# MissDeepCausal

## Causal inference from incomplete data using deep latent variable models


Julie Josse[(1,2)], Imke Mayer[(1,3)], Jean-Philippe Vert[(4)]

[(1)]École Polytechnique, [(2)]Inria XPOP, [(3)]École des Hautes Études en Sciences Sociales, [(4)]Google Brain


## Motivations

*"One of the big ironies of Big Data is that missing data play an even more important role."* – R. Samworth (2019)

- Goal: Infer causal effects of a treatment from – almost inevitably incomplete – observational data.
- Issue: Available methods [?] rely on the difficult *Unconfoundness with missing values* hypothesis or parametrics models.
- Assumption: Covariates are noisy proxies of true latent confounders, relationships can be nonlinear.
- Strategy: Couple VAE with missing values and double robust estimation.

## Framework

### Neyman-Rubin potential outcomes [?]

→ $W$ binary treatment, $(Y_i(w))_{w\in\{0,1\}}$ potential outcomes.

→ Average treatment effect (ATE):

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

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

→ Classical assumptions: SUTVA, overlap.

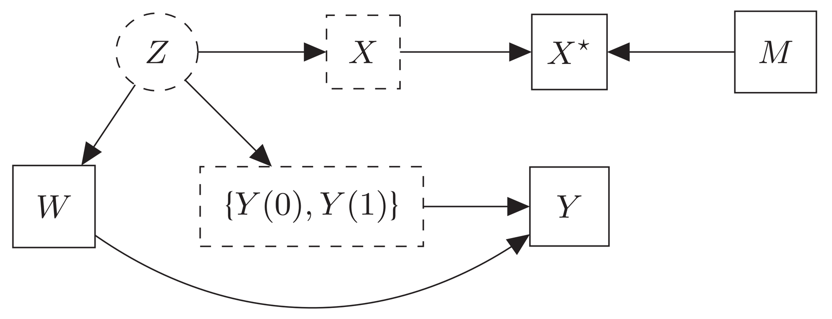
→ AIPW doubly robust estimator [?]:

$$\hat{\tau}_{DR} = \frac{1}{n}\sum_{i=1}^{n} \mu_1(X_i) - \mu_0(X_i) + W_i\frac{Y_i - \mu_1(X_i)}{e(X_i)} - (1 - W_i)\frac{Y_i - \mu_0(X_i)}{1 - e(X_i)} \tag{1}$$

### Missing values and latent confounders:

→ $\mathbf{M} \in \{0,1\}^{n\times p}$ mask, missing at random [?, MAR], $\mathbf{X}^* = \mathbf{X} \odot (1 - \mathbf{M}) + \text{NA} \odot \mathbf{M} \in \mathcal{X}^*$ observed covariates, $\mathcal{X}^* = (\mathbb{R} \cup \text{NA})^{n\times p}$.

→ **Unconfoundedness w.r.t. latent variables:** $\mathbf{Z} \in \mathbb{R}^{n\times d}$ latent confounders.




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

## Causal inference with missing values in the covariates

Assume an unbiased estimator $\hat{f}(Z)$ of $\mathbb{E}[Y(1) - Y(0)|Z]$ and access to the distribution $P(Z|X^*)$

### Latent variables estimation as a pre-processing step (MDC-process)

→ Heuristic nonlinear extension of [?]

→ Regression model: $Y = \tau W + Z\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$.

**MDC-process**

1. Estimate latent confounders with $\hat{Z}(x^*) = \mathbb{E}[Z|X^* = x^*]$.
2. Plug these $\hat{Z}(x^*)$ into regression model or define $\hat{\tau}_{process} = \mathbb{E}[f(\mathbb{E}[Z|X^*])]$.

### Multiple imputation strategy

→ Monte-Carlo approximation using posterior distribution $P(Z|X^*)$.

**MDC-MI**

1. Sample $(Z^{(j)})_{1\le j\le B}$ from $\hat{P}(Z|X^*)$.
2. For each sample $j$, compute estimate $\hat{\tau}^{(j)} = f(Z^{(j)})$.
3. Aggregate into final estimate: $\hat{\tau}_{MI} = \frac{1}{B}\sum_{j=1}^{B}\hat{\tau}^{(j)} \approx \mathbb{E}[\mathbb{E}[f(Z|X^*)]]$.

Estimation of and sampling from $P(Z|X^*)$:
1) Use missing data importance weight autoencoder [?, MIWAE]: imputation by a constant maximizes the ELBO.
2) Approximate with self-normalized importance sampling on variational distribution $Q(Z|X^*)$:

$$\mathbb{E}[s(Z)|X^*] \approx \sum_{l=1}^{L} w_l s(Z^{(l)}),\text{ where } w_l = \frac{r_l}{r_1 + \cdots + r_L} \text{ and } r_l = \frac{p(X^*|Z^{(l)})p(Z^{(l)})}{q(Z^{(l)}|X^*)} \text{ for any measurable function } s.$$

