Sequential Conditional Transport on Probabilistic Graphs for Interpretable Counterfactual Fairness

Agathe Fernandes Machado✉[1], Arthur Charpentier[1], and Ewen Gallic[2]

[1]Département de Mathématiques, Université du Québec à Montréal, Montréal, Québec, Canada
[2]Aix Marseille Univ, CNRS, AMSE, Marseille, France, Marseille, France

December 20, 2024

Abstract

In this paper, we link two existing approaches to derive counterfactuals: adaptations based on a causal graph, and optimal transport. We extend "Knothe's rearrangement" and "triangular transport" to probabilistic graphical models, and use this counterfactual approach, referred to as sequential transport, to discuss fairness at the individual level. After establishing the theoretical foundations of the proposed method, we demonstrate its application through numerical experiments on both synthetic and real datasets.

1 Introduction

Most applications concerning discrimination and fairness are based on "group fairness" concepts (as introduced in Hardt et al. (2016); Kearns and Roth (2019), or Barocas et al. (2023)). However, in many cases, fairness should be addressed at the individual level rather than globally. As claimed in Dwork et al. (2012), "*we capture fairness by the principle that any two individuals who are similar with respect to a particular task should be classified similarly.*" The concept of "counterfactual fairness" was formalized in Kusner et al. (2017), addressing questions such as "*had the protected attributes of the individual been different, other things being equal, would the decision had remain the same?*" Such a statement has clear connections with causal inference, as discussed in Pearl and Mackenzie (2018). Formally, consider observations $\{s_i, \boldsymbol{x}_i, y_i\}$, where s is a binary protected attribute (e.g., $s \in \{0, 1\}$), and $\boldsymbol{x}$ is a collection of legitimate features (possibly correlated with s). The model output is y, which is analyzed to address "algorithmic fairness" issues. Following Rubin (2005), let $y^\star(s)$ denote the potential outcome of y if s is seen as a treatment. With these notations, counterfactual fairness is achieved for individual $(s, \boldsymbol{x})$ if the average "treatment effect," conditional on $\boldsymbol{x}$ (or "CATE") is zero, i.e., $\mathbb{E}[Y^\star(1) - Y^\star(0)|\boldsymbol{X} = \boldsymbol{x}] = 0$. This quantity could be termed "*ceteris paribus* CATE" since all $\boldsymbol{x}$'s are supposed to remain unchanged for both treated and non-treated.

Agathe Fernandes Machado acknowledges that the project leading to this publication has received funding from OBVIA. Arthur Charpentier acknowledges funding from the SCOR Foundation for Science and the National Sciences and Engineering Research Council (NSERC) for funding (RGPIN-2019-07077). Ewen Gallic acknowledge funding from the French government under the "France 2030" investment plan managed by the French National Research Agency (reference: ANR-17-EURE-0020) and from Excellence Initiative of Aix-Marseille University – A*MIDEX.

Replication codes and companion e-book: https://github.com/fer-agathe/sequential_transport

✉Corresponding author: fernandes_machado.agathe@courrier.uqam.ca

Following Kilbertus et al. (2017), it is possible to suppose that the protected attribute s could actually affect some explanatory variables $\boldsymbol{x}$ in a non-discriminatory way. In Charpentier et al. (2023), the outcome y was "having a surgical intervention" during childbirth in the U.S., s was the mother's ethnic origin ("Black," or not) and $\boldsymbol{x}$ included factors such as "weight of the baby at birth." If Black mothers undergo less surgery because they tend to have smaller babies, there is no discrimination *per se*. At the very least, it should be fair, when assessing whether hospitals have discriminatory policies, to account for that difference in baby weights. Such a variable is named "revolving variable" in Kilbertus et al. (2017). Using heuristic notations, the "*ceteris paribus* CATE" $\mathbb{E}[Y^\star(1)|\boldsymbol{X}=\boldsymbol{x}] - \mathbb{E}[Y^\star(0)|\boldsymbol{X}=\boldsymbol{x}]$ should become a "*mutatis mutandis* CATE". For some individual $(s=0,\boldsymbol{x})$, this indicator would be $\mathbb{E}[Y^\star(1)|\boldsymbol{X}=\boldsymbol{x}^\star(1)] - \mathbb{E}[Y^\star(0)|\boldsymbol{X}=\boldsymbol{x}]$, as coined in Charpentier et al. (2023), to quantify discrimination, where fictitious individual $(s=1,\boldsymbol{x}^\star(1))$ is a "counterfactual" version of $(s=0,\boldsymbol{x})$.

Two recent approaches have been proposed to assess counterfactual fairness using this *mutatis mutandis* approach. On the one hand, Plečko and Meinshausen (2020) and Plečko et al. (2024) used causal graphs (DAGs) to construct counterfactuals and assess the counterfactual fairness of outcomes y based on variables $(s,\boldsymbol{x},y)$. In network flow terminology, s acts as a "source" (only outgoing flow, or no parents), while y is a "sink" (only incoming flow). On the other hand, Black et al. (2020), Charpentier et al. (2023) and De Lara et al. (2024) used optimal transport (OT) to construct counterfactuals. Moreover, using counterfactual reasoning to achieve fair machine learning (ML) models has also been notably studied (Ma et al., 2023; Robertson et al., 2024). For evaluation, while De Lara et al. (2024) provided a theoretical framework, its implementation is challenging (except in the Gaussian case), and usually hard to interpret. Here, we combine the two approaches, using OT within a causal graph structure. The idea is to adapt "Knothe's rearrangement" Bonnotte (2013), or "triangular transport" Zech and Marzouk (2022a,b), to a general probabilistic graphical model on $(s,\boldsymbol{x},y)$, rather than a simplistic $s \to x_1 \to x_2 \to \cdots \to x_d \to y$. The concept of "conditional OT" has been recently discussed in Bunne et al. (2022) and Hosseini et al. (2023), but here, instead of learning the causal graph, we assume a known causal graph and use it to construct counterfactual versions of individuals $(s_i,\boldsymbol{x}_i,y_i)$ to address fairness issues. Additionally, since we use univariate (conditional) transport, standard classical properties of univariate transport facilitates explanations (non-decreasing mappings, and quantile based interpretations).

