Universal Adaptability

A New Method to Draw Inference from Non-Probability Surveys and Other Data Sources

Christoph Kern
Department of Statistics, LMU Munich
christoph-kern@stat.uni-muenchen.de

Frauke Kreuter
Department of Statistics, LMU Munich
JPSM, University of Maryland

with Michael Kim, Shafi Goldwasser, Omer Reingold



Proceedings of the National Academy of Sciences 119(4)
doi.org/10.1073/pnas.2108097119

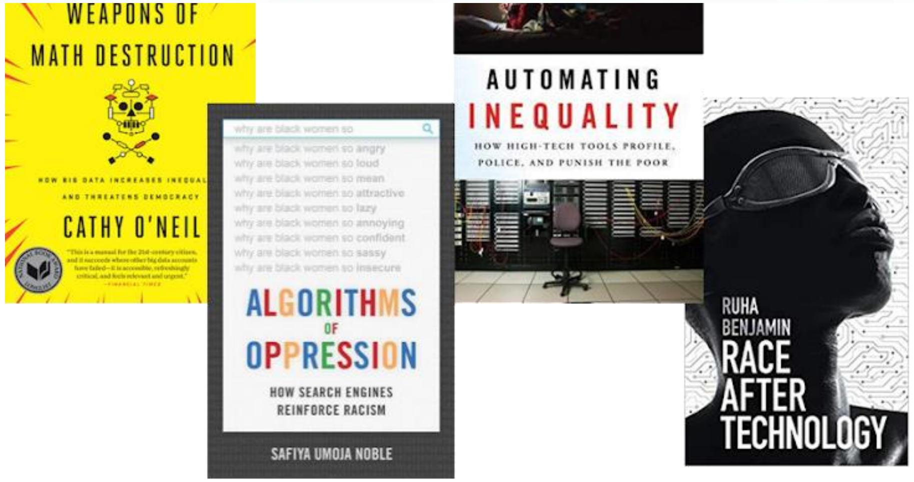
Overview

1. Algorithmic Fairness & Multicalibration
2. Inference Challenge:
    - Single source, many targets
    - ***Universal Adaptability***
3. MCBoost algorithm and applications
4. Expansion of MCBoost to CATE estimation

Algorithmic Fairness

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Buolamwini 2019



ts can

PRINT E-MAIL

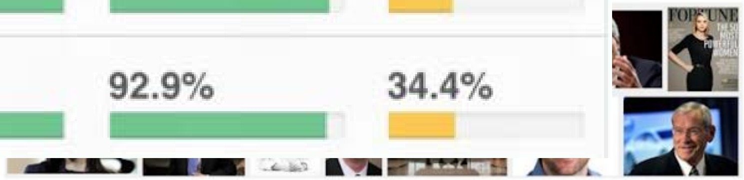


IMAGE: PERCENTAGE OF WOMEN IN TOP 100 GOOGLE IMAGE SEARCH RESULTS FOR CEO IS: 11 PERCENT. PERCENTAGE OF US CEOS WHO ARE WOMEN IS: 27 PERCENT. view more ›

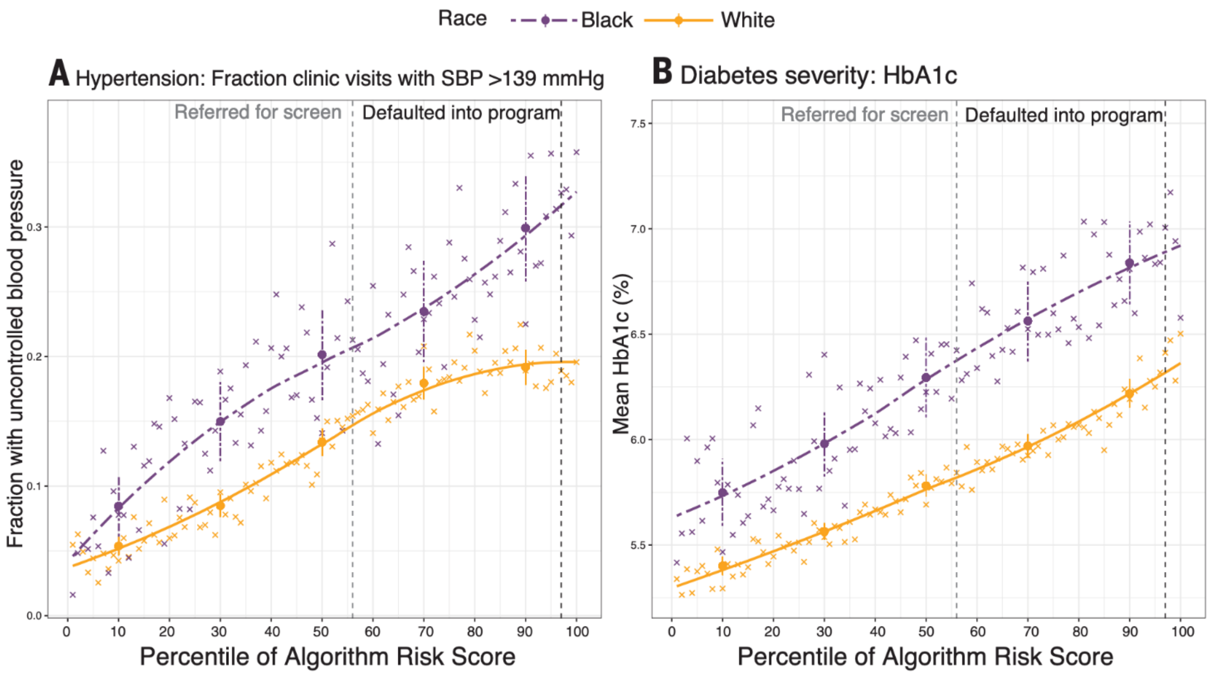
Amazon reportedly scraps internal AI recruiting tool that was biased against women

