OTClean: Data Cleaning for Conditional Independence Violations using Optimal Transport

Alireza Pirhadi
Mohammad Hossein Moslemi
Mostafa Milani

Western University · Canada

Babak Salimi
Alexander Cloninger

UC San Diego

SIGMOD 2024, Santiago, Chile

Content

- **Motivation:** Conditional Independence (CI) Constraints
  - Algorithmic Fairness and Data Cleaning
- **Problem Formulation**
  - Data Repair w.r.t. CI Constraints using Optimal Transport (OT)
- **Solutions**
  - Quadratic Constrained Linear Program (QCLP) Formulation
  - FastOTClean and Optimization
- **Experimental Results**

CI and Algorithmic Fairness

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

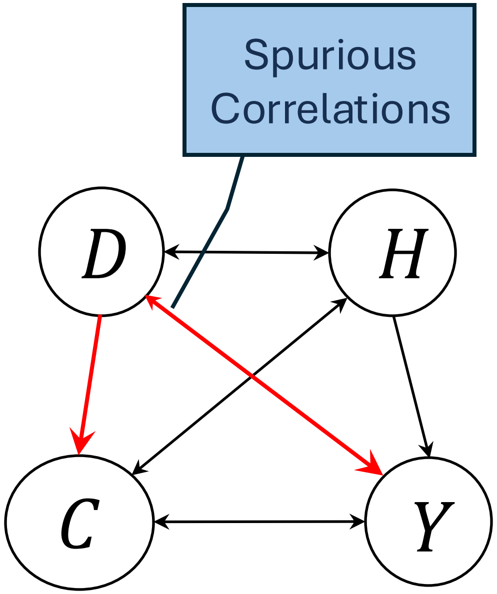
Science

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and **affecting millions of patients**, exhibits significant **racial bias**: **At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses** … **despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise** …

CI and Algorithmic Fairness

- **Demographic Info (D):** Contains Race, Gender, and...
- **Health Metrics (H):** Encompasses Medical history and...
- **Cost Data (C):** Consists of Financial records related to...
- **Label (Y):** Healthcare Needs

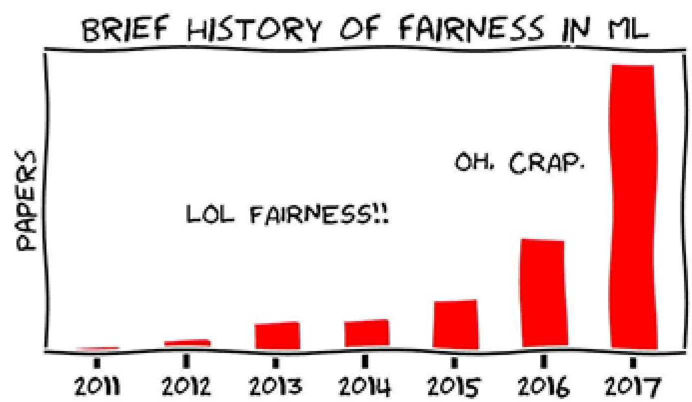
$$(D \perp C | H)$$



CI and Algorithmic Fairness (Another Example)

Priors (prior arrests or convictions), Age, Race, and Recidivism (label)

- **Statistical Parity:**
  Predictions $\perp$ Race
- **Equality of Odds:**
  Predictions $\perp$ Race | Recidivism
- **Conditional Statistical Parity:**
  Predictions $\perp$ Race | Prior



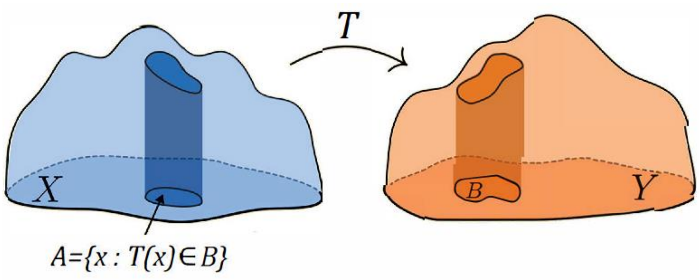
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

CI and Data Cleaning

- Sporous correlations in training data due to
  - Erroneous values and dirty data
  - Biases (sampling, measurement or other biases)
- Downstream ML models underperform since these sporous correlations do not exist at usage time

Optimal Transport (Monge Formulation)

- The most efficient way of transferring mass from a probability distribution P to another distribution Q



- Monge formulation:

Cost function

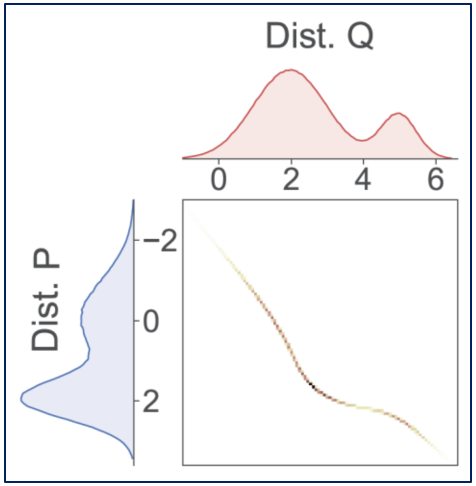
Transport Map

$$OT_{Monge}(P,Q) = \underset{T:\mathcal{X}\rightarrow\mathcal{Y}}{\operatorname{argmin}} \sum_{\mathrm{x_i}\in\mathcal{X}} \mathrm{c}(\mathrm{x_i}, \mathrm{T}(\mathrm{x_i}))$$

Figure from [Kolouri et al., SPM’17]

Optimal Transport (Kontorovich)

- Kantorovich formulation of OT uses a probabilistic map or “plan”



- Kantorovich formulation:

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

Transport Plan

Problem Formulation

- A CI Constraint σ: $X \perp\!\!\!\perp Y \mid Z$

$$P_{X,Y|Z}(\bar{x}, \bar{y}, \bar{z}) = P_{X|Z}(\bar{x}, \bar{z}) \times P_{Y|Z}(\bar{y}, \bar{z})$$

- CI Data cleaner

$\boldsymbol{D} \nvDash \boldsymbol{\sigma}$ **CI data cleaner $\boldsymbol{T^*}$** $\widehat{\boldsymbol{D}} = \boldsymbol{T^*}(\boldsymbol{D}) \vDash \boldsymbol{\sigma}$

Problem Formulation

- CI data cleaner T^*:

$$T^* = \underset{T:\mathcal{V}\rightarrow\mathcal{V}}{\operatorname{argmin}} \sum_{v_i \in D} \mathrm{c}(\mathrm{v_i}, \mathrm{T}(\mathrm{v_i})) \quad s.t. \quad T(D) \vDash \sigma$$

- Probabilistic optimal data cleaner π^*:

$$\pi^* = \underset{\pi}{\operatorname{argmin}} \sum_{v_i, v_j' \in \mathcal{V}} \mathrm{c}\left(\mathrm{v_i}, v_j'\right) \pi(\mathrm{v_i}, v_j')$$

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

First Solution: QCLP

- Quadratically Constrained Linear Program (QCLP)
  - **Objective function:**

$$\min_{\hat{\pi}} \sum_{i,j \in [1, d_{\mathcal{V}}]} \mathrm{c}(\mathrm{v_i}, v_j) \times \hat{\pi}_{i,j}$$

  - **Constraints:**
    1. Validity constraints:

$$\hat{\pi}_{i,j} \geq 0 \; \forall i \in [1, d_{\mathcal{V}}], j \in [1, d_{\mathcal{V}}]$$

    2. Marginal constraints:

$$\sum_{j \in [1, d_{\mathcal{V}}]} \hat{\pi}_{i,j} = P^D(v_i) \quad \forall i \in [1, d_{\mathcal{V}}]$$

    3. Independence constraints: ensures satisfaction of CI constraint σ

Second Solution: Fast Approximation

- Relaxed OT with Entropic Regularize:

The entropic regularization parameter

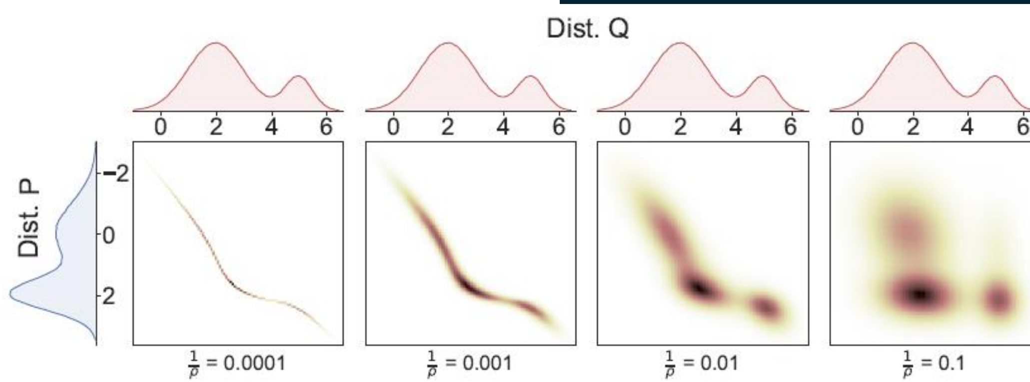
$$\underset{\pi}{\operatorname{argmin}} \sum_{\mathrm{x_i} \in \mathcal{X}, \mathrm{y_i} \in \mathcal{Y}} \mathrm{c}(\mathrm{x_i}, y_j)\pi(x_i, y_j) - \frac{1}{\rho} H(\pi)$$

$$+ \lambda\Big(D_{KL}(\pi(v'), Q) + D_{KL}\big(\pi(v), P^D\big)\Big) + \mu\delta_\sigma(Q)$$

The relaxation regularization coefficient

The CI constraint term

Entropic regularization allows **Sinkhorn algorithm**
[M. Cuturi, NIPS'13]



Second Solution: FastOTClean

Algorithm 2: FastOTClean: Fast Computation of Probabilistic Data Cleaner for Conditional Independence

Input: Database D, cost function c, and CI constraint $\sigma : X \perp\!\!\!\perp Y \mid Z$

Output: Transport plan (probabilistic data cleaner) π

$\mathbf{p} := vector(P^D)$; $\mathbf{C} := matrix(c)$;

Randomly initialize $\mathbf{q}$ ▹ An initial guess for Q

$\mathbf{u} := \mathbb{1}_{d_X}$; $\mathbf{v} := \mathbb{1}_{d_Y}$; $\mathbf{K} := e^{-\frac{\mathbf{C}}{\rho}}$; ▹ Sinkhorn Initialization

Initializations

while $\mathbf{q}$ *is not converged* **do** ▹ Sinkhorn iterations

  while $\mathbf{u}$ *and* $\mathbf{v}$ *are not converged* **do**

    $\mathbf{u} := (\mathbf{p} \oslash (\mathbf{K} \cdot \mathbf{v}))^{\frac{\rho\lambda}{\rho\lambda+1}}, \mathbf{v} := (\mathbf{q} \oslash (\mathbf{K} \cdot \mathbf{u}))^{\frac{\rho\lambda}{\rho\lambda+1}}$;

  $\pi = diag(\mathbf{u}) \cdot \mathbf{K} \cdot diag(\mathbf{v})$;

Sinkhorn iterations

  for *each* $z \in \mathcal{Z}$ **do**

    Initialize $\mathbf{W}_z$, $\mathbf{H}_z$ randomly.

    while $\mathbf{W}_z$ *and* $\mathbf{H}_z$ *are not converged* **do**

      Update $\mathbf{W}_z$ to minimize $D_{\mathrm{KL}}(\pi(X', Y', Z' = z) \mid \mathbf{W}_z \cdot \mathbf{H}_z^T)$ with $\mathbf{H}_z$ fixed

      Update $\mathbf{H}_z$ to minimize $D_{\mathrm{KL}}(\pi(X', Y', Z' = z) \mid \mathbf{W}_z \cdot \mathbf{H}_z^T)$ with $\mathbf{W}_z$ fixed

  Construct $\mathbf{q}$ using $\mathbf{W}_z$s and $\mathbf{H}_z$s computed in the previous step

