When causality meets optimal transport

Lucas De Lara

April 21, 2022

Institut de Mathématiques de Toulouse
Artificial and Natural Intelligence Toulouse Institute

Outline

Context and motivation

FlipTest [Black et al., 2020]

Fairness/XAI framework motivated by questions framed as

Had an individual been of a different protected status, would the model have treated them differently?

FlipTest [Black et al., 2020]

Fairness/XAI framework motivated by questions framed as

Had an individual been of a different protected status, would the model have treated them differently?

Relies on **optimal transport (OT)** rather than **structural causal models (SCM)** to compute **counterfactual counterparts**.

FlipTest [Black et al., 2020]

Fairness/XAI framework motivated by questions framed as

Had an individual been of a different protected status, would the model have treated them differently?

Relies on **optimal transport (OT)** rather than **structural causal models (SCM)** to compute **counterfactual counterparts**.

- OT matches two observable distributions (e.g., females to males)
- operations on an SCM enable to generate alternative individuals after a feature modification (e.g., change of sex)

Example: the Law dataset

Black and white students described by *(LSAT,GPA)*,

Example: the Law dataset

Black and white students described by *(LSAT,GPA)*, along with the white counterfactual counterparts of the black students.

Example: the Law dataset

Black and white students described by $(LSAT, GPA)$, along with the white counterfactual counterparts of the black students.

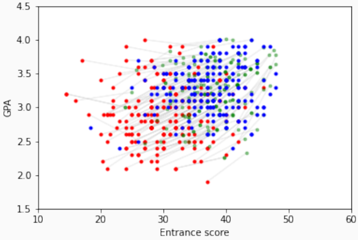

Figure 1: OT generated

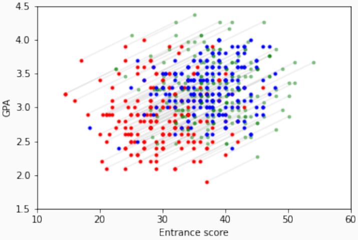

Figure 2: SCM generated

Example: the Law dataset

Black and white students described by $(LSAT, GPA)$, along with the white counterfactual counterparts of the black students.

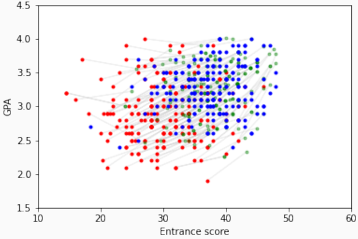

Figure 1: OT generated

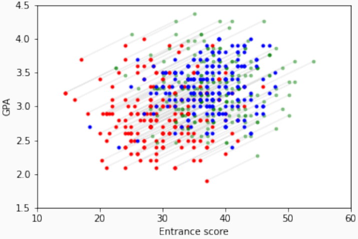

Figure 2: SCM generated

"FlipTest can give nearly identical results as causally generated counterfactuals." [Black et al., 2020]

Question

Are OT counterfactuals and SCM counterfactuals equal?

Question

Are OT counterfactuals and SCM counterfactuals equal?

Yes, under some specific assumptions.

Optimal transport

Random and deterministic couplings

P, Q two Borel probability distributions on $\mathbb{R}^d$

- $\Pi(P, Q)$ set of joint probability distributions with P and Q as first and second marginals.
- $\mathcal{T}(P, Q)$ set of measurable maps **pushing forward** P to Q

Random and deterministic couplings

P, Q two Borel probability distributions on $\mathbb{R}^d$

- $\Pi(P, Q)$ set of joint probability distributions with P and Q as first and second marginals.
- $\mathcal{T}(P, Q)$ set of measurable maps **pushing forward** P to Q

$$T \in \mathcal{T}(P, Q) \iff T_\sharp P = Q \iff (X \sim P \implies T(X) \sim Q).$$

Random and deterministic couplings

P, Q two Borel probability distributions on $\mathbb{R}^d$

- $\Pi(P, Q)$ set of joint probability distributions with P and Q as first and second marginals.
- $\mathcal{T}(P, Q)$ set of measurable maps **pushing forward** P to Q

$$T \in \mathcal{T}(P, Q) \iff T_\sharp P = Q \iff (X \sim P \implies T(X) \sim Q).$$

A coupling $\pi \in \Pi(P, Q)$ matches every instance from P to one or several instances from Q with probability weights.