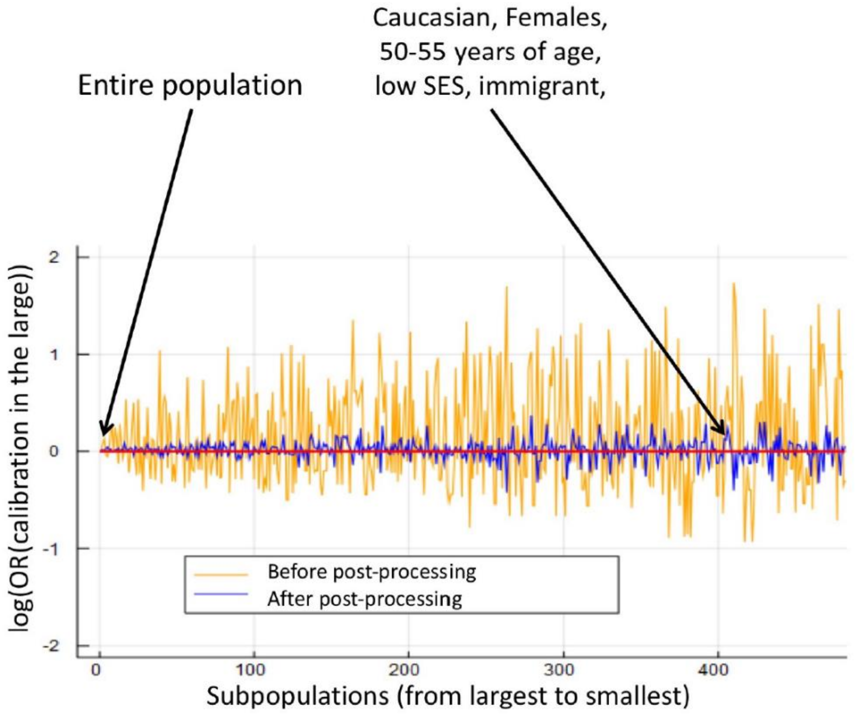
The secret program penalized applications that contained the word "women's"

By James Vincent on October 10, 2018 7:09 am

Miscalibration leads to unfair decisions

- Predictions mean different things in different groups





[Obermeyer, Powers, Vogeli, Mullainathan ’19]

[Barda et al. ’21]

Multicalibration

- Calibration for every "computationally-identifiable" group

Definition: For a class of functions C, a predictor $\tilde{p}$ is (C, α)-***multicalibrated***, if for every $c \in C$

$$\left| E\left[c(X) \cdot \left(Y - \tilde{p}(X) \right) \right] \right| \leq \alpha$$

[Hébert-Johnson, Kim, Reingold, Rothblum '18]

- Think of C as:
  - A collection of demographic subpopulations
  - A learnable hypothesis class (e.g., decision trees, linear functions, etc.)

Protecting subpopulations

- Multicalibration in prediction settings
  - Prediction/ imputation of citizenship, wage, record linkage…
    [Beck, Dumpert, Feuerhake ’18]

Guarantees for multiple subgroups, defined by complex intersections!

- Multicalibration in **estimation settings**
  - Estimation of mortality rates, voting or economic outcomes…

Guarantees for multiple target populations?

Inference Challenge

Goal: Given access to

- ***labeld*** source data $\{(X_i, Y_i)\} \sim s$ (with outcome)
- ***unlabeled*** target data $\{(X_i, ?\,)\} \sim t$

estimate average outcome Y in target.

Challenge: source/target populations differ in composition

→ Reweight source population to “look like” target population

Target-Specific Inference

- Fit propensity score $\sigma \in \Sigma$ to minimize estimation error

Propensity Score Reweighting:
Given a score $\sigma: \mathcal{X} \to [0,1]$, estimate $E[Y|Z=t]$ as

$$PS_{st}(\sigma) = E\left[\left(\frac{1-\sigma(X)}{\sigma(X)}\right) \cdot Y \middle| Z = s\right]$$

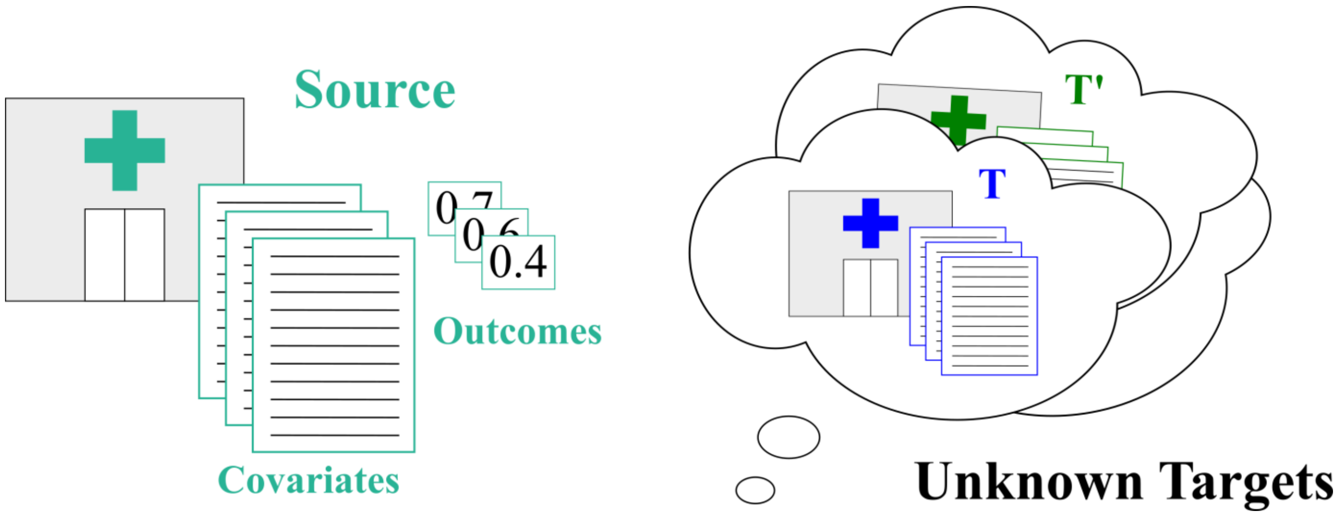
For a class of propensity scores Σ, we measure the estimation error as:

$$\text{error}(PS_{st}(\Sigma)) = \min_{\sigma \in \Sigma} |PS_{st}(e_{st}) - PS_{st}(\sigma)|$$

Multi-Target Challenge

Single source → many different targets!

- s: *large medical study run by Alpert Medical School*
- t: *different hospital populations across the country*



Multi-Target Challenge

Single source $\rightarrow$ many different targets!

- s: *large medical study run by Alpert Medical School*
- t: *different hospital populations across the country*

Challenge: Reweighting for every target is costly

Insight from study requires target-specific propensity score

Burden lies with target communities to reweight

Goal: Provide insights in a "universal" format

Reorient responsibility to reweight at the source

Universal Adaptability

- Set requirements for predictor trained on source to give well performing estimates on targets

Definition: For a fixed source s, and a class of propensity scores Σ, a predictor $\tilde{p}$ is (Σ, β)-***universally adaptable***, if for ***any*** target t,

$$\text{error}(\hat{\mu}_t(\tilde{p})) \leq \text{error}(PS_{st}(\Sigma)) + \beta$$

Multicalibration Guarantees Universal Adaptability

- Given a class of propensity scoring functions Σ and a class of propensity odds ratios $C(\Sigma)$

