Doubly robust treatment effect estimation with incomplete confounders

Treatment effect estimation of tranexamic acid on mortality for traumatic brain injury patients

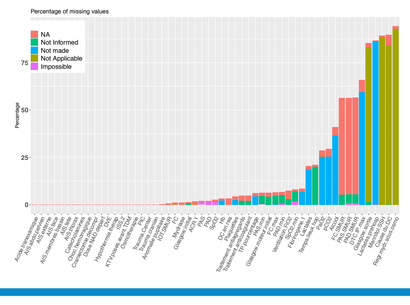
IMKE MAYER[1,2], JULIE JOSSE[2,3], STEFAN WAGER[4], TOBIAS GAUSS[5], JEAN-DENIS MOYER[5]

[1]École des Hautes Études en Sciences Sociales, [2]École Polytechnique, [3] Inria XPOP, [4] Stanford University, [5] Traumabase® Group

MOTIVATIONS

Estimate the effect of tranexamic acid (TA) on the in-ICU mortality among patients with traumatic brain injury (TBI), based on the observational database **Traumabase®**. This database includes 7,945 major trauma patients, of which 3,050 have traumatic brain injury, with 244 pre-hospital and hospital measurements. The data is **heterogeneous**, being composed of both quantitative or categorical variables. Major trauma is a public health challenge and a major source of mortality and handicap around the world.

Treatment effect (TE) estimation on **observational data** is challenging when the data contains **missing values**.

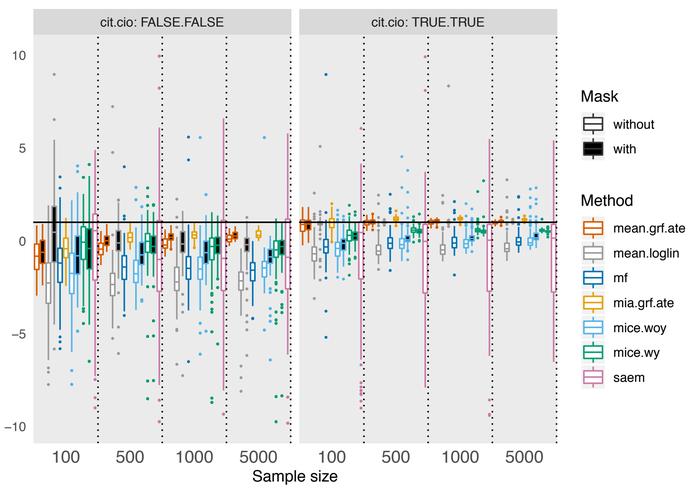


PROPOSAL

- Comparison of different TE estimators when covariates are partially observed, analysis of the bias.
- Proposition of new double robust TE estimator, based on random forests, handling incomplete confounders.
- Application to critical care patient data.

SIMULATIONS

- i.i.d observations from mixture model: $X \mid C = c \sim \mathcal{N}(\mu_c, \Sigma_c), X \in \mathbb{R}^{10}$
- Logistic-linear model for $W \in \{0,1\}, Y \in \mathbb{R}$, satisfying or not CIT/CIO.
- MNAR (NA in $X_1, \ldots, X_5$ depend on $X_6, \ldots, X_{10}$).
- True ATE: $\tau = 1$.
- DR estimator $\tau_{DR,*}$.
- (Generalized) propensity score (PS) estimation with missing values:
  - (a) imputation (**mean**, **mice**, LR matrix factorization) + logistic regression,
  - (b) logistic regression handling NAs (**SAEM**) [3],
  - (c) random forest with missing incorporated in attributes (**MIA**) or mean imputation.



FUTURE RESEARCH

- Prove consistency / double robustness of the proposed ATE estimator in cases other than MCAR (and for heterogeneous data).
- TBI is very heterogeneous in terms of clinical presentation, pathophysiology and outcome
  → heterogeneous TE estimation.
- Long-term objective: developing a decision support tool for clinical care management.
- Compare results to the soon to be published randomized controlled trial *CRASH-3* results [1].

