_i(0)$ . The individual causal effect of the treatment is  $\tau_i \triangleq Y_i(1) - Y_i(0)$  and the average treatment effect (ATE) is defined as

$$\tau \triangleq \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[\tau_i].$$

The ATE  $\tau$ , i.e., the link between  $W$  and  $Y$ , can be estimated from observational data by taking into account the confounding factors  $X$ , i.e., the common causes of  $W$  and  $Y$ . A popular estimator of  $\tau$  from observational data is the

so-called doubly robust estimator:

$$\begin{aligned} \hat{\tau}_{DR} \triangleq & \frac{1}{n} \sum_{i=1}^n \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \\ & + W_i \frac{Y_i - \hat{\mu}_1(X_i)}{\hat{e}(X_i)} - (1 - W_i) \frac{Y_i - \hat{\mu}_0(X_i)}{1 - \hat{e}(X_i)}, \end{aligned} \quad (1)$$

where  $\hat{\mu}_w(x)$  are regression estimates of the conditional response surfaces  $\mu_w(x) \triangleq \mathbb{E}[Y(w) | X = x]$ ,  $w \in \{0, 1\}$ , and  $\hat{e}(x)$  is an estimate of the propensity score  $e(x) \triangleq \mathbb{P}(W_i = 1 | X_i = x)$  (Rosenbaum & Rubin, 1983; Imbens & Rubin, 2015).

Standard results state that if either  $(\hat{\mu}_0, \hat{\mu}_1)$  or  $\hat{e}$  is correctly specified, then  $\hat{\tau}_{DR}$  is an unbiased estimator of  $\tau$  (Robins et al., 1994; Chernozhukov et al., 2018; Wager & Athey, 2018) under the following assumptions (Rosenbaum & Rubin, 1983): the *ignorability* or *unconfoundedness* assumption that states that all confounding factors are measured, i.e., conditionally on  $X$ , the treatment assignment is independent of the potential outcomes:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i | X_i, \quad \text{for all } i; \quad (2)$$

and the *overlap* assumption assuming the existence of some  $\eta > 0$  such that  $\eta < e(x) < 1 - \eta$ , for all  $x \in \mathcal{X}$ .

We now consider an extension to account for possible missing entries in the covariates. For that purpose, we denote the missingness pattern of the  $i$ -th sample as  $M_i \in \{0, 1\}^p$  such that  $M_{ij} = 0$  if  $X_{ij}$  is observed and  $M_{ij} = 1$  otherwise. The matrix of observed covariates can be written as  $X^* \triangleq X \odot (1 - M) + \text{NA} \odot M$ , with  $\odot$  the elementwise multiplication and  $1$  the matrix filled with 1, so that  $X^*$  takes its value in the half discrete space  $\mathcal{X}^* \triangleq (\mathbb{R} \cup \{\text{NA}\})^p$ . We model  $M_i$  as a random vector, and the possibility to infer causal effects with missing data now depends on additional assumptions on the joint law of  $(Y_i(0), Y_i(1), W_i, X_i, M_i)_{i=1, \dots, n}$ . Methods for causal inference with missing covariates can be classified into two categories.

**Unconfoundedness with missing values.** Rosenbaum & Rubin (1984) extend the unconfoundedness hypothesis (2) to missing values as

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i | X_i^*, \quad \text{for all } i. \quad (3)$$

This implies the assumption, illustrated in Figure 1, that if a covariate is not observed, it is not a confounder. In particular, observations can have different confounders depending on their pattern of missing data. They define the generalized propensity score as:

$$\forall x^* \in \mathcal{X}^*, \quad e^*(x^*) \triangleq \mathbb{P}(W_i = 1 | X_i^* = x^*), \quad (4)$$

which is a balancing score under (3). Consequently, an IPW estimator formed with estimators of  $e^*$  can be an unbiased estimator of the ATE with missing values. Nevertheless, this method relies both on the fact that the covariates  $X$  are the appropriate set of confounders, which can be questioned without missing data (Kallus et al., 2018), and requires certain expert input and reasoning to verify that for each observation, treatment assignment and/or outcome values depend only on observed values of the confounders (Blake et al., 2019; Mayer et al., 2019b). Note in particular, that it is not because the missing data in the covariates are completely at random (MCAR), i.e.,  $M \perp\!\!\!\perp X$ , that (3) is met. In practice, in addition, a difficulty with this approach is that estimating (4) requires fitting one model per pattern of missing values, which is unrealistic with classical tools (Miettinen, 1985; D’Agostino Jr & Rubin, 2000; D’Agostino Jr et al., 2001; Blake et al., 2019); Mayer et al. (2019b) address this problem using random forests adapted to covariates with missing values.

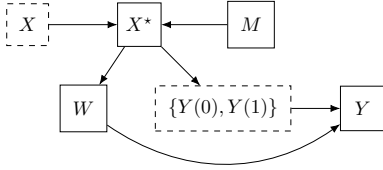


Figure 1. Unconfoundedness with missing values.  $X$  represents a the complete covariates, and  $M$  a missing data mechanism,  $X^*$  represents the observed incomplete covariates, confounding the treatment assignment. The formalism of Pearl (1995) and Richardson & Robins (2013) is used.

**Missingness mechanisms assumptions.** Multiple imputation is one of the most powerful approaches to estimate parameters and their variance from an incomplete data (Little & Rubin, 2002; van Buuren, 2018). Seaman & White (2014) show that when assuming (i) identifiability of the ATE in the complete case, (ii) missing at random (MAR) values given  $W$  and  $Y$ , (iii) correct specification of the propensity score with logistic regression and of the Gaussian distribution of covariates, then multiple imputation gives a consistent estimate for the ATE estimated with IPW. An extension to doubly robust estimation has been proposed by Mayer et al. (2019b).