Theorem: If $\tilde{p}$ is $(C(\Sigma), \alpha)$-multicalibrated over source s, then $\tilde{p}$ is (Σ, β)-universally adaptable for $\beta \leq \alpha + \delta_{st}(\Sigma)$.

where $\delta_{st}(\Sigma)$ captures how well Σ fits the true propensity score

MCBoost: Post-Processing for Multicalibration

R package – https://github.com/mlr-org/mcboost

Given:

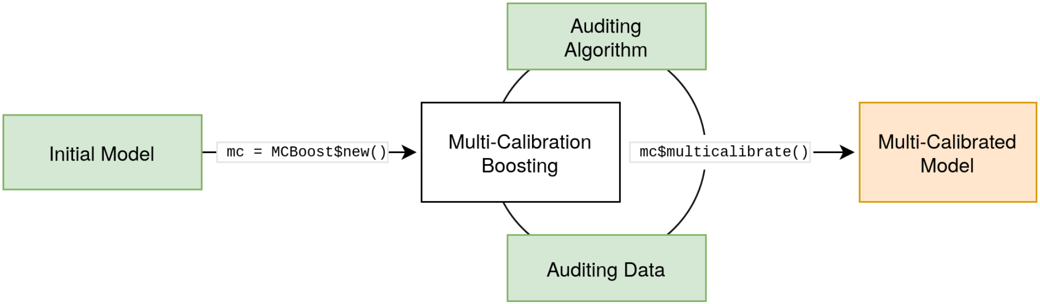
- Initial predictor $\tilde{p}$
- Validation data D
- An auditor to search for subpopulations c
  - Find largest residuals
  - e.g. ridge regression, decision tree (auditor defines collection C)

Repeat:

- Search over $c \in C$
- If $\left|E_{x \sim D}\left[c(x) \cdot \left(y - \tilde{p}(x)\right)\right]\right| > \alpha$
  - update as $\tilde{p}(x) \leftarrow \tilde{p}(x) - \eta \cdot c(x)$

Multi-Calibration Boosting for R (Pfisterer et al., 2021)

R package mcboost – `https://github.com/mlr-org/mcboost`



Mitigating Bias Across Subpopulations

Analogy between two goals

Fairness goal: protect subpopulations from miscalibrated predictions

Statistical goal: ensure unbiased estimates on downstream targets

Mitigating Bias Across Subpopulations

Analogy between two goals

Fairness goal: protect subpopulations from miscalibrated predictions

Statistical goal: ensure unbiased estimates on downstream targets

The role of post-processing for multicalibration

Identifies ***qualified minority*** subpopulations

Identifies ***potential shifts*** in covariate distribution

Empirical Evaluation

- Setting
  - Source: US National Health and Nutrition Examination Survey
  - Target: US National Health Interview Survey (weighted)
  - Estimate 15-year mortality rate across demographic groups
- Inference Methods
  - **IPSW-Overall**: Reweighting with global propensity scores (PS)
  - **IPSW-Subgroup**: Reweighting with subgroup-specific PS
  - **RF-Naive**: Mortality prediction with random forest
  - **RF-MCBoost**: Mortality prediction with multicalibrated RF

