# 1 Supplementary Material

## 1.1 Details of Simulation Framework Cont.

This section contains the levels of each knob that are used for each of the black box and DIY data sets across the 77 settings, as detailed in Appendix A.1. The R package at https://github.com/vdorie/aciccomp/tree/master/2016 recreates the simulated data by setting parameters according to the following enumeration.

Table 1: Simulation Settings

| # | Treatment Model | Trt % | Overlap | Response Model | Trt/Rsp Alignment | hetero-geneity | DIY # |
|---|---|---|---|---|---|---|---|
| 1 | linear | low | penalize | linear | high | high | 10 |
| 2 | polynomial | low | penalize | exponential | high | none | 1 |
| 3 | linear | low | penalize | linear | high | none | 9 |
| 4 | polynomial | low | full | exponential | high | high | 4 |
| 5 | linear | low | penalize | exponential | high | high | 15 |
| 6 | polynomial | low | penalize | linear | high | high | 2 |
| 7 | polynomial | low | penalize | exponential | high | high | 5 |
| 8 | polynomial | low | penalize | exponential | none | high | 13 |
| 9 | step | low | penalize | step | high | high | 8 |
| 10 | linear | low | penalize | exponential | low | high | 14 |
| 11 | polynomial | low | penalize | linear | low | high | 19 |
| 12 | polynomial | low | penalize | exponential | low | high | 12 |
| 13 | linear | high | penalize | exponential | high | high | 18 |
| 14 | polynomial | high | penalize | linear | high | high | 20 |
| 15 | polynomial | high | penalize | exponential | high | high | 6 |
| 16 | polynomial | high | penalize | exponential | none | high | 17 |
| 17 | step | high | penalize | step | high | high | 7 |
| 18 | linear | high | penalize | exponential | low | high | 3 |
| 19 | polynomial | high | penalize | linear | low | high | 16 |
| 20 | polynomial | high | penalize | exponential | low | high | 11 |
| 21 | polynomial | low | penalize | step | low | low | |
| 22 | polynomial | low | penalize | step | low | high | |
| 23 | polynomial | low | penalize | step | high | low | |
| 24 | polynomial | low | penalize | step | high | high | |
| 25 | polynomial | low | penalize | exponential | low | low | |
| 26 | polynomial | low | penalize | exponential | high | low | |
| 27 | polynomial | low | full | step | low | low | |
| 28 | polynomial | low | full | step | low | high | |
| 29 | polynomial | low | full | step | high | low | |
| 30 | polynomial | low | full | step | high | high | |
| 31 | polynomial | low | full | exponential | low | low | |
| 32 | polynomial | low | full | exponential | low | high | |
| 33 | polynomial | low | full | exponential | high | low | |
| 34 | polynomial | high | penalize | step | low | low | |
| 35 | polynomial | high | penalize | step | low | high | |
| 36 | polynomial | high | penalize | step | high | low | |
| 37 | polynomial | high | penalize | step | high | high | |
| 38 | polynomial | high | penalize | exponential | low | low | |
| 39 | polynomial | high | penalize | exponential | high | low | |
| 40 | polynomial | high | full | step | low | low | |
| 41 | polynomial | high | full | step | low | high | |
| 42 | polynomial | high | full | step | high | low | |
| 43 | polynomial | high | full | step | high | high | |
| 44 | polynomial | high | full | exponential | low | low | |
| 45 | polynomial | high | full | exponential | low | high | |
| 46 | polynomial | high | full | exponential | high | low | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 47 | polynomial | high | full | exponential | high | high |
| 48 | step | low | penalize | step | low | low |
| 49 | step | low | penalize | step | low | high |
| 50 | step | low | penalize | step | high | low |
| 51 | step | low | penalize | exponential | low | low |
| 52 | step | low | penalize | exponential | low | high |
| 53 | step | low | penalize | exponential | high | low |
| 54 | step | low | penalize | exponential | high | high |
| 55 | step | low | full | step | low | low |
| 56 | step | low | full | step | low | high |
| 57 | step | low | full | step | high | low |
| 58 | step | low | full | step | high | high |
| 59 | step | low | full | exponential | low | low |
| 60 | step | low | full | exponential | low | high |
| 61 | step | low | full | exponential | high | low |
| 62 | step | low | full | exponential | high | high |
| 63 | step | high | penalize | step | low | low |
| 64 | step | high | penalize | step | low | high |
| 65 | step | high | penalize | step | high | low |
| 66 | step | high | penalize | exponential | low | low |
| 67 | step | high | penalize | exponential | low | high |
| 68 | step | high | penalize | exponential | high | low |
| 69 | step | high | penalize | exponential | high | high |
| 70 | step | high | full | step | low | low |
| 71 | step | high | full | step | low | high |
| 72 | step | high | full | step | high | low |
| 73 | step | high | full | step | high | high |
| 74 | step | high | full | exponential | low | low |
| 75 | step | high | full | exponential | low | high |
| 76 | step | high | full | exponential | high | low |
| 77 | step | high | full | exponential | high | high |

## 1.2 Full list of metrics used to describe experimental settings

List of metrics used to describe the experiments. We denote [oracle] those metrics which are not accessible to a researcher in an ordinary observational study, but are available to us as the creators of the competition data set. We further denote by [knob]the oracle metrics that correspond to the explicit experimental settings we created, as described in Subsection 4.3. After each non-knob metric we list in parentheses which aspect of the experiment are they meant to measure.

- [oracle][knob]Degree of outcome nonlinearity: 0, 1, 2 for linear, non-linear, step-function response surface.
- [oracle][knob]Degree of treatment assignment nonlinearity: 0, 1, 2 for linear, non-linear, step-function treatment assignment model.
- [oracle][knob]Percentage of treated: setting more treated or more control units. Note that despite being an oracle metric, this is in fact easily estimated from data, and is *included as a non-oracle measure in the discussion.*
- [oracle][knob]Overlap: binary, whether there was or was not considerable overlap between treated and control.
- [oracle][knob]Alignment: 0,1,2 for no, low, and high alignment between the confounders pertaining to treatment assignment and confounders pertaining to outcome.
- [oracle][knob]Treatment effect heterogeneity: 0, 1, 2 for no, low or high treatment effect heterogeneity.
- [oracle]Correlation between true propensity score and the outcome (alignment).

- [oracle]Mean Mahalanobis distance to nearest counterfactual neighbor in the "ground truth" design matrix (overlap).
- [oracle]Euclidean norm of distance between mean of control and mean of treated in "ground truth" design matrix (balance).
- [oracle]Wasserstein distance [Villani, 2008, Cuturi, 2013] between treated and control using the "ground truth" design matrix (balance).
- [oracle]$R^2$ of the logit of the true propensity score regressed on the observable design matrix (treatment assignment nonlinearity).
- [oracle]$R^2$ of the true treatment effect regressed on the observable design matrix.
- [oracle]$R^2$ of the outcome regressed on the "ground truth" design matrix (outcome nonlinearity).
- [oracle]The ratio of the $R^2$ of the outcome regressed on the observable design matrix divided by the $R^2$ of the outcome regressed on the "ground truth" design matrix (outcome nonlinearity).
- [oracle]$R^2$ of the true potential outcome function $Y(0)$ on the control units ($Z = 0$) in the observable design matrix (outcome nonlinearity).
- [oracle]$R^2$ of the true potential outcome function $Y(0)$ on the control units ($Z = 0$) in the "ground truth" design matrix (outcome nonlinearity).
- [oracle]$R^2$ of the true potential outcome function $Y(1)$ on the treated units ($Z = 1$) in the observable design matrix (outcome nonlinearity).
- [oracle]$R^2$ of the true potential outcome function $Y(1)$ on the treated units ($Z = 1$) in the "ground truth" design matrix (outcome nonlinearity).
- [oracle]Standard deviation of the ground truth treatment effect function $\mathrm{E}\left[Y(1) - Y(0) \mid X\right]$ (treatment effect heterogeneity).
- $R^2$ of the observed outcome $Y$ on the observed design matrix (outcome nonlinearity).
- $R^2$ of fitting a propensity score model estimated with logistic regression on the observable design matrix (treatment assignment nonlinearity).
- $R^2$ between the estimated unit level treatment effect estimated by BART ($\mathrm{E}\left[\hat{Y}(1) - \hat{Y}(0) \mid X\right]$), and propensity scores estimated with logistic regression on the observable design matrix (alignment).
- Mean Mahalanobis distance to nearest counterfactual neighbor in the observable design matrix (overlap).
- Euclidean norm of distance between mean of control and mean of treated in the observable design matrix (balance).
- Wasserstein distance [Villani, 2008, Cuturi, 2013] between treated and control using the observable design matrix (balance).