Instead of assuming that confounders are observed directly, Kallus et al. (2018) consider a more general model where observed covariates  $X$  are noisy and/or incomplete proxies of the true latent confounders  $Z$ . More specifically, they assume a low-rank model for the covariates and estimate the latent variables from the incomplete confounders using matrix completion methods (Hastie et al., 2015; Josse et al., 2016). Then, under the linear regression model

$$Y_i = Z_i^T \alpha + \tau W_i + \varepsilon_i, \quad (5)$$

with random latent variables  $Z$ , missing values completely at random (MCAR) in  $X$ , unconfoundedness given  $Z$ , and some additional assumptions, they prove that regressing  $Y$  on  $\hat{Z}$  and  $W$  leads to a consistent ATE estimator. Both techniques, multiple imputation and matrix factorization, rely on parametric (and linear) frameworks.

### 3. MissDeepCausal

To avoid relying on the hypothesis of unconfoundedness with missing values (3) or being in the very parametric (and linear) framework of multiple imputation and matrix factorization, we propose MissDeepCausal, an approach based on deep latent variable models where the latent variables are assumed to be the confounders as represented in Figure 2.

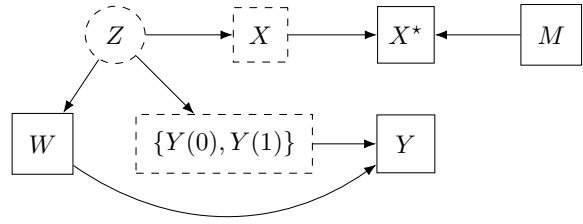


Figure 2. Graphical representation of the model underlying MissDeepCausal.  $Z$  represents the unobserved latent confounders of the treatment  $W$  and the effect  $Y$ .  $X$  represents a proxy for the confounders, and  $M$  a missing data mechanism;  $X^*$  represents the observed incomplete covariates.

Under this model, the unconfoundedness hypothesis (3) does *not* hold, so a standard treatment effect estimator using  $X^*$  as covariates would be biased. On the other hand, we can express the treatment effect conditioned on  $X^*$  as follows:

$$\begin{aligned} \mathbb{E}[Y(1) - Y(0) | X^*] &= \mathbb{E}[\mathbb{E}[Y(1) - Y(0) | Z, X^*] | X^*] \\ &= \mathbb{E}[\mathbb{E}[Y(1) - Y(0) | Z] | X^*]. \end{aligned}$$

Consequently, if we have an unbiased estimator  $\hat{f}(Z)$  of  $\mathbb{E}[Y(1) - Y(0) | Z]$ , the treatment effect conditioned on  $Z$ , and if we know  $P(Z | X^*)$ , the conditional distribution of  $Z$  given  $X^*$ , then we can derive the treatment effect conditioned on  $X^*$  by

$$\hat{g}(X^*) \triangleq \mathbb{E}[\hat{f}(Z) | X^*]. \quad (6)$$

Furthermore, by expressing the ATE as

$$\tau = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}[Y(1) - Y(0) | X^*]],$$

we can form an estimate of the ATE by  $\mathbb{E}[\hat{g}(X^*)]$ . We describe such an estimator in Section 3.2 below, which is reminiscent of multiple imputation techniques in the field of missing value imputation (Rubin, 1987).

Another strategy, described in Section 3.3, is to consider latent variables estimation as a pre-processing step prior to causal inference by computing

$$h(X^*) \triangleq f(\mathbb{E}[Z|X^*]); \quad (7)$$

this can be seen as a non-linear extension of Kallus et al. (2018). Both estimators require sampling from the posterior distribution  $P(Z|X^*)$ . Consequently, we first describe in Section 3.1 how to learn the joint distribution of  $(Z, X)$  from  $X^*$  using a variational autoencoder (VAE) with missing data, before turning to the details of each strategy.

### 3.1. Deep latent variable models with missing values

**Variational autoencoding** Deep latent variable models can be defined as follows. Let  $(X_i, Z_i)_{i \leq n}$  be  $n$  i.i.d. random variables such that

$$\begin{cases} Z_i \sim P(Z_i) \\ X_i \sim P_\theta(X_i|Z_i) = \Phi(X_i|f_\theta(Z_i)). \end{cases}$$

The prior distribution of the latent variables or *codes*  $Z_i \in \mathbb{R}^d$  is often isotropic Gaussian  $Z_i \sim \mathcal{N}(0_d, I_d)$ . The function  $f_\theta : \mathbb{R}^d \rightarrow H$  is a (deep) neural network called the *decoder* and  $\Phi(\cdot|\eta)_{\eta \in H}$  is a parametric observation model, which we take to be multivariate Gaussian. The inference of deep latent variable models can be achieved by maximizing evidence lower bounds of the likelihood, such as the variational autoencoder bounds.

With missing values, the appropriate quantity to target for inference on  $\theta$ , when the missing values mechanism can be ignored (Rubin, 1976; Little & Rubin, 2002), is the observed log-likelihood. Using Rubin (1976)’s notations, we define  $X_i = (X_{i,obs}, X_{i,mis})$  the partition of the data in realized observed and missing values given a specific realization of the pattern, it can be written as:

$$\begin{aligned} \ell(\theta) &\triangleq \sum_{i=1}^n \log p_\theta(X_{i,obs}) \\ &= \sum_{i=1}^n \log \int p_\theta(X_{i,obs}|Z_i) p(Z_i) dZ_i. \end{aligned}$$

The corresponding evidence lower bound (ELBO) is:

$$\begin{aligned} \mathcal{L}(\theta, \gamma) &\triangleq \sum_{i=1}^n \mathbb{E}_{Q_\gamma} [\ln P_\theta(X_{i,obs}|Z_i)] \\ &\quad - KL(Q_\gamma(Z_i|X_{i,obs}) \| P_\theta(Z_i)), \end{aligned}$$

with  $KL$  for the Kullback-Leibler divergence and the variational distribution

$$Q_\gamma(Z|X_{obs}) \triangleq \Psi(Z|g_\gamma(X_{obs})),$$

with  $\Psi(\cdot)$  the (parametric) variational distribution over  $\mathbb{R}^d$ . The function  $g_\gamma : \mathcal{X} \rightarrow \mathcal{K}$ , called the *encoder*, is

parametrized by a (deep) neural network whose weights are stored in  $\gamma \in \Gamma$ .