Random and deterministic couplings

P, Q two Borel probability distributions on $\mathbb{R}^d$

- $\Pi(P, Q)$ set of joint probability distributions with P and Q as first and second marginals.
- $\mathcal{T}(P, Q)$ set of measurable maps **pushing forward** P to Q

$$T \in \mathcal{T}(P, Q) \iff T_\sharp P = Q \iff (X \sim P \implies T(X) \sim Q).$$

A coupling $\pi \in \Pi(P, Q)$ matches every instance from P to one or several instances from Q with probability weights.

π is **deterministic** if it concentrates on the graph of a map $T \in \mathcal{T}(P, Q)$, formally $\pi = (I \times T)_\sharp P$.

Optimal transport [Villani, 2008]

- P, Q Borel probability distributions on $\mathbb{R}^d$
- $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ cost function, typically $c(x, x') := \|x - x'\|^2$

Optimal transport [Villani, 2008]

- P, Q Borel probability distributions on $\mathbb{R}^d$
- $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ cost function, typically $c(x, x') := \|x - x'\|^2$

OT looks for couplings in $\Pi(P, Q)$, or maps in $\mathcal{T}(P, Q)$, that are optimal in the sense of c.

Optimal transport [Villani, 2008]

- P, Q Borel probability distributions on $\mathbb{R}^d$
- $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ cost function, typically $c(x, x') := \|x - x'\|^2$

OT looks for couplings in $\Pi(P, Q)$, or maps in $\mathcal{T}(P, Q)$, that are optimal in the sense of c.

Figure 3: Illustration from David Alvarez-Melis and Nicolo Fusi

Optimal transport [Villani, 2008]

Monge problem:

$$\min_{T \in \mathcal{T}(P,Q)} \int c(x, T(x)) \mathrm{d}P(x)$$

Optimal transport [Villani, 2008]

Monge problem:

$$\min_{T \in \mathcal{T}(P,Q)} \int c(x, T(x)) \mathrm{d}P(x)$$

Kantorovich problem:

$$\min_{\pi \in \Pi(P,Q)} \iint c(x, x') \mathrm{d}\pi(x, x')$$

Optimal transport [Villani, 2008]

Monge problem:

$$\min_{T \in \mathcal{T}(P,Q)} \int c(x, T(x)) \mathrm{d}P(x)$$

Kantorovich problem:

$$\min_{\pi \in \Pi(P,Q)} \iint c(x, x') \mathrm{d}\pi(x, x')$$

People are interested in either

Optimal transport [Villani, 2008]

Monge problem:

$$\min_{T \in \mathcal{T}(P,Q)} \int c(x, T(x)) \mathrm{d}P(x)$$

Kantorovich problem:

$$\min_{\pi \in \Pi(P,Q)} \iint c(x, x') \mathrm{d}\pi(x, x')$$

People are interested in either

- the **value** of the minimum, to define **metrics** between distributions (e.g., Wasserstein distances)

Optimal transport [Villani, 2008]

Monge problem:

$$\min_{T \in \mathcal{T}(P,Q)} \int c(x, T(x)) \mathrm{d}P(x)$$

Kantorovich problem:

$$\min_{\pi \in \Pi(P,Q)} \iint c(x, x') \mathrm{d}\pi(x, x')$$

People are interested in either

- the **value** of the minimum, to define **metrics** between distributions (e.g., Wasserstein distances)
- or the **minimizers** of these programs, to define **matchings** between distributions (e.g., fairness, domain adaptation)

Example

Output of a POT solver for the Monge problem [Flamary et al., 2021]

Computed on 800/800 points

Represented on 200/200 points

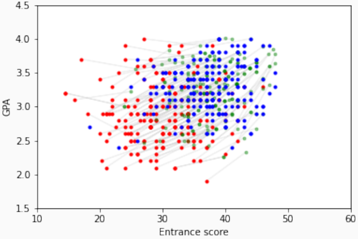

Figure 4: Estimated OT map

Structural counterfactuals

Pearl's causal framework [Pearl, 2009]

Exogenous $U = (U_1, U_2, \ldots)$

Immutable, prior knowledge

Endogenous

$V = (X_1, X_2, \ldots, X_d, S)$

Defined as

$V_i = G_i(V_{\mathrm{Endo}(i)}, U_{\mathrm{Exo}(i)})$

Pearl's causal framework [Pearl, 2009]

Endogenous

Exogenous $U = (U_1, U_2, \ldots)$

$V = (X_1, X_2, \ldots, X_d, S)$

Immutable, prior knowledge

Defined as

$V_i = G_i(V_{\text{Endo}(i)}, U_{\text{Exo}(i)})$

Solvability: There exists a solution map Γ such that $V = \Gamma(U)$

In particular $X = F(S, U_X)$

Do-intervention and counterfactuals

The operation $\mathrm{do}(S = s')$ forces the sensitive variable to take the fixed value s' while keeping the rest of the causal equations untouched.

Do-intervention and counterfactuals

The operation $\mathrm{do}(S = s')$ forces the sensitive variable to take the fixed value s' while keeping the rest of the causal equations untouched.

$$X = F(S, U_X) \xrightarrow{\mathrm{do}(S=s')} X_{S=s'} = F(s', U_X)$$

Do-intervention and counterfactuals

The operation $\mathrm{do}(S = s')$ forces the sensitive variable to take the fixed value s' while keeping the rest of the causal equations untouched.

$$X = F(S, U_X) \xrightarrow{\mathrm{do}(S=s')} X_{S=s'} = F(s', U_X)$$

The **counterfactual counterparts** of an instance $\{X = x, S = s\}$, Had S been equal to s' instead of s, are given by the distribution

$$\mathcal{L}(X_{S=s'} \mid X = x, S = s).$$

Do-intervention and counterfactuals

The operation $\mathrm{do}(S = s')$ forces the sensitive variable to take the fixed value s' while keeping the rest of the causal equations untouched.

$$X = F(S, U_X) \xrightarrow{\mathrm{do}(S=s')} X_{S=s'} = F(s', U_X)$$

The **counterfactual counterparts** of an instance $\{X = x, S = s\}$, Had S been equal to s' instead of s, are given by the distribution

$$\mathcal{L}(X_{S=s'} \mid X = x, S = s).$$

It can be generated by estimating and sampling from $\mathcal{L}(U_X \mid X = x, S = s)$.

Example

Structural equations [Kusner et al., 2017]:

$$X_1 = w_1 S + U_1$$
$$X_2 = w_2 S + U_2$$
$$U_1 \perp\!\!\!\perp U_2.$$

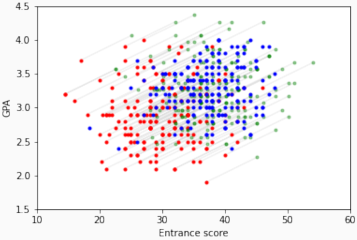

Figure 5: SCM counterfactuals

Example

Structural equations [Kusner et al., 2017]:

$$X_1 = w_1 S + U_1$$
$$X_2 = w_2 S + U_2$$
$$U_1 \perp\!\!\!\perp U_2.$$

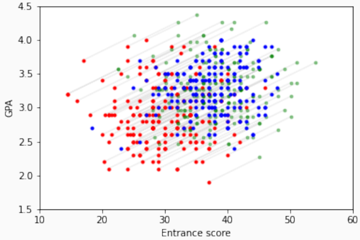

Figure 5: SCM counterfactuals

Obs 1: deterministic counterfactuals (i.e., one-to-one)

Obs 2: white counterfactuals seem to agree with white factuals

The mass-transportation viewpoint of structural counterfactuals

Counterfactual inference as mass transportation

The effect of $\mathrm{do}(S = s' | S = s)$ is fully characterized by the coupling

$$\pi^*_{\langle s' | s \rangle} := \mathcal{L}\left((X, X_{S=s'}) | S = s\right).$$

It assigns a probability to all the pairs (x, x') between an observable value x and a counterfactual counterpart x'.

Counterfactual inference as mass transportation

The effect of $\mathrm{do}(S = s' | S = s)$ is fully characterized by the coupling

$$\pi^*_{\langle s' | s \rangle} := \mathcal{L}\left((X, X_{S=s'}) | S = s\right).$$

It assigns a probability to all the pairs (x, x') between an observable value x and a counterfactual counterpart x'.

This coupling admits $\mu_s := \mathcal{L}(X | S = s)$ as first marginal and $\mu_{\langle s' | s \rangle} := \mathcal{L}(X_{S=s'} | S = s)$ as second marginal.

Counterfactual inference as mass transportation

The effect of $\mathrm{do}(S = s' | S = s)$ is fully characterized by the coupling

$$\pi^*_{\langle s'|s\rangle} := \mathcal{L}\left((X, X_{S=s'})|S = s\right).$$

It assigns a probability to all the pairs (x, x') between an observable value x and a counterfactual counterpart x'.

This coupling admits $\mu_s := \mathcal{L}(X|S = s)$ as first marginal and $\mu_{\langle s'|s\rangle} := \mathcal{L}(X_{S=s'}|S = s)$ as second marginal.

Remark: Therefore, $\pi^*_{\langle s'|s\rangle} \in \Pi(\mu_s, \mu_{\langle s'|s\rangle}) \neq \Pi(\mu_s, \mu_{s'})$.

The exogenous case

Assumption (RE):

1. S does not have endogenous parents
2. $U_S \perp\!\!\!\perp U_X$

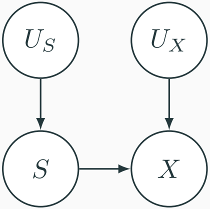

Figure 6: DAG satisfying (RE)

The exogenous case

Assumption (RE):

1. S does not have endogenous parents
2. $U_S \perp\!\!\!\perp U_X$

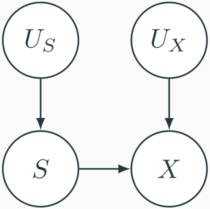

Figure 6: DAG satisfying (RE)

Proposition

If (RE) holds, then $S \perp\!\!\!\perp U_X$ and

$$\mu_{\langle s' | s \rangle} = \mu_{s'}$$

The exogenous case

Assumption (RE):

1. S does not have endogenous parents
2. $U_S \perp\!\!\!\perp U_X$

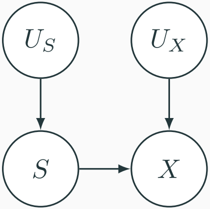

Figure 6: DAG satisfying (RE)

Proposition

If (RE) holds, then $S \perp\!\!\!\perp U_X$ and

$$\mu_{\langle s' | s \rangle} = \mu_{s'}$$

Consequence: $\pi^*_{\langle s' | s \rangle} \in \Pi(\mu_s, \mu_{s'})$.

The deterministic case

Reminder: $X = F(S, U_X)$

The deterministic case

Reminder: $X = F(S, U_X)$

Assumption (I): Knowing $S = s$, the model induces a one-to-one relationship between X values and U_X values:

The function $f_s := F(s, \cdot)$ is injective

The deterministic case

Reminder: $X = F(S, U_X)$

Assumption (I): Knowing $S = s$, the model induces a one-to-one relationship between X values and U_X values:

The function $f_s := F(s, \cdot)$ is injective

Proposition

If (I) holds, then μ_s-almost every instance x admits a unique counterfactual counterpart $x' = T^*_{\langle s' | s \rangle}(x)$ where

$$T^*_{\langle s' | s \rangle} := f_{s'} \circ f_s^{-1}.$$

Holds in every **additive model**, where U_X is additive in the causal equations

An example

Linear additive SCM:

$$
\begin{aligned}
S &= \dots \\
X &= MX + wS + b + U_X
\end{aligned}
$$

An example

Linear additive SCM:

$$S = \dots$$
$$X = MX + wS + b + U_X$$

Acyclicity implies that $I - M$ is invertible so that

$$X = (I - M)^{-1}(wS + b + U_X) =: F(S, U_X).$$

An example

Linear additive SCM:

$$S = \dots$$
$$X = MX + wS + b + U_X$$

Acyclicity implies that $I - M$ is invertible so that

$$X = (I - M)^{-1}(wS + b + U_X) =: F(S, U_X).$$

Consequently,

$$T^*_{\langle s' | s \rangle}(x) := x + (I - M)^{-1} w(s' - s).$$

Checkpoint

	¬(RE)	(RE)
¬(I)	$\pi^*_{\langle s' \mid s \rangle} \in \Pi(\mu_s, \mu_{\langle s' \mid s \rangle})$	$\pi^*_{\langle s' \mid s \rangle} \in \Pi(\mu_s, \mu_{s'})$
(I)	$T^*_{\langle s' \mid s \rangle \sharp} \mu_s = \mu_{\langle s' \mid s \rangle}$	$T^*_{\langle s' \mid s \rangle \sharp} \mu_s = \mu_{s'}$

Effect of $\mathrm{do}(S = s' \mid S = s)$

When quadratic optimal transport meets causality

Theorem [De Lara et al., 2021]

OT: $c(x, x') = \|x - x'\|^2$, X has a density and a finite second-order moment

Theorem [De Lara et al., 2021]

OT: $c(x, x') = \|x - x'\|^2$, X has a density and a finite second-order moment

$\implies$ unique solution $T_{\langle s'|s \rangle}$ to the OT problem

Theorem [De Lara et al., 2021]

OT: $c(x, x') = \|x - x'\|^2$, X has a density and a finite second-order moment

$$\implies \text{unique solution } T_{\langle s'|s\rangle} \text{ to the OT problem}$$

SCM: (RE) and (I) hold

Theorem [De Lara et al., 2021]

OT: $c(x, x') = \|x - x'\|^2$, X has a density and a finite second-order moment

$$\implies \text{unique solution } T_{\langle s'|s\rangle} \text{ to the OT problem}$$

SCM: (RE) and (I) hold

$$\implies \text{counterfactuals given by } T^*_{\langle s'|s\rangle\,\sharp}\mu_s = \mu_{s'}$$

Theorem [De Lara et al., 2021]

OT: $c(x, x') = \|x - x'\|^2$, X has a density and a finite second-order moment

$\implies$ **unique solution $T_{\langle s'|s \rangle}$ to the OT problem**

SCM: (RE) and (I) hold

$\implies$ **counterfactuals given by ${T^*_{\langle s'|s \rangle}}_\sharp \mu_s = \mu_{s'}$**

$T^*_{\langle s'|s \rangle} = T_{\langle s'|s \rangle} \iff f_{s'} \circ f_s^{-1}$ **is the gradient of a convex function**

Theorem [De Lara et al., 2021]

OT: $c(x, x') = \|x - x'\|^2$, X has a density and a finite second-order moment

$\implies$ unique solution $T_{\langle s'|s\rangle}$ to the OT problem

SCM: (RE) and (I) hold

$\implies$ counterfactuals given by $T^*_{\langle s'|s\rangle\sharp}\mu_s = \mu_{s'}$

$T^*_{\langle s'|s\rangle} = T_{\langle s'|s\rangle} \iff f_{s'} \circ f_s^{-1}$ is the gradient of a convex function

Condition satisfied in any linear additive model
(e.g., the Law dataset)

Quadratic optimal transport

Monotone measure-preserving map

If P and Q are absolutely continuous w.r.t. Lebesgue measure, then there exists a convex potential $\phi : \mathbb{R}^d \to \mathbb{R}$ such that $\nabla\phi_\sharp P = Q$. The map $T := \nabla\phi$ is unique P-almost everywhere.