## 2 Submissions and Acknowledgements

We would like to thank the following people for taking the time to submit. Affiliations are those of the first author at the time of submission and methods are in no particular order. Some methods were submitted as being representative in their fields and may not reflect their submitter's beliefs for best practices.

## Do-It-Yourself Methods

| **Method** | Author | Institution |
|---|---|---|
| IPTW estimator | Chanmin Kim | Deptartment of Biostatistics, Harvard University |
| MITSS | Liangyuan Hu and Chenyang Gu | Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai |
| Bayes LM | Christoph Kurz | German Research Center for Environmental Health, Helmholtz Zentrum Mnchen |
| Regression Trees | Dave Harris | Department of Wildlife Ecology and Conservation, University of Florida |
| Calibrated IPW | Gi-Soo Kim | Department of Statistics, Seoul National University |
| DR w/GBM + MDIA and Ad Hoc | John Lockwood | Educational Testing Service |
| Weighted GP | Junfeng Wen and Russ Grenier | Department of Computing Science, University of Alberta |
| LAS Gen GAM | Leonid Liu and Annie Wang | Analyst Institute |
| GLM-Boost | Manuel Huber | German Research Center for Environmental Health, Helmholtz Zentrum Mnchen |
| Manual | Mao Hu | Acumen, LLC |
| RBD TwoStepLM | Peng Ding | Department of Statistics, University of California Berkeley |
| ProxMatch | Hui Fen Tan, David Miller, and James Savage | Department of Statistics, Cornell University |
| VarSel NN | Zhipeng Hou and Bryan Keller | Teacher's College, Columbia University |

### Black Box Methods

| **Method** | Author | Institution |
|---|---|---|
| MHE Algorithm | Peter Aronow | Department of Political Science, Yale University |
| BART | Douglas Galagate and separately Nicole Bohme Carnegie | Department of Math, University of Maryland and Zilber School of Public Health, University of Wisconsin-Milwaukee |
| teffects methods | Seth Lirette | Center of Biostatistics and Bioinformatics, University of Mississippi Medical Center |
| LASSO+CBPS | James Pustejovsky | Department of Computer Science, Brandeis University |
| calCause | Chen Yanover, Omer Weissbrod, Michal Ozery-Flato, Tal El-Hay, Assaf Gottlieb and Yishai Shimoni | IBM Research - Haifa |
| BalanceBoost | Qingyuan Zhao | Department of Statistics, Stanford University |
| Tree Strat and Adj. Tree Strat | Stefan Wager | Department of Statistics, Stanford University |
| h2o Ensemble | Hyunseung Kang | Graduate School of Business, Stanford University |
| CBPS | Yongnam Kim | Department of Educational Psychology, University of Wisconsin Madison |
| SL+TMLE | Susan Gruber and Mark van der Laan | T.H. Chan School of Public Health, Harvard University |

## References


Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.