To take into account missing values in deep latent variable models, Mattei & Frellsen (2019) suggest the missing data importance weight autoencoder bound (MIWAE) approach. They use a simple variational family where they impute the missing entries with a constant and show that using this class of distributions, it maximizes a lower bound of the observed log-likelihood. Specifically, they replace  $Q_\gamma$  with

$$Q_\gamma(Z|X_{obs}) = \Psi(Z|g_\gamma(\iota(X_{obs}))),$$

where  $\iota$  is an imputation function chosen beforehand that transforms  $X_{obs}$  into a complete input vector  $\iota(X_{obs}) \in \mathcal{X}$ .

**Self-normalized importance sampling** To estimate and sample from  $P(Z|X^*)$ , we use the missing data importance weight autoencoder bound (MIWAE) approach of Mattei & Frellsen (2019), which is summarized above. They use a simple variational family where they impute the missing entries with a constant and show that using this class of distributions, it maximizes a lower bound of the observed log-likelihood. Note that their approach requires the classical missing at random (MAR) (Rubin, 1976) assumption to ignore the missing values mechanism when maximizing the observed likelihood for the VAE inference.

In the MIWAE approach, the variational distribution  $Q_\gamma(Z|X^*)$  plays a central role but is not necessarily a good surrogate for the posterior distribution  $P_\theta(Z|X^*)$ . To sample from the true posterior distribution, we resort to importance sampling techniques using the variational distribution  $Q_\gamma$  for proposal. More precisely, we can define, for any measurable function  $s$ ,

$$\begin{aligned} \mathbb{E}[s(Z)|X^*] &= \int s(Z) p_\theta(Z|X^*) dZ \\ &= \frac{1}{p(X^*)} \int s(Z) \frac{p_\theta(X^*|Z)p(Z)}{q_\gamma(Z|X^*)} q_\gamma(Z|X^*) dZ. \end{aligned}$$

This quantity can be estimated using self-normalized importance sampling with:

$$\mathbb{E}[s(Z)|X^*] \approx \sum_{l=1}^L w_l s(Z^{(l)}),$$

$$\text{where } w_l \triangleq \frac{r_l}{r_1 + \dots + r_L}, \text{ with } r_l \triangleq \frac{p_\theta(X^*|Z^{(l)})p(Z^{(l)})}{q_\gamma(Z^{(l)}|X^*)}. \quad (8)$$

Equation (8) is used in our second strategy described in Section 3.3, while for our first strategy (described in Section 3.2) we sample  $L$  samples  $Z^{(1)}, \dots, Z^{(L)}$  according to  $Q_\gamma(Z|X^*)$ , compute the weights as in (8) and re-sample  $B \ll L$  with probability proportional to the weights.



### 3.2. MissDeepCausal with multiple imputation (MDC-MI)

MDC-MI uses the importance sampling strategy presented in Section 3.1, to compute an approximation of (6) by Monte-Carlo as follows. First, we draw  $B$  i.i.d. samples  $(Z^{(j)})_{1 \leq j \leq B} \in \mathbb{R}^{n \times d}$  from the posterior distribution  $P(Z|X^*)$ . On each sample, we evaluate the function  $f$  and aggregate the results:  $\hat{g}^{(B)}(X^*) = \frac{1}{B} \sum_{j=1}^B f(Z^{(j)})$ . This approach can be viewed as a multiple imputation method, which consists in generating different imputed data sets by drawing the missing values from their posterior distribution given observed values, then estimating the parameters of interest on each imputed data set and aggregating the results according to Rubin’s rules (Rubin, 1987) to obtain a final estimate for the quantity of interest. Here we consider the samples  $Z^{(j)}$  of the latent variables and apply the doubly robust estimator from (1) on each table  $Z^{(j)}$ :

$$\begin{aligned} \hat{\tau}^{(j)} = & \frac{1}{n} \sum_{i=1}^n \left( \hat{\mu}_1^{(j)}(Z_i^{(j)}) - \hat{\mu}_0^{(j)}(Z_i^{(j)}) \right) \\ & + W_i \frac{Y_i - \hat{\mu}_1^{(j)}(Z_i^{(j)})}{\hat{e}^{(j)}(Z_i^{(j)})} \\ & - (1 - W_i) \frac{Y_i - \hat{\mu}_0^{(j)}(Z_i^{(j)})}{1 - \hat{e}^{(j)}(Z_i^{(j)})}, \end{aligned} \quad (9)$$

and get the final estimate for the causal effect by computing the mean of the estimators i.e.  $\hat{\tau} = \frac{1}{B} \sum_{j=1}^B \hat{\tau}^{(j)}$ . The doubly robust estimator from (1) is asymptotically normal (under some mild assumptions) (Wager & Athey, 2018) which is required for the aggregation in multiple imputation procedures (Rubin, 1987). Note that this multiple imputation strategy additionally allows to reflect the variability due to the missing values in the variance estimation of the estimator  $\hat{\tau}$ .

### 3.3. MissDeepCausal with latent variables estimation as a pre-processing step (MDC-process)

We also propose MDC-process as a non-linear extension of Kallus et al. (2018), where we estimate  $h(X^*)$  defined in (7). For that purpose, we first approximate the expectation of the posterior distribution

$$\hat{Z}(x^*) \triangleq \mathbb{E}[Z|X^* = x^*] \quad (10)$$

to get estimates for the latent confounders. In a second step, we use them under the regression model (5) and accordingly regress the observed outcome  $Y$  on the estimated latent factors  $\hat{Z}(x^*)$  and the treatment assignment  $W$  to obtain an estimation of the treatment effect. This strategy is a heuristic extension of Kallus et al. (2018) to a non-linear case in the sense that the latent variables encode non-linear relationship between covariates.

