Doubly robust treatment effect estimation with incomplete confounders

Imke Mayer[1,2], Julie Josse[2,3], Stefan Wager[4]

[1] Centre d'Analyse et de Mathématique Sociales, École des Hautes Études en Sciences Sociales, 75006 Paris, France.
[2] Centre de Mathématiques Appliquées, École Polytechnique, 91120 Palaiseau, France.
[3]XPOP project-team, INRIA Saclay, 91120 Palaiseau, France.
[4]Graduate School of Business, Stanford University, CA 94305, USA.

Abstract

In healthcare and social sciences research, prospective observational studies are frequent, relatively easily put in place (compared to experimental randomized trial studies for instance) and can allow for different kinds of posterior analyses such as causal inferences. Average treatment effect (ATE) estimation for instance is possible through the use of propensity scores which allow to correct for treatment assignment biases in the non-randomized study design. However, a major caveat of large observational studies is their complexity and incompleteness: the covariates are often taken at different levels and stages, they can be heterogeneous – categorical, discrete, continuous – and almost inevitably contain missing values. The problem of missing values in causal inference has long been ignored and only recently gained some attention due to the non-negligible impacts in terms of bias induced by complete case analyses and misspecified imputation models. We discuss conditions under which causal inference can be possible despite missing confounder values, namely unconfoundedness on the observed values; we propose two alternative ATE estimators which directly account for the missing values, the first is built on logistic-linear specification and observed likelihood, appropriate for data *missing at random*, while the second uses semi-parametric estimation based on random forests with the great advantage of handling data *missing not at random*. We compare these two estimators to different methods proposed in the past to deal with missing confounder values. We assess the performance of our estimators on a large prospective database containing detailed information about over 20,000 severely traumatized patients in France. Using the proposed ATE estimators and this database we study the effect on mortality of tranexamic acid administration to patients with traumatic brain injury in the context of critical care management.

Keywords: missing data, causal inference, potential outcomes, observational data, propensity score estimation, causal forest, major trauma, critical care management

1 Introduction

1.1 Hemorrhagic shock and traumatic brain injury in critical care management

Our work is motivated by a prospective observational database, the Traumabase®, that currently includes around 20,000 major trauma patients with 244 pre-hospital and hospital measurements[1]. This data is heterogeneous, being composed of both quantitative and categorical variables and it contains an important fraction of missing values in many of these variables. In this context we are interested in estimating the effect of tranexamic acid (TA), an antifibrinolytic agent that limits excessive bleeding, on mortality among patients with traumatic brain injury during their stay in the intensive care unit (ICU), based on the observational database. It is currently recommended to administer TA to patients with hemorrhagic shock but in practice it is difficult to determine whether a patient has an hemorrhagic shock in an early

[1]Major trauma is defined as any injury that potentially causes prolonged disability or death and it is a public health challenge and a major source of mortality and handicap around the world (Hay et al., 2017).

phase (Pommerening et al., 2015), leading to an important fraction of false positives, which allows us to study the effect of TA on traumatic brain injury patients, not necessarily having an hemorrhagic shock[2].

As in almost all areas of empirical research, the Traumabase also presents missing data as shown in Figure 1. There are various reasons why missing data may occur, including non-response, unavailability of measurements, and lost data. Straightforward application of causal inference methods in the presence of missing data is not possible and naive approaches such as complete-case analysis are known to heavily bias the treatment effect estimations (Bartlett et al., 2015). The Traumabase encodings of missing values is potentially rich in information since some missing values encodings such as *impossible* implicitly contain the information that a measurement could not be made due to the circumstances (potentially a *missing at random* case), whereas *not applicable* suggests that the measurement is not relevant and therefore not really missing

A detailed list of the confounders and predictors of the outcome, in-ICU mortality, that were predetermined by a Delphi method Jones and Hunter (1995) is given in the appendix.

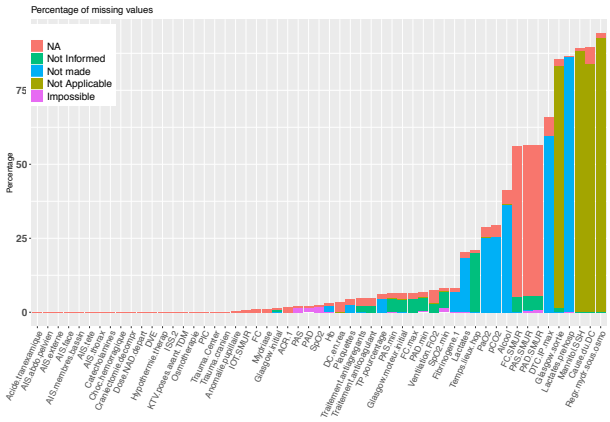


Figure 1: Percentage of missing values for each variable of the Traumabase (subset of variables relevent for traumatic brain injury). Different encodings of missing values: *NA* (not available), *not informed*, *not made*, *not applicable*, *impossible*.

1.2 Causal inference from observational data

Causal inference questions arise in many domains (socio-economy, politics, psychology, medicine, etc.) and are of the form "given the circumstances, what action should be taken to achieve a certain goal". The action could be the administration of a drug and its effect on the patient's health, or a marketing strategy for product placement and its effect on the consumer's purchase behaviour, etc. The notion of causal inference has not been addressed until the middle of the last century and has often been confounded with the notion of causality, a concept which cannot be of interest in statistics (Hernán and Robins, 2019). The causal inference formalism allows one to study questions like the one given above as a common estimation problem.It is commonly admitted that the gold standard for treatment effect estimation is a randomized controlled trial (RCT) that allows to estimate the average effect of a treatment, an intervention or a policy on a well defined population of interest. For instance, in pharmaceutical and medical research RCTs are compulsory for the authorization of new drugs or other treatments. However RCTs are generally very expensive in terms of time and financial costs and in recent years many RCTs lead to unsatisfactory results. Furthermore in some areas such as economics or political sciences, it is often impossible to implement an RCT to assess the effectiveness of a given intervention or policy, for instance the impact of a minimum wage policy on employment.

Once we have an understanding of causal relations between variables we can attempt to use this knowledge to make predictions, for instance treatment prescriptions, as opposed to ordinary predictions obtained with (supervised) learning algorithms applied directly on the data. Indeed probabilistic inference focuses on predicting consequences of observations by modeling the data distribution. Causal inference models the mechanism that generates the data and allows to predict results of interventions. But, as pointed out as the fundamental problem of causal inference by Holland (1986), we want to estimate

[2]The beneficial effect of TA on hemorrhagic shock patients has been demonstrated with the CRASH-2 trial Shakur et al. (2010).

something that we never observe since we never see the counterfactuals for a same individual at a same time (induced by different treatments or policies).

Despite this fundamental problem, there exists a multiplicity of well studied methods to efficiently and consistently estimate causal effects in different scenarios. One scenario that has not addressed intensively in the past is the case of missing confounder values in the context of causal inference. In this work we propose and compare several methods to handle missing values in the confounders, i.e. covariates that are associated both with treatment assignment and outcome, we discuss the underlying assumptions of these methods and assess them in simulations and finally we apply these methods to answer the medical question in the context of critical care management introduced above.

1.3 Notation

We consider a sample of n observations drawn from a population of size N. An observation i comprises a vector of covariates, $X^i = [X_1^i, \ldots, X_p^i]^T$, a treatment assignment variable W_i and an outcome Y_i. Depending on the context, the covariates for observation i will either be denoted by X^i or $X_{i,\cdot}$ (the i-th row from the covariate matrix $\boldsymbol{X} = [X_1, \ldots, X_p] \in \mathbb{R}^{n \times p}$). In the following expectations and probabilities will refer to the distribution induced by the random sampling from the population, or by the (conditional) random assignment of the treatment.

1.4 Definitions and assumptions

In light of our goal of performing causal analyses, we consider the potential outcomes framework from the Neyman-Rubin causal model (Rubin, 1974; Imbens and Rubin, 2015) and define potential outcomes $Y_i(w)$ for observation i and treatment $w \in \text{Span}(W)$[3]. In case of binary treatment assignment, for instance *treatment vs. control* or *treatment A vs. treatment B*, this leads to two potential outcomes, in some cases also referred to as counterfactuals, $Y_i(1)$ and $Y_i(0)$. The observed outcome for unit i is then defined as $Y_i \triangleq W_i Y_i(1) + (1 - W_i) Y_i(0)$. If not stated otherwise we always consider the binary treatment case and refer to individuals having $W_i = 1$ as *treated* and to those having $W_i = 0$ as *control*.

To assess the effect of a treatment we are interested in the individual treatment effect, which is defined for unit i as $\tau_i \triangleq Y_i(1) - Y_i(0)$ but which, by definition, is never observed. Faced with this impossibility to observe the quantity of interest τ_i, other treatment effect quantities are considered: averages of τ_i over different subsets of the original sample, for instance the average treatment effect, ATE, is defined as

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

The average treatment effect corresponds to the effect of switching every individual from one group to the other. There are similar quantities of interest such as the average treatment effect on the treated (ATT) or on the control (ATC) which allow to slightly relax some of the assumptions introduced below.

We distinguish two cases of data settings: *experimental data* from randomized controlled trials (RCT) where the covariate distributions (before treatment) between treated and control are identical and we know the law of the treatment assignment random variable. In general this is considered to be the gold standard for causal inference. However, in practice, such RCTs can come at high operational costs or can even be impossible in some domains such as social or political sciences. The second setting is different and is referred to as *observational data*: treated and control groups do not necessarily have the same distribution (before treatment) since the treatment assignment is not independent of the covariates and the potential outcomes. This invokes the notion of *confoundedness* which means that treatment assignment is not random due to the presence of confounding factors X as illustrated in Figure 2.

The *ignorability*, *unconfoundedness* or *exogeneity* assumption (the terms are equally used in the literature) states that all confounding factors are measured[4], i.e. conditionally on $\boldsymbol{X}$, the treatment assignment is independent of the potential outcomes. In other words, there is no unobserved confounding variable U in Figure 2. Formally this means

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

This assumption is necessary for identifiability of $\mathbb{E}[Y_i(w)]$, $w \in \{0, 1\}$; indeed the inherent problem of missing values due to counterfactuals can be handled under this assumption since $\mathbb{E}[Y_i(w) \,|\, W = w, X] = \mathbb{E}[Y_i(w) \,|\, X]$, $w \in \{0, 1\}$.

[3] Note that there are other approaches to causal inference than the potential outcome framework that have been proposed such as causal Bayesian networks and structural causal models.

[4] Assuming a constant treatment effect, unconfoundedness is equivalent to independence of treatment assignement and noise ε conditional on X, see Imbens (2004).

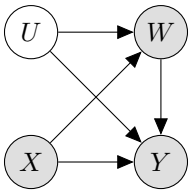


Figure 2: Observational data model with observed (X) and unobserved (U) confounding factors. We are interested in estimating the link between W and Y. We need to take into account the confounders, i.e. the common causes of W and Y.

We also make another standard assumption in the causal inference in the Neyman-Rubin framework (Imbens and Rubin, 2015) is the *Stable Unit Treatment Value Assumption* (SUTVA, Rubin (1978); Cox (1958)). This assumption translates into two aspects: the outcome of unit i is independent of the treatment assignment of other units and the treatment is stable, i.e. there are no multiple versions of the treatment which could lead to different outcomes. For instance if the treatment is *surgery* then we assume that the result of the surgery does not depend on the surgeon who operated the patient. Formally, the SUTVA assumption is:

$$Y_i = Y_i(W_i) \tag{3}$$
$$Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0). \tag{4}$$

Finally, an important assumption is that of *probabilistic treatment assignment*[5], i.e. if we define the propensity score $e(x) \triangleq \mathbb{P}(W_i = 1 \,|\, X_i = x)$ (Rosenbaum and Rubin, 1983; Imbens and Rubin, 2015), then we assume

$$0 < e(x) < 1 \quad \text{for all } x \in \mathcal{X}. \tag{5}$$

A well known and important result related to the unconfoundedness assumption is that if condition (2) holds then we also have

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \,|\, e(X_i) \qquad \text{for all } i, \tag{6}$$

which implies that instead of having to control for all covariates X_i one can limit oneself to controlling for $e(X_i)$[6]. Indeed Rosenbaum and Rubin (1983) show that the propensity score is a balancing score: it balances the two groups in terms of covariate distribution.

$$\mathbb{P}(X, W \,|\, e(X)) = \mathbb{P}(X \,|\, e(X)) \mathbb{P}(W \,|\, e(X)). \tag{7}$$

The proof holds in a couple of lines:

Proof. We have $\mathbb{P}(X, W \,|\, e(X)) = \mathbb{P}(X \,|\, e(X)) \mathbb{P}(W \,|\, X, e(X)) = \mathbb{P}(X \,|\, e(X)) \mathbb{P}(W \,|\, X)$ where the first equality always holds and the second one is given by the fact that $e(X)$ is a function of X, therefore conditioning on X is equivalent to conditioning on $X, e(X)$. Next we note that $\mathbb{P}(W = 1 \,|\, X) = e(X)$ by definition and $\mathbb{P}(W = 1 \,|\, e(X)) = \mathbb{E}[W \,|\, e(X)] = \mathbb{E}[\mathbb{E}[W \,|\, X] \,|\, e(X)] = \mathbb{E}[e(X) \,|\, e(X)] = e(X)$, which concludes the proof. □

Implicit assumptions in most works on causal inference are those of *perfect compliance*, i.e. every individual assigned to a group effectively belongs to this group (this is not always true, for instance in social sciences the perfect compliance assumption is often found to be violated), and of sufficient information, i.e. the data contains sufficient information to estimate the target causal effect, otherwise we cannot hope performing any causal inference.

For aspects of sensitivity analysis, especially for the unconfoundedness and probabilistic treatment assumptions, we refer to Rosenbaum (2010).

[5] Note that the coupling of probabilistic assignment and unconfoundedness assumptions is also referred to as *strongly ignorability* assumption (Rosenbaum and Rubin, 1983).

[6] This result can be extended to multivalued treatment, see Imbens (2000).

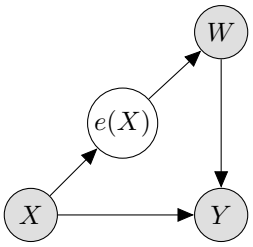


Figure 3: Observational data model with observed (and no latent) factors. We are interested in estimating the link between W and Y. The propensity score $e(X)$ breaks the path between confounders X and treatment.

1.5 RCT

If we assume that the treatment assignment is random such that the covariate distributions between treated and control are identical, then a simple difference of means is a consistent estimator of the ATE. More specifically,

$$\hat{\tau}_{DM} \triangleq \frac{1}{|\{i:\, W_i = 1\}|} \sum_{\{i:\, W_i=1\}} Y_i - \frac{1}{|\{i:\, W_i = 0\}|} \sum_{\{i:\, W_i=0\}} Y_i \tag{8}$$

is a consistent estimator of $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$. This is justified with the fact that treatment assignment W is independent from the potential outcomes $\{Y(1), Y(0)\}$ and thus $\mathbb{E}[Y_i(w) \,|\, W_i = w] = \mathbb{E}[Y_i(w)]$ for $w \in \{0, 1\}$ which motivates the use of the difference in means estimator. Note that the propensity score in a randomized experiment is constant by definition, i.e. every unit has the same probability of treatment assignment, independently from its covariates X. An improvement in terms of variance reduction can be achieved by regressing the outcome on the covariates.

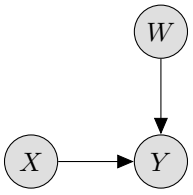


Figure 4: Randomized controlled trial data model with observed covariates. We are interested in estimating the link between W and Y.

1.6 Short review of existing methods for causal inference on observational data

Since we are interested in estimating the causal effects in observational data we cannot apply the above method since this would lead to inconsistent estimates due to confoundedness. But the basic idea behind the following methods is to emulate one or several RCTs, i.e. for a given $x \in \mathcal{X}$ we would like to estimate $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$ by a simple estimate as in the RCT case. For a more detailed review of existing literature on treatment effect estimation we refer to Imbens (2004); Lunceford and Davidian (2004).

1.6.1 Matching

Matching may be the most intuitive way to handle observational data. Adopting the perspective of Ho et al. (2007), matching methods can be considered as nonparametric data pre-processing methods, and therefore a possible preliminary step to statistical estimation.The choice of the latter can however be impacted by the chosen matching method. We refer the reader to Iacus et al. (2012) for a detailed review of existing matching methods (e.g. *one-to-one exact*, *exact matching*, *approximate*, *propensity score*, *coarsened exact* matching). In a nutshell, the aim of matching is to establish independence between the covariates X and treatment assignment W, by balancing the covariate distributions in both groups (without using the outcome variable Y). For instance one can adopt a nearest neighbor approach, i.e.

given a distance metric and an observation X^i of treatment group w, search for the nearest observation X^j with $W_j \neq w$.

Whatever matching strategy is chosen, the resulting balance quality can be assessed by different means. Ideally one compares the joint covariate distribution in both groups after matching but this becomes hard in high-dimensional settings. In this case comparing summary statistics such as mean differences, variance ratios or empirical CDF or different tests (t, F, Kolmogorov-Smirnov) can provide some information on the adequacy of the chosen matching strategy. Another possibility to measure covariate balance is to use the multivariate standardized bias introduced by Rosenbaum and Rubin (1985) which is a summary measure of (im)balance across all covariates.