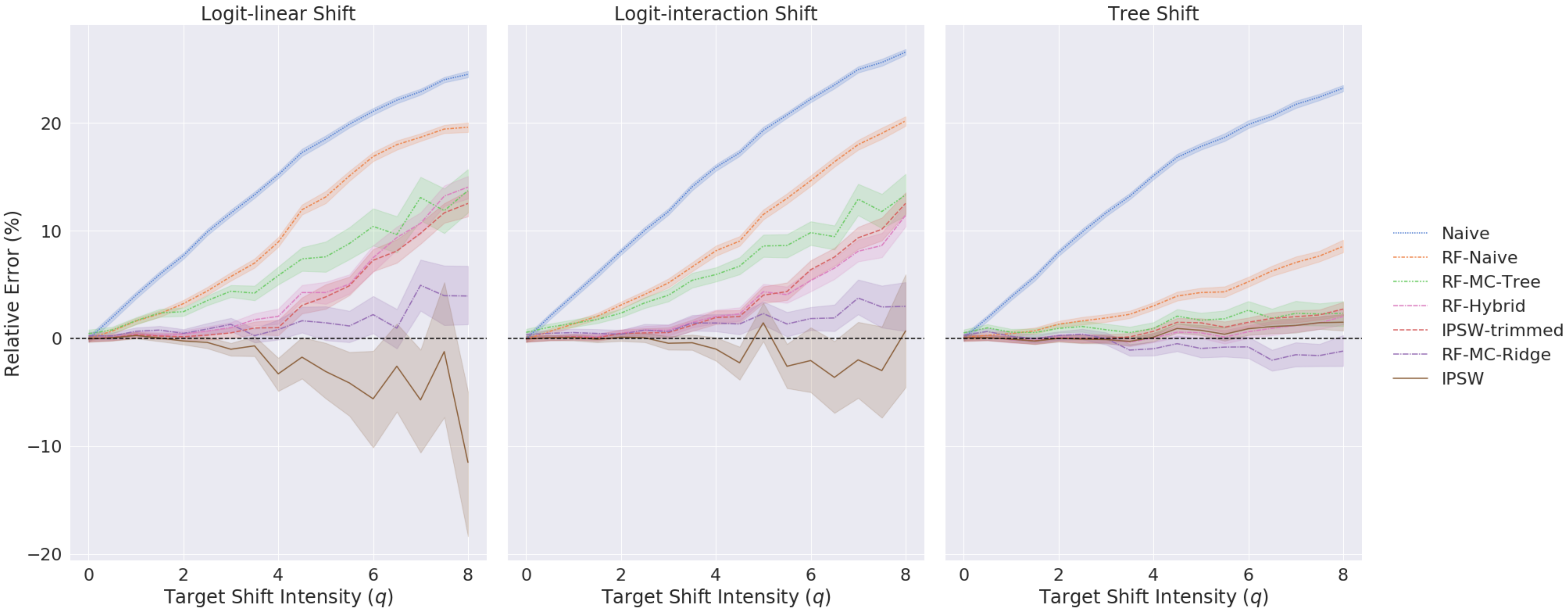
Empirical Evaluation – Results

	IPSW		RF	
	Overall	Subgroup	Naive	MC-Boost
Overall	2.37 (13.5%)	—	1.11 (6.3%)	**0.52 (3.0%)**
Male	2.51 (13.4)	0.91 (4.9)	-0.34 (1.8)	**0.11 (0.6)**
Female	2.40 (14.6)	3.99 (24.2)	2.43 (14.8)	**0.90 (5.4)**
Age 18-24	**0.00 (0.1)**	-0.39 (17.5)	6.03 (270.2)	1.76 (79.0)
Age 25-44	**-0.20 (5.2)**	-0.41 (10.6)	0.82 (21.2)	0.66 (17.2)
Age 45-64	-0.75 (4.2)	-0.41 (2.3)	0.86 (4.8)	-0.29 (1.6)
Age 65-69	-4.23 (9.3)	-5.23 (11.5)	**-3.52 (7.7)**	**-1.99 (4.4)**
Age 70-74	-1.36 (2.3)	**0.47 (0.8)**	-3.02 (5.0)	**0.61 (1.0)**
Age 75+	3.53 (4.1)	2.85 (3.3)	0.51 (0.6)	2.19 (2.5)
White	3.53 (18.9)	0.75 (4.0)	1.03 (5.5)	0.69 (3.7)
Black	-4.00 (21.1)	**-0.48 (2.5)**	**-0.66 (3.5)**	**-0.52 (2.7)**
Hispanic	1.73 (17.0)	**0.48 (4.7)**	2.91 (28.6)	1.55 (15.2)
Other	**-0.02 (0.2)**	-3.54 (39.5)	3.52 (39.3)	-2.06 (23.0)

Semi-synthetic Simulation

- Setting
  - A “non-probability” sample, D_{np}, based on 31,319 online opt-in panel interviews
  - A “reference population”, D_p, with 20,000 observations that combines information from high quality surveys
  - Estimate voting rates for the 2014 midterm election across *different degrees of covariate shift*
    1. We estimate the propensity score between D_{np} and D_p using different techniques (**Logit-linear**, **Logit-interaction**, **Tree**)
    2. For each propensity model, we generate synthetic data of various shift intensity (q) by sampling from D_{np} with weights

Semi-synthetic Simulation – Results



Summary and Takeaways

Multicalibration

Algorithmic fairness useful beyond "fairness"

Universal Adaptability

Valid inferences across a rich class of targets

General Result

Multicalibration persists under covariate shift

Can we robustify conditional average treatment effect (CATE) estimation via multi-calibration?

CATE Estimation

Setup

- Covariates X
- Treatment $T \in \{0, 1\}$
- (Potential) outcomes $Y(T)$

Estimand of interest

- *Conditional average treatment effect* (CATE)

$$\tau(X) = \mathrm{E}\left[Y(1) - Y(0) \mid X\right]$$

Assumptions

- Unconfoundedness
- Consistency, SUTVA, overlap

CATE Estimation

CATE learner

- The *T-learner* differences treatment-conditional outcome regressions

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

$$\mu_t(x) = \mathrm{E}[Y \mid X = x, T = t]$$

- X-learner (Künzel et al., 2019), R-learner (Nie and Wager, 2020), causal forests (Wager and Athey, 2018)

Performance assessment

- MSE of the CATE

$$\mathrm{E}[(\hat{\tau}(X) - \tau(X))^2]$$

- Bias under a different distribution $X \sim Q$

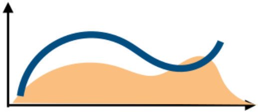
$$\mathrm{E}_Q[(\hat{\tau}(X) - \tau(X))]$$

Setting 1: <u>External shift</u>

Unconfounded data

$$\mu_1(X) - \mu_0(X)$$

`MCBoost Auditing`



CATE

$\mathbb{E}[Y(1) - Y(0) \mid X]$

P(X) covariate density

Unknown P'(X)
Deployment distributions
Unknown at test time

Meta-Algorithm

Algorithm 1 Multi-accuracy for CATE estimation for unknown covariate shifts

1: Input: (X, T, Y) unconfounded data, $\mathcal{F}$ auditor function class, $\mathcal{G}$ function class for outcome functions

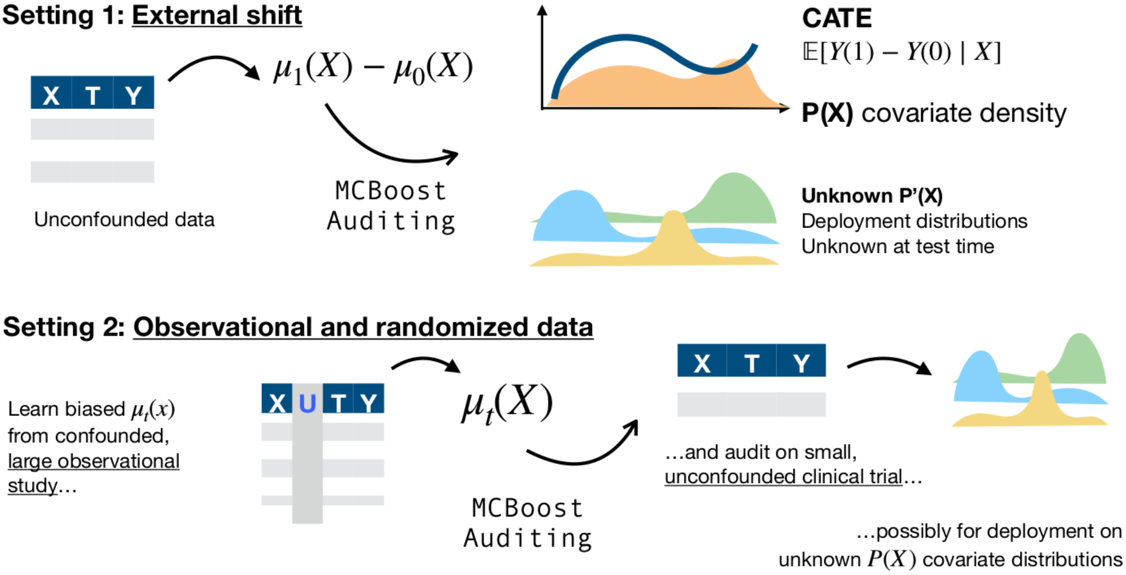
2: Fit treatment-conditional outcome functions from the observational dataset:

$$\hat{\mu}_t(x) \leftarrow \arg\min_{g \in \mathcal{G}} \mathrm{E}[(g - Y)^2 \mid T = t], \text{ for } t \in \{0, 1\}$$

3: Post-process $\hat{\mu}_t(X)$ for $t \in \{0, 1\}$ by multi-accuracy: Find $\tilde{\mu}_t(x)$, for $t \in \{0, 1\}$ s.t.

$$\max_{f \in \mathcal{F}} |\mathrm{E}[f(X) \cdot (Y - \tilde{\mu}(X)) \mid T = t]| \leq \alpha.$$