An alternative, still heuristic, approach is to use the estimated latent confounders from (10) as inputs for standard techniques to estimate the average treatment effect. More precisely, for the doubly robust estimator (1), we replace the estimates for the propensity score with estimates for

$$\tilde{e}(z) = \mathbb{P}(W_i = 1 | \hat{Z}_i(x^*) = z),$$

and similarly for the conditional response surfaces.

However, note that this latter strategy would require  $\hat{Z}(x^*)$  from (10) to be a confounder instead of  $Z$  as it is assumed (see Figure 2).

## 4. Simulation study

### 4.1. Methods

We compare the following methods to handle missing values (the following acronyms are identical to the method labels used in Figures 3–7):

- MissDeepCausal:
  - MDC.process: using the estimations of the latent variables either in a regression adjustment estimator or in a double robust estimator as presented in Section 3.3;
  - MDC.mi: using the doubly robust estimator MDC-mi Section 3.2.

We extended the publicly available code of Mattei & Frellsen (2019) to implement both methods. Throughout all experiments, we fix  $L = 10,000$  for the importance sampling weights. We choose hyperparameters,  $\sigma_{prior}^2$  (variance of the prior on  $Z$ ) and  $d_{miwae}$  (dimension of estimated latent space), by cross-validation. We vary the number of draws  $B$  from the posterior for the MDC.mi approach from 50 to 500 (results only reported for  $B = 500$ ).

- MI: the multiple imputation approach as suggested in Mattei & Mealli (2009) and Seaman & White (2014). We generate 20 imputations, using the python implementation in the scikit-learn (Pedregosa et al., 2011).
- MF: the matrix factorization approach of (Kallus et al., 2018) based on nuclear norm penalty (python implementation inspired by the R package `softImpute` (Hastie & Mazumder, 2015)). The dimension of the latent space is chosen via cross-validation on the nuclear norm penalty parameter.

### 4.2. Settings

Under the latent confounding assumption (corresponding to the graphical model in Figure 2), we generate covariates according to two models:

- LRMF: The covariates are generated from a low-rank matrix factorization model as in [Kallus et al. \(2018\)](#).
- DLVM: The covariates are generated from a deep latent variable model as in [Kingma & Welling \(2014\)](#).  $Z_i \sim \mathcal{N}_d(0, 1)$ , covariates  $X_i$  are sampled from  $\mathcal{N}_p(\mu_{(Z)}, \Sigma_{(Z)})$ , where  $(\mu_{(Z)}, \Sigma_{(Z)}) = (V \tanh(UZ + a) + b, \text{diag}\{\exp(\eta^T \tanh(UZ + a) + \delta)\})$  with  $U, V, a, b, \delta, \eta$  drawn from standard Gaussian distributions and uniform distributions.

We define treatment and outcome models with a logistic-linear model as follows:  $\text{logit}(e(Z_i)) = \alpha^T Z_i$ . and  $Y_i \sim \mathcal{N}((\beta^T Z_i + \tau W_i, \sigma^2))$ . We add an additive noise term in the outcome model such that the signal-to-noise ratio (SNR) is either 5 or 10 (results are only reported for the case  $SNR = 10$ ).

Missing values are generated completely at random (MCAR), i.e.,  $\mathbb{P}(M_{ij} = 1) = \rho$ ,  $\forall i, \forall j$ , with  $\rho \in \{0, 0.3, 0.5, 0.9\}$  and we consider the following problem dimensions:  $n \in \{1000, 10000\}$ ,  $p \in \{10, 100, 1000\}$ , and  $d \in \{2, 10\}$ . Results are averaged over 30 replications for each setting. We only report results for  $n = 10,000$ , experiments with other choices of parameters are reported in the Supplementary Material. Throughout all experiments the true ATE  $\tau$  is fixed at 1.<sup>1</sup>

### 4.3. Results

#### 4.3.1. REGRESSION ADJUSTMENT

First, we assess the quality of our heuristic described in Section 3.3 concerning the non-linear extension of [Kallus et al. \(2018\)](#). An estimation of  $\tau$  is obtained by regressing the observed outcomes  $Y$  on the estimations of the latent factors  $Z$  (for `MDC.process`, `MF`) and on the imputed data  $X_{imp}$  (for `MI`).

Figures 3 and 4 show that our proposed method, `MDC.process` tends to slightly outperform all other methods when the covariates are generated according to a DLVM model. As expected the performances of all the methods decrease when the percentage of missing values increase, and both `MF` and `MDC process` better recover the latent structure when  $p$  is larger. Additionally we find that when the data is generated under the LRMF model, then our method performs as well as the initial proposal of [Kallus et al. \(2018\)](#) (results are reported in the Supplementary Material).

#### 4.3.2. DOUBLY ROBUST ESTIMATION

Now we turn to the more flexible framework which does not assume linear relationships (5) between the outcome and

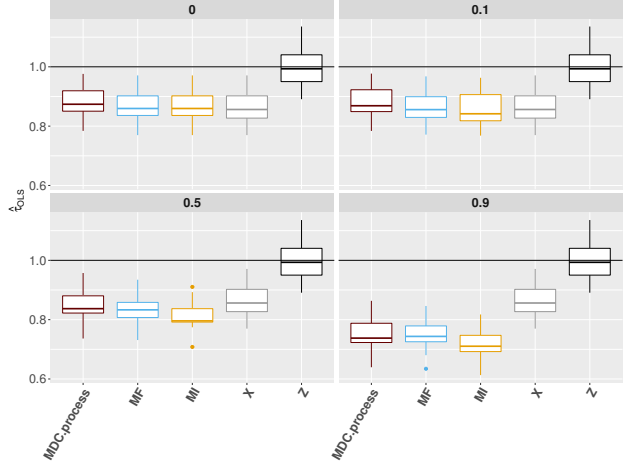


Figure 3. Estimated ATE via regression adjustment for varying amount of missing values; covariates generated from a DLVM, (logistic-)linear model specification for  $(e, \mu_0, \mu_1)$ ; results with  $Z$  results are obtained using the true confounders  $Z$ .  $(n, p, d) = (10000, 100, 2)$ .

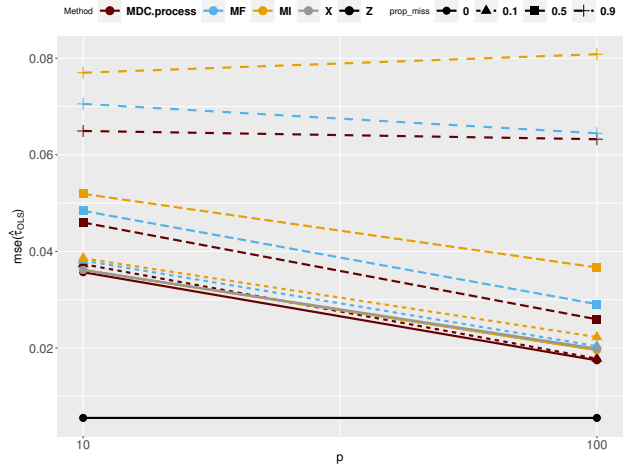


Figure 4. MSE of estimated ATE via regression adjustment for varying  $p$ ; covariates generated from a DLVM, (logistic-)linear model specification for  $(e, \mu_0, \mu_1)$ ; results with  $Z$  results are obtained using the true confounders  $Z$ .  $(n, d) = (10000, 2)$ .

the confounders. We consider the doubly robust estimator (1) with the (imputed) covariates  $X$  for `MI` and with the estimation of the latent variables  $Z$  for `MF` and `MDC`.

To estimate the regression surfaces  $(\mu_1, \mu_0)$  and the propensity score  $e$  required for the doubly robust estimator (1), we use a logistic-linear model, either with or without additional  $\ell_2$  regularization.

Figure 5<sup>2</sup> illustrates that even when the latent variables are

<sup>1</sup>Our code for these experiments is available at <https://github.com/imkemayer/MissDeepCausal>.

<sup>2</sup>The multiple imputation approach fails due to memory satura-

generated from matrix factorization, our approaches based on the VAE with missing values lead to unbiased estimates. We note as well that all methods perform similarly, independently of the number of observed covariates  $p$  (results for the other values of  $p$  are in the Supplementary Material).

Figures 6 and 7<sup>3</sup> show that as expected, due to the flexibility of MissDeepCausal, the suggested approaches better handle highly non-linear relationships between the latent confounders and the observed (incomplete) covariates. MDC methods are the only ones achieving no bias or small bias under this non-linear model. This is all the more true as the number of variables  $p$  is large compared to the dimension of the latent space  $d$ . The matrix factorization approach fails in this setting to recover the confounders  $Z$ .

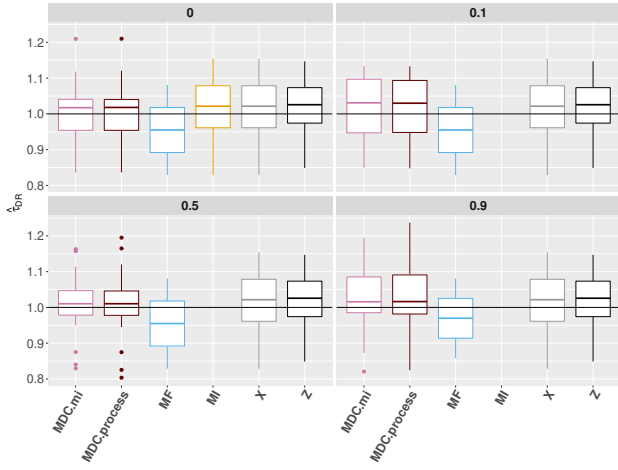


Figure 5. Estimated ATE via parametric doubly robust estimation for varying amount of missing values; covariates generated from a LRMF, (logistic-)linear model specification for  $(e, \mu_0, \mu_1)$ ; results with  $Z$  are obtained using the true confounders  $Z$ .  $(n, p, d) = (10\,000, 1\,000, 2)$ .

#### 4.4. IHDP data

We assess our methodology on the Infant Health and Development Program (IHDP) benchmark data (Hill, 2011). The original data comes from a randomized control trial where the aim was to assess the impact of visits by specialists on children’s test scores. There are six quantitative and 19 binary variables, recorded for 985 individuals. Hill (2011) transformed the original experimental data into observational data by selecting a nonrandom subset among the treated, stratified along an ethnicity variable, which leads to two unbalanced treatment groups. In total there are 139 treated and 608 control observations in the new

<sup>3</sup>Again the multiple imputation approach fail due to memory saturation.

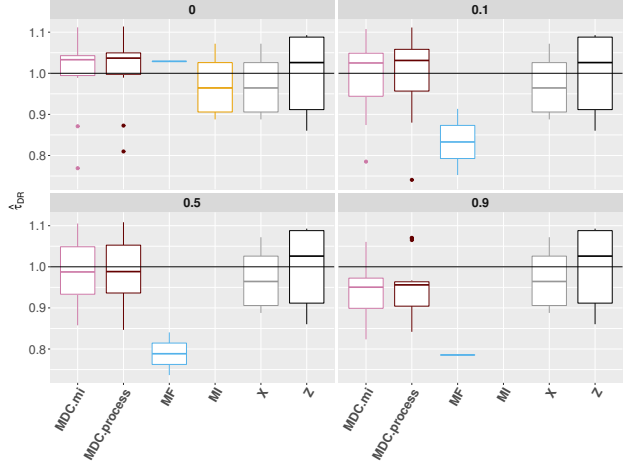


Figure 6. Estimated ATE via parametric doubly robust estimation for varying amount of missing values; covariates generated from a DLVM, (logistic-)linear model specification for  $(e, \mu_0, \mu_1)$ ; results with  $Z$  are obtained using the true confounders  $Z$ .  $(n, p, d) = (10\,000, 1\,000, 2)$ .