1.6.2 Stratification

In a nutshell, stratification methods generalize matching to subpopulations. It allows to "match" subpopulations in treated and control with similar covariate distributions. Stratification on the propensity scores allows not only to balance treated and control groups but, as a consequence of Rosenbaum and Rubin (1984, Theorem 1), it also allows to use F tests (on each covariate) to approximately assess the adequacy of the propensity model. However one drawback of stratification is that there is potential for remaining heterogeneity within strata leading to biased treatment effect estimations.

1.6.3 Regression adjustment

A more direct solution to estimate τ can be to define it as parameter of a regression model: $\mathbb{E}[Y \mid X, W] = \beta_0 + X\beta_1 + \tau W$. However, as pointed out by Rubin (1979), such regression adjustments are sensitive to model mis-specification if the two groups differ considerably in the covariates. In such a case of insufficient overlap between treated and control the regression involves extrapolation of treated and control in the different regions (Lunceford and Davidian, 2004).

1.6.4 Instrumental variables

An alternative approach to causal inference which does not require the somewhat strong and also untestable unconfoundedness assumption is the use of *instrumental variables*. An instrumental variable is defined as a variable that influences the treatment variable but that has no impact on the outcome conditionally on the treatment. For a detailed introduction to instrumental variables we refer the reader to Angrist et al. (1996).

1.6.5 Weighting methods

Weighting methods are used in observational studies for estimating the effect of a treatment or an intervention but also in surveys for estimating the mean of an outcome variable in the presence of unit nonresponse and there exists a broad literature on weighting methods (see for instance Imbens and Rubin (2015); Lunceford and Davidian (2004)). The goal of weighting is twofold: balance the empirical distributions of the observed covariates (to remove biases due to observed confounders or recover the observed structure of the target population) and to yield stable estimates of the parameters of interest (very large weights may overly influence the results and highly variable weights produce results with high variance (Little and Rubin, 2002)).

Inverse probability of treatment weighting (IPW) Originally proposed by Horvitz and Thompson (1952) in the context of survey theory in finite settings, the IPW estimator was re-defined in a more general context by Rosenbaum (1987). It is closely related to the difference of means estimator but with the main difference that the observations are weighted by the inverse of the propensity score, the probability of treatment given the covariates,

$$\hat{\tau}_{IPW_0} \triangleq \frac{1}{n}\sum_{i=1}^{n}\frac{W_i Y_i}{\hat{e}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}. \tag{9}$$

Assuming consistent propensity score estimations, this estimator $\hat{\tau}_{IPW}$ is an unbiased estimator of the ATE. Indeed it uses the fact that under SUTVA and unconfoundedness we have

$$\mathbb{E}\left[\frac{W_i Y_i}{e(X_i)}\right] = \mathbb{E}\left[\frac{W_i Y_i(1)}{e(X_i)}\right] \tag{10}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}_{\{W_i=1\}} Y_i(1)}{e(X_i)} | X_i, Y_i(1)\right]\right] \tag{11}$$

$$= \mathbb{E}\left[\frac{Y_i(1)}{e(X_i)}\mathbb{E}\left[\mathbb{1}_{\{W_i=1\}} | X_i, Y_i(1)\right]\right] \tag{12}$$

$$= \mathbb{E}[Y_i(1)]. \tag{13}$$

And similarly we also get $\mathbb{E}\left[\frac{(1-W_i)Y_i}{1-e(X_i)}\right] = \mathbb{E}[Y_i(0)]$.

In practice a normalized version of (9), derived in the context of survey sampling by Hájek (1971), is used since precision is often enhanced if using weighted averages for the two groups as pointed out by Kang et al. (2007), i.e.

$$\hat{\tau}_{IPW} \triangleq \left(\sum_{i=1}^{n} \frac{W_i}{\hat{e}(X_i)}\right)^{-1} \sum_{i=1}^{n} \frac{W_i Y_i}{\hat{e}(X_i)} - \left(\sum_{i=1}^{n} \frac{1-W_i}{1-\hat{e}(X_i)}\right)^{-1} \sum_{i=1}^{n} \frac{(1-W_i)Y_i}{1-\hat{e}(X_i)}, \tag{14}$$

which is justified by the fact that $\mathbb{E}\left[\frac{W_i}{e(X_i)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{W_i}{e(X_i)} | X_i\right]\right] = \mathbb{E}\left[\frac{e(X_i)}{e(X_i)}\right] = 1$.

If the propensity score is not known, a popular choice for estimating it is to use logistic regression: it allows to estimate the probabilities of treatment from the observed covariates and then one inverts these probabilities to calculate the weights. However it does not aim explicitly at covariate balance or at restraining the variability of the weights. Weights can therefore vary substantially and lead to instability in the estimates (Robins and Wang, 2000; Kang et al., 2007). A key difficulty is that estimating the treatment effect then involves dividing by either $e^{-X_i\beta}$ or $1 - e^{-X_i\beta}$. Hence small inaccuracies in the estimation of β can have large effects on the subsequent estimators, especially when the propensity score $e(x) = \mathbb{P}(W_i = 1 \,|\, X_i = x)$ can be close to zero and one; this problem can be even more important in high-dimensional settings where perfect separation can lead to estimates $\hat{e}(x)$ that are exactly zero or one (Hill et al., 2011).

If the propensity model is correctly specified, i.e. the distribution model for treatment assignment given the covariates, then it is correct to have highly variable weights; however this is hard to determine in practice. This explains the common practice of trimming extreme weights, but this is often done in an arbitrary way that introduces bias in the estimates (see Crump et al. (2009) for discussions of different methods and Li et al. (2018) for an alternative to trimming, *overlap weights*).

Another solution that is available in certain settings is to learn the weights (or the probabilities of treatment) with a nonparametric (machine learning) approach to obtain weights that are less sensitive to model misspecification. More specifically if the propensities are a more complex function of the covariates than logistic-linear, for instance involving nonlinearities, learning a richer propensity score model can be advantageous. Note that independently of the choice for estimating the propensity scores, if the estimations are consistent then the resulting ATE estimator is more efficient than the estimator using the true propensity score as shown by Hirano et al. (2003). Intuitively this can be explained by the motivation of estimating the propensity scores: it aims at recovering the assignment "policy" which lead to the observed samples and the estimations might account for additional variance in the sample which is not accounted for in the true propensity model. If the predictions are too accurate, i.e. if one achieves (almost) perfect separation, even on some held out test data, then this strongly suggests that the probabilistic treatment assumption might not be met.

Note that we can obtain the normalized IPW estimator (14) by a weighted simple linear regression of the outcome on the treatment variable, $Y_i \sim W_i$, with weights $\frac{W_i}{e(X_i)} + \frac{1-W_i}{1-e(X_i)}$.

Balancing and overlap weights Balancing weights can be viewed as a generalization of the above inverse propensity weighting. Observations are weighted such that their distribution approximates the distribution of a predefined target population and the potential outcomes are averaged over this target distribution. For instance in the case of inverse propensity weighting the target population is the entire population of treated and control. Overlap weights as defined in Li et al. (2018) are a special case of the class of balancing weights where each unit is weighted proportionally to its probability of being assigned to the opposite group. This weighting targets the subpopulation of units which receive either

treatment in substantial proportions. This new class of weights is motivated by the remark that rather than prioritizing good covariate balance between groups over generalizability to a recognizable target population one should rather investigate the "optimal" subpopulation for which the causal effect can be estimated with smallest variance (Crump et al., 2009).

Let $f(x)$ be the marginal density of the covariates X w.r.t. some base measure μ. The densities in each group, $f_1(x)$ and $f_0(x)$ are proportional to $f(x)e(x)$ and $f(x)(1-e(x))$ respectively. Given a target distribution $f(x)h(x)$, e.g., the distribution of the overall population or of the population of the treated, the idea is to use the weights $\frac{h(x)}{e(x)}$ and $\frac{h(x)}{1-e(x)}$ that allow to balance the covariate distributions towards the target distribution. In case of the overlap weights, $h(x) = e(x)(1-e(x))$ and this places more emphasis on units with propensity score close to 0.5. In a medical context these units could be seen as patients with ambiguous profiles which lead to an absence of consensus between experts. An attractive aspect of overlap weights is their small-sample exact balance property. More precisely they lead to exact balance in the means of any covariate between treated and control groups.

Another line of research for balancing weights can be found in Zubizarreta (2015). The idea is to formulate the problem of finding balancing weights with small variance in a convex constrained optimization problem.

Imai and Ratkovic (2013) derive a *covariate balancing propensity score* (CBPS) by focussing on robust propensity score estimation (instead of robust propensity score matching or weighting). This approach exploits both aspects of the propensity score, namely the covariate balancing property and its definition as conditional probability of treatment.

1.6.6 Doubly robust methods

The previously mentioned CBPS comes along with a property qualified as *double robustness*. Indeed it can be seen as a special case of the work of Robins et al. (1994) on regression with incomplete outcomes that lead to several doubly robust estimators. Another doubly robust estimator is the *augmented inverse propensity weighted estimator* (AIPW).

$$\hat{\tau}_{DR} \triangleq \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + W_i\frac{Y_i - \hat{\mu}_1(X_i)}{\hat{e}(X_i)} - (1-W_i)\frac{Y_i - \hat{\mu}_0(X_i)}{1-\hat{e}(X_i)}, \tag{15}$$

where $\hat{\mu}_t(x)$ are regression estimators of the conditional response surfaces $\mathbb{E}[Y_i(w) \mid X_i = x]$, $w \in \{0,1\}$, the regression being defined separately on each of the two groups. Despite its initial derivation in the context of regression if some of the outcomes are missing, the link to causal inference can be easily established by viewing each potential outcome as a separate case of this problem. For instance the outcomes $Y_i(1)$ are only observed for the treated and not for the control. The probability of observing $Y_i(1)$ given the covariates X_i is exactly the propensity score for observation i. Then consistently estimating the conditional response surfaces of the potential outcomes $\mathbb{E}[Y_i(1)|X_i]$ and $\mathbb{E}[Y_i(0)|X_i]$ allows to consistently estimate $\tau(x)$.

A common description of the AIPW estimator is that it tries to estimate two different nuisance components, i.e. the outcome model and the propensity model; it then achieves consistency if either of these components is itself estimated consistently, and efficiency if both components are estimated at fast enough rates. In order to show the double robustness of $\hat{\tau}_{DR}$, let us rewrite it by rearranging the terms:

$$\begin{aligned}\hat{\tau}_{DR} &= \frac{1}{n}\sum_{i=1}^{n}\frac{W_iY_i}{\hat{e}(X_i)} - \frac{W_i - \hat{e}(X_i)}{\hat{e}(X_i)}\hat{\mu}_1(X_i) - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-W_i)Y_i}{1-\hat{e}(X_i)} + \frac{W_i - \hat{e}(X_i)}{1-\hat{e}(X_i)}\hat{\mu}_0(X_i)\\ &=: \hat{\mu}_{1,DR} - \hat{\mu}_{0,DR}.\end{aligned}$$

First note that by the law of large numbers, $\hat{\mu}_{1,DR}$ and $\hat{\mu}_{0,DR}$ respectively estimate $\mathbb{E}[Y_i(1)] + \eta_1$ and $\mathbb{E}[Y_i(0)]+\eta_0$ where η_1 is given by $\eta_1 \triangleq \mathbb{E}\left[\frac{W_i - e(X_i)}{e(X_i)}(Y_i(1) - \mu_1(X_i))\right]$ and $\eta_0 \triangleq \mathbb{E}\left[\frac{W_i - e(X_i)}{1-e(X_i)}(Y_i(0) - \mu_0(X_i))\right]$. Indeed we have that

$$\begin{aligned}\mathbb{E}\left[\frac{W_iY_i}{e(X_i)} - \frac{W_i - e(X_i)}{e(X_i)}\mu_1(X_i)\right] &= \mathbb{E}\left[\frac{W_iY_i(1)}{e(X_i)} - \frac{W_i - e(X_i)}{e(X_i)}\mu_1(X_i)\right]\\ &= \mathbb{E}[Y_i(1)] + \mathbb{E}\left[\frac{W_i - e(X_i)}{e(X_i)}(Y_i(1) - \mu_1(X_i))\right],\end{aligned}$$

where the first equality results from SUTVA: $W_iY_i = W_i(W_iY_i(1) + (1 - W_i)Y_i(0)) = W_iY_i(1) + W_i(1 - W_i)Y_i(0)$. And similar for the derivation of η_0.

The double robustness can easily be shown by considering these two terms:

- If the propensity model $e(x)$ is correctly specified but the outcome model $(\mu_0(x), \mu_1(x))$ is mis-specified we have

$$\begin{aligned}\eta_1 &= \mathbb{E}\left[\mathbb{E}\left[\frac{W_i - e(X_i)}{e(X_i)}(Y_i(1) - \mu_1(X_i))\right] \mid Y_i(1), X_i\right] \\ &= \mathbb{E}\left[\frac{\mathbb{E}[W_i \mid Y_i(1), X_i] - e(X_i)}{e(X_i)}(Y_i(1) - \mu_1(X_i))\right] \\ &= \mathbb{E}\left[\frac{\mathbb{E}[W_i \mid X_i] - e(X_i)}{e(X_i)}(Y_i(1) - \mu_1(X_i))\right] = 0.\end{aligned}$$

  We use the unconfoundedness assumption to go from the second to the third line and the definition of the propensity score for the last equality.

- If the propensity model $e(x)$ is mis-specified but the outcome model $(\mu_0(x), \mu_1(x))$ is correctly specified we have

$$\begin{aligned}\eta_1 &= \mathbb{E}\left[\mathbb{E}\left[\frac{W_i - e(X_i)}{e(X_i)}(Y_i(1) - \mu_1(X_i))\right] \mid W_i, X_i\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{W_i - e(X_i)}{e(X_i)}\left(Y_i(1) - \mathbb{E}[Y_i \mid W_i = 1, X_i]\right) \mid W_i, X_i\right]\right] \\ &= \mathbb{E}\left[\frac{W_i - e(X_i)}{e(X_i)}\left(\mathbb{E}[Y_i(1) \mid W_i, X_i] - \mathbb{E}[Y_i \mid W_i = 1, X_i]\right)\right] \\ &= \mathbb{E}\left[\frac{W_i - e(X_i)}{e(X_i)}\left(\mathbb{E}[Y_i(1) \mid X_i] - \mathbb{E}[Y_i(1) \mid X_i]\right)\right] = 0,\end{aligned}$$

  where we use SUTVA and unconfoundedness to go from the third to the fourth line.

Analogously we obtain in both cases of mis-specification that $\eta_0 = 0$, proving the double robustness of $\hat{\tau}_{DR}$.

This AIPW estimator and other doubly robust variants attain the semiparametric efficiency bound for ATE estimation. This Cramer-Rao type bound is derived in Hahn (1998) for non-parametric average treatment effect estimation. Chernozhukov et al. (2018) detail the sufficient conditions for consistent semiparametric ATE estimation, namely overlap, sup-norm consistency, a $o(n^{-1})$ risk decay and cross-fitting:

$$\sup_{x \in \mathcal{X}} \left|\hat{\mu}_{(w)}(x) - \mu_{(w)}(x)\right| \xrightarrow{p} 0 \qquad \sup_{x \in \mathcal{X}} |\hat{e}(x) - e(x)| \xrightarrow{p} 0, \tag{16}$$

and

$$\mathbb{E}_{\hat{\mu}_{(w)}, X}\left[\left(\hat{\mu}_{(w)}\right)(X) - \mu_{(w)}(X)\right)^2\right] \mathbb{E}_{\hat{e}, X}\left[(\hat{e}(X) - e(X))^2\right] = o_P\left(\frac{1}{n}\right). \tag{17}$$

Under (16), (17) and the overlap assumption introduced earlier, the corresponding ATE estimators are guaranteed to be $\sqrt{n}$-consistent if $\hat{\mu}_{(w)}$ and $\hat{e}$ are estimated using *cross-fitting* (also called *sample splitting*). This last key element for consistency of such semiparametric estimators has been pointed out independently by Athey et al. (2019) and Chernozhukov et al. (2018).

1.7 Review of missing values mechanisms and methods

Missing values arise in a variety of domains and many methods – more or less generic – have been proposed to handle missing values for different purposes, yet they generally share a common framework that formalizes the missing values problem (Rubin, 1976). We will present this framework and then we will briefly discuss the main methods that allow statistical analysis with missing values. For a more detailed review of all existing methods we refer the interested reader to Little and Rubin (2002) and to the `R-miss-tastic`[7] platform on missing values methods and workflows, gathering references, lectures and R-tutorials on this topic.

[7] `https://rmisstastic.netlify.com`

1.7.1 Missing values mechanisms

The definitions of missing values mechanisms proposed by Rubin (1976) are based on realizations x_i of random variables X_i with distribution in $\mathbb{R}^p$.

Response (or equivalently missingness) information is encoded in binary variables $R_i \in \mathbb{R}^p$ such that $r_i = 1$ if x_i is observed and $r_i = 0$ otherwise. Additionally we denote by x_i^{obs} and x_i^{mis} the observed and the missing values of x_i. For simplicity, we will assume that the realizations $(x_i, r_i)_{1 \leq i \leq n}$, $n \in \mathbb{N}$, are i.i.d. samples from a distribution in the family $\mathcal{P} \triangleq \{p_\theta(x) q_\phi(r \mid x) : \theta \in \Theta, \phi \in \Phi\}$. Hence q_ϕ characterizes the missingness mechanism. Statistical inference is generally about estimating the parameter θ, a possible approach under some regularity assumptions and assuming fully observed x_i is maximum likelihood estimation: $\hat{\theta} \triangleq \arg\max_\theta \mathcal{L}(\theta)$ where $\mathcal{L}(\theta) \triangleq \prod_{i=1}^n p_\theta(x_i)$ is the likelihood. In order to perform maximum likelihood estimation on incomplete data x_i some assumptions on the mechanism q_ϕ have to be made as can be seen when writing out the full likelihood $\mathcal{L}_{full}$, which is obtained by integrating over the missing values:

$$\mathcal{L}_{full}(\theta, \phi) \triangleq \prod_{i=1}^{n} \int_{\mathcal{X}^{mis}} q_\phi(r_i|x_i) p_\theta(x_i) dx_i^{mis} \tag{18}$$

Since the parameter ϕ and a modeling of the missingness mechanism are generally not of interest, a more practical quantity, the observed likelihood $\mathcal{L}_{obs}$, can be derived, assuming that the missingness is *ignorable* (Little and Rubin, 2002). Ignorability requires functional independence of the two parameters θ and ϕ and that the missingness mechanism is either *missing completely at random* (MCAR) or *missing at random* (MAR). The former means that the missingness mechanism is independent of the data x, whereas the latter states that the missingness only depends on the observed values x^{obs}. More formally, given r and $x = (x^{obs}, x^{mis})$,

$$(MCAR) \quad \forall\, \phi, \forall\, x' = (x'^{obs}, x'^{mis}),\ q_\phi(r|x') = q_\phi(r|x) = q_\phi(r) \tag{19}$$

$$(MAR) \quad \forall\, \phi, \forall\, x'^{mis} \text{ such that } x' = (x^{obs}, x'^{mis}),\ q_\phi(r|x') = q_\phi(r|x) \tag{20}$$

$$\mathcal{L}_{obs}(\theta) \triangleq \prod_{i=1}^{n} q_\phi(r_i|x_i^{obs}) \int_{\mathcal{X}^{mis}} p_\theta(x_i) dx_i^{mis}. \tag{21}$$

This reduction to the observed likelihood is not possible if the missingness mechanism is nonignorable. In this case the mechanism is qualified as *missing not at random* (MNAR) and it formally states that the mechanism does not satisfy (19) or (20), in other words the missingness is allowed to depend on the missing values themselves. A classical example that illustrates well this case is the well known fact that very wealthy – but also very poor – people tend to keep their earnings secret, so in a survey they would leave out questions related to their earnings leading to missing values that are therefore missing not at random.

1.7.2 Missing values handled in the analysis: EM

The initial question being *how to perform statistical analyses with missing values* or more precisely in the parametric setting stated above *how to estimate the parameter* θ, a first solution is to adapt existing statistical methods to take into account the presence of missing values. As seen above under ignorability, i.e. assuming MCAR or MAR mechanism, the parameter θ can be estimated by maximum observed likelihood estimation. However, since the expression of $\mathcal{L}_{obs}$ involves integrating over all possible missing values, a direct maximization of $\mathcal{L}_{obs}$ is generally intractable but a well known solution to this is the *Expectation-Maximization* algorithm (EM) proposed by Dempster et al. (1977). It assumes that the joint distribution of missing and observed variables, $p_\theta(x) = p_\theta(x^{obs}, x^{mis})$ is explicitly known and it aims at maximizing the observed log-likelihood ℓ_{obs},

$$\ell_{obs}(\theta) = \log \left(\prod_{i=1}^{n} q_\phi(r_i|x_i^{obs}) \int_{\mathcal{X}^{mis}} p_\theta(x_i) dx_i^{mis} \right) = \sum_{i=1}^{n} \log \left(\int_{\mathcal{X}^{mis}} p_\theta(x_i) dx_i^{mis} \right) + \log q_\phi(r_i|x_i^{obs}), \tag{22}$$

where the last term is constant in θ, hence it can be dropped for finding (or approximating) the value θ that maximizes the observed log-likelihood.

The EM algorithm is an iterative algorithm starting at some initial $\theta^{(0)} \in \Theta$. Using Jensen's inequality, it consists in alternately taking the expectation of the complete-data log-likelihood $\ell(\theta; x^{obs}, x^{mis}) \triangleq \log p_\theta(x^{obs}, x^{mis})$ with respect to the conditional distribution of missing covariates parameterized by $\theta^{(t)}$ at step t and then finding $\theta^{(t+1)}$ by maximizing this expectation in θ:

$$\begin{aligned} \text{E(xpectation) step:} \quad Q(\theta|\theta^{(t)}) &\triangleq \sum_{i=1}^{n} \mathbb{E}[\ell(\theta; x_i^{obs}, x_i^{mis})|X_i^{obs} = x_i^{obs}; \theta^{(t)}] \\ &= \int \ell(\theta; x_i^{obs}, x_i^{mis}) p_{\theta^{(t)}}(x_i^{mis}|x_i^{obs}) dx_i^{mis}, \end{aligned} \tag{23}$$

$$\text{M(aximization) step:} \quad \theta^{(t+1)} \in \arg\max_\theta Q(\theta|\theta^{(t)}). \tag{24}$$

An important property of this algorithm is that the sequence $(\theta^{(t)})_{t\geq 0}$ is guaranteed to increase the observed log-likelihood $\ell_{obs}(\theta^{(t)})$, however there is no guarantee for convergence towards a global maximum.

A supplemented EM algorithm (SEM) allows to estimate the variance of the resulting maximum likelihood estimate $\hat{\theta}_{MLE}$ (Meng and Rubin, 1991). Alternatively one can use Louis' formula to estimate $Var(\hat{\theta}_{MLE})$ (Louis, 1982).

1.7.3 Missing values handled in pre-processing: imputation

A drawback of the expectation-maximization algorithm is its lack of genericness: the E and M steps have to be derived for every statistical method and these derivations can involve complicated or intractable terms hindering the implementation of a computationally efficient estimation algorithm. Since most of the existing statistical methods are designed for complete data another idea consists in *imputing*, i.e. filling in, the missing values to recover a complete dataset (Rubin, 1987). There exist several approaches to impute the data with "plausible" values: assuming a known joint distribution of the data, *joint modeling* consists in exploiting this knowledge to impute the missing values based on the observed values (Little and Rubin, 2002). Other approaches are based on low-rank modeling of the data (Hastie et al., 2015; Josse et al., 2011) or on fully conditional specification (FCS) (van Buuren, 2018; Stekhoven and Bühlmann, 2012). Following the trend of deep learning now there also exist imputation methods based on generative adversarial networks (Yoon et al., 2018) and denoising autoencoders (Gondara and Wang, 2018).

If the goal is to perform statistical inferences, then a single imputation, i.e. replacing each missing value with one plausible value to get a single completed dataset, is not sufficient to take into account the additional variability due to missing values and therefore a multiple imputation (MI) strategy Rubin (1987) is adopted with the imputation methods listed above. The principle of MI is proposing M different (plausible) values for each missing value. The variability across these imputations reflects the variance of the imputation of the missing values. Statistical analyses are then carried out separately over the M resulting imputed datasets and the M estimations $(\theta_m)_{1\leq m\leq M}$ are combined according to Rubin's rules Rubin (1987) to obtain a single estimate $\hat{\theta}$ with a well estimated variance, i.e. taking into account the additional uncertainty due to the missing values.

1.7.4 Missing values in the context of supervised learning

Previously we discussed estimation problems in the presence of missing values, i.e. the estimation of some parameter $\theta \in \Theta$. However, if the goal is to make predictions about a response variable y given the information x there exist other approaches to handle missing values in x that are not about accurate imputation of x or good parameter estimation. For instance, random trees (Breiman et al., 1984) are nonparametric models that aim at estimating discriminative models, allowing to predict y given x. An appealing property of tree-based models is their ability to handle semi-continuous variables therefore allowing for missing values in the data. One solution that takes into account the missingness in the discriminative model estimation is *missing incorporated in attributes* (MIA) (Twala et al., 2008; Josse et al., 2019). It allows optimal splits along the observed variables X^{obs} and the response pattern R. Another, conceptually even simpler approach for prediction with incomplete data is mean imputation which is consistent, provided that one uses a learning algorithm with infinite learning capacity (Josse et al., 2019).

1.8 Related work

The literature on consistent ATE estimation is well known and has explored different types of problems, and we briefly described the main methods in Section 1.6 but do not apply in a straightforward manner since we are faced with missing values in the confounders. The problem of missing values (in the covariates) in the context of causal inference however has only been partially explored, with focus on propensity score methods. In a nutshell there are two lines of works based on different types of assumptions: the first adapts the initial ignorability/unconfoundedness assumptions such that treatment assignment is ignorable given only the observed covariates and the missingness pattern; the second uses standard missing values mechanism assumptions such as MAR and studies different multiple imputation strategies.

1.8.1 Inverse probability of treatment weighting

Generalized propensity score The starting point for the first strategy is given in Rosenbaum and Rubin (1984) who propose a generalized propensity score analysis which uses a *pattern mixture model* (or *missingness pattern approach*, MPA) for estimating propensity scores. They define the *generalized propensity score* e^* by conditioning treatment assignment on the covariates $X \in \mathcal{X}$ (where $\dim(\mathcal{X}) = p$) and the response pattern $R \in \{0,1\}^p$. According to this response pattern one can express X as $X = (X^{obs}, X^{mis})$ where $X^{obs} = \{X_j : R_j = 1\}$. Alternatively one defines a new variable $X^* \in \{\mathbb{R}, *\}^{n \times p}$ such that $X^*_{ij} = X_{ij}$ if $R_{ij} = 1$ and $X^*_{ij} = *$ otherwise. With these notations the generalized propensity score can be written as

$$e^*(X^{obs}, R) \triangleq \mathbb{P}(W = 1 \,|\, X^{obs}, R) = \mathbb{P}(W = 1 \,|\, X^*) = e^*(X^*) \tag{25}$$

This definition allows to balance treatment and control groups in the case of missing values as it is shown in Rosenbaum and Rubin (1984), assuming that the data is unconfounded given X^{obs} and R, i.e.

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \,|\, X_i^{obs}, R_i, \tag{26}$$

Note that the generalized propensity score only allows to balance the observed covariates, i.e. for observations with the same response pattern, balance is achieved on the observed covariates but not necessarily on the X^{mis}.

Proof. To prove the balancing property of the generalized propensity score, we note that the distribution of W is fully specified by its mean. Therefore we need to prove that:

$$\mathbb{E}[W_i|\{Y_i(0), Y_i(1)\}, X_i^*] = \mathbb{E}[W_i|X_i^*] \Rightarrow \mathbb{E}[W_i|\{Y_i(0), Y_i(1)\}, e^*(X_i^*)] = \mathbb{E}[W_i|e^*(X_i^*)]$$

a) By the law of total expectation we have:

$$\mathbb{E}[W_i \,|\, e^*(X_i^*)] = \mathbb{E}[\mathbb{E}[W_i|X_i^*, e^*(X_i^*)] \,|\, e^*(X_i^*)] = \mathbb{E}[\mathbb{E}[W_i|X_i^*] \,|\, e^*(X_i^*)] = e^*(X_i^*)$$

b) And again using the law of total expectation we have the following:

$$\begin{aligned}
\mathbb{E}[W_i \,|\, \{Y_i(0), Y_i(1)\}, e^*(X_i^*)] &= \mathbb{E}[\mathbb{E}[W_i \,|\, \{Y_i(0), Y_i(1)\}, X_i^*, e^*(X_i^*)] \,|\, \{Y_i(0), Y_i(1)\}, e^*(X_i^*)] \\
&= \mathbb{E}[\mathbb{E}[W_i|\{Y_i(0), Y_i(1)\}, X_i^*] \,|\, \{Y_i(0), Y_i(1)\}, e^*(X_i^*)] \\
&= \mathbb{E}[\mathbb{E}[W_i \,|\, X_i^*] \,|\, \{Y_i(0), Y_i(1)\}, e^*(X_i^*)] \qquad \text{(assuming (27))} \\
&= \mathbb{E}[e^*(X_i^*) \,|\, \{Y_i(0), Y_i(1)\}, e^*(X_i^*)] = e^*(X_i^*)
\end{aligned}$$

□

A major drawback of MPA is its applicability when there are many different patterns and the data is not sufficiently large to estimate one propensity model per pattern.

Missingness pattern approach with smoothing One possible workaround to handle the complexity of the MPA consists in smoothing the model by using all available observations for estimating the model for a given pattern and only use the estimated propensity scores for units with this exact pattern, e.g., D'Agostino Jr et al. (2001). For instance, assume $p = 4$ and missing values in X_2, X_3 and X_4, in order to estimate the propensity model for the pattern $r = (1, 0, 1, 0)$, this approach uses all observations X^k for which the first and the third value are observed but not necessarily with pattern $R^k = r$. The estimated model corresponding to this pattern is then used exclusively on observations with the exact pattern $R^i = r$.

Another solution makes use of the missing indicator method Miettinen (1985) and is discussed in D'Agostino Jr and Rubin (2000) for propensity score estimation in view of treatment effect estimation. They propose to model the joint distribution (X, W, R) by log-linear specification, more specifically using the general location model (Olkin et al., 1961). However, if there are many qualitative variables (even if each of them has only few categories) such an approach is hardly practicable.

These methods handle missing values indifferently from the missing data mechanism. However a drawback of these approaches is that they assume that for each missingness pattern the propensity score only depends on the observed covariates, in other words if a covariate is not observed then it is not a confounder.

Underlying unconfoundedness assumptions Recently, Blake et al. (2019) have proposed a detailed discussion about the plausibility and means of verification of the underlying assumptions of the approaches proposed for instance in Rosenbaum and Rubin (1984); D'Agostino Jr and Rubin (2000). Indeed the necessary assumption evoked above (26) can be formally written as

$$\begin{aligned} &\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \,|\, (X_i^{obs}, X_i^{mis}), R_i, \\ &\begin{cases} \text{CIT:} & W_i \perp\!\!\!\perp X_i^{mis} \,|\, X_i^{obs}, R_i \\ \textbf{or} & \\ \text{CIO:} & Y_i(t) \perp\!\!\!\perp X_i^{mis} \,|\, X_i^{obs}, R_i \quad \text{for } w \in \{0,1\}. \end{cases} \end{aligned} \tag{27}$$

Blake et al. (2019) refer to these assumptions as Missingness Strongly Ignorable Treatment Assignment[8], Conditionally Independent Treatment and Conditionally Independent Outcomes, respectively. As pointed out already by Mattei and Mealli (2009) the CIT and CIO resemble the missing data mechanism assumptions (MCAR, MAR, MNAR) but they are not about the missing data mechanism but rather about the relationships between missing confounder values and the treatment or outcome. In order to consistently estimate the ATE by IPW using generalized propensity scores, the first and either one of the second and third must be satisfied. Note that one can equivalently assume (26). Blake et al. (2019) argue that causal diagrams build upon experts prior knowledge and d-separation arguments allow to test these assumptions in practice but they do not give an empirical sensitivity study of these assumptions. Note that none of the previously cited works discusses extensions to double robustness.

Mattei and Mealli (2009) point out that the additional problem of missing values in the confounders demands for either stronger assumptions on the treatment assignment mechanism, i.e. the unconfoundedness and probabilistic treatment assignment, or additional assumptions on the missing data mechanism. They compare several approaches based on the above assumptions to handle missing values in causal inference: complete case analysis, generalized propensity score estimation and multiple imputation. For multiple imputation they assume *latent ignorability of the assignment mechanism* (Frangakis and Rubin, 1999), i.e.

$$\mathbb{P}(W \,|\, X, R, Y(1), Y(0)) = \mathbb{P}(W \,|\, X), \tag{28}$$

and that the data is MAR given W and Y, that is $\mathbb{P}(R \,|\, X, W, Y(1), Y(0)) = \mathbb{P}(R \,|\, X^{obs}, W, Y)$ (where we recall that $Y = WY(1) + (1-W)Y(0)$ so that Y is the observed part of $(Y(1), Y(0))$). Seaman and White (2014) derive consistency for the multiple imputation approach as they prove that using multiple imputations for propensity score estimation gives a consistent treatment effect estimator if there are infinitely many imputations. More specifically assuming strong ignorability and covariates that are MAR given W and Y and correct model specification of the propensity model and of the distribution of covariates X given W and Y, $f(X \,|\, W, Y; \psi)$, they prove, by using estimating equations arguments, that improper multiple imputation (using the maximum observed likelihood estimator $\hat{\psi}$) and proper multiple imputation for ATE estimation are consistent for $M = \infty$. In both cases, the imputation is carried out using the covariates X, the treatment assignment W and the outcome Y. They also discuss the inclusion of the missingness indicator R which can reduce bias in some cases but which can also introduce bias in other cases.

Indeed, when imputing on all available information, including treatment and outcome, Mattei and Mealli (2009) empirically obtain that multiple imputation dominates the other methods they test. They suggest this behavior being partly due to the fact that multiple imputation allows to use different models for imputation and propensity score estimation, for instance the use of Y and W in the imputation. However the latent ignorability of the assignment assumption is stronger than the modified unconfoundedness (26).

