Estimation double robuste d'effet du traitement avec facteurs confondants incomplets

Imke Mayer [1] & Julie Josse [2] & Jean-Pierre Nadal [1] & Stefan Wager [3] & Tobias Gauss [4] & Jean-Denis Moyer [4]

[1] *CAMS, EHESS, 54 boulevard Raspail, 75014 Paris, imke.mayer@ehess.fr, jpnadal@ehess.fr*
[2] *CMAP, X-Paris-Saclay, 91128 Palaiseau Cedex, julie.josse@polytechnique.edu*
[3] *Stanford Graduate School of Business, 655 Knight Way, Stanford, CA 94305, swager@stanford.edu*
[4] *Traumabase® Group, Hôpital Beaujon, APHP, France, tgauss@protonmail.com, jean-denis.moyer@aphp.fr*

Résumé. Dans la recherche en santé et en sciences sociales, les études observationnelles prospectives sont fréquentes, relativement faciles à mettre en place (contrairement aux études expérimentales d'essais randomisés qui sont parfois même impossible à réaliser) et peuvent permettre différents types d'analyses postérieures telles que des inférences causales. L'estimation de l'effet moyen du traitement (en anglais *average treatment effect*, ATE), par exemple, est possible grâce à l'utilisation de scores de propension qui permettent de corriger les biais d'affectation du traitement dus à de la confusion, i.e. la présence de facteurs liés à la fois à l'affectation du traitement et à la variable d'intérêt. Cependant, un problème majeur avec des grandes études observationnelles est leur complexité et leur caractère souvent incomplet : les covariables sont souvent prises à différents niveaux et stades, elles peuvent être hétérogènes – catégorielles, discrètes, continues – et contiennent presque inévitablement des valeurs manquantes. Le problème des valeurs manquantes dans l'inférence causale a longtemps été ignoré et n'a regagné l'attention que récemment en raison des impacts non négligeables en termes de puissance et de biais induits par des analyses de cas complètes et des modèles d'imputation mal spécifiés. Nous discutons des conditions dans lesquelles une inférence causale peut être possible malgré la présence de valeurs manquantes dans les facteurs confondants, nous comparons différentes méthodes proposées dans le passé pour traiter les valeurs confondantes manquantes et proposons deux estimateurs ATE double robustes qui rendent directement compte des valeurs manquantes. Nous évaluons la performance de nos estimateurs sur une base de données prospective considérable contenant des informations détaillées sur environ 20 000 patients poly-traumatisés graves en France. À l'aide des estimateurs d'ATE proposés et de cette base de données, nous étudions l'effet sur la mortalité de l'administration de l'acide tranexamique aux patients présentant un choc hémorragique.

Mots-clés. données manquantes, inférence causale, réponses potentielles, données observationnelles, estimation du score de propension, prise en charge de poly-traumatisés graves

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 power and bias induced by complete case analyses and misspecified imputation models. We discuss conditions under which causal inference can be possible despite missing confounder values, we compare different methods proposed in the past to deal with missing confounder values and propose two doubly robust ATE estimators which directly account for the missing values. We assess the performance of our estimators on a large prospective database containing detailed information about nearly 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 hemorrhagic shock in the context of critical care management.

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

1 Context and motivation

1.1 Hemorrhagic shock 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. 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. 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). In this context we are interested in estimating the effect of tranexamic acid, an antifibrinolytic agent that limits excessive bleeding, on the in-ICU mortality among patients with hemorrhagic shock, based on the observational database.

As in almost all areas of empirical research, the Traumabase also presents missing data. 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).

1.2 Causal inference

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 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 previously 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 authorisation of new drugs or other treatments. However RCTs are generally very expensive in terms of time and financial costs. 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. But, as identified as the fundamental problem of causal inference by (Holland, 1986), we want to estimate 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 exist a multiplicity of well studied methods to efficiently and consistently estimate causal effects in different scenarios. One scenario that has only rarely been addressed rigorously in the past is the case of missing confounder values. 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 with the goal to apply these methods to answer the medical question in the context of critical care management introduced above.