CAUSAL INFERENCE WITH MISSING VALUES IN THE COVARIATES

Assumptions:

→ **Rubin's potential outcome framework**: W binary treatment, $(Y_i(t))_{w\in\{0,1\}}$ potential outcomes.

$$\tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \quad \textbf{(ATE)},$$

$\mathbf{X} = (\mathbf{X}^{obs}, \mathbf{X}^{mis}) \in \mathbb{R}^{n\times p}$ completely observed confounders, $e(x) = \mathbb{P}(W = 1 \mid X = x)$ propensity score, $\mu_w(x) = \mathbb{E}[Y(w)|X = x]$ conditional response surface.

→ **Missing values**: $\mathbf{R} \in \{0,1\}^{n\times p}$ response indicator matrix, $\tilde{\mathbf{X}} = \mathbf{X} \odot \mathbf{R} + \texttt{NA}(1 - \mathbf{R}) \in (\mathbb{R} \cup \texttt{NA})^{n\times p}$ observed confounders, $e^*(x, r) = \mathbb{P}(W = 1 \mid X^{obs} = x, R = r)$ generalized propensity score [7].

→ Classical causal inference assumptions: SUTVA, unconfoundedness, overlap.

→ Additional assumptions due to missingness:

- unconfoundedness*: $Y_i(t) \perp\!\!\!\perp W_i \mid X_i, R_i \quad t \in \{0,1\}$
- CIT or CIO: $W_i \perp\!\!\!\perp X_i^{mis} \mid X_i^{obs}, R_i$ or $Y_i(t) \perp\!\!\!\perp X_i^{mis} \mid X_i^{obs}, R_i \quad t \in \{0,1\}$

Method

→ Doubly robust treatment effect estimator $\hat{\tau}_{DR,*}$:

$$\hat{\tau}_{DR,*} = \frac{1}{n}\left(\sum_{i=1}^{n} \hat{\mu}_1(\tilde{X}_i) - \hat{\mu}_0(\tilde{X}_i) + W_i \frac{Y_i - \hat{\mu}_1(\tilde{X}_i)}{\hat{e^*}(\tilde{X}_i)} - (1 - W_i)\frac{Y_i - \hat{\mu}_0(\tilde{X}_i)}{1 - \hat{e^*}(\tilde{X}_i)}\right)$$

Propensity model (e^*) correctly specified:

$$\mathbb{E}\left[1 - \frac{W_i}{e^*(\tilde{X}_i)} \mid X_i^{obs}, R_i\right] = 0$$

$\Rightarrow \hat{\tau}_{DR,*}$ & $\hat{\tau}_{IPW,*}$ are consistent.

Outcome model (μ) correctly specified:

$$\mathbb{E}\left[Y_i - \mu_1(\tilde{X}_i) | W_i = 1, X_i^{obs}, R_i\right] = 0$$

$\Rightarrow \hat{\tau}_{DR,*}$ is consistent.

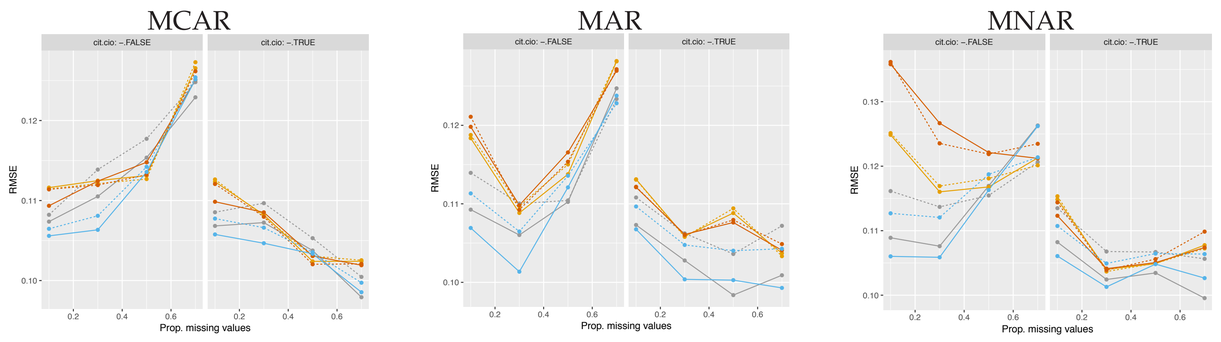
→ Parametric or nonparametric estimation of $\mu_t(\cdot)$ and $e(\cdot)$ → **interpretability of $\hat{\tau}_{\mathbf{DR}}$** is the same.