## IHDP data [?]

| % NA | Method | Δ OLS | Δ DR$_{rf}$ |
|---|---|---|---|
| 0 | $X$ (complete data) | $0.72 \pm 0.02$ | $0.20 \pm 0.01$ |
| | $MF$ | $0.56 \pm 0.03$ | $0.16 \pm 0.01$ |
| | $MDC.process$ | $0.51 \pm 0.03$ | $0.19 \pm 0.03$ |
| | $MDC.mi$ | $0.47 \pm 0.03$ | $0.14 \pm 0.02$ |
| | $CEVAE(X)$ | $0.34 \pm 0.02$ | |
| 10 | $MICE$ | $0.85 \pm 0.02$ | $0.24 \pm 0.01$ |
| | $MIA.GRF$ | – | $0.23 \pm 0.01$ |
| | $MF$ | $0.50 \pm 0.03$ | $0.15 \pm 0.01$ |
| | $MDC.process$ | $0.42 \pm 0.02$ | $0.16 \pm 0.02$ |
| | $MDC.mi$ | $0.35 \pm 0.02$ | $0.13 \pm 0.02$ |
| | $CEVAE(X_{imp})$ | $0.31 \pm 0.01$ | |
| 30 | $MICE$ | $1.20 \pm 0.02$ | $0.32 \pm 0.01$ |
| | $MIA.GRF$ | – | $0.17 \pm 0.01$ |
| | $MF$ | $0.39 \pm 0.02$ | $0.17 \pm 0.01$ |
| | $MDC.process$ | $0.37 \pm 0.02$ | $0.15 \pm 0.02$ |
| | $MDC.mi$ | $0.30 \pm 0.02$ | $0.13 \pm 0.01$ |
| | $CEVAE(X_{imp})$ | $0.38 \pm 0.02$ | |
| 50 | $MICE$ | $1.54 \pm 0.03$ | $0.42 \pm 0.01$ |
| | $MIA.GRF$ | – | $0.19 \pm 0.01$ |
| | $MF$ | $0.28 \pm 0.01$ | $0.21 \pm 0.02$ |
| | $MDC.process$ | $0.24 \pm 0.01$ | $0.21 \pm 0.02$ |
| | $MDC.mi$ | $0.18 \pm 0.01$ | $0.22 \pm 0.03$ |
| | $CEVAE(X_{imp})$ | $0.38 \pm 0.02$ | |

Mean absolute error Δ (with standard error) across 1000 simulations. OLS: estimator obtained by regression, DR: doubly robust estimator. $X_{imp}$: mean imputed $X^*$. $MIA.GRF$: causal forest extension handling incomplete covariates [?].

## Future Research

- Handling missing not at random type data (MNAR).
- Heterogeneous treatment effect estimation.

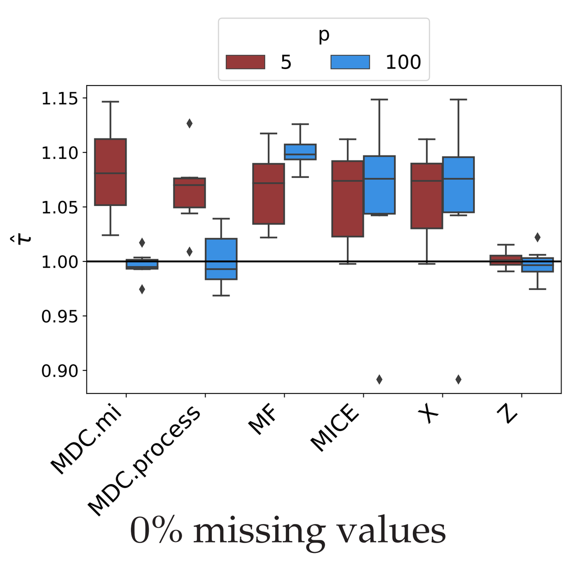
## Simulations

→ Varying number of covariates ($p$), fixed small number of latent variables ($d = 3$) and $n = 10000$.

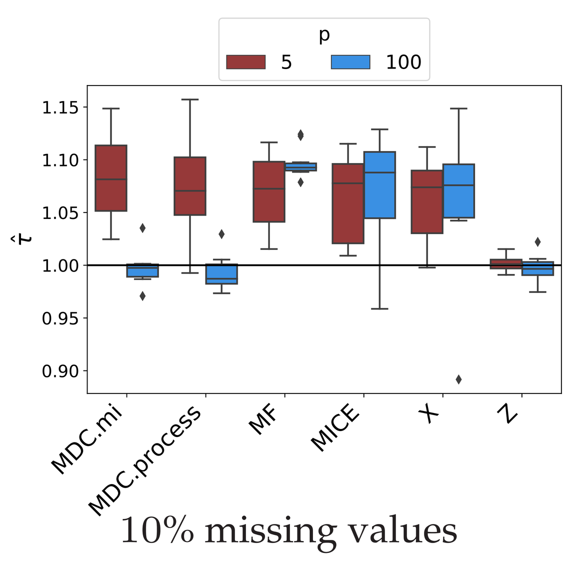
→ Deep latent variable data generating model [?].

→ Choice for $f$: doubly robust estimator (??).

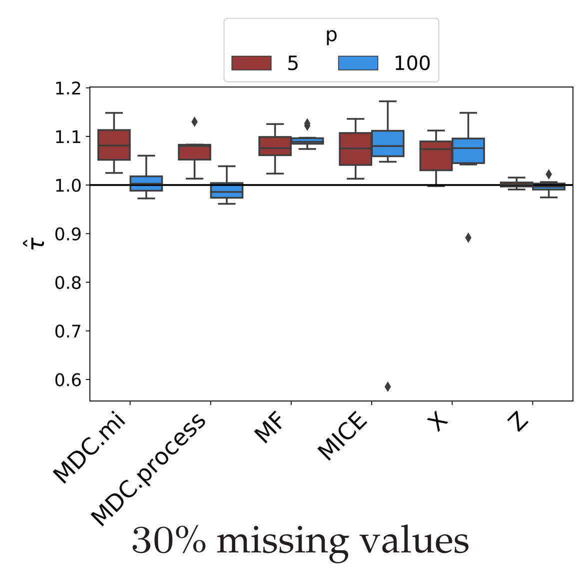
→ Comparison methods: $f$ on estimated linear latent factors [?, MF], $f$ on multiply imputed $X^*$ (MICE).




0% missing values




10% missing values




30% missing values

## 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.


For more details and results of **MissDeep-Causal**, read our full paper:

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