1.3 Definitions and assumptions

In light of our goal of performing causal analyses, we consider the potential outcomes framework from the Rubin causal model (Rubin, 1974) and define potential outcomes $Y_i(t)$ for observation i and treatment $t \in \text{Span}(T)$. 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 $Y_i \triangleq T_i Y_i(1) + (1 - T_i)Y_i(0)$. In what follows we consider the binary treatment case and refer to individuals having $T_i = 1$ as *treated* and to those having $T_i = 0$ 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 estimated 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.

The *ignorability* or *unconfoundedness* assumption states that all confounding factors are measured, i.e. conditionally on covariates $X_i \in \mathcal{X}$, the treatment assignment is independent of the potential outcomes. Formally this means

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

Under this assumption we can hope to identify $\mathbb{E}[Y_i(t)]$, $t \in \{0, 1\}$, from the data despite the inherent problem of missing values due to counterfactuals since $\mathbb{E}[Y_i(t) \,|\, T_i = t, X_i] = \mathbb{E}[Y_i(t) \,|\, X_i]$, $t \in \{0, 1\}$. We also need to assume the *Stable Unit Treatment Value Assumption* (SUTVA) (Rubin, 1978). Finally we assume *probabilistic treatment assignment*, i.e. if we define the propensity score $e(x) \triangleq \mathbb{P}(T_i = 1 \,|\, X_i = x)$ (Rosenbaum and Rubin, 1983), then we assume $0 < e(x) < 1$, for all $x \in \mathcal{X}$. A well known and important result is the balancing property of the propensity score, that is if condition (2) holds then we can control for $e(X_i)$ in order to balance the covariates distributions.

1.4 Consistent treatment effect estimators in the complete case

We distinguish two cases of data settings: experimental data from an 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 this case one can consistently estimate τ by a difference in means estimator:

$$\hat{\tau}_{DM} \triangleq \frac{1}{|\{i \ : \ T_i = 1\}|} \sum_{i \ : \ T_i = 1} Y_i - \frac{1}{|\{i \ : \ T_i = 0\}|} \sum_{i \ : \ T_i = 0} Y_i. \tag{3}$$

The second setting is conceptually 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. Assumption (2) allows however to overcome this issue by adjusting for the nonrandom treatment assignment and balancing the covariates distributions. The emulation of an RCT from observational data can be approached in different ways (Imbens, 2004): (1) matching or stratifying the observations on their pre-treatment covariates, (2) matching or stratifying the observations on their propensity score, (3) inverse propensity weighting.

Matching and stratification can be considered as a nonparametric data pre-processing and allows to balance the observations of the two groups, provided that the covariate space is small enough and that there are sufficiently many observations in both groups. Inverse propensity weighting re-weights every observation by the inverse of its propensity score in order to balance

the distributions in the two groups, leading to a difference in weighted means estimator $\hat{\tau}_{IPW}$. The quality of this reweighting depends solely on the quality of the propensity score estimation and in order to reduce the sensitivity of the corresponding treatment effect estimator $\hat{\tau}_{IPW}$ to model misspecification, other estimators have been proposed that additionally model the outcome Y and combine the weighting and outcome estimates (Robins et al., 1994):

$$\hat{\tau}_{DR} \triangleq \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + T_i\frac{Y_i - \hat{\mu}_1(X_i)}{\hat{e}(X_i)} - (1-T_i)\frac{Y_i - \hat{\mu}_0(X_i)}{1-\hat{e}(X_i)}. \tag{4}$$

Such estimators belong to the class of *doubly robust* treatment effect estimators and have the appealing property of being consistent as long as at least one of the two estimations, the propensity scores or the outcomes, is consistent.