→ Nonparametric estimation using **random forests** to handle heterogeneous data and missing values consistently under MCAR [4].

FIRST RESULTS

On IHDP data [2]:

→ Simulated observational data from original experimental data
→ 6 quant. variables, quant. outcome, binary treatment
→ MCAR and MNAR
→ Simulate Y w/ and w/o CIO.
→ Same methods as in Simulations part.



⇒ Empirically, importance of CIO assumption increases with the amount of missing values, for all mechanisms.

On Traumabase:

→ 12 identified confounders (continuous & discrete & categorical).

→ 3169 patients with traumatic brain injury.

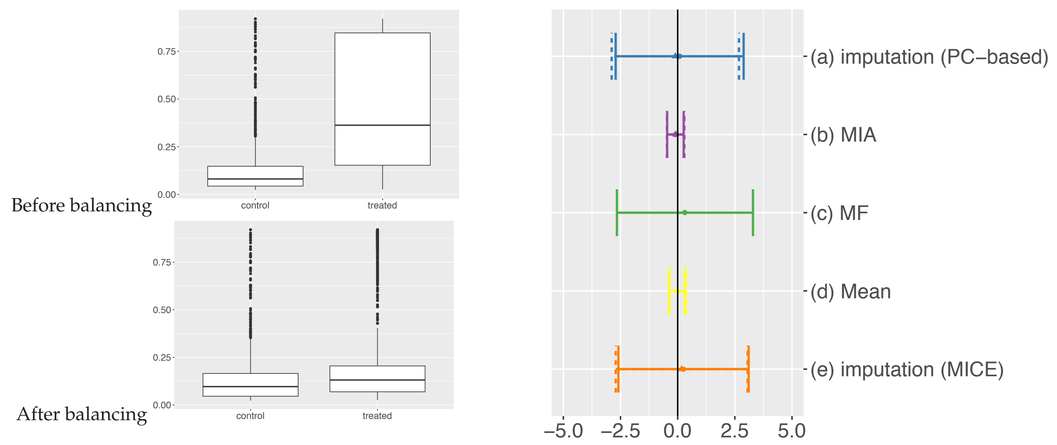
→ 12% treated patients.

→ 0% - 23% of missing values (in confounders).

→ Fully observed treatment and outcome.

→ PS and outcome regression using random forests with sample splitting and cross-splitting (R-package `grf`)

→ 5 estimation approaches:
(a) Imputation (pca-based)
(b) Missing Incorporated in Attribute
(c) Low-rank approximation [5]
(d) Mean imputation
(e) Imputation (mice)



- Difference in percentage points between mortality rates in treatment and control groups.
- **No evidence for rejecting null hypothesis of no effect of TA on in-ICU mortality among TBI patients**.
- Next: different TE w.r.t. severity of TBI and extra-cranial lesions?

REFERENCES

[1] Y. Dewan, E. O. Komolafe, J. H. Mejía-Mantilla, P. Perel, I. Roberts, and H. Shakur. Crash-3-tranexamic acid for the treatment of significant traumatic brain injury: study protocol for an international randomized, double-blind, placebo-controlled trial. *Trials*, 13(1):87, 2012.

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

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

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

[5] N. Kallus, X. Mao, and M. Udell. Causal inference with noisy and missing covariats via matrix factorization. *arXiv preprint*, 2018.

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

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

See also **R-miss-tastic**, a unified platform on missing values methods and workflows: