Simulation-based studies related to the G-computation for causal inference: an overview of recent results

Yohann Foucher (1), Arthur Chatton (1,2), Maxime Léger (1,3), Florent Le Borgne (1,2)

1. INSERM UMR 1246 - SPHERE, Nantes University 2. IDBC - Groupe A2com. 3. Angers Hospital University

Yohann.Foucher@univ-nantes.fr

June 22, 2022

UNIVERSITÉ DE NANTES

SPHERE
methodS in Patient-centered outcomes & HEalth ResEarch

Plan

UNIVERSITÉ DE NANTES

Average causal effect

- Let A denote the binary treatment ($A = 1$ for treatment and 0 otherwise).
- Let Y denote the binary outcome ($Y = 1$ for event and 0 otherwise).
- Let $Z = (Z_1, Z_2, ..., Z_k)$ denote the set of the k baseline covariates.
- Let $Y(1)$ and $Y(0)$ be the two potential outcomes under the treatment and the control, respectively.
- The average causal effect is :

$$ACE = E[Y(1) - Y(0)]$$

- It represents the mean difference between the outcomes of individuals if they had been treated or untreated.

UNIVERSITÉ DE NANTES

Three categories of methods for estimating the ACE

- The regression of the treatment allocation to obtain propensity scores (PS) :
  - Inverse Probabiliy Weighting (**IPW**)
  - Full matching (**FM**)
  - Etc.
- The regression of the outcome for G-computation (**GC**)
- The targeted maximum likelihood estimator (**TMLE**) as a doubly robust estimation which combines the outcome and treatment regressions
- Etc.

UNIVERSITÉ DE NANTES

Literature related to GC is less prolific compared to PS-based methods

- Suppose (Y_i, A_i, Z_i) a dataset of n independent realisations of (Y, A, Z).
- The first step of GC is to fit $f(Y|A, Z)$
- This outcome model is frequently referred to the Q-model.
- The second step consists in predicting the two potential outcomes for each individual i : $\hat{Y}_i(1) = \hat{f}(Y|1, Z_i)$ and $\hat{Y}_i(0) = \hat{f}(Y|0, Z_i)$
- The average causal effect is then estimated by

$$A\hat{C}E = n^{-1} \sum_i \left[\hat{Y}_i(1) - \hat{Y}_i(0) \right]$$

UNIVERSITÉ DE NANTES

Three simulation-based were performed in the context of binary outcome and binary treatment

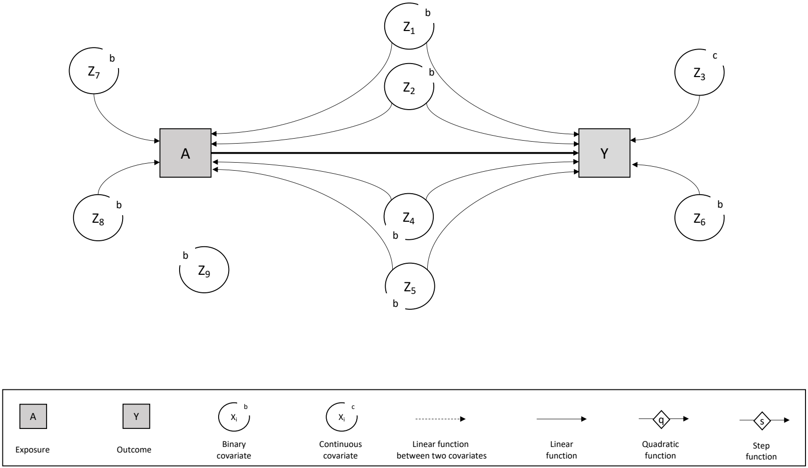
1. Which covariates should be considered in GC (true confounders, those causing the outcome, etc.) ?
2. What is the robustness of GC to a near-violation of the positivity assumption ?
3. What are the performances of GC associated with machine learning (ML) ?

UNIVERSITÉ DE NANTES

Plan

UNIVERSITÉ DE NANTES

Simulated data



- 4 sample sizes ($n = 100, 300, 500, 2000$)
- 2 treatment effect (H_1 versus H_0)

Compared methods

We compared several methods (all based on logistic regression) :

- GC : variance obtained by parametric simulations.
- IPTW : stabilized weights and robust sandwich-type variance estimator.
- FM : robust sandwich-type variance estimator.
- TMLE : variance obtained by efficient influence curve.

We compared different sets of covariates :

- those causing outcome $(Z_1, Z_2, Z_3, Z_4, Z_5, Z_6)$.
- those causing treatment $(Z_1, Z_2, Z_4, Z_5, Z_7, Z_8)$.
- those causing outcome and treatment (true confounders : Z_1, Z_2, Z_4, Z_5).
- all the covariates.

Results for $n = 100$ under H_1

n	method	selection strategy	mean bias π_0	mean bias π_1	mean bias $\Delta\pi$	mean bias log OR	log OR MSE	log OR MSE*	log OR VEB (%)	log OR coverage (%)	log OR power (%)
100	GC	outcome	0.000	−0.001	−0.001	0.012	0.526	0.716	−6.2	94.1	17.7
		treatment	0.002	−0.001	−0.003	0.006	0.580	0.786	−5.7	94.1	14.0
		common	0.002	−0.001	−0.003	0.006	0.552	0.735	−4.2	94.8	15.1
		entire	−0.001	−0.001	−0.001	0.013	0.558	0.768	−8.8	93.3	16.9
	IPTW	outcome	0.000	−0.001	−0.001	0.008	0.578	0.727	10.8	97.3	7.8
		treatment	−0.000	−0.001	−0.001	0.000	0.716	0.837	−1.2	95.1	9.8
		common	0.002	−0.001	−0.003	0.003	0.587	0.743	6.6	96.8	8.8
		entire	−0.003	−0.001	0.002	0.005	0.741	0.838	−1.5	95.2	9.6
	TMLE	outcome	−0.001	−0.001	0.000	0.002	0.694	0.794	30.0	95.7	5.8
		treatment	0.000	−0.001	−0.001	−0.020	0.876	0.955	183.3	98.8	1.0
		common	−0.000	−0.001	−0.001	−0.001	0.702	0.794	10.4	95.3	7.3
		entire	−0.003	−0.001	0.001	−0.013	0.886	0.953	412.2	98.8	0.5
	FM	outcome	−0.004	−0.001	0.003	0.022	0.665	0.787	−16.7	90.1	18.9
		treatment	−0.006	−0.001	0.004	0.017	0.822	0.911	−32.3	81.3	25.2
		common	−0.001	−0.001	−0.000	0.010	0.653	0.795	−15.3	91.0	17.5
		entire	−0.008	−0.001	0.006	0.022	0.842	0.921	−33.8	80.3	26.7

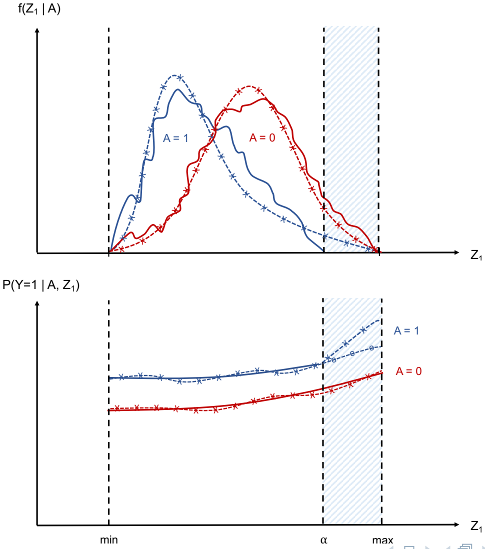
- No bias for GC, IPTW, TMLE.
- The highest power was for GC with the covariates causing outcome.

UNIVERSITÉ DE NANTES

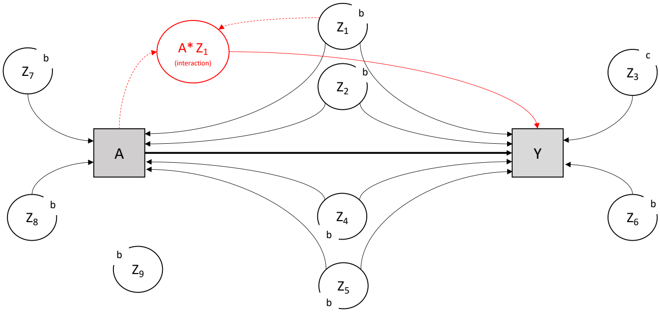
Plan

UNIVERSITÉ DE NANTES

Extrapolation issue from the Q-model

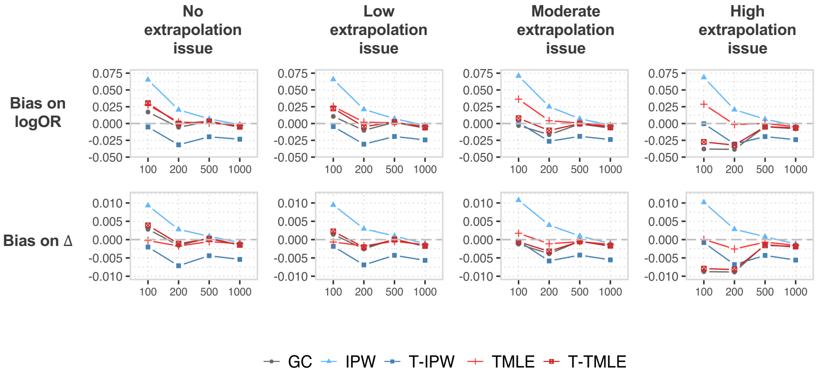


Simulated data



- Z_1 was generated with a 10% prevalence. The near-violation concerned :
  - 0% of the datasets for $n \geq 500$.
  - 1.3% for $n = 200$ subjects.
  - 14.1% for $n = 100$ subjects.
- The extrapolation issue was proportional to the interaction level.

Results



- T-IPW and T-TMLE are the truncated IPW and TMLE with bounds at the 10th and 90th percentiles.
- GC and TMLE were the most robust methods
  - only high extrapolation issue lead to substantial bias.