[8]This is the *missingness* equivalent to the weak unconfoundedness assumption which, in the complete case, is sufficient to obtain unbiased treatment effect estimators Imbens (2000).

As a sidenote, we point out that multiple imputation can also be combined with (propensity score) matching to handle missing values in the confounders. Since this work's focus is on weighting methods rather than the matching approach we refer the interested reader to Hill (2004).

1.8.2 Latent confounders

In another line of work, Kallus et al. (2018) use a low-rank assumption on the covariate matrix to derive a consistency result in the case of linear regression model, i.e. $\mathbb{E}[Y|X,W] = X\alpha + \tau W$. They assume a low-rank model for the covariates, namely that the observed covariates X are noisy and/or incomplete proxies of the true latent confounders U: $X = UV' + \eta$ where η contains i.i.d sub-Gaussian error terms with zero mean and known variance. Hence they assume a linear regression model on (Y,U,W): $Y_i = U_i^T\alpha + \tau W_i + \varepsilon_i$. Furthermore they assume the MCAR mechanism for X. Under this model and appropriate assumptions on α, $\|U\|$, the relationship between U and W and the estimation of $col(U)$ by $col(\hat{U})$ (where $\hat{U}$ is obtained by low-rank matrix factorization of X) and assuming unconfoundedness, they prove that the resulting ATE estimator obtained by regressing Y on $\hat{U}$ and W is consistent. The strong model assumptions for this result are a limiting factor for extension to more complex scenarios, however they show that this low-rank matrix factorization approach also performs well empirically in conjunction with other ATE estimators (IPW, double-robust, propensity score matching).

2 Treatment effect estimation in the presence of missing values

As discussed in the previous sections there exist "good" treatment effect estimators in the sense that they are consistent and have a well-understood asymptotic sampling distribution (for instance a normal distribution). These estimators are valid in the complete case. The literature on treatment effect estimation in the presence of missing values is much less developed and except a few recent results (Rosenbaum and Rubin, 1984; Seaman and White, 2014; Kallus et al., 2018) there is a lack of theoretical results on statistical performance of treatment effect estimators in the presence of missing data, even though in most real-world applications datasets are incomplete. Note that consistency results are even more important in causal inference since there is no possibility of test-set validation due to the problem formulation in the potential outcome framework (as opposed to prediction tasks where we have a labelled test set to evaluate the learned model).

We propose two ATE estimators, both in an IPW and a double robust version, based on different assumptions and definitions of double robustness as we will detail below. Our methods differ from previous solutions in several ways:

- they allow for doubly robust estimation whereas previous solutions only consider regression adjustment (Kallus et al., 2018) or IPW (Seaman and White, 2014),
- they are based on consistency results that allow for theoretical guarantees of our methods in a wide range of scenarios,
- they can easily be implemented and allow a simple handling of MNAR in practice as opposed to a recent treatment effect estimation for confounders with MNAR missing values based on integral equations Yang et al. (2017).

2.1 Parametric IPW and double robust treatment effect estimation

Building on a recent result of efficient logistic regression in the presence of missing values (Jiang et al., 2018), we propose to adopt a parametric approach to ATE estimation: assuming a logistic-linear model for treatment and outcome model, parametrized respectively by some $\beta \in \mathbb{R}^p$ and $(\gamma_{(1)}, \gamma_{(0)}) \in \mathbb{R}^{2p}$, we leverage existing expectation maximization algorithms that allow for estimation of logistic and linear regression models in the presence of ignorable missing values. As announced in the review on missing values (see Section 1.7), expectation maximization allows to take into account missing values in a statistical model but it has to be derived separately for every model and it is not always applicable in practice due to computational complexity. For logistic regression, a stochastic approximation EM (SAEM) algorithm has recently been proposed that allows to efficiently estimate and predict from a logistic model (Jiang et al., 2018). Under the assumption that the data is MAR, we make use of this method to estimate the generalized propensity score e^* which directly leads to a first ATE estimator handling missing values and that is based on inverse propensity weighting:

$$\hat{\tau}_{EM,IPW} \triangleq \left(\sum_{i=1}^{n} \frac{W_i}{\hat{e^*}(X_i^*;\hat{\beta}_{SAEM})}\right)^{-1} \sum_{i=1}^{n} \frac{W_i Y_i}{\hat{e^*}(X_i^*;\hat{\beta}_{SAEM})} - \left(\sum_{i=1}^{n} \frac{1-W_i}{1-\hat{e^*}(X_i^*;\hat{\beta}_{SAEM})}\right)^{-1} \sum_{i=1}^{n} \frac{(1-W_i)Y_i}{1-\hat{e^*}(X_i^*;\hat{\beta}_{SAEM})}. \tag{29}$$

Under the modified unconfoundedness assumptions (27) discussed previously in Section 1.8.1 and assuming that the missing values are MAR, this estimator is a consistent estimator of the ATE τ.

In order to reduce sensitivity to model specification an augmentation of this estimator that leverages a priori on the outcome model $\mu_{(w)}$ can be defined as in the complete case (see Section 1.6.6). Here again we make use of an existing expectation maximization approach to estimate the parameter γ of a linear regression model in the presence of missing values. Under similar assumptions as for the IPW estimator (29) the resulting ATE estimator is consistent if either the treatment or the outcome model are correctly specified.

$$\hat{\tau}_{EM,DR} \triangleq \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(X_i^*;\hat{\gamma}_{(1),EM}) - \hat{\mu}_0(X_i^*;\hat{\gamma}_{(0),EM}) + W_i \frac{Y_i - \hat{\mu}_1(X_i^*;\hat{\gamma}_{(1),EM})}{\hat{e^*}(X_i;\hat{\beta}_{SAEM})} - (1-W_i)\frac{Y_i - \hat{\mu}_0(X_i^*;\hat{\gamma}_{(0),EM})}{1-\hat{e^*}(X_i^*;\hat{\beta}_{EM})}, \tag{30}$$

Since these estimators are based on expectation maximization algorithms we can directly derive the following consistency result for average treatment effect estimation under logistic-linear model specification.

Proposition 2.1. *Assume (a) multivariate Gaussian covariates $(X_1, \ldots, X_p)$, (b) missing values generated under MCAR, (c) modified unconfoundedness* (27) *and (d) logistic specification for the treatment assignment W and linear specification for the potential outcomes $(Y(0), Y(1))$. Then $\hat{\tau}_{EM,IPW}$ and $\hat{\tau}_{EM,DR}$ consistently estimate the average treatment effect τ. Furthermore $\hat{\tau}_{EM,DR}$ satisfies the double robustness property.*

Proof. From Jiang et al. (2018) we deduce that under the assumptions (a)-(d) the SAEM algorithm consistently estimates β, hence we obtain a consistent estimate of the risk $\mathbb{E}[W \mid X_{obs}]$. Under MCAR, this risk corresponds to the generalized propensity score e^*. Similarly, the regression vectors $\gamma_{(0)}$ and $\gamma_{(1)}$ from the two response surface models $\{\mu_w = \mathbb{E}[Y(w)|X]\}_{w\in\{0,1\}}$ can be consistently estimated by expectation maximization, assuming conditions (a)-(d). The consistency of $\hat{\tau}_{EM,IPW}$ and $\hat{\tau}_{EM,DR}$ follows immediately, using the balancing property of the generalized propensity score proven in Section 1.8.1.

The proof of the double robustness of $\hat{\tau}_{EM,DR}$ is analogous to the one in the complete case (see Section 1.6.6), by again additionally using the balancing property of the generalized propensity score. □

2.2 Nonparametric IPW and double robust treatment effect estimation

A drawback of the estimators proposed in the previous paragraphs is their strong model dependence. Indeed, in the complete case it has been demonstrated that model mis-specification can lead to large bias in the treatment effect estimation, for both the IPW and the double robust estimators (Kang et al., 2007).

Nonparametric estimation is a possible solution in this case, provided that the sample sizes are sufficiently large, and we therefore propose an IPW and a doubly robust estimator based on random forests that allow to incorporate missing values information, i.e. an implicit encoding of the response pattern R. Indeed, this approach leverages the ability of random forests to handle the half-discrete nature of variables with missing values, i.e. if $X \in \mathbb{R}$ has missing values, than X^* is defined over $(\mathbb{R}, *)$. Twala et al. (2008) propose a *missing incorporated in attributes* (MIA) approach to handle missing values in tree-based models and Josse et al. (2019) provide theoretical and empirical proof of the consistency of such an approach. This approach has the additional appeal of handling both ignorable and nonignorable missing values since the MIA approach allows tree-splits along the response patterns. On a practical side we choose to implement the MIA approach by replacing each missing value by two "surrogates", $-\infty$

and $+\infty$. This allows splits along the missingness pattern R if the missingness is informative and therefore selects patterns that are important for predicting the treatment assignment (and also the outcome) instead of adjusting one model per pattern as in the MPA approach.

Since we drop model specifications for this approach the consistency result for the associated IPW estimator $\hat{\tau}_{MIA,IPW}$ is based on consistent estimation of e^* (Josse et al., 2019) and for $\hat{\tau}_{MIA,DR}$ we additionally refer to semi-parametric estimation results for treatment effect estimation (Chernozhukov et al., 2018). As we define $\hat{\tau}_{MIA,IPW}$ and $\hat{\tau}_{MIA,DR}$ on the generalized propensity score we assume the same modified unconfoundedness assumptions (27) as in the parametric case. The expressions of $\hat{\tau}_{MIA,IPW}$ and $\hat{\tau}_{MIA,DR}$ are similar to those of $\hat{\tau}_{EM,IPW}$ and $\hat{\tau}_{EM,DR}$, the only difference relying in the estimation of e^* and $(\mu_{(w)})_{w\in\{0,1\}}$:

$$\begin{aligned}\hat{\tau}_{MIA,IPW} \triangleq &\left(\sum_{i=1}^{n}\frac{W_i}{\hat{e^*}(X_i^*)}\right)^{-1}\sum_{i=1}^{n}\frac{W_iY_i}{\hat{e^*}(X_i^*)}\\ &-\left(\sum_{i=1}^{n}\frac{1-W_i}{1-\hat{e^*}(X_i^*)}\right)^{-1}\sum_{i=1}^{n}\frac{(1-W_i)Y_i}{1-\hat{e^*}(X_i^*)},\end{aligned} \tag{31}$$

$$\begin{aligned}\hat{\tau}_{MIA,DR} \triangleq &\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(X_i^*)-\hat{\mu}_0(X_i^*)\\ &+W_i\frac{Y_i-\hat{\mu}_1(X_i^*)}{\hat{e^*}(X_i)}-(1-W_i)\frac{Y_i-\hat{\mu}_0(X_i^*)}{1-\hat{e^*}(X_i^*)},\end{aligned} \tag{32}$$

where $\hat{e^*}$ is obtained by random forest classification with MIA splits and $(\hat{\mu}_{(w)})_{w\in\{0,1\}}$ by random forest regressions with the MIA criterion.

Lemma 2.1 (Theorem 3, Josse et al. (2019))**.** *Let $X \in \mathbb{R}^p$ be a random vector with continuous density and let the response $Y \triangleq f(X)+\varepsilon$ such that $\|f\|_\infty < \infty$, $\mathbb{E}[\varepsilon]=0$ and $\varepsilon \perp\!\!\!\perp (X, R_1, \ldots, R_j)$. Assume missing values occur for variables $X_1, \ldots, X_j$ such that $(R_1, \ldots, R_j) \perp\!\!\!\perp (X_1, \ldots, X_j) \,|\, (X_{j+1}, \ldots, X_p)$ and such that for every pattern $r \in \{0,1\}^j \times \{1\}^{p-j}$, the functions $(x_{j+1}, \ldots, x_p) \mapsto \mathbb{P}(R = r \,|\, X_{j+1} = x_{j+1}, \ldots, X_p = x_p)$ are continuous. If the missing values in $(X_1, \ldots, X_j)$ are imputed with the respective means $(\mathbb{E}[X_1], \ldots, \mathbb{E}[X_j])$ and if a learning algorithm that is universally consistent when trained on a fully observed data set, then the prediction is equal to the Bayes function almost everywhere.*

This result motivates the use of decision trees on the mean imputed data to learn a prediction model for Y given X. The CART algorithm (classification and regression trees, Breiman et al. (1984)) is one of the most popular choices among tree-based methods. However an alternative splitting criterion, *missing incorporated in attribute* (Twala et al., 2008), is preferable in the case of incomplete data as shows the following result.

Lemma 2.2 (Proposition 2, Josse et al. (2019))**.** *Let the data generating process satisfy the assumptions of Lemma 2.1 and additionally assume missing values are generated under an MCAR mechanism. Then the "missing incorporated in attribute" approach (MIA) improves the prediction accuracy with respect to the CART approach.*

Proposition 2.2. *Under the assumptions of Lemma 2.2 and assuming the modified unconfoundedness (27), $\hat{\tau}_{MIA,IPW}$ and $\hat{\tau}_{MIA,DR}$ consistently estimate the average treatment effect τ.*

Proof. Using Lemmata (2.1) and (2.2) and the completely random response pattern, we obtain consistent estimates of $\mathbb{E}[Y(1)|X_{obs}] = \mathbb{E}[Y(1)|X^*]$, $\mathbb{E}[Y(0)|X^*]$ and of $e^*(X_{obs}) = e^*(X^*)$. Under the modified unconfoundedness assumption, it directly follows that $\hat{\tau}_{MIA,IPW}$ and $\hat{\tau}_{MIA,DR}$ are consistent estimates of τ. □

Remark 1. *Mixed data: Another important advantage of random forests, besides the accounting for (informative) missing values, is their ability to handle mixed data, i.e. both quantitative and qualitative variables in a same data set.*

2.3 Simulation study

We assess the performance of the previously introduced treatment effect estimators in different scenarios, modifying the confounders' correlation structure and the missingness mechanism. We focus on the doubly robust version of the proposed estimators and refer to the Appendix A.2 for results with the IPW estimators. For comparison, for the imputation approach we choose imputation by conditional equations (ICE) (van Buuren, 2018) that imputes using conditional models. We also compare our approach to the recent solution proposed by Kallus et al. (2018) based on low-rank matrix factorization.

All simulations are implemented in `R` R Core Team (2018), for the parametric $\hat{\tau}_{EM}$ we use the R-package `misaem`, for the nonparametric $\hat{\tau}_{MIA}$ we use the R-package `grf`, in particular its function `average_treatment_effect`, for generalized random forests (Athey et al., 2019), for multiple imputation we use the R-package `mice`, finally for the matrix factorization approach we adapt the implementation of the Kallus et al. (2018) based on the R-package `softImpute`.[9] We also report results for the alternative nonparametric strategy consisting in performing mean imputation the missing values and using random forests to estimate $e^*(\cdot)$ and $(\mu_{(w)}(\cdot))_{w\in=\{0,1\}}$.

2.3.1 Parametric double robustness

We illustrate the importance of the modified unconfoundedness assumption (26) for the first estimator $\hat{\tau}_{EM}$ using the following setting: we generate normally distributed confounders $X_{i\cdot} = [X_{i1} \ \dots \ X_{ip}]^T \sim \mathcal{N}(\mathbf{1}, \Sigma)$, $i \in \{1, \dots, n\}$ for $p = 10$, where $\Sigma = I - \rho(I-1)$, with $\rho \in \{0.3, 0.6\}$, $\mathbf{X} = [X_{1\cdot} \ \dots \ X_{p\cdot}]^T \in \mathbb{R}^{n\times p}$. We generate missing values either under MCAR (i.e. $\mathbb{P}(R_{ij} = 1) = 1 - \mathcal{B}(\eta)$ such that on average we have ηnp missing values), under MAR (we generate missing values in $X_{\cdot,1:5}$ depending on the fully observed $X_{\cdot,6:10}$ such that we get around $\eta np/2$ missing values) or under MNAR (missing values in $X_{\cdot,1:5}$ are generated depending on the quantiles of $X_{\cdot,1:5}$ such that there are about $\eta np/2$ missing values). In the results presented here we fix $\eta = 0.5$ and repeat each experiment 100 times. The results, estimations of the true ATE, are summarized in the form of boxplots in the Figures 5. To simulate treatment and outcome under the CIT/CIO assumption we proceed as follows:

- **CIT**: $W \sim X \odot R$ (where $R_{ij} = \mathbb{1}_{\{X_{ij} \ is \ observed\}}$ and $\odot$ = Hadamard product).
  Example: for fixed $\alpha \in \mathbb{R}^4$ and $\tau \in \mathbb{R}$:
  $r^i = (1,1,0,0,0,1,0,0,0,1) \Rightarrow logit(\mathbb{P}(W^i = 1 | X^i_{obs} = x^i_{obs}, R^i = r^i)) = \alpha_0 + \alpha_1 x^i_1 + \alpha_2 x^i_2 + \alpha_6 x^i_6 + \alpha_{10} z^i_{10}$
  $r^j = (0,1,0,0,0,0,0,0,0,0) \Rightarrow logit(\mathbb{P}(W^j = 1 | X^j_{obs} = x^j_{obs}, R^i = r^j)) = \alpha_0 + \alpha_2 x^j_2$

- **CIO**: $Y \sim X \odot R$.
  Example: for fixed $\beta \in \mathbb{R}^4$ and $\tau \in \mathbb{R}$:
  $r^i = (1,1,0,0,0,1,0,0,0,1) \Rightarrow \mathbb{E}(Y^i | X^i_{obs} = x^i_{obs}, R^i = r^i, W^i = w^i) = \beta_0 + \beta_1 x^i_1 + \beta_2 x^i_2 + \beta_6 x^i_6 + \beta_{10} x^i_{10} + \tau w^i$
  $r^j = (0,1,0,0,0,0,0,0,0,0) \Rightarrow \mathbb{E}(Y^j | X^j_{obs} = x^j_{obs}, R^i = r^j, W^j = w^j) = \beta_0 + \beta_2 x^j_2 + \tau w^j$

- **¬CIT**: $logit(\mathbb{P}(W^i = 1 | X^i = x^i)) = \alpha_0 + \alpha^T x^i$.

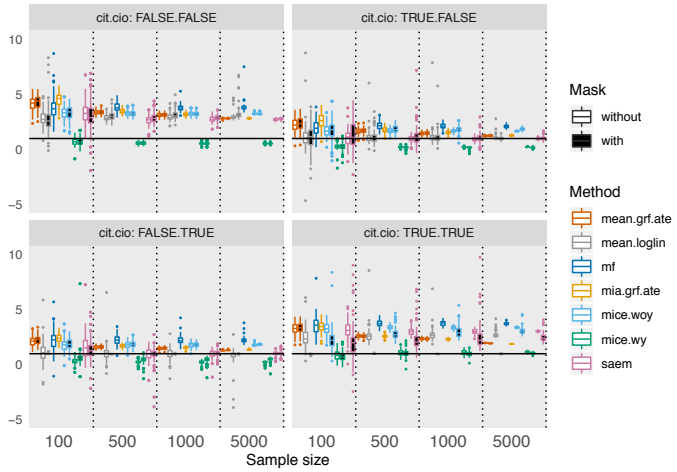
- **¬CIO**: $\mathbb{E}(Y^i | X^i = x^i, W^i = w^i) = \beta_0 + \beta^T x^i + \tau w^i$.

When assessing the sensitivity of the different proposed methods with respect to the CIT/CIO assumption, we observe on the Figures 5 that, under all three missing values mechanisms, if CIT/CIO is violated, then all methods give biased results whereas under CIT/CIO some become unbiased. Indeed our $\hat{\tau}_{EM}$ converges fast to the true τ, provided that either CIT or CIO are satisfied. A rather surprising result that still needs further investigations to be fully understood is the increase in bias of almost all methods when both CIT and CIO are satisfied simultaneously. It would have seemed more likely to observe a behavior at least as good as when only one assumption is met. Indeed, both assumptions being satisfied simultaneously means that for an observation i, the missing values are neither predictive of treatment nor of the outcome, except for the MNAR case where the missingness depends on the missing values, hence the pattern is informative. An improvement to the propensity and outcome estimation, hence also in the final treatment effect estimation, can be achieved by adding the missingness pattern (or *mask*) to the data: the models are fit on the stacked matrices $[\tilde{X} \ R]$, where $\tilde{X}$ is either imputed using an imputation

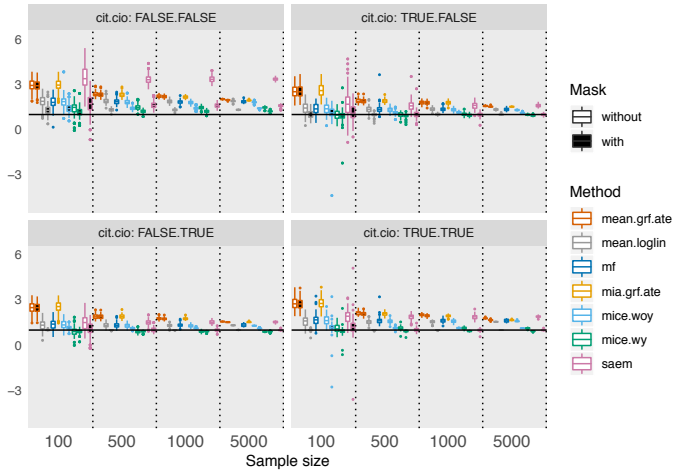
[9] The code for reproducing the experiments presented in this work is available online at `https://github.com/imkemayer/causal-inference-missing`.

model, or by mean- or zero-imputation, depending on the method. This modified model fitting does not change the results for the tree-based estimator $\hat{\tau}_{MIA}$ (not reported on the figures), which corroborates the fact that MIA for random forests allows to detect interactions between the data and the missingness pattern.

Note that the multiple imputation method behaves similarly in all settings, which is expected as it does not rely on the CIT/CIO assumption. Hence it performs better than the other methods when CIT/CIO is violated. We also observe that adding the outcome Y to the imputation model improves the estimation result; this approach has been recommended when dealing with incomplete confounders (Mattei and Mealli, 2009; Leyrat et al., 2019). The good performance of multiple imputation in the MNAR case might be due to the strong correlation structure of the covariates X.



(a) MCAR (with 50% missing values in $X_{\cdot,1:10}$)



(b) MNAR (with 50% missing values in $X_{\cdot,1:5}$)

Figure 5: Estimated average treatment effect $\hat{\tau}$ by double robust approach. Strongly correlated confounders. *mean.loglin*: mean imputation and logistic/linear regression on imputed data; *mice*: multiple imputation ($M = 10$) and standard complete case estimators on imputed data (woY: without outcome in the imputation model; wY: with outcome in the imputation model); *mia+grf*: generalized random forest propensity and outcome estimation with MIA; *saem*: EM estimation for propensity and outcome models; *mf*: low-rank matrix factorization and standard complete case estimators on estimated factors; $Mask$: logistic/linear regressions on imputed data concatenated with the mask R; (black solid line: true treatment effect τ; 100 simulations for sample sizes $n \in \{100, 500, 1000, 5000\}$).

2.3.2 Nonparametric double robustness

Next we assess the performance of the second estimator, $\hat{\tau}_{MIA}$ and its sensitivity to the CIT/CIO assumption. We consider two data generating models: a latent class model and a standard hierarchical

data-generating model.

Latent class model We consider a Gaussian mixture model, i.e. we first generate class labels C from a multinomial distribution with three categories. Then the confounders of observation i $X_{i\cdot}$ are sampled from the corresponding class distribution, i.e. $X_{i\cdot} \sim \mathcal{N}(\mu(c_i), \Sigma(c_i)) \,|\, C_i = c_i$. Missing values are generated either under MCAR (as in the previous case) or under MNAR (where we generate missing values in $X_{\cdot,1:5}$ depending only on $X_{\cdot,1:5}$). Treatment and outcome are again defined via the logistic-linear model in the following way: $logit(e(X_{i\cdot})) = (\alpha(C_i))^T X_{i\cdot}$. This allows to add an additional interaction between treatment and the latent class. Analogously, the outcome is defined as $Y_i \sim \mathcal{N}((\beta(C_i))^T X_{i\cdot} + \tau W_i, \sigma^2)$.

Figures 6 show that with one exception, all methods fail if the CIT/CIO assumption is violated, independently from the missingness mechanism. However, if CIT/CIO is satisfied (i.e. if we have $CIT \vee CIO = true$) then our estimator $\hat{\tau}_{MIA}$ converges to the true value τ while the other methods remain biased.

If the data contains missing values generated under MNAR, the multiple imputation using chained equations (mice) fails to impute the data and therefore we do not report ATE estimations based on multiple imputations in this case.

Hierarchical data-generating model An alternative to defining a Gaussian mixture model, i.e. a small finite set of classes $c \in \mathcal{C} \triangleq \{1, \ldots, N_c\}$ and corresponding parameters $(\mu(c), \Sigma(c))_{c\in\mathcal{C}}$, is to use a simplified shallow version of a *deep latent variable model* (DLVM, Kingma and Welling (2014)): instead of sampling the classes from a finite set $\mathcal{C}$, one samples codes C from a normal distribution $\mathcal{N}_d(0, 1)$. Covariates X^i are then sampled from $\mathcal{N}_p(\mu(c), \Sigma(c)) \,|\, C^i = c$, where

$$(\mu(c), \Sigma(c)) = (V \tanh(Wc + a) + b, \exp(\gamma^T(Wc + a) + \delta)I_p).$$

In our simulations we fix $d = 3$. Missing values are generated as for the latent classes model. Treatment and outcome are again defined via the logistic-linear model in the following way: $logit(e(X_{i\cdot})) = \alpha^T X_{i\cdot}$ and $Y_i \sim \mathcal{N}((\beta^T X_{i\cdot} + \tau W_i, \sigma^2)$, i.e. without an additional class interaction in the treatment or outcome model.

A first remark is that even in the MCAR case, the parametric observed-likelihood based approach, namely the estimator $\hat{\tau}_{EM}$, fails under DLVM. Indeed, while satisfying the necessary normality assumption, the observations Z^i are not i.i.d. due to their (nonlinear) dependence on the (latent) codes C^i. The multiple imputation method yields some biased estimations in the MCAR case but fails completely in the MNAR case. The forest-based estimator $\hat{\tau}_{MIA}$ is converging to the true value τ in all cases, provided the CIT/CIO assumption is met.

Further experiments under a different generative data model, a standard hierarchical data-generating model (Kingma and Welling, 2014) can be found in the Appendix.

2.3.3 Joint conclusion on the simulation study

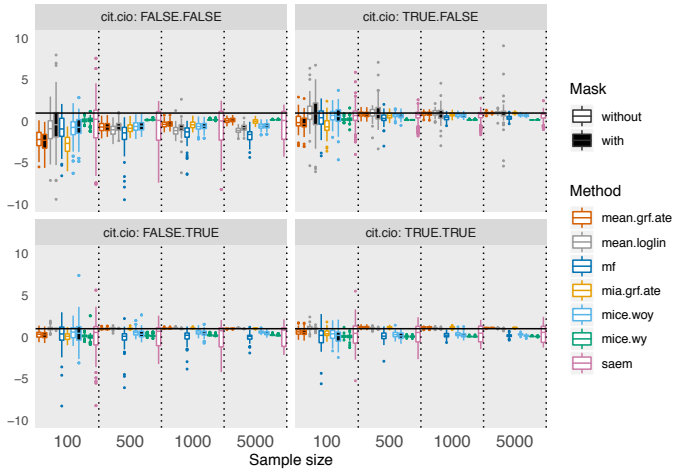
The results from this first simulation study can be summarized in several general observations:

- Multiple imputation using all available information (X^{obs}, W, Y) generally performs well, independently from the CIT/CIO assumption. However in most scenarios there is a small remaining bias, even for large sample sizes, but from the theorem from Seaman and White (2014) on multiple imputation with $M = \infty$ imputations it is expected that an increase of the number of imputations should decrease this remaining bias.

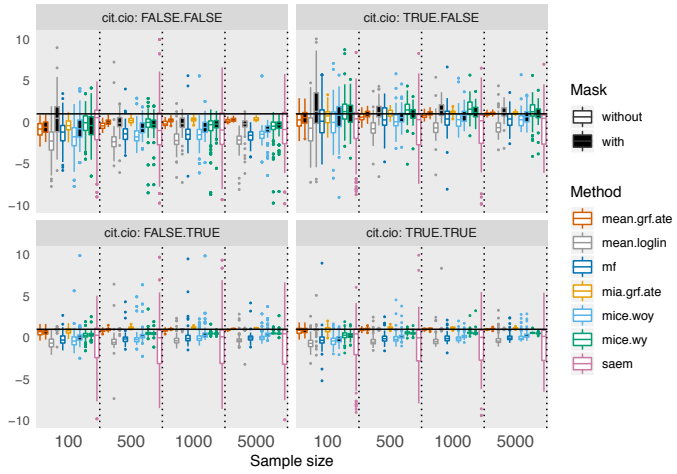
- The tree-based estimation, using the MIA criterion or mean imputation, generally performs at least as good as multiple imputation but yields unbiased results if the CIT/CIO assumption is satisfied.

- Mean imputation and concatenation of the imputed data with the response pattern, followed by logistic regression for W and linear regression for Y leads to unbiased estimates provided that CIT/CIO is satisfied. Otherwise this approach is outperformed by the competing methods.

- The EM-based estimator performs well under correct specification (multivariate Gaussian confounders, logistic treatment assignment, linear outcome, ignorable missing data mechanism, CIT/CIO satisfied) and adding the response pattern to the initial data matrix yields unbiased estimates even if the missing data mechanism is not ignorable (MNAR). It fails however in the cases where the data is not Gaussian but generated using a hierarchical structure.



(a) MCAR (with 50% missing values in $X_{\cdot,1:10}$)



(b) MNAR (with 50% missing values in $X_{\cdot,1:5}$)

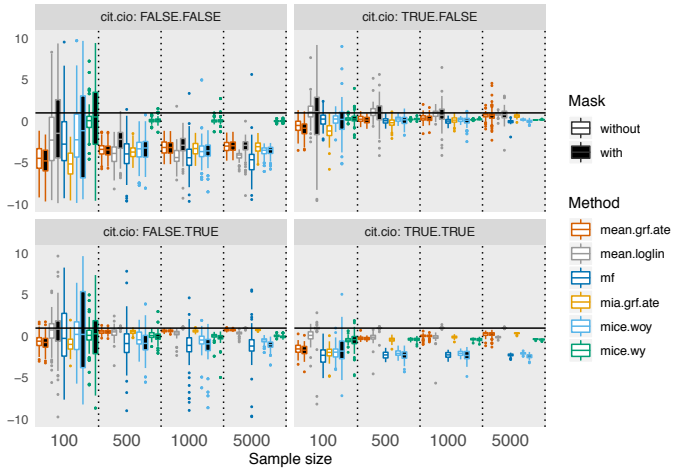
Figure 6: Estimated average treatment effect $\hat{\tau}$. Latent classes model for confounders. *mean.loglin.woR*: mean imputation and logistic/linear regression on imputed data (without taking into account the mask R); *mean.loglin.wR*: mean imputation and logistic/linear regression on imputed data, stacked with the mask R; *mice*: multiple imputation ($M = 10$) and standard complete case estimators on imputed data (*woY*: without outcome in the imputation model; *wY*: with outcome in the imputation model); *mia+grf*: generalized random forest propensity and outcome estimation with MIA; *saem*: EM estimation for propensity and outcome models; *mf*: low-rank matrix factorization and standard complete case estimators on estimated factors; (black solid line: true treatment effect τ; 100 simulations for sample sizes $n \in \{100, 500, 1000, 5000\}$).

Remark 2. *In the case where both CIT and CIO are simultaneously satisfied some doubly robust estimators perform worse than when exactly one of the two conditions is satisfied.*

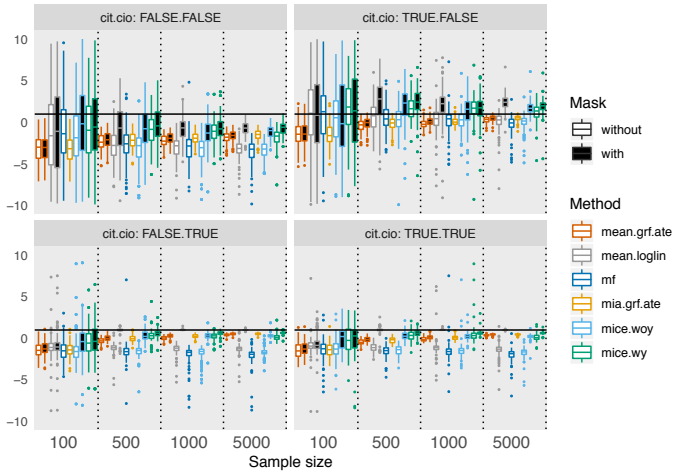
2.4 Simulation study on IHDP data

Before we turn to the open medical question presented in the introduction we first apply and compare our estimation methods on a data set from the Infant Health and Development Program (IHDP)[10]. We follow (Hill, 2011) for the emulation of an observational data set from the original experimental data[11]. The experimental data set is composed of six quantitative and 19 binary variables, recorded for 985 individuals. In order to "transform" the experimental data into observational data, (Hill, 2011) proposed to select 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

[10] `https://github.com/vdorie/npci/tree/master/examples/ihdp_sim/data`
[11] We use and adapt the corresponding code from V. Dorie: `https://github.com/vdorie/npci/`.