Main Contributions

- We use multivariate transport theory for constructing counterfactuals, as suggested in De Lara et al. (2024), and connect it to quantile preservation on causal graphs from Plečko and Meinshausen (2020) to develop a sequential transport methodology that aligns with the underlying DAG of the data.
- Sequential transport, using univariate transport maps, provides closed-form solutions for deriving counterfactuals. This allows for the development of a data-driven estimation procedure that can be applied to new out-of-samples observations without recalculating, unlike multivariate OT with non-Gaussian distributions.
- The approach's applicability is demonstrated through numerical experiments on both synthetic data and case studies, highlighting the interpretable analysis of individual counterfactual fairness when using sequential transport.

Section 2 introduces various concepts used in probabilistic graphical models from a causal perspective. Section 3, revisits classical OT covering both univariate and multivariate cases. Sequential transport is covered in Section 4. Section 5 discusses counterfactual fairness. Illustration with real data are provided in Section 6.

2 Graphical Model and Causal Network

2.1 Probabilistic Graphical Model

Following standard notations in probabilistic graphical models (see Koller and Friedman (2009) or Barber (2012)), given a random vector $\boldsymbol{X} = (X_1, \cdots, X_d)$, consider a directed acyclic graph (DAG) $\mathcal{G} = (V, E)$, where $V = \{x_1, x_2, \cdots, x_d\}$ are the vertices (corresponding to each variable), and E are directed edges, such that $x_i \to x_j$ means "variable x_i causes variable x_j," in the sense of Susser (1991). The joint distribution of $\boldsymbol{X}$ satisfies the (global) Markov property w.r.t. $\mathcal{G}$:

$$\mathbb{P}[x_1, \cdots, x_d] = \prod_{j=1}^{d} \mathbb{P}[x_j | \text{parents}(x_j)],$$

where $\text{parents}(x_i)$ are nodes with edges directed towards x_i, in $\mathcal{G}$. Watson et al. (2021) suggested the causal graph in Figure 1 for the German Credit dataset, where s is the "sex" (top left) and y is the "default" indicator (right). Observe that variables $\boldsymbol{x}_j$ are here sorted. As discussed in Ahuja et al. (1993), such a causal graph imposes some ordering on variables. In this "topological sorting," a vertex must be selected before its adjacent vertices, which is feasible because each edge is directed such that no cycle exists in the graph. In our analysis, we consider a network $\mathcal{G}$ on variables $\{s, \boldsymbol{x}, y\}$ where s is the sensitive attribute, acting as a "source" (only outgoing flow, or no parents) while y is a "sink" (only incoming flow, i.e., $y \notin \text{parents}(x_i), \forall i$).

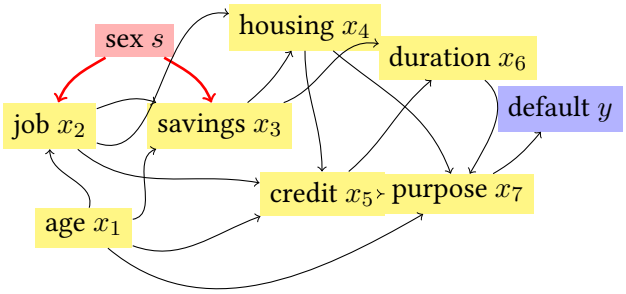


Figure 1: Causal graph in the German Credit dataset from Watson et al. (2021), or DAG.

2.2 Causal Networks and Linear Structural Models

Wright (1921, 1934) used directed graphs to represent probabilistic cause-and-effect relationships among a set of variables and developed path diagrams and path analysis. Simple causal networks can be visualized on top of Figure 2. On the left is a simple model where the "cause" C directly causes ($\to$) the "effect" E. On the right, a "mediator" X is added. There is still the direct impact of C on E ($C \to E$), but there is also a mediated indirect impact ($C \to X \to E$).