4: Return $\tilde{\tau}(x) = \tilde{\mu}_1(x) - \tilde{\mu}_0(x)$


Setting 1: External shift
X T Y
$\mu_1(X) - \mu_0(X)$
MCBoost Auditing
Unconfounded data
CATE
$\mathbb{E}[Y(1) - Y(0) \mid X]$
P(X) covariate density
Unknown P'(X)
Deployment distributions
Unknown at test time
Setting 2: Observational and randomized data
Learn biased $\mu_t(x)$ from confounded, large observational study…
X U T Y
$\mu_t(X)$
MCBoost Auditing
X T Y
…and audit on small, unconfounded clinical trial…
…possibly for deployment on unknown $P(X)$ covariate distributions

Meta-Algorithm

Algorithm 2 Multi-accuracy for CATE estimation for calibrating CATE on small Randomized Controlled Trial data

1: Input: $\mathcal{D}_{\text{obs}} = (X, T, Y)$ confounded observational data, $\mathcal{D}_{\text{rct}} = (X, T, Y)$ unconfounded randomized data, $\mathcal{F}$ auditor function class, $\mathcal{G}$ function class for outcome functions

2: Fit treatment-conditional outcome functions from the observational dataset:

$$\hat{\mu}_t(x) \leftarrow \arg\min_{g \in \mathcal{G}} \mathrm{E}_{\text{obs}}[(g - Y)^2 \mid T = t], \text{ for } t \in \{0, 1\}$$

3: Apply $\mathrm{MCBoost}$ to $\hat{\mu}_t(x), t \in \{0, 1\}$ using $\mathcal{D}_{\text{rct}}$ as validation set

4: Return $\tilde{\tau}(x) = \tilde{\mu}_1(x) - \tilde{\mu}_0(x)$

Multi-Accurate CATE Estimates

Characteristics of multi-accurate CATE T-learner

1. "Do-no-harm" property w.r.t. MSE
2. Bias guarantees under unknown shifts

Proposition

Let $\mathcal{F} = \mathcal{C} \times \mathcal{H}$ where $\mathcal{C}$ indexes subgroups and $\mathcal{H}$ is a collection of test functions. Then multi-accuracy of the T-learner CATE estimate $\tilde{\tau}(X)$ implies that, for all distributions Q such that the likelihood ratios $\frac{dQ_0}{dP_0}, \frac{dQ_1}{dP_1} \in \mathcal{H}$,

$$\mathrm{E}_Q[\tilde{\tau}(X)c(X)] - (\mathrm{E}_Q[Yc(X) \mid T = 1] - \mathrm{E}_Q[Yc(X) \mid T = 0]) \leq 2\alpha, \forall c \in \mathcal{C}$$

Simulation Setup

1. Simulate data (X, T, Y)
   - Given propensity score and outcome functions with different degrees of complexity
2. Sample with weights to introduce distribution shift
   - Based on external shift function and different shift intensities

Setting 1

- External shift, only observational data
- No unobserved confounding

$$(X_{train}, T_{train}, Y_{train}) \sim \mathcal{D}_{os}$$
$$(X_{audit}, T_{audit}, Y_{audit}) \sim \mathcal{D}_{os}$$
$$(X_{test}, T_{test}, Y_{test}) \sim \mathcal{D}_{os-shift}$$

Setting 2

- Observational data, small (shifted) RCT
- Unobserved confounding in obs. data

$$(X_{train}, T_{train}, Y_{train}) \sim \mathcal{D}_{os}$$
$$(X_{audit}, T_{audit}, Y_{audit}) \sim \mathcal{D}_{rct}$$
$$(X_{test}, T_{test}, Y_{test}) \sim \mathcal{D}_{os}$$

Simulation Setup

Setting 1

- **CForest-OS**
  - Causal forest trained in the observational training data
- **T-learner-OS**
  - T-learner using random forest trained in the observational training data
- **T-learner-MC-Ridge**
  - T-learner using random forest in the observational training data is post-processed with MCBoost using ridge regression in the auditing data
- CForest-wOS
- T-learner-wOS

Setting 2

- **CForest-OS**
  - Causal forest trained in the observational training data
- **T-learner-OS**
  - T-learner using random forest trained in the observational training data
- **T-learner-MC-Tree**
  - T-learner using random forest in the observational training data is post-processed with MCBoost using decision trees in the RCT
- CForest-RCT, CForest-wRCT
- T-learner-RCT, T-learner-wRCT