UNIVERSITÉ DE NANTES

Plan

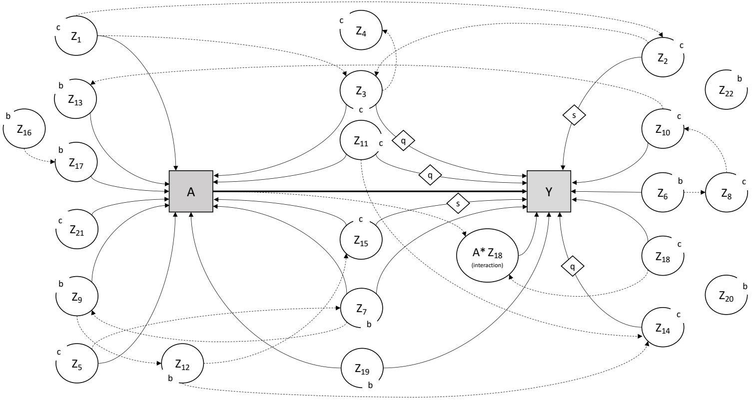
UNIVERSITÉ DE NANTES

Method : a super learner (SL) applicable in practice with small sample size and computation time

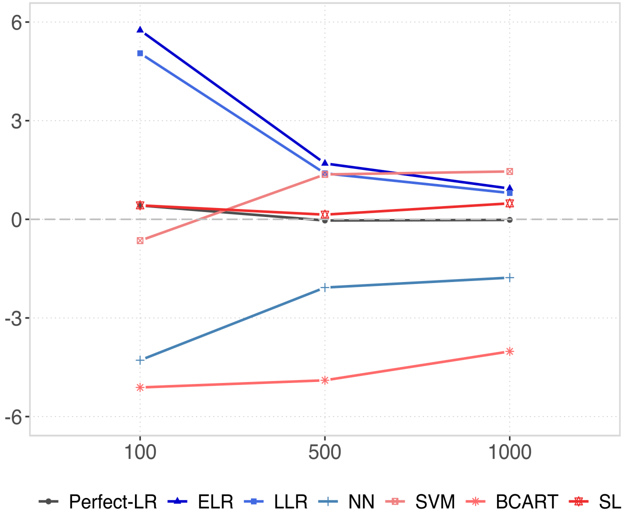
- The SL consists in averaging the predictions obtained from the four approaches :
  - **Lasso logistic regression.** A was forced. All the possible interactions between A and covariates Z were tested. B-splines for the quantitative covariates Z.
  - **Elasticnet logistic regression** with similar assumptions.
  - **Neural network** with one hidden layer.
  - **Support vector machine** with a radial basis function kernel.
- The tunning parameters and the weights were obtained by maximizing the average AUC of a ten-fold cross-validation.
- The variance was obtained by bootstrap cross validation.
- The tuning parameters being estimated once on the entire sample.

UNIVERSITÉ DE NANTES

Simulated data

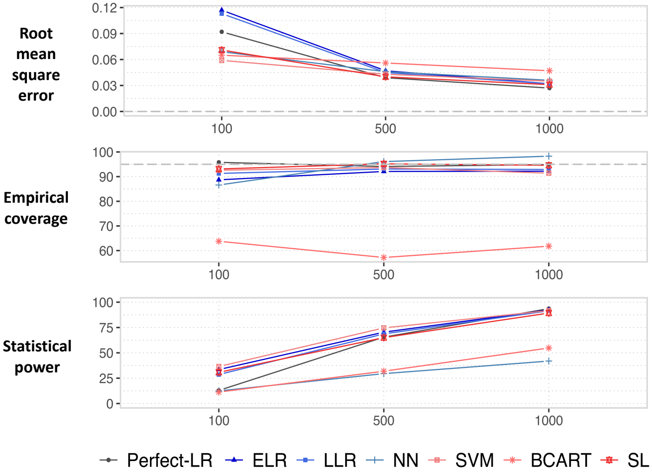


Results in terms of Mean Bias (MB).



UNIVERSITÉ DE NANTES

Results in terms of power



UNIVERSITÉ DE NANTES

Plan

1. Introduction
2. Covariates to consider in GC
3. GG with positivity near-violation
4. GC with Machine Learning
5. Conclusions

UNIVERSITÉ DE NANTES

Conclusions

The main advantages of the GC :

- The GC is simple to implement.
- The GC is a powerful method, especially when considering the covariates causing the outcome.
- The GC is quite robust to the positivity near violation, except for high extrapolation issues.
- The proposed SL allows to prevent the Q-model misspecification.
- The automatic algorithm allows bootstrapping the entire estimation procedure, including the Q-model construction, in the variance estimation.

The main limitation of the GC :

- It does not constitute a doubly robust estimator.

UNIVERSITÉ DE NANTES

References

1. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets : a comparative simulation study. **Chatton** et al. Sci Rep. 2020.

2. G-computation and machine learning for estimating the causal effects of binary exposure statuses on binary outcomes. **Le Borgne** et al. Sci Rep. 2021

3. Causal inference in case of near-violation of positivity : comparison of methods. **Léger** et al. In revision.

UNIVERSITÉ DE NANTES