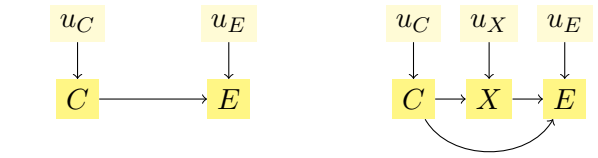
2.2.1 Intervention in a Linear Structural Model

In a simple causal graph, with two nodes, C (the cause) and E (the effect), the causal graph is $C \to E$, and the mathematical interpretation can be summarized in two (linear) assignments:

$$\begin{cases} C = a_c + U_C \\ E = a_e + b_e C + U_E, \end{cases} \tag{1}$$

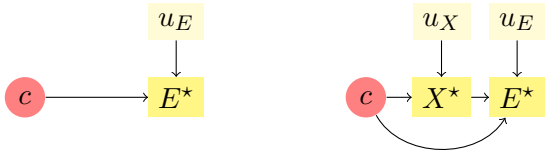
where U_C and U_E are independent Gaussian random variables. That causal graph can be visualized in Figure 2, and its corresponding structural causal model (SCM) described in Equation 1

illustrates the causal relationships between variables, as in Pearl (2000). Suppose here that C is a binary variable, taking values in $\{c_0, c_1\}$. Given an observation (c_0, e), the "counterfactual outcome" if the cause had been set to c_1 (corresponding to the intervention in Figure 3), would be $e + b_e(c_1 - c_0)$. Following Pearl (2009), one can also introduce the "twin network" corresponding to a mirrored version of the initial causal graph in the counterfactual world. Plečko and Meinshausen (2020) coined this approach "fair-twin projection" when C is a binary sensitive attribute.



$$\begin{cases} C = a_c + U_C \\ E = a_e + b_e C + U_E \end{cases} \qquad \begin{cases} C = a_c + U_C \\ X = a_x + b_x C + U_X \\ E = a_e + b_e C + \gamma_e X + U_E \end{cases}$$

Figure 2: Linear Structural Causal Model – observation.



$$\begin{cases} C = c \ \ (\text{or } \mathrm{do}(C = c)) \\ E_c^\star = a_e + b_e c + U_E \end{cases} \qquad \begin{cases} C = c \ \ (\text{or } \mathrm{do}(C = c)) \\ X_c^\star = a_x + b_x c + U_X \\ E_c^\star = a_e + b_e c + \gamma_e X_c^\star + U_E \end{cases}$$

Figure 3: Linear Structural Causal Model – intervention.

2.3 Non-Linear Structural Models

2.3.1 Presentation of the Model

More generally, consider a non-Gaussian and nonlinear structural model, named "non-parametric structural equation model" (with independent errors) in Pearl (2000),

$$\begin{cases} C = h_c(U_C) \\ E = h_e(C, U_E), \end{cases} \tag{2}$$

where $u \mapsto h_c(\cdot, u)$ and $u \mapsto h_e(\cdot, u)$ are strictly increasing in u; U_C and U_E are independent, and, without loss of generality, supposed to be uniform on $[0, 1]$. For a rigorous mathematical framework for non-linear non-Gaussian structural causal models, see Bongers et al. (2021) or Shpitser et al. (2022).

2.3.2 Connections With Conditional Quantiles

Consider now some general DAG, $\mathcal{G}$, on $\boldsymbol{X} = (X_1, \cdots, X_d)$, supposed to be absolutely continuous. With previous notations, $X_i = h_i(\text{parents}(X_i), U_i)$, a.s., for all variables, representing

the structural equations. We can write this compactly as $\boldsymbol{X} = h(\text{parents}(\boldsymbol{X}), \boldsymbol{U})$, a.s., by considering h as a vector function. Solving the structural model means finding a function g such that $\boldsymbol{X} = g(\boldsymbol{U})$, a.s. To illustrate, consider a specific i, and $X_i = h_i(\text{parents}(X_i), U_i)$. If $\text{parents}(X_i) = \boldsymbol{x}$ is fixed, define $h_{i|\boldsymbol{x}}(u) = h_i(\boldsymbol{x}, u)$. Let U be a uniform random variable, and let $F_{i|\boldsymbol{x}}$ be the cumulative distribution of $h_{i|\boldsymbol{x}}(U)$, $F_{i|\boldsymbol{x}}(x) = \mathbb{P}[h_{i|\boldsymbol{x}}(U) \leq x]$. Since X_i is absolutely continuous, $F_{i|\boldsymbol{x}}$ is invertible, and $F^{-1}_{i|\boldsymbol{x}}$ is a conditional quantile function (conditional on $\text{parents}(X_i) = \boldsymbol{x}$). Let $V = F_{i|\boldsymbol{x}}(h_{i|\boldsymbol{x}}(U))$, then $X_i = F^{-1}_{i|\boldsymbol{x}}(V)$ and V is uniformly distributed on $[0, 1]$. This means that $x_i = h_{i|\boldsymbol{x}}(u_i)$ corresponds to the quantile of variable X_i, conditional on the values of its parents, $\text{parents}(X_i)$, with probability level u_i. In the observational world, u_i represents the (conditional) probability level associated with observation x_i, and its counterfactual counterpart is $x_i^\star$ corresponding to the (conditional) quantile associated with the same probability level u_i.

This representation has been used in Plečko and Meinshausen (2020) and Plečko et al. (2024), where $X_i = F^{-1}_{i|\boldsymbol{x}}(V)$ is simply the probabilistic representation of "quantile regression," as introduced by Koenker and Bassett Jr (1978) (and further studied in Koenker (2005) and Koenker et al. (2017)). This could be extended to "quantile regression forests," as in Meinshausen and Ridgeway (2006), or any kind of ML model, as Cannon (2018) or Pearce et al. (2022). Observe that Ma and Koenker (2006) considered some close "recursive structural equation models," characterized by a system of equations where each endogenous variable is regressed on other endogenous and exogenous variables in a hierarchical manner. They used some sequential quantile regression approach to solve those recursive SEMs. An alternative we consider here is to use the connection between quantiles and OT (discussed in Chernozhukov et al. (2013) or Hallin and Konen (2024)) to define some "conditional transport" that relates to those conditional quantiles.

3 Optimal Transport

Given two metric spaces $\mathcal{X}_0$ and $\mathcal{X}_1$, consider a measurable map $T : \mathcal{X}_0 \to \mathcal{X}_1$ and a measure μ_0 on $\mathcal{X}_0$. The push-forward of μ_0 by T is the measure $\mu_1 = T_\#\mu_0$ on $\mathcal{X}_1$ defined by $T_\#\mu_0(B) = \mu_0\big(T^{-1}(B)\big)$, $\forall B \subset \mathcal{X}_1$. For all measurable and bounded $\varphi : \mathcal{X}_1 \to \mathbb{R}$,

$$\int_{\mathcal{X}_1} \varphi(x_1) T_\#\mu_0(\mathrm{d}x_1) = \int_{\mathcal{X}_0} \varphi\big(T(x_0)\big)\mu_0(\mathrm{d}x_0).$$

For our applications, if we consider measures $\mathcal{X}_0 = \mathcal{X}_1$ as a compact subset of $\mathbb{R}^d$, then there exists T such that $\mu_1 = T_\#\mu_0$, when μ_0 and μ_1 are two measures, and μ_0 is atomless, as shown in Villani (2003) and Santambrogio (2015). In that case, and if we further suppose that measures μ_0 and μ_1 are absolutely continuous, with densities f_0 and f_1 (w.r.t. Lebesgue measure), a classical change of variable expression can be derived. Specifically, the previous integral

$$\int_{\mathcal{X}_1} \varphi(\boldsymbol{x}_1) f_1(\boldsymbol{x}_1)\mathrm{d}\boldsymbol{x}_1$$

is simply (if ∇T is the Jacobian matrix of mapping T):

$$\int_{\mathcal{X}_0} \varphi\big(T(\boldsymbol{x}_0)\big)\ \underbrace{f_1(T(\boldsymbol{x}_0))\det \nabla T(\boldsymbol{x}_0)}_{=f_0(\boldsymbol{x}_0)}\ \mathrm{d}\boldsymbol{x}_0.$$

Out of those mappings from μ_0 to μ_1, we can be interested in "optimal" mappings, satisfying Monge problem, from Monge (1781), i.e., solutions of

$$\inf_{T_\#\mu_0=\mu_1} \int_{\mathcal{X}_0} c\big(\boldsymbol{x}_0, T(\boldsymbol{x}_0)\big)\mu_0(\mathrm{d}\boldsymbol{x}_0),$$

for some positive ground cost function $c : \mathcal{X}_0 \times \mathcal{X}_1 \to \mathbb{R}_+$.

In general settings, however, such a deterministic mapping T between probability distributions may not exist (in particular if μ_0 and μ_1 are not absolutely continuous, with respect to Lebesgue measure). This limitation motivates the Kantorovich relaxation of Monge's problem, as considered in Kantorovich (1942),

$$\inf_{\pi \in \Pi(\mu_0, \mu_1)} \int_{\mathcal{X}_0 \times \mathcal{X}_1} c(\boldsymbol{x}_0, \boldsymbol{x}_1) \pi(\mathrm{d}\boldsymbol{x}_0, \mathrm{d}\boldsymbol{x}_1),$$

with our cost function c, where $\Pi(\mu_0, \mu_1)$ is the set of all couplings of μ_0 and μ_1. This problem focuses on couplings rather than deterministic mappings It always admits solutions referred to as OT plans.

3.1 Univariate Optimal Transport

Suppose here that $\mathcal{X}_0 = \mathcal{X}_1$ is a compact subset of $\mathbb{R}$. The optimal Monge map $T^\star$ for some strictly convex cost c such that $T^\star_\# \mu_0 = \mu_1$ is $T^\star = F_1^{-1} \circ F_0$, where $F_i : \mathbb{R} \to [0, 1]$ is the cumulative distribution function associated with μ_i, $F_i(x) = \mu_i((-\infty, x])$, and F_i^{-1} is the generalized inverse (corresponding to the quantile function), $F_i^{-1}(u) = \inf \left\{ x \in \mathbb{R} : F_i(x) \geq u \right\}$. Observe that $T^\star$ is an increasing mapping (which is the univariate definition of being the gradient of a convex function, from Brenier (1991)). This is illustrated in Figure 4, with a Gaussian case on the left ($T^\star$ affine), and general densities on the right.

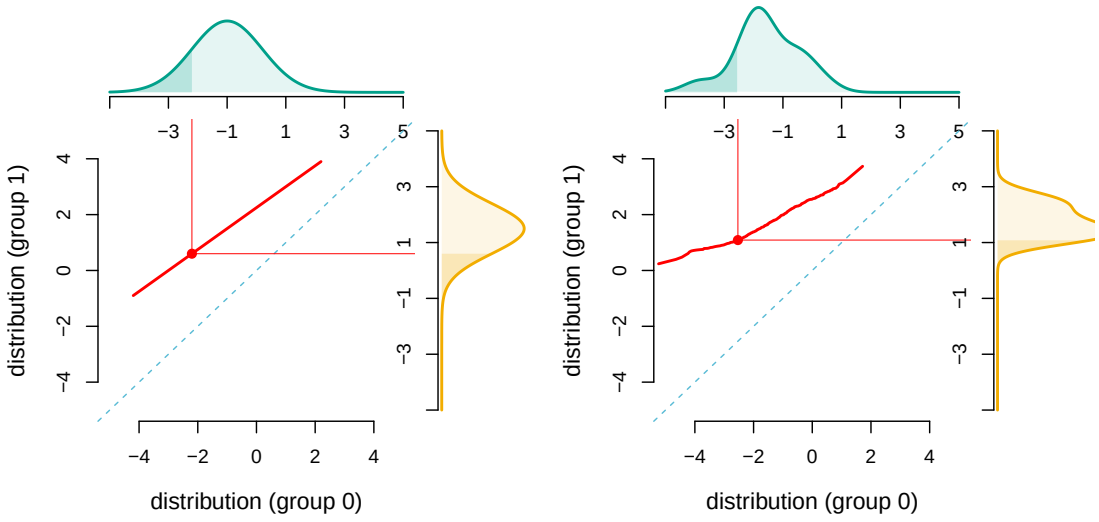


Figure 4: Univariate OT, with Gaussian distributions (left) and general marginal distributions (right). The transport curve ($T^\star$) is shown in red.

3.2 Multivariate Optimal Transport

In a multivariate setting, when $\mathcal{X}_0 = \mathcal{X}_1$ is a compact subset of $\mathbb{R}^d$, from Brenier (1991), with a quadratic cost, the optimal Monge map $T^\star$ is unique, and it is the gradient of a convex mapping $\psi : \mathbb{R}^d \to \mathbb{R}^d$, $T^\star = \nabla \psi$. Therefore, its Jacobian matrix $\nabla T^\star$ is nonnegative and symmetric. More generally, with strictly convex cost in $\mathbb{R}^d \times \mathbb{R}^d$, the Jacobian matrix $\nabla T^\star$, even if not necessarily nonnegative symmetric, is diagonalizable with nonnegative eigenvalues, as proved in Cordero-Erausquin (2004) and Ambrosio et al. (2005). Unfortunately, it is generally difficult

to give an analytic expression for the optimal mapping $T^\star$, unless additional assumptions are made, such as assuming that both distributions are Gaussian, as in Appendix A.

4 Sequential Transport

4.1 Knothe-Rosenblatt Conditional Transport

As explained in Villani (2003); Carlier et al. (2010); Bonnotte (2013), the Knothe-Rosenblatt (KR) rearrangement is directly inspired by the Rosenblatt chain rule, from Rosenblatt (1952), and some extensions obtained on general measures by Knothe (1957). Using notations of Section 2.3 in Santambrogio (2015), let $\mu_{0:d}$ denote the marginal d-th measure, $\mu_{0:d-1|d}$ the conditional $d-1$-th measure (given x_d), $\mu_{0:d-2|d-1,d}$ the conditional $d-2$-th measure (given x_{d-1} and x_d), etc. Suppose that the μ_0-conditionals, corresponding to measures $\mu_{0:d}$, $\mu_{0:d-1|d}$, etc., are atomless (satisfied as soon as μ_0 is absolutely continuous with respect to the Lebesgue measure). For the first two,

$$\begin{aligned}\mu_0(\mathbb{R}^{d-1}\times \mathrm{d}x_d) &= \mu_{0:d}(\mathrm{d}x_d)\\ \mu_0(\mathbb{R}^{d-2}\times \mathrm{d}x_{d-1}\times \mathrm{d}x_d) &= \mu_{0:d}(\mathrm{d}x_d)\mu_{0:d-1|d}(\mathrm{d}x_{d-1}|x_d)\end{aligned}$$

and iterate. Define conditional (univariate) cumulative distribution functions:

$$\begin{cases} F_{0:d}(x_d) = \mu_{0:d}((-\infty, x_d]) = \mu_0(\mathbb{R}^{d-1}\times(-\infty,x_d])\\ F_{0:d-1|d}(x_{d-1}|x_d) = \mu_{0:d-1|d}((-\infty,x_{d-1}]|x_d),\end{cases}$$

etc. And similarly for μ_1. For the first component, let $T^\star_{\underline{d}}$ denote the monotone nondecreasing map transporting from $\mu_{0:d}$ to $\mu_{1:d}$, defined as $T^\star_{\underline{d}}(\cdot) = F^{-1}_{1:d}(F_{0:d}(\cdot))$. For the second component, let $T^\star_{\overline{d-1}}(\cdot|x_d)$ denote the monotone nondecreasing map transporting from $\mu_{0:d-1|d}(\cdot|x_d)$ to $\mu_{1:d-1|d}(\cdot|T^\star_{\underline{d}}(x_d))$, $T^\star_{\overline{d-1}}(\cdot|x_d) = F^{-1}_{1:d-1|d}(F_{0:d-1|d}(\cdot|x_d)|T^\star_{\underline{d}}(x_d))$. We can then repeat the construction, and finally, the KR rearrangement is

$$T_{\overline{kr}}(x_1,\cdots,x_d) = \begin{pmatrix} T^\star_{\underline{1}}(x_1|x_2,\cdots,x_d)\\ T^\star_{\underline{2}}(x_2|x_3,\cdots,x_d)\\ \vdots\\ T^\star_{\overline{d-1}}(x_{d-1}|x_d)\\ T^\star_{\underline{d}}(x_d)\end{pmatrix}.$$

As proved in Santambrogio (2015) and Carlier et al. (2010), $T_{\overline{kr}}$ is a transportation map from μ_0 to μ_1, in the sense that $\mu_1 = T_{\overline{kr}\#}\mu_0$. Following Bogachev et al. (2005) and Backhoff et al. (2017), $T_{\overline{kr}}$ is the "monotone upper triangular map" uniquely defined when the μ_1-conditionals are atomless for a chosen coordinate order. Bogachev et al. (2005) defined the "monotone lower triangular map,"

$$T_{\underline{kr}}(x_1,\cdots,x_d) = \begin{pmatrix} T^\star_{\underline{1}}(x_1)\\ T^\star_{\underline{2}}(x_2|x_1)\\ \vdots\\ T^\star_{\underline{d-1}}(x_{d-1}|x_1,\cdots,x_{d-2})\\ T^\star_{\underline{d}}(x_d|x_1,\cdots,x_{d-1})\end{pmatrix}.$$

The map $x_i \mapsto T^\star_{\underline{i}}(x_i|x_1,\cdots,x_{i-1})$ is monotone (nondecreasing) for all $(x_1,\cdots,x_{i-1}) \in \mathbb{R}^{i-1}$. Further, by construction, this KR transport map has a triangular Jacobian matrix $\nabla T_{\underline{kr}}$

with nonnegative entries on its diagonal, making it suitable for various geometric applications. However, this mapping does not satisfy many properties; for example, it is not invariant under isometries of $\mathbb{R}^d$ as mentioned in Villani (2009). Carlier et al. (2010) proved that the KR transport maps could be seen as limits of quadratic OTs. A direct interpretation is that this iterative sequential transport can be seen as "marginally optimal." Some explicit formulas can be obtained in the Gaussian case, as discussed in Appendix A.

4.2 Sequential Conditional Transport on a Probabilistic Graph

The "monotone lower triangular map," introduced in Bogachev et al. (2005) could be used when dealing with time series, since there is a natural ordering between variables, indexed by the time, as discussed in Backhoff et al. (2017) or Bartl et al. (2021). In the general non-temporal case of time series X_t, it is natural to extend that approach to acyclical probabilistic graphic models, following Cheridito and Eckstein (2023). Instead of two general measures μ_0 and μ_1 on $\mathbb{R}^d$, we use only measures "factorized according to $\mathcal{G}$," some probabilistic graphical model, as defined in Lauritzen (2020).

Definition. *Consider some acyclical causal graph $\mathcal{G}$ on $(s, \boldsymbol{x})$ where variables are topologically sorted, where $s \in \{0, 1\}$ is a binary variable, defining two measures μ_0 and μ_1 on $\mathbb{R}^d$, by conditioning on $s = 0$ and $s = 1$, respectively, factorized according to $\mathcal{G}$. Define*

$$T^\star_{\mathcal{G}}(x_1, \cdots, x_d) = \begin{pmatrix} T^\star_1(x_1) \\ T^\star_2(x_2|\text{ parents}(x_2)) \\ \vdots \\ T^\star_{d-1}(x_{d-1}|\text{ parents}(x_{d-1})) \\ T^\star_d(x_d|\text{ parents}(x_d)) \end{pmatrix}.$$

This mapping will be called "sequential conditional transport on the graph $\mathcal{G}$, "or shortly "sequential transport."[1]

A classical algorithm for topological sorting is Kahn (1962)'s "Depth First Search" (DFS), and other algorithms are discussed in Section 20.4 in Cormen et al. (2022). For the causal graphs of Figure 5:

$$T^\star_{\mathcal{G}}(x_1, x_2) = \begin{pmatrix} T^\star_1(x_1) \\ T^\star_2(x_2|x_1) \end{pmatrix}, \text{ for Figure 5a,}$$

$$T^\star_{\mathcal{G}}(x_1, x_2) = \begin{pmatrix} T^\star_1(x_1|x_2) \\ T^\star_2(x_2) \end{pmatrix}, \text{ for Figure 5b.}$$

In that simple case, for Figure 5a, we recognize the "monotone lower triangular map," and the "monotone upper triangular map," for 5b (see Section 4.1). Finally, for the causal graph on the German Credit dataset of Figure 1, variables are sorted, and

$$T^\star_{\mathcal{G}}(x_1, \cdots, x_7) = \begin{pmatrix} T^\star_1(x_1) \\ T^\star_2(x_2|x_1) \\ T^\star_3(x_3|x_1, x_2) \\ T^\star_4(x_4|x_2, x_3) \\ T^\star_5(x_5|x_1, x_2, x_4) \\ T^\star_6(x_6|x_3, x_5) \\ T^\star_7(x_7|x_1, x_4, x_5, x_6) \end{pmatrix}.$$

[1] Given the topological order of the graph and assuming the μ_0, μ_1-conditionals are atomless, the existence and unicity of the sequential transport map are guaranteed, as it involves fewer conditioning variables compared to the KR transport map.

Alternatively, using the “monotone lower triangular map" for the German Credit dataset to compute counterfactuals suggests that the assumed DAG contains more edges than the DAG illustrated in Figure 5. In this case, the edges are specified as $E = \{(i,j) \in V^2 : i < j\}$, with $V = \{s, x_1, x_2, \ldots, x_7\}$. Notably, multivariate OT corresponds to a fully connected graph, with $E = \{(i,j) \in V^2 : i \neq j\}$ Cheridito and Eckstein (2023). The impact of edge mispecifications on sequential transport is examined in Appendix E. If the graph is entirely unknown, one could infer it as discussed in Zheng et al. (2018); Yu et al. (2019); Cai et al. (2023), or use Bayes' rule to compute posterior DAGs, incorporating uncertainty quantification for counterfactuals Toth et al. (2022).