(a) MCAR (with 50% missing values in $X_{\cdot,1:10}$)

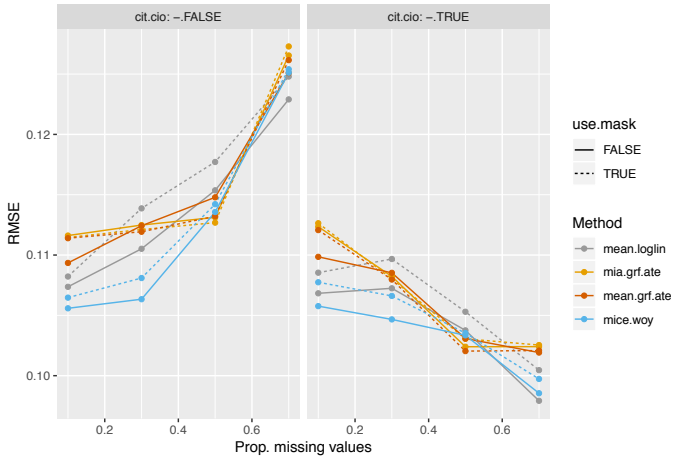


(b) MNAR (with 50% missing values in $X_{\cdot,1:5}$)

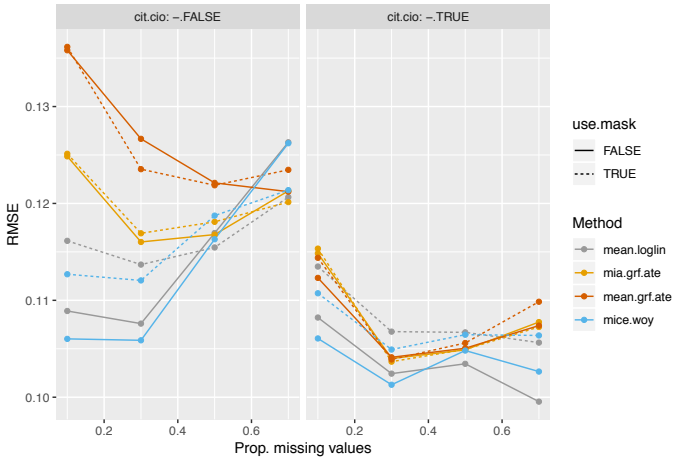
Figure 7: Estimated average treatment effect $\hat{\tau}$. **Hierarchical data-generating model** for confounders. *mean.loglin.woR*: mean imputation and logistic/linear regression on imputed data (without taking into account the mask R); *mean.loglin.wR*: mean imputation and logistic/linear regression on imputed data, stacked with the mask R; *mice*: multiple imputation ($M = 10$) and standard complete case estimators on imputed data (*woY*: without outcome in the imputation model; *wY*: with outcome in the imputation model); *mia+grf*: generalized random forest propensity and outcome estimation with MIA; *mf*: low-rank matrix factorization and standard complete case estimators on estimated factors; (black solid line: true treatment effect τ; 100 simulations for sample sizes $n \in \{100, 500, 1000, 5000\}$).

new data set. For our comparison study we only focus on the quantitative variables and simulate the response surfaces μ_w and potential outcomes $Y(w) \sim \mathcal{N}(\mu_w, 1)$, $w \in \{0, 1\}$ according to scenario A in (Hill, 2011), but using only these six quantitative variables. As in the previous simulation study we modify this simulation step to obtain two scenarios: one with the CIO assumption satisfied and another where it is violated.

When comparing the different methods w.r.t. the CIO assumption in Figure 8, it consistently appears that the methods perform better when the CIO assumption is satisfied, i.e. the potential outcomes only depend on the observed variables and the response pattern. The difference in performance increases as the amount of missing values increases. This is true even in the MCAR case, provided the proportion of missing values is sufficiently large (more than 50% of missing values). Interestingly, the tree-based methods (`mia.grf.ate` and `mean.grf.ate` on the graphs) improve as the number of missing values increases, with one exception: in the MCAR setting and with violated CIO assumption. Note that in all these simulations the multiple imputation that makes use of the outcome information Y performs pooly even for small amounts of missing values. The RMSE obtained with this method are outside the plotted



(a) MCAR



(b) MNAR

Figure 8: RMSE of estimated average treatment effect $\hat{\tau}_{DR}$ on IHDP dataset. *mean.loglin*: mean imputation and logistic/linear regression on imputed data; *mice*: multiple imputation ($M = 10$) and standard complete case estimators on imputed data (woY: without outcome in the imputation model; wY: with outcome in the imputation model); *mia+grf*: generalized random forest propensity and outcome estimation with MIA; *saem*: EM estimation for propensity and outcome models; *mf*: low-rank matrix factorization and standard complete case estimators on estimated factors; $Mask$: logistic/linear regressions on imputed data concatenated with the mask R; (200 simulations for varying proportion of missing values).

range in Figure 8 and they increase as the proportion of missing values increases. Similar behavior is observed for the IPW estimators, see Appendix B for more details.

3 Application on observational critical care management data

As announced in the introduction we apply our methods to clinical data from a French observational database on major trauma patients. The medical question we aim to answer is whether administrating the drug *tranexamic acid* (TA) has an effect on in-ICU mortality for patients with traumatic brain injury.

3.1 Data and causal DAG

Out of the 20,000 currently available patient records we consider a subset of 7,040 observations that have been validated by the medical expert team after a first pre-treatment of a subset of 7,495 observations available at the beginning of this study. The pre-treatment consisted in identifying outliers clearly due to erroneous inputs and recoding missing values that are not really missing (for instance the variable

informing previous pregnancies is evidently consistently missing, or ideally set to false, for male patients; or the measured reaction to a specific treatment is evidently missing if the treatment has not been administrated, etc.)[12]. Out of these 7,040 patients, 3,050 are identified as having a traumatic brain injury (defined by the medical expert team as either the presence of brain lesion visible on the first computed tomography (CT) scan – which is generally taken within the first three hours after the accident – or as a head AIS score[13] greater or equal 2).

The treatment of interest, tranexamic acid, is an antifibrinolyic agent limiting excessive bleeding and it is currently used in patients suspected of developing an hemorrhagic shock, a state in which the body is no longer able to provide vital organs with sufficient quantities of dioxygen to sustain them. The average cost of a dose of tranexamic acid lies below 10€ and the drug is generally available immediately after the arrival of the medical first responders team at the place of the accident. After a large clinical trial demonstrating the effectiveness of this treatment in limiting the risk of hemorrhagic shock Shakur et al. (2010) it is now recommended to be administer this drug to patients at risk of developing an hemorrhagic shock. In the pre-treated subset of patients with major trauma (counting 7,040 observations) 603 patients have an hemorrhagic shock and 1,060 patients receive the treatment, corresponding to 8.6% and 15% respectively. Among the patients with traumatic brain injury, 307 patients have an hemorrhagic shock and 543 receive the treatment, corresponding to 10% and 18% respectively. These numbers illustrate the difficulty of correctly diagnosing the risk of hemorrhagic shock in an early phase Pommerening et al. (2015) and the current practice of administrating the drug in case of doubt. This difficulty leads to a assignment heterogeneity in the data that is in our favor in the sense that it potentially allows us to study the effect of the drug on patients with traumatic brain injury but without an hemorrhagic shock. Indeed we are interested in the direct causal effect of the treatment on mortality among traumatic brain injury patients, excluding the indirect causal effect passing through the mediator variable that indicates hemorrhagic shock.

In order to clarify the causal question given the data, we first establish a causal graph in order to summarize the a priori on existing confounding and the CIT/CIO assumption and to highlight the causal question, as suggested, for instance, by Lederer et al. (2019); Blake et al. (2019). The causal graph in Figure 9 is the result of a two-step Delphi procedure (Jones and Hunter, 1995) consisting first in a selection of covariates related to either treatment or outcome or both and second in a classification of these covariates into confounders and predictors of only treatment or outcome. The Delphi procedure consists in consulting a panel of experts, in this case six anesthetists and resuscitators specialized in critical care, to identify a stable set of quantities relevant for the question at hand. The absence of an exact timestamp for the drug administration is compensated by the fact that it is always given within the first three hours from the accident and that the treatment does not have an immediate effect on variables such as blood pressure, hemoglobin level or the Glasgow Coma Scale (GCS) which are measured at various moments within the first three hours.

From this graph it becomes clear as well that a method that incorporates a model of the outcome as a function of the identified potential predictors (red and blue nodes on the graph) might achieve more precise results than a method that uses the observed outcome directly. The large number of predictors of the outcome is due both to the medical complexity of traumatic brain injury and to the ambiguous treatment target: the assignment is made in the context of hemorrhagic shock but since recently there is some evidence that there might also be a beneficial effect in the context of traumatic brain injury (Hijazi et al., 2015).

3.2 Results

When adjusting for confounding using the identified confounders (pink nodes on the graph in Figure 9), using additional predictors for the outcome model (blue nodes on the graph in Figure 9), we obtain the following estimations of the direct causal effect of tranexamic acid on in-ICU mortality among traumatic brain injury patients:

Before analyzing the results in Figure 10 we briefly describe the estimation approaches chosen.

- **Imputation**: we impute the data with two popular imputation methods: (a) using a factorial analysis of mixed data approach (Escofier, 1979) that is implemented in the `missMDA` R-package

[12]The code for pre-treatment and for estimating the treatment effect on this data are available at `https://github.com/imkemayer/causal-inference-missing`.

[13]The head Abbreviated Injury Score indicates, on a scale from one to six, the severity of the most severe observed brain lesion. This score is defined in the context of the Abbreviated Injury Scale proposed by the American Association for Automotive Medicine. See Appendix C.1 or `https://www.aaam.org/abbreviated-injury-scale-ais/` for more information.

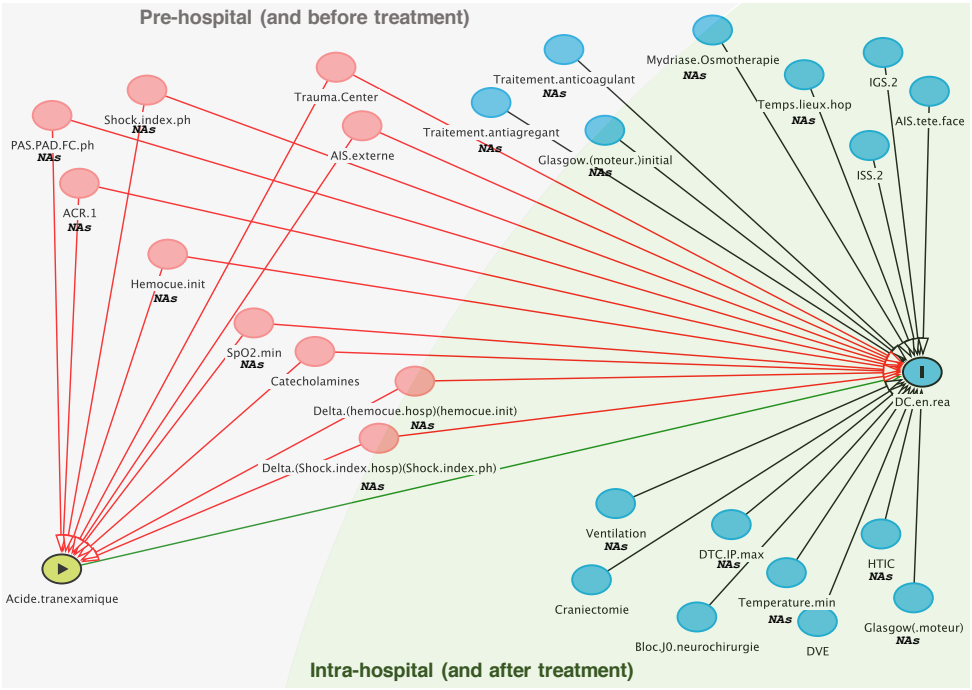


Figure 9: Causal graph representing treatment, outcome, confounders and other predictors of outcome (Figure generated using DAGitty (Textor et al., 2011); `NAs` indicates variables that still have missing values after pre-treatment).

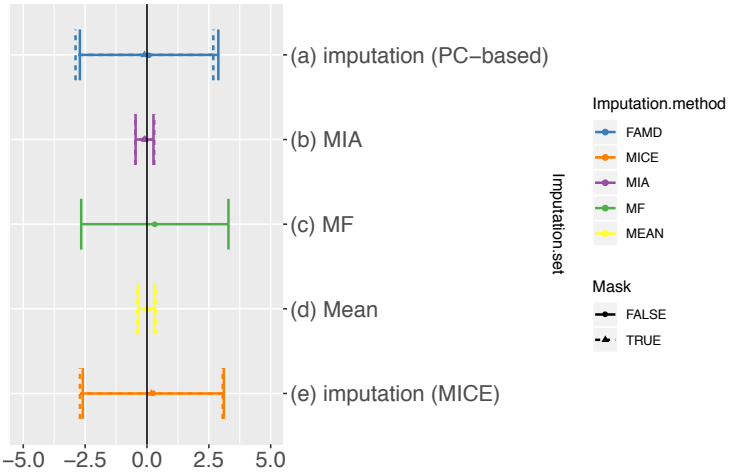


Figure 10: Double robust ATE estimations on Traumabase data (x-axis: $100 \times (\hat{\tau} \pm 1.96\hat{\sigma})$).

(Josse et al., 2016). It is the equivalent of principal component analysis for mixed data – we have both quantitative and categorical variables in our data set – and therefore relies on a similar low-rank assumption; (b) using conditional models and the corresponding imputation by chained equations algorithm implemented in the `mice` R-package (van Buuren, 2018). After (single) imputation (using the outcome in the imputation model), the propensity and the outcome model are fitted using random forests, more precisely the ATE is estimated using the `average_treatment_effect` function of the `grf` R-package.

- **MIA**: this approach corresponds to the $\hat{\tau}_{MIA,DR}$ estimator and is implemented in the same way as in the simulation study.
- **Mean**: this approach corresponds to a variant of the $\hat{\tau}_{MIA,DR}$, using mean imputation instead of the MIA encoding detailed earlier.

- **MF**: this approach corresponds to the low-rank matrix factorization pre-processing method proposed by (Kallus et al., 2018). Once the factors – the estimated confounders – are obtained, the propensity and the outcome model are fitted using random forests, more precisely the ATE is estimated using the `average_treatment_effect` function of the `grf` R-package.

Unlike the simulations of the previous paragraph, the real-world medical data is more complicated and some concessions have to be made to apply the previously discussed method. We quickly mention some of the choices made for this analysis:

- A current limitation of the SAEM for logistic regression is that it has only been derived for continuous (and normally distributed) variables. Hence we do not report results for the EM-based ATE estimator for this medical application.
- Due to an important number of outliers in the variable *Temps.lieux.hop* that are related with inconsistent units of the recorded values and with patient transfers from one hospital to another, we chose to drop this variable in our analyses since, according to the practitioners, its predictive power does not outweigh the potential issues related to inconsistent recording of this variable.
- The imputation based on low-rank assumptions currently does allow for specification of a range for some variables (for instance we know that by definition the Glasgow coma scale is between 3 and 15. See Appendix C.1 for a complete listing.). Hence, after imputation some variables have values outside of the pre-defined range but we choose to use these imputed values and do not truncate them manually.

Note that apart from the issue with the variable *Temps.lieux.hop*, the estimation via random forest and MIA (or mean imputation) does not require substantial pre-processing of the data and is therefore straightforward, once the MIA recoding and the random forest are implemented. A remaining issue might consist in the overlap assumption which is generally difficult to assess in most medical applications and which might be slightly violated due in part to the heterogeneity of patient profiles. A solution to handle such weak overlap is the use of overlap weights (Li et al., 2018) and we give the results using this alternative to inverse propensity weights in Appendix C.2.

A first observation on the results reported in Figure 10 is the concordance of the different estimators: none of the estimation strategies allows to reject the null hypothesis of no treatment effect. However the estimated variances differ between the imputation based methods (when counting the matrix factorization method as being close to an imputation method) and the MIA-tree based estimators. In order to analyze and comment on these differences, further investigations about the variance estimation with incomplete confounders are required. Finally, adding the response pattern (or mask) to the data matrix does not seem to lead to major changes in the estimations.

4 Discussion and perspectives

Double robustness under modified unconfoundedness assumptions The main contribution of this work is the first consistency result for doubly robust average treatment effect estimation with incomplete confounders. Our empirical study corroborates the necessity of the CIT/CIO assumptions and the validity of our extension to doubly robust treatment effect estimation with incomplete confounders under these assumptions. More specifically, if one can exclude smooth regression functions for the treatment assignment and the outcome model, such as logistic and linear models, then our tree-based estimator $\hat{\tau}_{MIA,DR}$ and its mean-imputation variant can be suited candidates for average treatment effect estimation with incomplete confounders.

It remains to study consistency rates and variances for the doubly robust estimators; indeed in the medical application illustrated above the variance is likely to be underestimated due to the uncertainty in the missing values that has not been accounted for in this work. This future study on variance estimates accounting for variability due to missing values will potentially be based on the set of theoretical results on random forests given by (Wager and Athey, 2018).

Furthermore, simulations studies on mixed confounders are necessary to corroborate the effectiveness of the proposed estimators in scenarios with this additional – yet in practice often encountered – aspect.

Heterogeneous treatment effects and policy learning Instead of estimating the average treatment effect τ, one could be interested in the conditional average treatment effect function $\tau(x)$ for several

reasons. For instance one might be interested in estimating how treatment effects vary across subpopulations, or assessing whether there is heterogeneity in the population w.r.t. a given treatment. Such questions anticipate problems of learning decision rules that exploit treatment effect heterogeneity (Wager and Athey, 2018) or of identifying subgroups (Athey and Imbens, 2016), for instance the problem of finding the best policy to assign individuals to a treatment. Note that thanks to unconfoundedness we can derive a consistent estimator for $\tau(x)$ as long as we know the propensity scores $e(x)$ for all $x \in \mathcal{X}$:

$$\mathbb{E}\left[\frac{W_i Y_i}{e(x)} - \frac{(1-W_i)Y_i}{1-e(x)} | X_i = x\right] = \tau(x). \tag{33}$$

At a high-level, heterogeneous treatment effect estimation is achieved by averaging nearby observations where the notion and estimation of neighborhood differs from one method to the other. For instance, Athey and Imbens (2016) define causal trees that allow to construct valid confidence intervals for average treatment effects for every leaf of the tree (containing a subgroup of the original data), as the leaves are built such that every leaf can be thought of as an RCT. The advantage of trees is that they are able to identify the dimensions in the feature space that actually matter for the neighborhood estimation, avoiding the curse of dimensionality of classical nearest neighbors methods. Another line of work offers causal inference via lasso-like methods in sparse high-dimensional linear settings (see for instance Imai and Ratkovic (2013)).