Simulation Results – Setting 1

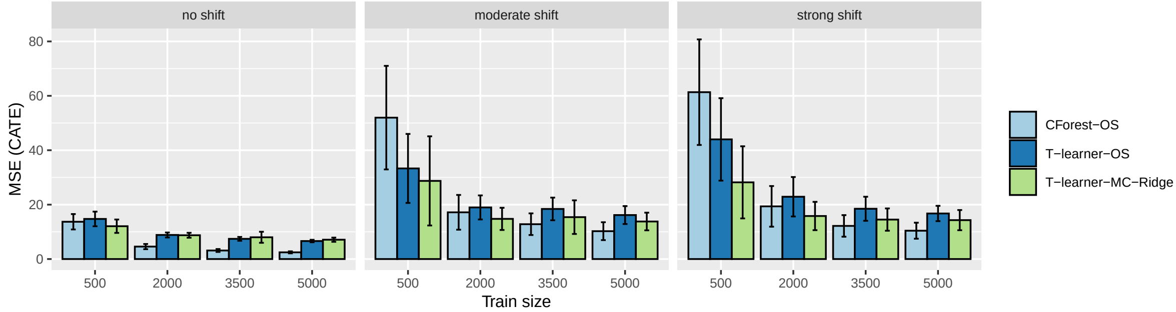


Figure: Average MSE of CATE estimation by shift intensity and training set size for post-processed (multi-calibrated) T-learners and benchmark methods in simulation studies (external shift)

Simulation Results – Setting 2

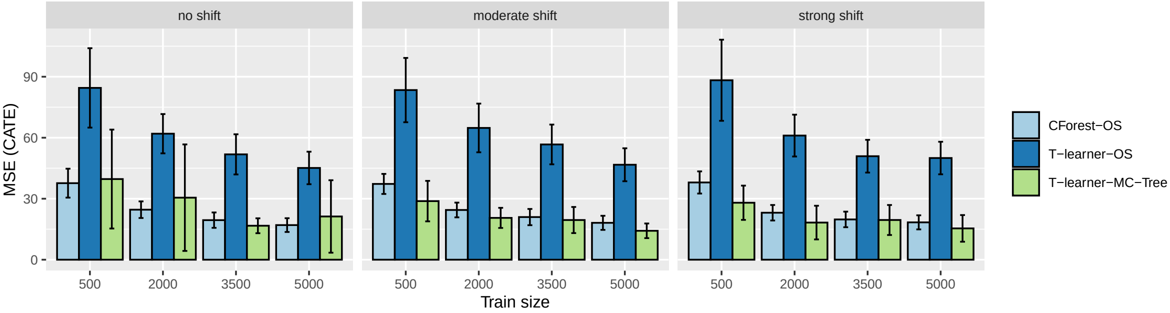


Figure: Average MSE of CATE estimation by shift intensity and training set size for post-processed (multi-calibrated) T-learners and benchmark methods in simulation studies (observational data with RCT)

Discussion

Approach

- Robustify CATE T-learners to unknown shifts via MCBoost post-processing
- Utilize multi-accuracy to jointly learn from observational data and RCT

Results

- General improvements in bias and MSE in simulations
- Multi-CATE is robust, but not efficient

Extensive related work

- Our focus: Show utility of “off-the-shelf” application of multi-accuracy in CATE estimation domain

References

Barda, N., Yona, G., Rothblum, G.N., Greenland, P., Leibowitz, M., Balicer, R., Bachmat, E., Dagan. N. (2021). Addressing bias in prediction models by improving subpopulation calibration. *Journal of the American Medical Informatics Association 28*(3), 549-558.

Beck, M., Dumpert, F., Feuerhake, J. (2018). *Machine Learning in Official Statistics*. https://arxiv.org/abs/1812.10422.

Buolamwini, T., Gebru T. (2018). Gender shades: Intersectional accuracy dispartities in commercial gender classification. FACCT Conference

Hebert-Johnson, U., Kim, M., Reingold, O. , Rothblum, G. (2018). Multicalibration: Calibration for the (Computationally-Identifiable) Masses. *Proceedings of the 35th International Conference on Machine Learning 80*, 1939-1948.

Kim, M.P., Kern, C., Goldwasser, S., Kreuter, F., Reingold, O. (2022). Universal Adaptability: Target-Independent Inference that Competes with Propensity Scoring. *Proceedings of the National Academy of Sciences of the United States of America (PNAS), 119*(4).

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*(6464), 447-453.

References

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.

Nie, X. and Wager, S. (2020). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.

Pfisterer, F., Kern, C., Dandl, S., Sun, M., Kim, M. P., and Bischl, B. (2021). mcboost: Multi-calibration boosting for r. *Journal of Open Source Software*, 6(64):3453.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment e ects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.