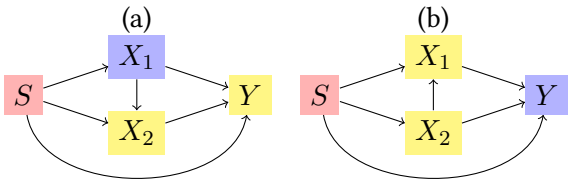


Figure 5: Two simple causal networks, with two legitimate mitigating variables, x_1 and x_2.

For the fairness application in the next section, s is treated as a “source” with no parents, allowing it to be the first vertex in the topological ordering of the network on $(s, \boldsymbol{x})$. The counterfactual value is then derived by propagating “downstream” in the causal graph as s changes from 0 to 1

4.3 Algorithm

Algorithm 1 Sequential transport on causal graph

Require: graph $\mathcal{G}$ on $(s, \boldsymbol{x})$, with adjacency matrix $\boldsymbol{A}$
Require: dataset $(s_i, \boldsymbol{x}_i)$ and one individual $(s = 0, \boldsymbol{a})$
Require: bandwidths $\boldsymbol{h}$ and $\boldsymbol{b}_j$'s
  $(s, \boldsymbol{v}) \leftarrow \boldsymbol{A}$ the topological ordering of vertices (DFS)
  $T_s \leftarrow$ identity
  for $j \in \boldsymbol{v}$ **do**
    $\boldsymbol{p}(j) \leftarrow$ parents(j)
    $T_j(\boldsymbol{a}_{\boldsymbol{p}(j)}) \leftarrow (T_{\boldsymbol{p}(j)_1}(\boldsymbol{a}_{\boldsymbol{p}(j)}), \cdots, T_{\boldsymbol{p}(j)_{k_j}}(\boldsymbol{a}_{\boldsymbol{p}(j)}))$
    $(x_{i,j|s}, \boldsymbol{x}_{i,\boldsymbol{p}(j)|s}) \leftarrow$ subsets when $s \in \{0,1\}$
    $w_{i,j|0} \leftarrow \phi(\boldsymbol{x}_{i,\boldsymbol{p}(j)|0}; \boldsymbol{a}_{\boldsymbol{p}(j)}, \boldsymbol{b}_j)$ (Gaussian kernel)
    $w_{i,j|1} \leftarrow \phi(\boldsymbol{x}_{i,\boldsymbol{p}(j)|1}; T_j(\boldsymbol{a}_{\boldsymbol{p}(j)}), \boldsymbol{b}_j)$
    $\hat{f}_{h_j|s} \leftarrow$ density estimator of $x_{\cdot,j|s}$, weights $w_{\cdot,j|s}$.
    $\hat{F}_{h_j|s}(\cdot) \leftarrow \int_{-\infty}^{\cdot} \hat{f}_{h_j|s}(u)\mathrm{d}u$, c.d.f.
    $\hat{Q}_{h_j|s} \leftarrow \hat{F}^{-1}_{h_j|s}$, quantile
    $\hat{T}_j(\cdot) \leftarrow \hat{Q}_{h_j|1} \circ \hat{F}_{h_j|0}(\cdot)$
  end for
  $\boldsymbol{a}^\star \leftarrow (T_1(\boldsymbol{a}_1), \cdots, T_d(\boldsymbol{a}_d))$
  return $(s = 1, \boldsymbol{a}^\star)$, counterfactual of $(s = 0, \boldsymbol{a})$

Algorithm 1 describes this sequential approach, which can be illustrated using the DAG in Figure 5a, as shown in Figure 6. The preliminary step is to determine the topological order of

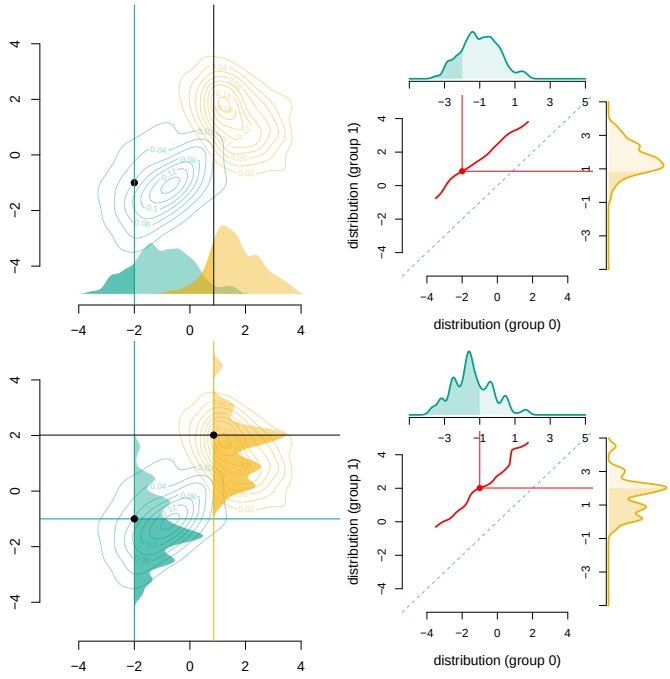


Figure 6: Illustration of Algorithm 1 for DAG in Figure 5a, with simulated data; first step at the top, second step at the bottom. The red square represents the multivariate OT of the bottom-left point.