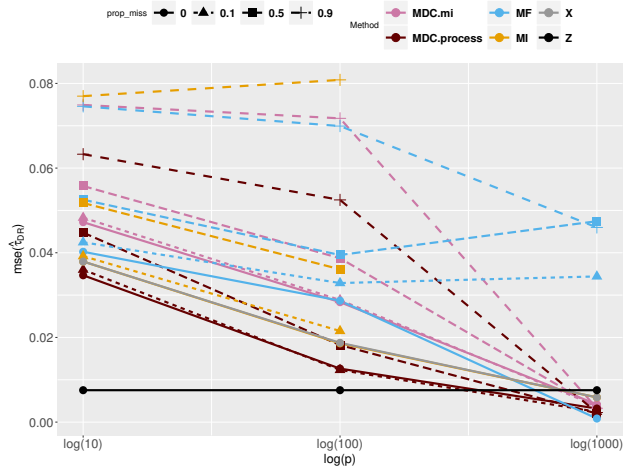


Figure 7. MSE of estimated ATE via parametric doubly robust estimation for varying  $p$ ; covariates generated from a DLVM, (logistic-)linear model specification for  $(e, \mu_0, \mu_1)$ ; results with  $Z$  are obtained using the true confounders  $Z$ .  $(n, d) = (10\,000, 2)$ .

data set. Then, keeping fixed the treatment variable, simulated data are obtained by generating new potential outcomes. More precisely, we follow the scenario “B” of (Hill, 2011), i.e.,  $Y(0) \sim \mathcal{N}(\mu_0, 1)$  and  $Y(1) \sim \mathcal{N}(\mu_1, 1)$ , with  $(\mu_0, \mu_1) = (\exp(X + W)\beta, X\beta - \omega)$  where  $\omega$  is chosen to get an average treatment effect  $\tau$  equal to 4.<sup>4</sup> After simulating the outcomes, we add missing values to the 25

<sup>4</sup>We use and adapt the corresponding code from V. Dorie: <https://github.com/vdorie/npci/>.

covariates, assuming an MCAR mechanism. For the MI-WAE part of our MDC methods, we select the parameters  $\sigma_{prior}$  and  $d_{miwae}$  by 5-fold cross-validation.

In addition to comparing the estimators considered in this paper that handle missing data, we also add two other approaches: the CEVAE estimator detailed in Louizos et al. (2017) as a baseline and the MIA.GRF estimator proposed in Mayer et al. (2019b). Note that CEVAE does not deal with missing values so that we replace the missing values by the mean of the variables. The CEVAE estimator is based on the difference between the two conditional expectations. The MIA.GRF estimator targets (4) and the generalized response surface analogue. It is based on estimation using random forests where missing values are encoded with *missing incorporated in attributes* such that the splitting rules in the random forests exploit the missingness pattern (Twala et al., 2008; Josse et al., 2019). We use the R package `grf` (Tibshirani et al., 2018) for the complete case and the implementation provided by Mayer et al. (2019b) for the incomplete case<sup>5</sup>.

Finally, we additionally apply a nonparametric doubly robust estimator, denoted by  $DR_{rf}$ , on the approximated confounders (resp. imputed covariates) based on (generalized) random forests (Athey et al., 2019). For this part we use the implementation of the R package `grf` (Tibshirani et al., 2018).

For comparability with previous experiments on these data, we report the in-sample mean absolute error, i.e. the mean absolute difference between the estimated ATE and the sample ATE (by construction of the data we know the exact values of  $\mu_{(1)}(X_i)$  and  $\mu_{(0)}(X_i)$  for all  $i$ ):  $\Delta = |\hat{\tau} - \frac{1}{n} \sum \mu_{(1)}(X_i) - \mu_{(0)}(X_i)|$ .

Table 1 shows that the doubly robust estimators (either in the parametric regression,  $DR_{log-lin}$ , or the random forest form  $DR_{rf}$ ) systematically outperform the corresponding OLS estimator which highlights that the linear model is not appropriate, at least that it is not linear in the covariates  $X$ . Indeed, we know that the outcome is simulated as a nonlinear function of the (complete) covariates  $X$ , whereas the treatment assignment is taken from the (de-randomized) experiment and can therefore well depend on latent variables. The results of MissDeepCausal are competitive with other approaches and greatly improve on CEVAE and MI. Its performances when used with the double robust estimators are stable with respect to the percentage of missing values.

<sup>5</sup><https://github.com/imkemayer/causal-inference-missing>

Table 1. Methods on the IHDP benchmark data. Mean absolute error  $\Delta$  (with standard error) across simulations on all the data points (in-sample error). OLS corresponds to the estimator obtained by regression and DR to the doubly robust estimator(s).

