OTClean: Data Cleaning for Conditional Independence Violations using Optimal Transport

Alireza Pirhadi Mohammad Hossein Moslemi Mostafa Milani Babak Salimi Alexander Cloninger

Overview and Motivation

- Repairing data with respect to Conditional Independence (CI) violations
- **Conditional Independence**

$$X \perp\!\!\!\perp Y \mid Z$$

- CIs are closely related to database dependencies such as **MVDs** [Wong et al. IEEE TSMC'00]
- Ensuring CI helps in developing ML models that are robust and unbiased

- **Our approach:** OTClean uses Optimal Transport (OT) to clean data by enforcing CI constraints

Two applications

1. **Data Cleaning Application:**
   - Erroneous values, biases, and inadequate preprocessing can lead to violations of CIs
   - These violations result in biased and inaccurate ML models
2. **Fairness Application:**
   - Medical expenses must be independent of demographic information (race and gender) given health records

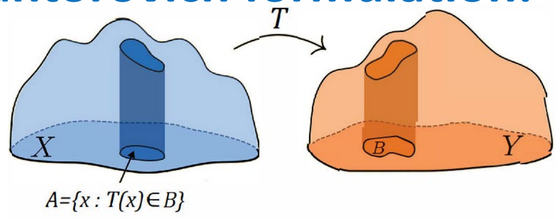
Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2,*], Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5,*,†]
+ See all authors and affiliations

Science 25 Oct 2019:
Vol. 366, Issue 6464, pp. 447-453
DOI: 10.1126/science.aax2342

Science

- **Optimal Transport and Kantorovich formulation:**



Cost function

$$OT(P,Q) = \underset{\pi \in \Pi(P,Q)}{\operatorname{argmin}} \sum_{x_i \in \mathcal{X}, y_i \in \mathcal{Y}} c(x_i, y_j)\pi(x_i, y_j)$$

Transport Plan

Problem Formulation

- **CI data cleaner:**

$D \not\models \sigma$ → CI data cleaner T^* → $\hat{D} = T^*(D) \models \sigma$

- Probabilistic Optimal Data Repair:

$$\pi^* = \underset{\pi}{\operatorname{argmin}} \sum_{v_i, v'_j \in \mathcal{V}} c(v_i, v'_j)\pi(v_i, v'_j)$$

$$s.t.\ \pi(v) = P^D, \pi(v') \models \sigma$$

Solution: Fast Approximation of OT

- Relaxed OT with Entropic Regularize

The entropic regularization parameter

$$\underset{\pi}{\operatorname{argmin}} \sum_{x_i \in \mathcal{X}, y_i \in \mathcal{Y}} c(x_i, y_j)\pi(x_i, y_j) \; -\frac{1}{\rho}H(\pi) + \lambda\left(D_{KL}(\pi(v'), Q) + D_{KL}(\pi(v), P^D)\right) + \mu\delta_\sigma(Q)$$

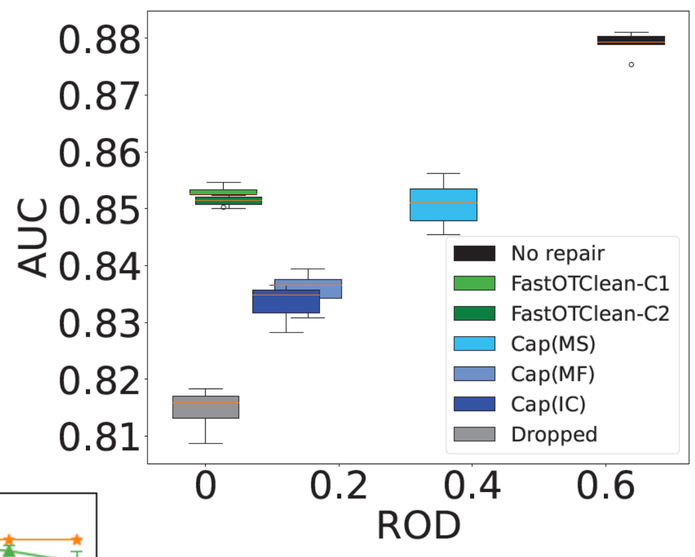
The CI constraint term

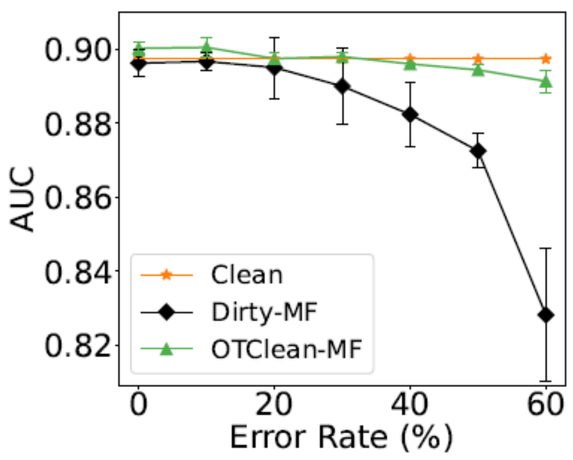
The relaxation regularization coefficient

Experiments

- Algorithmic fairness

CI: sex ⊥ income | occ., educ., hours per week





- Data cleaning

CI: door ⊥ cond. | safety, maint. cost, buying price

References:

- Cuturi, M. (2013). "Sinkhorn distances: Lightspeed computation of optimal transport." NIPS
- Salimi, B., et al. (2019). "Interventional Fairness: Causal Database Repair for Algorithmic Fairness." SIGMOD.
- Wong, S. K. M., et al. (2000). "On the implication problem for probabilistic conditional independency." IEEE Transactions on Systems, Man, and Cybernetics.