Quadratic optimal transport

Monotone measure-preserving map

If P and Q are absolutely continuous w.r.t. Lebesgue measure, then there exists a convex potential $\phi : \mathbb{R}^d \to \mathbb{R}$ such that $\nabla\phi_\sharp P = Q$. The map $T := \nabla\phi$ is unique P-almost everywhere.

Optimal transport map

If P and Q are absolutely continuous w.r.t. Lebesgue measure and have finite second order moments, then there exists a unique solution to

$$\min_{\pi \in \Pi(P,Q)} \iint \|x - x'\| \mathrm{d}\pi(x, x'),$$

which is $\pi := (I \times T)_\sharp P$ where T is "the" monotone measure-preserving map from P to Q.

Nonlinear nonadditive positive example

SCM:

$$
\begin{aligned}
X_1 &= \alpha(S)U_1 + \beta_1(S) \\
X_2 &= -\alpha(S)\ln^2\left(\frac{X_1 - \beta_1(S)}{\alpha(S)}\right)U_2 + \beta_2(S) \\
S &= U_S \perp\!\!\!\perp (U_1, U_2)
\end{aligned}
$$

Nonlinear nonadditive positive example

SCM:

$$
\begin{aligned}
X_1 &= \alpha(S)U_1 + \beta_1(S) \\
X_2 &= -\alpha(S)\ln^2\left(\frac{X_1 - \beta_1(S)}{\alpha(S)}\right)U_2 + \beta_2(S) \\
S &= U_S \perp\!\!\!\perp (U_1, U_2)
\end{aligned}
$$

Counterfactuals:

$$
T^*_{\langle s' | s \rangle}(x) = \frac{\alpha(s')}{\alpha(s)}x + [\beta(s') - \beta(s)]
$$

Nonlinear nonadditive positive example

SCM:

$$\begin{aligned} X_1 &= \alpha(S)U_1 + \beta_1(S) \\ X_2 &= -\alpha(S)\ln^2\left(\frac{X_1 - \beta_1(S)}{\alpha(S)}\right)U_2 + \beta_2(S) \\ S &= U_S \perp\!\!\!\perp (U_1, U_2) \end{aligned}$$

Counterfactuals:

$$T^*_{\langle s'|s\rangle}(x) = \frac{\alpha(s')}{\alpha(s)}x + [\beta(s') - \beta(s)]$$

If $\alpha(\cdot) > 0$, this is the gradient of a convex function.

Conclusion

And so what?

And so what?

OT counterfactuals and SCM counterfactuals share a common **mass-transportation** formalism, and can even coincide, making them natural surrogate

And so what?

OT counterfactuals and SCM counterfactuals share a common **mass-transportation** formalism, and can even coincide, making them natural surrogate

Practical interest: For feasibility reasons, use OT solutions instead of SCMs in counterfactual frameworks (see [Black et al., 2020] and [De Lara et al., 2021] for applications)

And so what?

OT counterfactuals and SCM counterfactuals share a common **mass-transportation** formalism, and can even coincide, making them natural surrogate

Practical interest: For feasibility reasons, use OT solutions instead of SCMs in counterfactual frameworks (see [Black et al., 2020] and [De Lara et al., 2021] for applications)

Theoretical interest: Reformulating counterfactual reasoning as a mass transportation problem allows new results and proofs (see [De Lara et al., 2021])

Final word

Optimal transport (a statistical tool) meets causality (under some assumptions)

Black, E., Yeom, S., and Fredrikson, M. (2020).
Fliptest: Fairness testing via optimal transport.
FAT* '20, page 111–121.

De Lara, L., González-Sanz, A., Asher, N., Risser, L., and Loubes, J.-M. (2021).
Transport-based counterfactual models.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021).
POT: Python optimal transport.
Journal of Machine Learning Research, 22(78):1–8.

References ii

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017).
Counterfactual fairness.
In *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076. Curran Associates, Inc.

Pearl, J. (2009).
Causality.
Cambridge university press.

Villani, C. (2008).
Optimal Transport: Old and New.
Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin.
OCLC: ocn244421231.