| % NA | Method              | OLS             | $\Delta$<br>$DR_{log-lin}$ | $DR_{rf}$       |
|------|---------------------|-----------------|----------------------------|-----------------|
| 0    | $X$ (complete data) | $0.72 \pm 0.02$ | $0.13 \pm 0.00$            | $0.20 \pm 0.01$ |
|      | $MF$                | $0.56 \pm 0.03$ | $0.14 \pm 0.01$            | $0.16 \pm 0.01$ |
|      | $MDC_{process}$     | $0.51 \pm 0.03$ | $0.15 \pm 0.01$            | $0.19 \pm 0.03$ |
|      | $MDC_{mi}$          | $0.47 \pm 0.03$ | $0.16 \pm 0.01$            | $0.14 \pm 0.02$ |
|      | $CEVAE(X)$          | $0.34 \pm 0.02$ |                            |                 |
| 10   | $MI$                | $0.85 \pm 0.02$ | $0.16 \pm 0.00$            | $0.24 \pm 0.01$ |
|      | $MIA.GRF$           | —               | —                          | $0.23 \pm 0.01$ |
|      | $MF$                | $0.50 \pm 0.03$ | $0.15 \pm 0.01$            | $0.15 \pm 0.01$ |
|      | $MDC_{process}$     | $0.42 \pm 0.02$ | $0.15 \pm 0.01$            | $0.16 \pm 0.02$ |
|      | $MDC_{mi}$          | $0.35 \pm 0.02$ | $0.17 \pm 0.01$            | $0.13 \pm 0.02$ |
|      | $CEVAE(X_{imp})$    | $0.31 \pm 0.01$ |                            |                 |
| 30   | $MI$                | $1.20 \pm 0.02$ | $0.30 \pm 0.00$            | $0.32 \pm 0.01$ |
|      | $MIA.GRF$           | —               | —                          | $0.17 \pm 0.01$ |
|      | $MF$                | $0.39 \pm 0.02$ | $0.16 \pm 0.01$            | $0.17 \pm 0.01$ |
|      | $MDC_{process}$     | $0.37 \pm 0.02$ | $0.16 \pm 0.01$            | $0.15 \pm 0.02$ |
|      | $MDC_{mi}$          | $0.30 \pm 0.02$ | $0.18 \pm 0.01$            | $0.13 \pm 0.01$ |
|      | $CEVAE(X_{imp})$    | $0.38 \pm 0.02$ |                            |                 |
| 50   | $MI$                | $1.54 \pm 0.03$ | $0.46 \pm 0.01$            | $0.42 \pm 0.01$ |
|      | $MIA.GRF$           | —               | —                          | $0.19 \pm 0.01$ |
|      | $MF$                | $0.28 \pm 0.01$ | $0.20 \pm 0.01$            | $0.21 \pm 0.02$ |
|      | $MDC_{process}$     | $0.24 \pm 0.01$ | $0.20 \pm 0.01$            | $0.21 \pm 0.02$ |
|      | $MDC_{mi}$          | $0.18 \pm 0.01$ | $0.22 \pm 0.01$            | $0.22 \pm 0.03$ |
|      | $CEVAE(X_{imp})$    | $0.38 \pm 0.02$ |                            |                 |

## 5. Conclusion

In this work we have investigated the problem of treatment effect estimation with incomplete covariates. This problem of missing values is highly relevant for modern causal inference as it is exacerbated with high dimensional data. Yet most causal inference techniques do not address this issue; and complete case analysis, in addition to leading to potentially inconsistent causal effects estimators, is not an option anymore. We have proposed MissDeepCausal which borrows the strength of deep latent variable models to retrieve the latent confounders from incomplete covariates encoding complex non-linear relationships. We use a modular approach in the style of Bayesian propensity based methods for treatment effect estimation (Zigler, 2016), where the latent variables are used as inputs for doubly robust estimators. We suggest a multiple imputation strategy that allows to fully exploit the posterior distribution of the latent variables. Numerical results are very encouraging insofar as we obtain best relative performance in terms of bias and MSE whether the underlying model is well or badly specified compared to current state of the art. Open challenges include heterogeneous treatment effect estimation with missing values as well as the ambitious task of handling missing not at random type data.

## References

- Athey, S., Tibshirani, J., Wager, S., et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.



- Blake, H. A., Leyrat, C., Mansfield, K., Seaman, S., Tomlinson, L., Carpenter, J., and Williamson, E. Propensity scores using missingness pattern information: a practical guide. *arXiv preprint*, 2019.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.
- D’Agostino Jr, R., Lang, W., Walkup, M., Morgan, T., and Karter, A. Examining the impact of missing data on propensity score estimation in determining the effectiveness of self-monitoring of blood glucose (smbg). *Health Services and Outcomes Research Methodology*, 2(3-4): 291–315, 2001. doi: 10.1023/A:102037541.
- D’Agostino Jr, R. B. and Rubin, D. B. Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, 95(451):749–759, 2000. doi: 10.2307/2669455.
- Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. A survey of learning causality with data: Problems and methods. *arXiv preprint arXiv:1809.09337*, 2019.
- Hastie, T. and Mazumder, R. *softImpute: Matrix Completion via Iterative Soft-Thresholded SVD*, 2015. R package version 1.4.
- Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research*, 16: 3367–3402, 2015.
- Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.
- Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.
- Iacus, S. M., King, G., and Porro, G. Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24, 2012. doi: 10.1093/pan/mpr013.
- Imbens, G. W. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.
- Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.
- Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029, 2016.
- Josse, J. and Reiter, J. P. Introduction to the special section on missing data. *Statist. Sci.*, 33(2):139–141, 05 2018. doi: 10.1214/18-STS332IN.
- Josse, J., Sardy, S., and Wager, S. denoiser: A package for low rank matrix estimation. *Journal of Statistical Software*, 2016.
- Josse, J., Prost, N., Scornet, E., and Varoquaux, G. On the consistency of supervised learning with missing values. *arXiv preprint*, 2019.
- Kallus, N., Mao, X., and Udell, M. Causal inference with noisy and missing covariates via matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6921–6932, 2018.
- Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.
- Kuroki, M. and Pearl, J. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.
- Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*. Wiley, 2002. ISBN 0471183865. doi: 10.2307/1533221.
- Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.
- Lunceford, J. K. and Davidian, M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004. doi: 10.1002/sim.1903.
- Mattei, A. and Mealli, F. Estimating and using propensity score in presence of missing background data: an application to assess the impact of childbearing on wellbeing. *Statistical Methods and Applications*, 18(2):257–273, 2009. doi: 10.1007/s10260-007-0086-0.
- Mattei, P.-A. and Frellsen, J. MIWAE: Deep generative modelling and imputation of incomplete data sets. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4413–4423, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