In light of our medical application, heterogeneous treatment effect estimation is of particular interest because of the existing heterogeneity among traumatic brain injury patients in terms of clinical presentation, pathophysiology and outcome. It is even more relevant since to this date there does not exist any general classification of patients with traumatic brain injury. Hence a causal inference approach allowing classification w.r.t. treatment heterogeneity for any given treatment is of interest for practitioners in critical care management.

Acknowledgement We thank Jean-Pierre Nadal for fruitful discussion, Julie Tibshirani for the suggestion to implement quickly MIA, and Tobias Gauss and Jean-Denis Moyer for the medical insight into traumatic brain injury and hemorrhagic shock. We acknowledge funding from the EHESS PhD fellowship.

Appendix

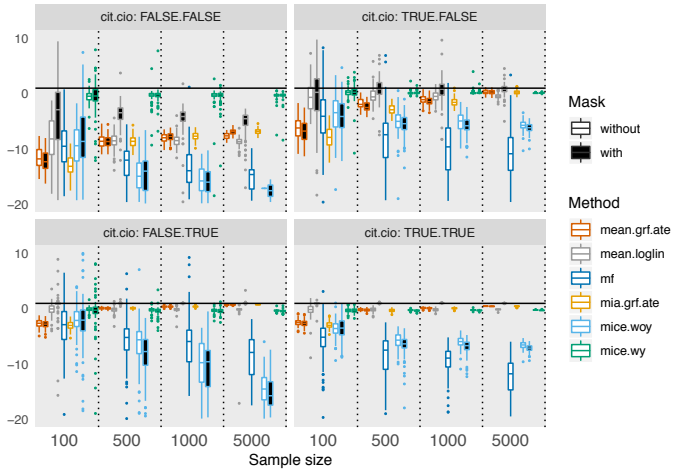
A Simulation study on synthetic data

A.1 Simulation results for non-parametric estimator under an hierarchical data-generating model

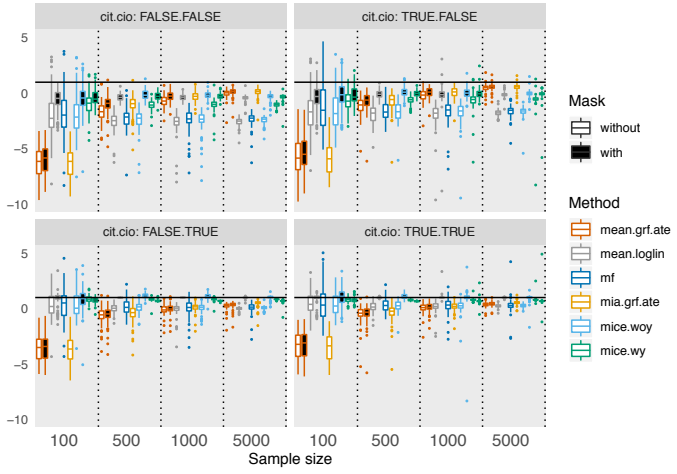
The hierarchical data-generating model used in Section 2.3.2 can be modified in order to allow for correlation between covariates by defining the code-depending Gaussian parameters as

$$(\mu(c), \Sigma(c)) = (U(V \tanh(Wc + a) + b), U \exp(\gamma^T (Wc + a) + \delta) I_p U^T),$$

for some randomly generated orthonormal matrix U.



(a) MCAR (with 50% missing values in $X_{\cdot,1:10}$)



(b) MNAR (with 50% missing values in $X_{\cdot,1:5}$)

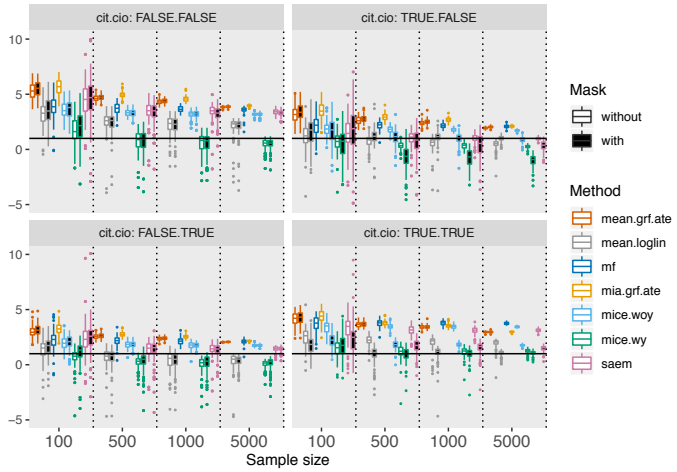
Figure 11: Estimated average treatment effect $\hat{\tau}$. **Hierarchical data-generating model with dense covariance matrices** for confounders. *mean.loglin.woR*: mean imputation and logistic/linear regression on imputed data (without taking into account the mask R); *mean.loglin.wR*: mean imputation and logistic/linear regression on imputed data, stacked with the mask R; *mice*: multiple imputation ($M = 10$) and standard complete case estimators on imputed data (woY: without outcome in the imputation model; wY: with outcome in the imputation model); *mia+grf*: generalized random forest propensity and outcome estimation with MIA; *mf*: low-rank matrix factorization and standard complete case estimators on estimated factors; (black solid line: true treatment effect τ; 100 simulations for sample sizes $n \in \{100, 500, 1000, 5000\}$).

A.2 Simulation results with IPW estimators

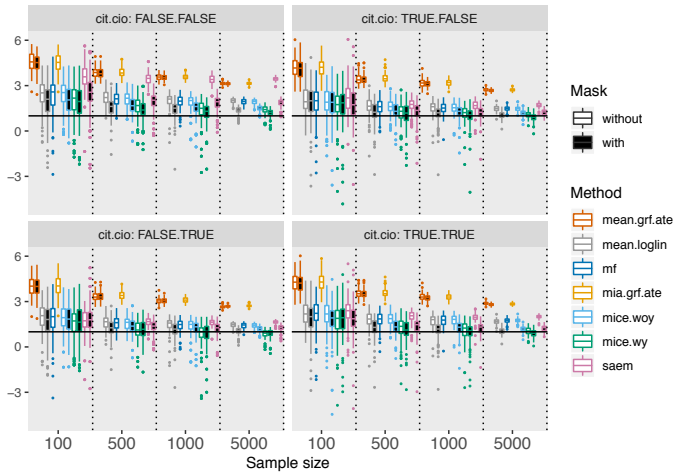
The data generating models and different scenarios for the missingness mechanisms are kept the same, but the estimations of the ATE (boxplots summarizing the $N = 100$ experiments) that are reported in Figures 12, 13, 14 and 15 are obtained by inverse-propensity weighting (IPW), i.e. the outcome Y is used as it and is not modeled by some function of the covariates X.

A.2.1 Parametric setting

The behavior of the different compared methods is similar to their doubly robust version (see Section 2.3.1), namely the sensitivity to the CIT/CIO assumptions and w.r.t. the missingness mechanism. The convergence of the non-parametric estimator $\hat{\tau}_{MIA,IPW}$ is slow but this is not surprising since random forests are generally not suited for fitting linear functions and in the parametric setting we consider only linear models.



(a) MCAR (with 50% missing values in $Z_{\cdot,1:10}$)

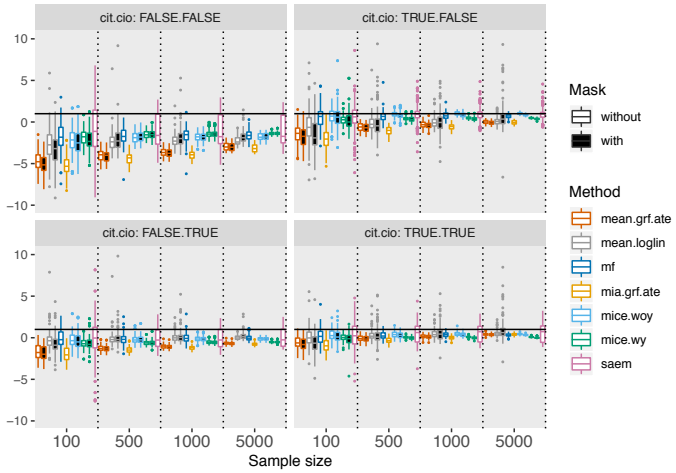


(b) MNAR (with 50% missing values in $Z_{\cdot,1:5}$)

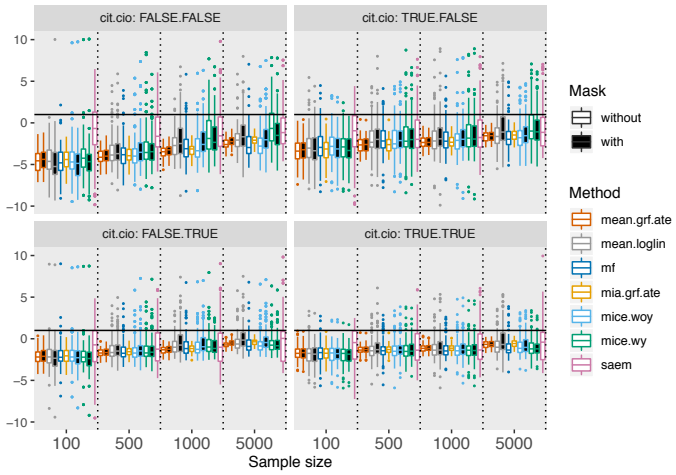
Figure 12: Estimated average treatment effect $\hat{\tau}$ by **inverse propensity weighting**. Strongly correlated confounders. *mean.loglin*: mean imputation and logistic/linear regression on imputed data; *mice*: multiple imputation ($M = 10$) and standard complete case estimators on imputed data (woY: without outcome in the imputation model; wY: with outcome in the imputation model); *mia+grf*: generalized random forest propensity and outcome estimation with MIA; *saem*: EM estimation for propensity and outcome models; *mf*: low-rank matrix factorization and standard complete case estimators on estimated factors; $Mask$: logistic/linear regressions on imputed data concatenated with the mask R; (black solid line: true treatment effect τ; 100 simulations for sample sizes $n \in \{100, 500, 1000, 5000\}$).

A.2.2 Nonparametric setting

The IPW estimators exhibit a similar behavior w.r.t. the CIT/CIO assumptions and the missingness mechanisms but – in general – perform less well than their corresponding doubly robust counterpart on this synthetic data.



(a) MCAR (with 50% missing values in $X_{\cdot,1:10}$)

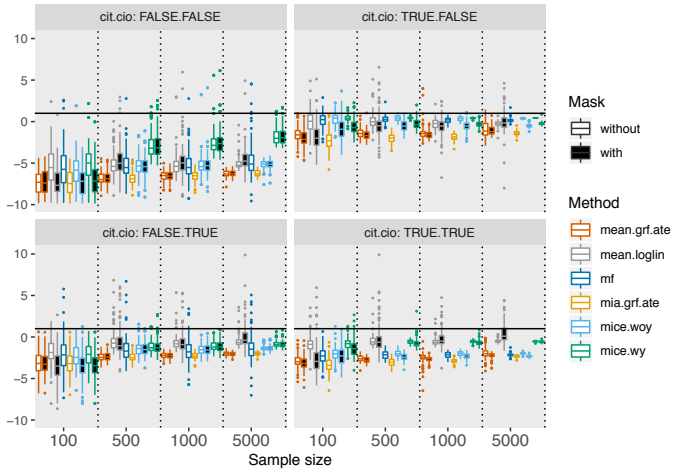


(b) MNAR (with 50% missing values in $X_{\cdot,1:5}$)

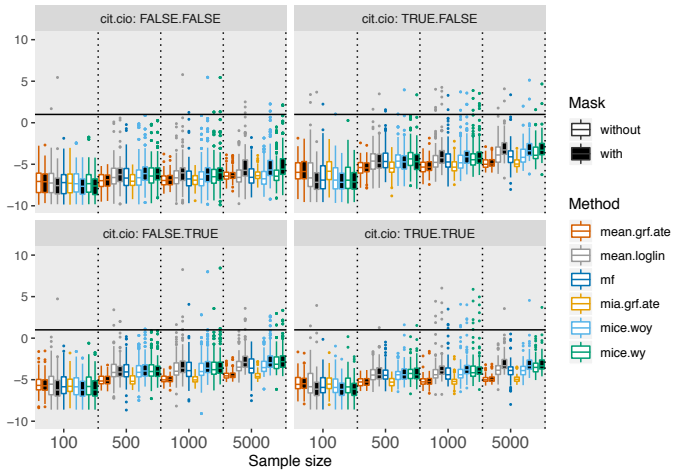
Figure 13: Estimated average treatment effect $\hat{\tau}$ by **inverse propensity weighting**. **Latent classes model** for confounders. *mean.loglin*: mean imputation and logistic/linear regression on imputed data; *mice*: multiple imputation ($M = 10$) and standard complete case estimators on imputed data (woY: without outcome in the imputation model; wY: with outcome in the imputation model); *mia+grf*: generalized random forest propensity and outcome estimation with MIA; *saem*: EM estimation for propensity and outcome models; *mf*: low-rank matrix factorization and standard complete case estimators on estimated factors; $Mask$: logistic/linear regressions on imputed data concatenated with the mask R; (black solid line: true treatment effect τ; 100 simulations for sample sizes $n \in \{100, 500, 1000, 5000\}$).

B Simulation study on IHDP data

The conclusions drawn for the doubly robust estimators can also be applied to the inverse propensity weighted estimators as we see in Figure 16.



(a) MCAR (with 50% missing values in $X_{\cdot,1:10}$)



(b) MNAR (with 50% missing values in $X_{\cdot,1:5}$)

Figure 14: Estimated average treatment effect $\hat{\tau}$ by **inverse propensity weighting**. **Hierarchical data-generating model** for confounders. *mean.loglin*: mean imputation and logistic/linear regression on imputed data; *mice*: multiple imputation ($M = 10$) and standard complete case estimators on imputed data (woY: without outcome in the imputation model; wY: with outcome in the imputation model); *mia+grf*: generalized random forest propensity and outcome estimation with MIA; *saem*: EM estimation for propensity and outcome models; *mf*: low-rank matrix factorization and standard complete case estimators on estimated factors; $Mask$: logistic/linear regressions on imputed data concatenated with the mask R; (black solid line: true treatment effect τ; 100 simulations for sample sizes $n \in \{100, 500, 1000, 5000\}$).

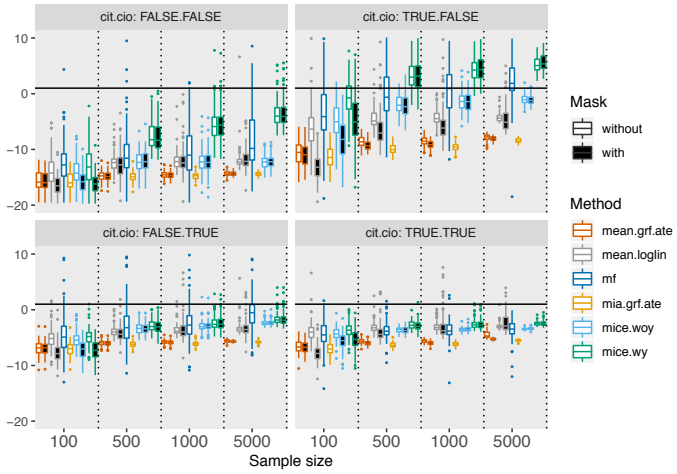
C Details on the medical application (Traumabase)

C.1 Definition of the variables of the Traumabase used in the analysis

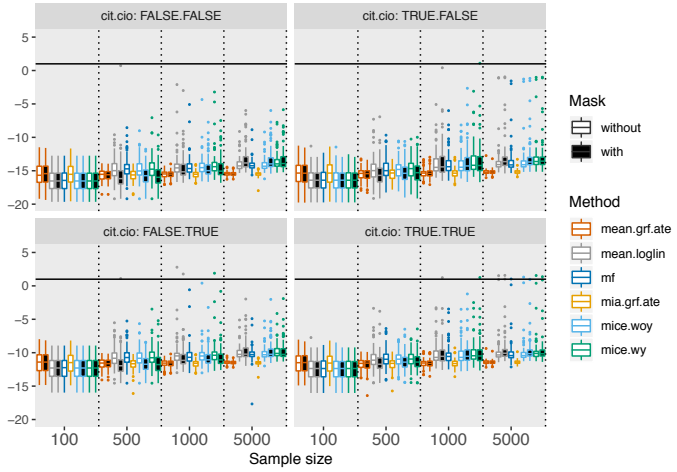
Here we give the names and short descriptions of the variables we use in our causal analysis. The moment at which the variable is first available is given in parentheses (ph = pre-hospital phase, h = hospital phase).

