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

<table>
<thead>
<tr>
<th>Gender Classifier</th>
<th>Darker Male</th>
<th>Darker Female</th>
<th>Lighter Male</th>
<th>Lighter Female</th>
<th>Largest Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>94.0%</td>
<td>79.2%</td>
<td>100%</td>
<td>98.3%</td>
<td>20.8%</td>
</tr>
<tr>
<td>FACE**</td>
<td>99.3%</td>
<td>65.5%</td>
<td>99.2%</td>
<td>94.0%</td>
<td>33.8%</td>
</tr>
<tr>
<td>IBM</td>
<td>88.0%</td>
<td>65.3%</td>
<td>99.7%</td>
<td>92.9%</td>
<td>34.4%</td>
</tr>
</tbody>
</table>

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 $\hat{p}$ is $(C, \alpha)$-**multicalibrated**, if for every $c \in C$

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

[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

- *labeled* 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

\[\rightarrow\] 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 $\sigma: X \rightarrow [0,1]$, estimate $E[Y|Z = t]$ as

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

For a class of propensity scores $\Sigma$, 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 $\rightarrow$ many different targets!

- $s$: large medical study run by Alpert Medical School
- $t$: different hospital populations across the country
Multi-Target Challenge

Single source → 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 $\Sigma$, a predictor $\tilde{\rho}$ is $(\Sigma, \beta)$-**universally adaptable**, if for any target $t$,

$$\text{error}(\hat{\mu}_t(\tilde{\rho})) \leq \text{error}(PS_{st}(\Sigma)) + \beta$$
Multicalibration Guarantees
Universal Adaptability

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

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

where $\delta_{st}(\Sigma)$ captures how well $\Sigma$ fits the true propensity score
MCBoost: Post-Processing for Multicalibration

Given:

- Initial predictor $\hat{p}$
- 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 $|E_{x \sim D}[c(x) \cdot (y - \hat{p}(x))]| > \alpha$
  - update as $\hat{p}(x) \leftarrow \hat{p}(x) - \eta \cdot c(x)$
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

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

Logit-linear Shift

Logit-interaction Shift

Tree Shift

Relative Error (%) vs Target Shift Intensity (q)

Lines represent different methods:
- Naive
- RF-Naive
- RF-MC-Tree
- RF-Hybrid
- IPSW-trimmed
- RF-MC-Ridge
- IPSW
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)*

$$
\tau(X) = E [Y(1) - Y(0) \mid X]
$$

Assumptions
- Unconfoundedness
- Consistency, SUTVA, overlap
CATE Estimation

CATE learner

- The \textit{T-learner} differences treatment-conditional outcome regressions

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

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

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

Performance assessment

- MSE of the CATE

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

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

\[
\mathbb{E}_Q[(\hat{\tau}(X) - \tau(X))]
\]
**Setting 1: External shift**

\[ \mu_1(X) - \mu_0(X) \]

Unconfounded data

MCBoost Auditing

**CATE**

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

\[ P(X) \text{ covariate density} \]

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:

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

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.

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

4: Return \(\tilde{\tau}(x) = \tilde{\mu}_1(x) - \tilde{\mu}_0(x)\)
Setting 1: External shift

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

Unconfounded data

MCBoost Auditing

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...

$\mu_t(X)$

MCBoost Auditing

...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: $D_{\text{obs}} = (X, T, Y)$ confounded observational data, $D_{\text{rct}} = (X, T, Y)$ 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:

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

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

4: Return $\tilde{\tau}(x) = \tilde{\mu}_1(x) - \tilde{\mu}_0(x)$
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 \( \frac{dQ_0}{dP_0}, \frac{dQ_1}{dP_1} \in \mathcal{H} \),

\[
E_Q[\tilde{\tau}(X)c(X)] - (E_Q[Yc(X) | T = 1] - E_Q[Yc(X) | T = 0]) \leq 2\alpha, \forall c \in \mathcal{C}
\]
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

\[
(X_{\text{train}}, T_{\text{train}}, Y_{\text{train}}) \sim D_{\text{os}}
\]
\[
(X_{\text{audit}}, T_{\text{audit}}, Y_{\text{audit}}) \sim D_{\text{os}}
\]
\[
(X_{\text{test}}, T_{\text{test}}, Y_{\text{test}}) \sim D_{\text{os}} - \text{shift}
\]

Setting 2
- Observational data, small (shifted) RCT
- Unobserved confounding in obs. data

\[
(X_{\text{train}}, T_{\text{train}}, Y_{\text{train}}) \sim D_{\text{os}}
\]
\[
(X_{\text{audit}}, T_{\text{audit}}, Y_{\text{audit}}) \sim D_{\text{rct}}
\]
\[
(X_{\text{test}}, T_{\text{test}}, Y_{\text{test}}) \sim D_{\text{os}}
\]
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