2 Extension to the incomplete confounders case

2.1 Additional assumptions

In order to be able to infer treatment effects, we need to adjust, or rather augment, the initial unconfoundedness assumption such that treatment assignment is unconfounded given only the observed covariates and the missingness pattern. (Rosenbaum and Rubin, 1984) sketch a generalized propensity score analysis which uses a *missingness pattern approach* for estimating treatment effects. They define the *generalized propensity score* e^* by conditioning treatment assignment T on the covariates $X \in \mathcal{X}$ (wheredim$(\mathcal{X}) = p$) and the response pattern $R \in \{0,1\}^p$ which is defined as $R_j \triangleq \mathbb{1}_{\{X_j \, is \, observed\}}$. According to this response pattern we can express X as $X = (X^{obs}, X^{mis})$ where $X^{obs} \triangleq \{X_j : R_j = 1\}$. With these notations the generalized propensity score can be written as

$$e^*(X^{obs}, R) \triangleq \mathbb{P}(T = 1 \,|\, X^{obs}, R) \tag{5}$$

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 treatment assignment is unconfounded given X^{obs} and R, i.e. $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \,|\, X_i^{obs}, R_i$. 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}.

We develop our method building on remarks of (Blake et al., 2019) who discuss this approach, the missingness pattern approach (MPA), and the missing indicator approach as a special case, and the plausibility of their underlying assumptions.

$$\text{unconfoundedness}^*: \quad T \perp\!\!\!\perp Y(t) \,|\, (X^{obs}, X^{mis}), R \quad \text{for } t \in \{0,1\} \tag{6}$$

$$\text{CIT}: \quad T \perp\!\!\!\perp X^{mis} \,|\, X^{obs}, R \tag{7}$$

$$\text{CIO}: \quad Y(t) \perp\!\!\!\perp X^{mis} \,|\, X^{obs}, R \quad \text{for } t \in \{0,1\}, \tag{8}$$

which are missingness unconfoundedness, Conditionally Independent Treatment and Conditionally Independent Outcomes, respectively. These assumptions describe the relationships between missing confounder values and the treatment or outcome. If either the CIT or CIO assumption holds, X does not confound the relationship between treatment and outcome when it is missing.

2.2 Existing estimators

Several approaches to treatment effect estimation with incomplete confounders have been proposed in the last few decades: multiple imputation using all observed information, including treatment assignment and outcome (Leyrat et al., 2019), inverse propensity weighting with propensity scores estimated using the missingness pattern approach (Rosenbaum and Rubin, 1984; D'Agostino Jr and Rubin, 2000), under a latent confounders assumption a matrix factorization pre-processing recovering the latent confounders from the incomplete covariates (Kallus et al., 2018).

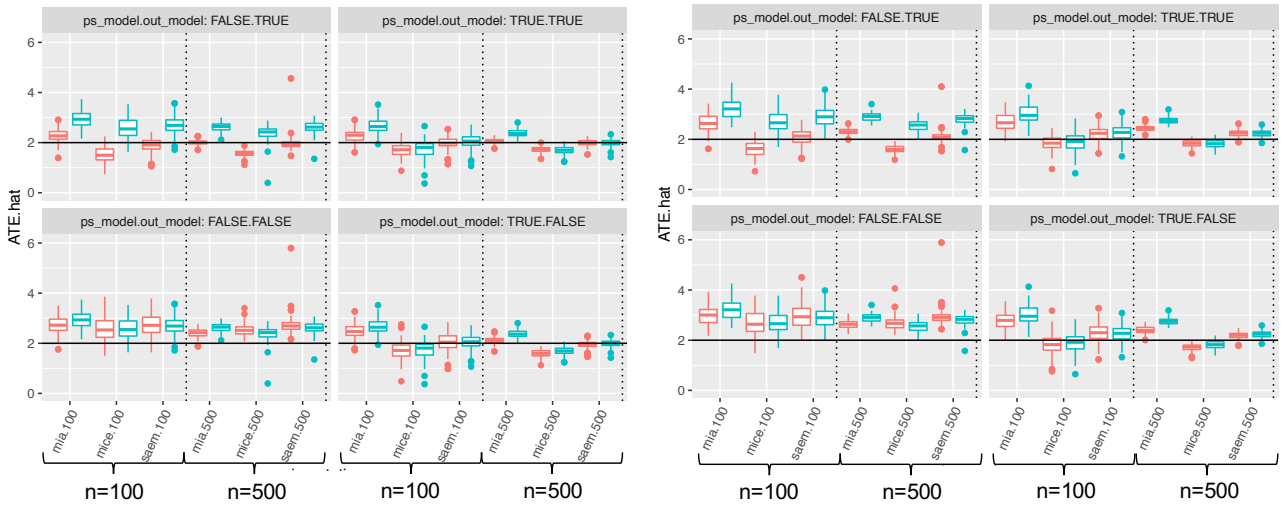
2.3 Doubly robust estimators

Building on two recent results of consistent predictions in the presence of missing values (Jiang et al., 2018; Josse et al., 2019), we propose two doubly robust treatment effect estimators that consistently estimate $e(\cdot)$ (or $e^*(\cdot)$) and $\{\mu_t(\cdot)\}_{t\in\{0,1\}}$ in different ways, one is based on expectation maximization (Dempster et al., 1977) that fits a logistic and a linear model on the incomplete data the other one uses random trees and a *missing incorporated in attributes* approach, MIA (Twala et al., 2008). The former is valid under normality assumptions and the MAR missingness mechanism (Little and Rubin, 2002) and allows to efficiently estimate a logistic regression using a stochastic approximation expectation maximization (SAEM) algorithm. The latter is a nonparametric approach using random trees that allow to incorporate missing values information, i.e. an implicit encoding of the response pattern R. This is due to the random trees' ability to handle the half-discrete nature of variables with missing values. This second estimator is more flexible in the sense that it does not build on any assumption about the missingness mechanism or a model specification of the propensity score and the outcome but only on the assumptions discussed earlier.

2.4 Simulation results

We assess the performance of the previously introduced treatment effect estimators in different scenarios, modifying the confounders' correlation structure and the missingness mechanism. We follow the strategy of (Kang et al., 2007) for assessing the estimators' sensitivity to model misspecification, both for propensity and outcome models. Results for strongly correlated confounders are similar to those for weakly correlated confounders. We therefore only report the results for the former. The same holds for results for MAR and MCAR and we only report results for the MAR scenario. The results reported in Figure 1 suggest that under the above

assumptions, the different $\hat{\tau}_{IPW}$ and $\hat{\tau}_{DR}$ estimators (using either SAEM or MIA) have the same sensitivity to model misspecification as in the complete case of (Kang et al., 2007), i.e. $\hat{\tau}_{IPW}$ is consistent if and only if $\hat{e}(\cdot)$ is consistent and $\hat{\tau}_{DR}$ is consistent if $\hat{e}(\cdot)$ or $\{\hat{\mu}_t(\cdot)\}_{t\in\{0,1\}}$ is consistent. However, if both CIT and CIO are violated, none of the proposed methods achieves unbiased results.



(a) $(CIT, CIO) = (True, False)$ (b) $(CIT, CIO) = (False, False)$

Figure 1: Strongly correlated confounders, MAR mechanism. *mice*: imputation and standard complete case estimators on imputed data; *mia*: random forest propensity and outcome estimation with MIA; *saem*: EM estimation for propensity and outcome models; (red: $\hat{\tau}_{DR}$; turquoise: $\hat{\tau}_{IPW}$; black solid line: true treatment effect τ; 200 simulations for sample sizes $n \in \{100, 500\}$).

3 Discussion and perspectives

Our empirical study corroborates the necessity of the CIT/CIO assumptions and our conjecture of doubly robust treatment effect estimation with incomplete confounders under these assumptions.

After having theoretically proven the double robustness of our proposed estimators we will apply them on the Traumabase to answer the medical question raised in the introduction.

We conjecture that the CIT/CIO assumptions can be made implicit by assuming a latent confounder model. More specifically, instead of assuming unconfoundedness given only the observed X^{obs} and R, we assume unconfoundedness given a set of latent confounders, similar to the setting of (Kallus et al., 2018). Once this relationship established, we claim that we do not need recovering these latent confounders since we demonstrated the consistency of our proposed estimators under the CIT/CIO assumptions.

Bibliography

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.

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*.

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.

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.

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.

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

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

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

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

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

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.

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.

Little, R. J. and Rubin, D. B. (2002). *Statistical analysis with missing data*. J. Wiley.

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.

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.

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

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

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.