- Mayer, I., Josse, J., Tierney, N., and Vialaneix, N. R-miss-tastic: a unified platform for missing values methods and workflows. *arXiv preprint arXiv:1908.04822*, 2019a.
- Mayer, I., Wager, S., Gauss, T., Moyer, J.-D., and Josse, J. Doubly robust treatment effect estimation with missing attributes. *arXiv preprint arXiv:1910.10624*, 2019b.
- Miettinen, O. S. *Theoretical epidemiology: principles of occurrence research*. John Wiley & Sons, New York, 1985. ISBN 0827343132.
- Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. pp. 1278–1286, 2014.
- Richardson, T. S. and Robins, J. M. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. Technical report, Center for Statistics and the Social Sciences, University of Washington, 2013.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. doi: 10.1080/01621459.1994.10476818.
- Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.
- Rosenbaum, P. R. and Rubin, D. B. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984. doi: 10.2307/2288398.
- Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.
- Rubin, D. B. Inference and missing data. *Biometrika*, 63(3): 581–592, 1976.
- Rubin, D. B. *Multilpe Imputation for Nonresponse in Surveys*. Wiley, Hoboken, NJ, USA, 1987. ISBN 9780471655740.
- Seaman, S. and White, I. Inverse probability weighting with missing predictors of treatment assignment or missingness. *Communications in Statistics-Theory and Methods*, 43(16):3499–3515, 2014. doi: 10.1080/03610926.2012.700371.
- Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3076–3085. JMLR. org, 2017.
- Tibshirani, J., Athey, S., Wager, S., Friedberg, R., Miner, L., and Wright, M. *grf: Generalized Random Forests (Beta)*, 2018. R package version 0.10.2.
- Twala, B., Jones, M., and Hand, D. J. Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, 29(7):950–956, 2008.
- van Buuren, S. *Flexible Imputation of Missing Data*. Chapman and Hall/CRC, Boca Raton, FL, 2018.
- Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.
- Yang, S., Wang, L., and Ding, P. Causal inference with confounders missing not at random. *Biometrika*, 2019.
- Yoon, J., Jordon, J., and van der Schaar, M. Ganite: Estimation of individualized treatment effects using generative adversarial nets. 2018.
- Zigler, C. M. The central role of bayes theorem for joint estimation of causal effects and propensity scores. *The American Statistician*, 70(1):47–54, 2016.

## A. Supplementary results

In this supplementary material we provide additional results for our simulation study. Namely, Figures S.8 and S.9 are complementary to Figures 4 and 7 respectively where instead of varying the dimension of the ambient space  $p$ , we vary the number of observations  $n$ . Figure S.10 illustrates the result of similar performance of all methods in the case of data generated under a LRMF model mentioned in Section 4.

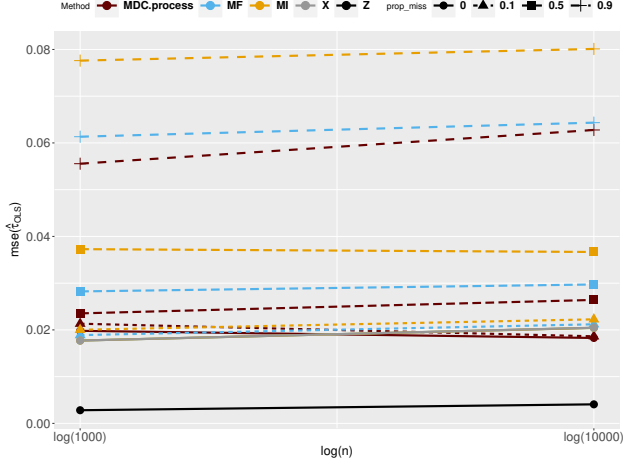


Figure S.8. MSE of estimated ATE via regression adjustment for varying  $n$ ; covariates generated from a DLVM, (logistic-)linear model specification for  $(e, \mu_0, \mu_1)$ ; results with  $Z$  results are obtained using the true confounders  $Z$ .  $(p, d) = (100, 2)$ .

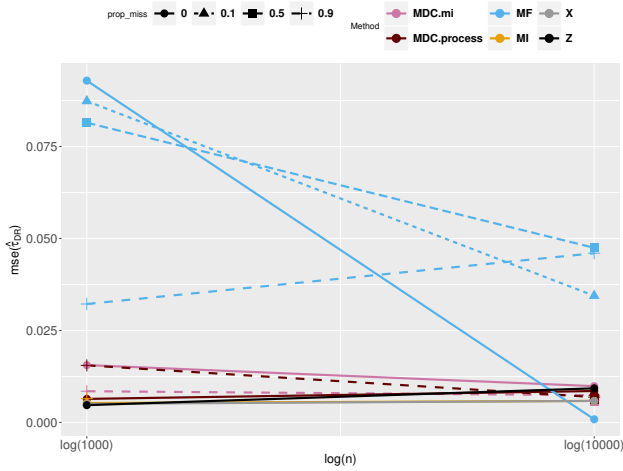


Figure S.9. MSE of estimated ATE via parametric doubly robust estimation for varying  $n$ ; covariates generated from a DLVM, (logistic-)linear model specification for  $(e, \mu_0, \mu_1)$ ; results with  $Z$  are obtained using the true confounders  $Z$ .  $(p, d) = (1000, 2)$ .

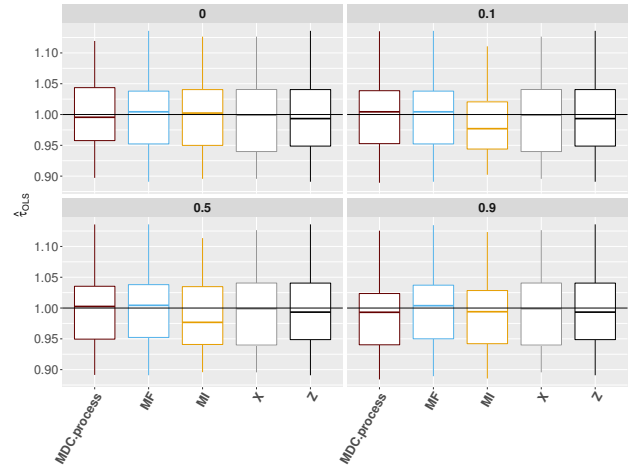


Figure S.10. Estimated ATE via regression adjustment for varying amount of missing values; covariates generated from a LRMF, (logistic-)linear model specification for  $(e, \mu_0, \mu_1)$ ; results with  $Z$  results are obtained using the true confounders  $Z$ .  $(n, p, d) = (10000, 100, 2)$ .