List of confounders:

- *Trauma.Center* (categorical): name of the trauma center. (ph/h)
- *PAS.ph*, *PAD.ph*, *FC.ph* (continuous): systolic and diastolic arterial pressure and heart rate during pre-hospital phase. (ph)



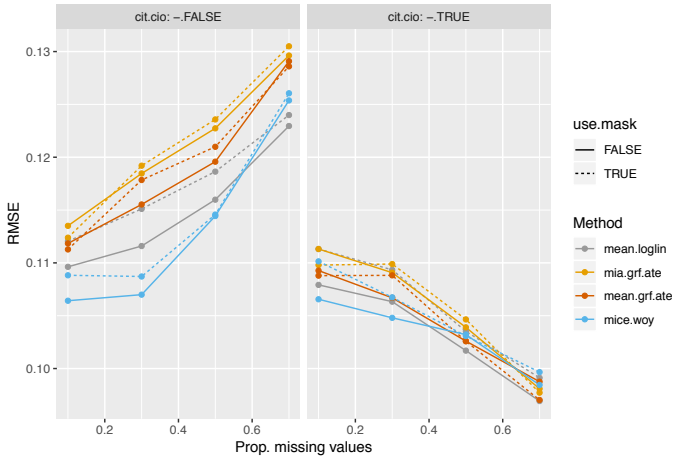
(a) MCAR (with 50% missing values in $X_{\cdot,1:5}$)



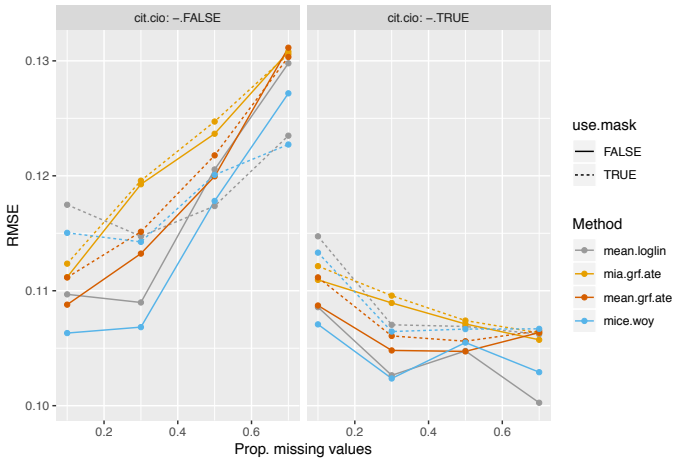
(b) MNAR (with 50% missing values in $X_{\cdot,1:5}$)

Figure 15: Estimated average treatment effect $\hat{\tau}$ by **inverse propensity weighting**. **Hierarchical data-generating modelwith dense covariance matrices** for confounders. *mean.loglin*: mean imputation and logistic/linear regression on imputed data; *mice*: multiple imputation ($M = 10$) and standard complete case estimators on imputed data (woY: without outcome in the imputation model; wY: with outcome in the imputation model); *mia+grf*: generalized random forest propensity and outcome estimation with MIA; *saem*: EM estimation for propensity and outcome models; *mf*: low-rank matrix factorization and standard complete case estimators on estimated factors; $Mask$: logistic/linear regressions on imputed data concatenated with the mask R; (black solid line: true treatment effect τ; 100 simulations for sample sizes $n \in \{100, 500, 1000, 5000\}$)..

- *ACR.1* (categorical): cardiac arrest during pre-hospital phase. (ph)
- *Hemocue.init* (continuous): prehospital capillary hemoglobin concentration (the lower, the more the patient is probably bleeding and in shock); hemoglobin is an oxygen carrier molecule in the blood. (ph)
- *SpO2.min* (continuous): peripheral oxygen saturation, measured by pulse oxymetry, to estimate oxygen content in the blood ($95 - 100\%$: considered normal; $< 90\%$ critical and associated with considerable trauma, danger and mortality). (ph)
- *Catecholamines* (continuous): treatment in case of physical or emotional stress increasing heart rate, blood pressure, breathing rate, muscle strength and mental alertness. (ph)
- *Shock.index.ph* (continuous): ratio of heart rate and systolic arterial pressure during pre-hospital phase. (ph)



(a) MCAR



(b) MNAR

Figure 16: RMSE of estimated average treatment effect $\hat{\tau}_{IPW}$ on IHDP dataset. *mean.loglin*: mean imputation and logistic/linear regression on imputed data; *mice*: multiple imputation ($M = 10$) and standard complete case estimators on imputed data (woY: without outcome in the imputation model; wY: with outcome in the imputation model); *mia+grf*: generalized random forest propensity and outcome estimation with MIA; *saem*: EM estimation for propensity and outcome models; *mf*: low-rank matrix factorization and standard complete case estimators on estimated factors; $Mask$: logistic/linear regressions on imputed data concatenated with the mask R; (200 simulations for varying proportion of missing values).

- *AIS.externe* (discrete, range: $[0, 6]$): Abbreviated Injury Score for external injuries, here it is assumed to be a proxy of information available/visible during pre-hospital phase. (ph/h)
- *Delta.shock.index* (continuous): Difference of shock index between arrival at the hospital and arrival on the scene. (h)
- *Delta.hemocue* (continuous): Difference of hemoglobin level between arrival at the hospital and arrival on the scene. (h)

List of predictors of mortality and that are not associated with treatment assignment

- *Traitement.anticoagulant* (categorical): oral anticoagulant therapy. (ph)
- *Traitement.antiaggregant* (categorical): anti-platelet therapy. (ph)
- *Glasgow(.initial)* (discrete, range: $[3, 15]$): Initial Glasgow Coma Scale (GCS) on arrival on scene of enhanced care team and on arrival at the hospital ($GCS = 3$: deep coma; $GCS = 15$: conscious and alert). (ph & h)

- *Glasgow.moteur(.initial)* (discrete, range: $[1, 6]$): Initial Glasgow Coma Scale motor score ($GCS.motor = 1$: no response; $GCS.motor = 6$: obeys command/purposeful movement). (ph & h)
- *Mydriase* (categorical): pupil dilation indicating brain herniation. (ph & h)
- *Osmotherapie, Regression.mydriase.sous.osmotherapie* (categorical): administration of osmotherapy to alleviate compression of the brain (either Mannitol or hypertonic saline solution); change of pupil anomaly after administration of osmotherapy. (ph & h)
- *Temps.lieux.hop* (continuous): total duration of prehospital care team engaged (arrival on scene to arrival at hospital). (h)
- *Ventilation* (discrete, range: $[0, 5]$): inspired concentration of oxygen on ventilatory support (the higher the more critical; $Ventilation = 0$: no ventilatory support). (h)
- *Temperature.min* (continuous): Minimal body temperature. (h)
- *DTC.IP.max* (continuous): pulsatility index (PI) measured by echodoppler sonographic examen of blood velocity in cerebral arteries ($PI > 1.2$: indicates altered blood flow maybe due to traumatic brain injury). (h)
- *HTIC* (categorical): at least one episode of increased intracranial pressure; mainly in traumatic brain injury; usually associated with worse prognosis. (h)
- *DVE* (categorical): external ventricular drainage (EVD); mean to drain cerebrospinal fluid to reduce intracranial pressure. (h)
- *Craniectomie* (categorical): decompressive craniectomy, surgical intervention to reduce intracranial hypertension. (h)
- *Bloc.J0.neurochirurgie* (categorical): neurosurgical intervention performed on day of admission. (h)
- *AIS.tete, AIS.face* (discrete, range: $[0, 6]$): Abbreviated Injury Score, describing and quantifying facial and head injuries ($AIS = 0$: no injury; the higher the more critical).(h)
- *ISS* (discrete, range: $[0, 108]$): Injury Severity Score, sum of squares of top three AIS scores. (h)
- *IGS.II* (continuous): Simplified Acute Physiology Score. (h)

C.2 ATE estimation on the Traumabase using overlap weights

An often raised concern with many medical observational data sets is the potential violation of the overlap assumption. For instance some patients might never get the treatment due to infrastructural circumstances or due to recommendations followed strictly by the entire medical staff. The overlap assumption however is needed for consistency of the treatment effect estimations and states that every patient has a non-zero probability of being in either treated or control group. Another way of describing this assumption is that the treatment groups are sufficiently comparable, otherwise the attempt of drawing causal inferences is doomed to failure from the beginning.

Given the important level of heterogeneity among trauma patients, especially among patients with traumatic brain injury, and the multi-level and multi-actor nature of the data, it cannot be ruled out that the treatment groups have only small overlap. When considering standardized mean differences of the confounding variables between treatment and control groups in Figure 17 it appears indeed that certain features such as the hemoglobine level differ considerably between the two groups. As detailed in Section 1.6, a possible solution to deal with this potential situation is the use of overlap weights instead of the inverse propensity weights (Li et al., 2018). However, in our case, when using the corresponding modified estimands and estimators, i.e. the average treatment effect on the overlap population, the results reported in Figure 18 are very similar to those from the normal average treatment effect estimation on the entire population (Figure 10) and lead to the same conclusion about the treatment effect.

Figure 17: Unadjusted standardized mean differences of the confounding variables.

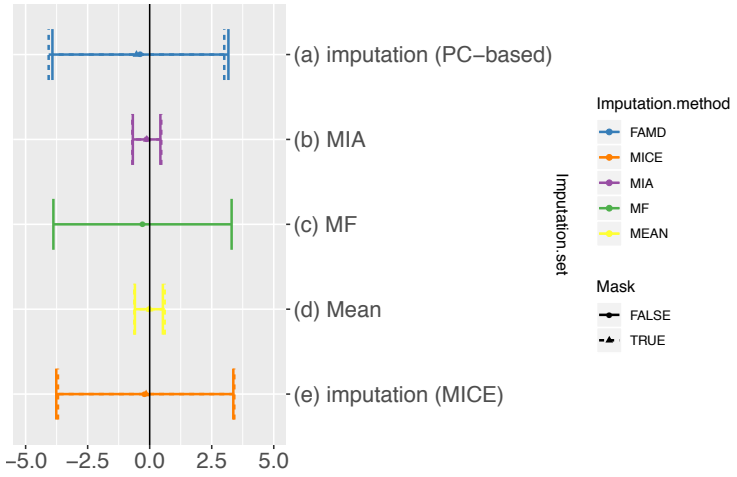


Figure 18: Double robust ATE estimations on overlap population on Traumabase data (x-axis: $100 \times (\hat{\tau} \pm 1.96\hat{\sigma})$).

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455. 6

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360. 26

Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178. 9, 17

Bartlett, J. W., Harel, O., and Carpenter, J. R. (2015). Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *American journal of epidemiology*, 182(8):730–736. 2

Blake, H. A., Leyrat, C., Mansfield, K., Seaman, S., Tomlinson, L., Carpenter, J., and Williamson, E. (2019). Propensity scores using missingness pattern information: a practical guide. *arXiv preprint*. 13, 23

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees. 11, 16

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J.

(2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68. 9, 16

Cox, D. R. (1958). *Planning of experiments*. Wiley & Sons, New York. 4

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199. 7, 8

D'Agostino Jr, R., Lang, W., Walkup, M., Morgan, T., and Karter, A. (2001). 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. 12

D'Agostino Jr, R. B. and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, 95(451):749–759. 13

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22. 10

Escofier, B. (1979). Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. *Cahiers de l'Analyse des Données*, 4(2):137–146. 23

Gondara, L. and Wang, K. (2018). Mida: Multiple imputation using denoising autoencoders. In Phung, D., Tseng, V., Webb, G., Ho, B., Ganji, M., and Rashidi, L., editors, *Proceedings of the 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2018)*, Lecture Notes in Computer Science, pages 260–272. Springer International Publishing. 11

Hájek, J. (1971). Comment on "an essay on the logical foundations of survey sampling, part one" by d. basu. *Foundations of Statistical Inference*, page 236. 7

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015). Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402. 11

Hay, S. I., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abdulkader, R. S., Abdulle, A. M., Abebo, T. A., Abera, S. F., et al. (2017). Global, regional, and national disability-adjusted life-years (dalys) for 333 diseases and injuries and healthy life expectancy (hale) for 195 countries and territories, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet*, 390(10100):1260–1344. 1

Hernán, M. A. and Robins, J. M. (2019). *Causal Inference*. Chapman & Hall/CRC. 2

Hijazi, N., Fanne, R. A., Abramovitch, R., Yarovoi, S., Higazi, M., Abdeen, S., Basheer, M., Maraga, E., Cines, D. B., and Higazi, A. A.-R. (2015). Endogenous plasminogen activators mediate progressive intracerebral hemorrhage after traumatic brain injury in mice. *Blood*, 125(16):2558–2567. 23

Hill, J. (2004). Reducing bias in treatment effect estimation in observational studies suffering from missing data. Technical report, Institute for Social and Economic Research and Policy, Columbia University. 14

Hill, J., Weiss, C., and Zhai, F. (2011). Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research*, 46(3):477–513. 7

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240. 20, 21

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189. 7

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15(3):199–236. 5

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960. 2

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685. 6

Iacus, S. M., King, G., and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24. 5

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470. 8, 26

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710. 4, 13

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29. 3, 5

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press. 3, 4, 6

Jiang, W., Josse, J., and Lavielle, M. (2018). Logistic regression with missing covariates–parameter estimation, model selection and prediction. *arXiv preprint*. 14, 15

Jones, J. and Hunter, D. (1995). Consensus methods for medical and health services research. *BMJ: British Medical Journal*, 311(7001):376. 2, 23

Josse, J., Husson, F., et al. (2016). missmda: a package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31. 24

Josse, J., Pagès, J., and Husson, F. (2011). Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification*, 5(3):231–246. 11

Josse, J., Prost, N., Scornet, E., and Varoquaux, G. (2019). On the consistency of supervised learning with missing values. *arXiv preprint*. 11, 15, 16

Kallus, N., Mao, X., and Udell, M. (2018). Causal inference with noisy and missing covariates via matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6921–6932. 14, 17, 25

Kang, J. D., Schafer, J. L., et al. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539. 7, 15

Kingma, D. P. and Welling, M. (2014). Stochastic gradient vb and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*. 19

Lederer, D. J., Bell, S. C., Branson, R. D., Chalmers, J. D., Marshall, R., Maslove, D. M., Ost, D. E., Punjabi, N. M., Schatz, M., Smyth, A. R., et al. (2019). Control of confounding and reporting of results in causal inference studies. guidance for authors from editors of respiratory, sleep, and critical care journals. *Annals of the American Thoracic Society*, 16(1):22–28. 23

Leyrat, C., Seaman, S. R., White, I. R., Douglas, I., Smeeth, L., Kim, J., Resche-Rigon, M., Carpenter, J. R., and Williamson, E. J. (2019). Propensity score analysis with partially observed covariates: How should multiple imputation be used? *Statistical methods in medical research*, 28(1):3–19. 18

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400. 7, 25, 33

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley. 6, 9, 10, 11

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233. 11

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960. 5, 6

Mattei, A. and Mealli, F. (2009). 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. 13, 18

Meng, X. L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association*, 86(416):899–909. 11

Miettinen, O. S. (1985). *Theoretical epidemiology: principles of occurrence research*. John Wiley & Sons, New York. 13

Olkin, I., Tate, R. F., et al. (1961). Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics*, 32(2):448–465. 13

Pommerening, M. J., Goodman, M. D., Holcomb, J. B., Wade, C. E., Fox, E. E., del Junco, D. J., Brasel, K. J., Bulger, E. M., Cohen, M. J., Alarcon, L. H., et al. (2015). Clinical gestalt and the prediction of massive transfusion after trauma. *Injury*, 46(5):807–813. 2, 23

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 17

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866. 8

Robins, J. M. and Wang, N. (2000). Inference for imputation estimators. *Biometrika*, 87(1):113–124. 7

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394. 6

Rosenbaum, P. R. (2010). *Design of observational studies*, volume 10. Springer. 4

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55. 4

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524. 6, 12, 13, 14

Rosenbaum, P. R. and Rubin, D. B. (1985). The bias due to incomplete matching. *Biometrics*, 41(1):103–116. 6

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701. 3

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592. 9, 10

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58. 4

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a):318–328. 6

Rubin, D. B. (1987). *Multlipe Imputation for Nonresponse in Surveys*. Wiley, Hoboken, NJ, USA. 11

Seaman, S. and White, I. (2014). Inverse probability weighting with missing predictors of treatment assignment or missingness. *Communications in Statistics-Theory and Methods*, 43(16):3499–3515. 13, 14, 19

Shakur, H., Roberts, I., Bautista, R., Caballero, J., Coats, T., Dewan, Y., El-Sayed, H., Gogichaishvili, T., Gupta, S., Herrera, J., et al. (2010). Crash-2 trial collaborators. effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant haemorrhage (crash-2): a randomised, placebo-controlled trial. *Lancet*, 376(9734):23–32. 2, 23

Stekhoven, D. J. and Bühlmann, P. (2012). Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118. 11

Textor, J., Hardt, J., and Knüppel, S. (2011). Dagitty: a graphical tool for analyzing causal diagrams. *Epidemiology*, 22(5):745. 24

Twala, B., Jones, M., and Hand, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, 29(7):950–956. 11, 15, 16

van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC, Boca Raton, FL. 11, 17, 24

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242. 25, 26

Yang, S., Wang, L., and Ding, P. (2017). Causal inference with confounders missing not at random. *arXiv preprint arXiv:1702.03951*. 14

Yoon, J., Jordon, J., and van der Schaar, M. (2018). GAIN: Missing data imputation using generative adversarial nets. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5689–5698, Stockholmsmässan, Stockholm Sweden. PMLR. 11

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922. 8