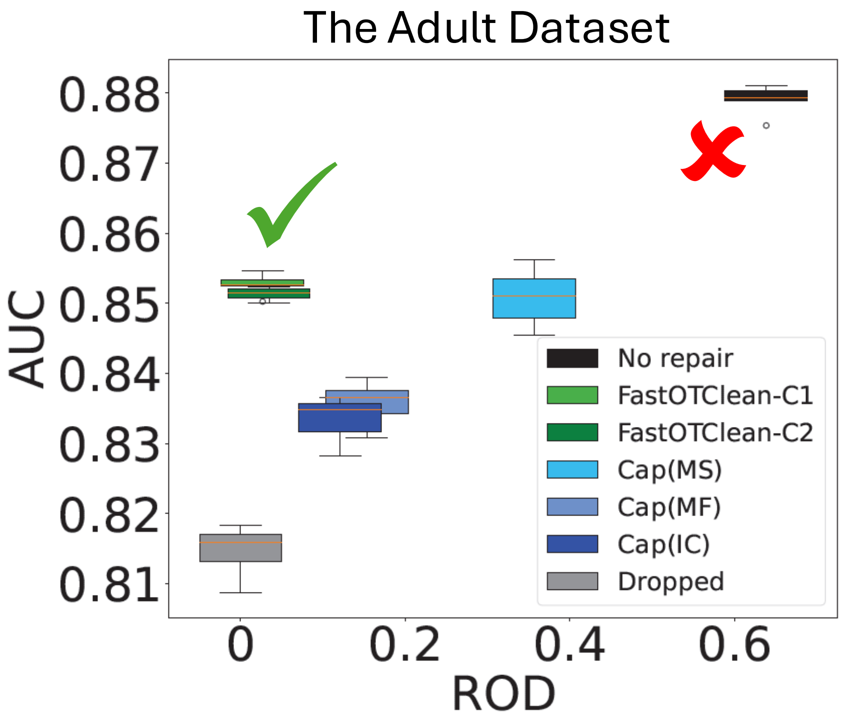
CI constraints and alternating updates

return π;

Experimental Results (Fairness)

- Datasets: Adult and COMPAS
- OTClean with two cost functions

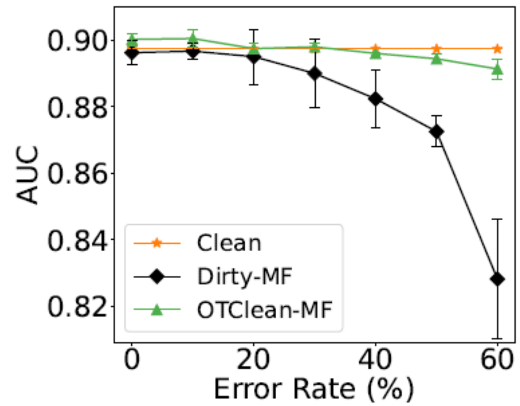
- Baselines:
  - Capuchin (Cap): three variations [Salimi et al., SIGMOD'19]
  - No repair
  - Dropped
- Performance Measures:
  - AUC for accuracy
  - ROD (Ratio of Observation Discrimination) for fairness



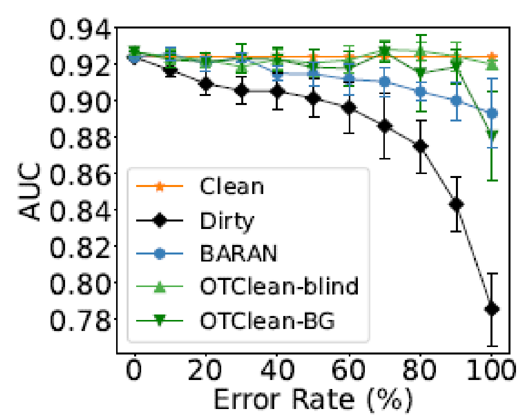
Experimental Results (Data Cleaning)

- Datasets: Car and Boston
- Attribute noise and missing values (MAR and MNAR)

- Performance Measures:
  - AUC for accuracy
- Baselines
  - Clean data
  - Dirty data
  - BARAN (attribute noise)
  - Knn, MF, etc. (missing values)



MAR
Most Frequent
Boston



Attribute Noise
Car

References

- Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science, 2019
- S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. IEEE Signal Processing Magazine, 34(4):43–59, 2017
- M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, NIPS, 2013
- B Salimi, A. L. Rodriguez, B. Howe, D. Suciu, Interventional Fairness: Causal Database Repair for Algorithmic Fairness, SIGMOD, 2019