the causal graph. In Figure 5a the order is $(s, (x_1, x_2))$. The first step is to estimate densities $\widehat{f}_{1|s}$ of x_1 in the two groups (s being either 0 or 1) as shown in the top left of Figure 6. Next, numerical integration and inverse are used to compute the cumulative distributions $\widehat{F}_{1|s}$ and quantile functions $\widehat{Q}_{1|s}$. To compute the counterfactual for $(s = 0, \boldsymbol{a})$, $a_1^\star$ is calculated as $\hat{T}_1(a_1)$, where $\hat{T}_1(\cdot) = \hat{Q}_{1|1} \circ \hat{F}_{1|0}(\cdot)$. The second step involves considering the second variable in the topological order, conditional on its parents. Suppose x_2 is the second variable, and for illustration that x_1 is the (only) parent of x_2. The densities $\widehat{f}_{2|s}$ of x_2 are then estimated in the two groups, conditional on their parents: either conditional on $x_1 = a_1$ (subgroup $s = 0$), or conditional on $x_1 = a_1^\star$ (subgroup $s = 1$). This is feasible since all transports of parents were computed in an earlier step. This can be visualized in the bottom left of Figure 6. As in the previous step, the conditional cumulative distributions $\widehat{F}_{2|s}$ and conditional quantile functions $\widehat{Q}_{2|s}$ (conditional on the parents) are computed. Then $a_2^\star$ is determined as $\hat{T}_2(a_2)$ where $\hat{T}_2(\cdot) = \hat{Q}_{2|1} \circ \hat{F}_{2|0}(\cdot)$. This process is repeated until all variables have been considered. At the end, starting from an individual with features $\boldsymbol{x} = \boldsymbol{a}$, in group $s = 0$, the counterfactual version in group $s = 1$ is obtained, with transported features, *mutatis mutandis*, $\boldsymbol{a}^\star$. As the number of parents per variable in the DAG increases, calculating conditional distributions for a variable becomes complex and less robust. Handling categorical variables in counterfactuals is detailed in Appendix B, enabling the application of sequential transport to datasets like `adult income` and COMPAS in Appendix D.

5 Interpretable Counterfactual Fairness

5.1 Individual Counterfactual Fairness

5.1.1 General Context

Following Dwork et al. (2012), a fair decision means that "similar individuals" are treated similarly. As discussed in the introduction, Kusner et al. (2017) and Russell et al. (2017) considered a "counterfactual fairness" criterion. Based on the approach discussed above, it is possible to quantify unfairness, for a single individual, of a model m, trained on features $(s, \boldsymbol{x})$ to predict an outcome y. If $y \in \{0, 1\}$ is binary, then m represents the underlying score, corresponding to the conditional probability that $y = 1$.

5.1.2 Illustration With Simulated Data

Consider the causal graphs in Figure 5, with one sensitive attribute s, two legitimate features x_1 and x_2 and one outcome y. Here, y is the score obtained from a logistic regression, specifically,

$$m(x_1, x_2, s) = \big(1 + \exp\big[-\big((x_1 + x_2)/2 + \mathbf{1}(s = 1)\big)\big]\big)^{-1}.$$

Iso-scores can be visualized at the top of Figure 7, with group 0 on the left, 1 on the right. Consider an individual $(s, x_1, x_2) = (s = 0, -2, -1)$ in group 0, with a score of 18.24% (bottom left of Figure 7). Using Algorithm 1, its counterfactual counterpart $(s = 1, x_1^\star, x_2^\star)$ can be constructed. The resulting score varies depending on the causal assumption. The score would be 40.95% assuming the causal graph of Figure 5a, and 54.06% assuming causal graph 5b. In the first case, the *mutatis mutandis* difference $m(s = 1, x_1^\star, x_2^\star) - m(s = 0, x_1, x_2)$, i.e., $+22.70\%$, is:

$$\begin{aligned} & m(s = 1, x_1, x_2) - m(s = 0, x_1, x_2) && : -10.65\% \\ + \;& m(s = 1, x_1^\star, x_2) - m(s = 1, x_1, x_2) && : +17.99\% \\ + \;& m(s = 1, x_1^\star, x_2^\star) - m(s = 1, x_1^\star, x_2) && : +15.37\%. \end{aligned}$$

The first term is the *ceteris paribus* difference, the second one the change in x_1 and the third one the change in x_2, conditional on the change in x_1. If, instead, we assume the causal graph of Figure 5b, the score of the same individual would become 54.06% and the *mutatis mutandis* difference $m(s = 1, x_1^\star, x_2^\star) - m(s = 0, x_1, x_2)$, i.e., $+35.82\%$, is:

$$\begin{aligned} & m(s = 1, x_1, x_2) - m(s = 0, x_1, x_2) && : -10.66\% \\ + \;& m(s = 1, x_1, x_2^\star) - m(s = 1, x_1, x_2) && : +16.07\% \\ + \;& m(s = 1, x_1^\star, x_2^\star) - m(s = 1, x_1, x_2^\star) && : +30.41\%. \end{aligned}$$

At the bottom right of Figure 7, the *mutatis mutandis* impact on the scores can be visualized.

5.2 Global Fairness Metrics

Instead of focusing on a single individual, it is possible to quantify the fairness of a model m on a global scale. For example, the Demographic Parity criterion can be extended to Counterfactual Demographic Parity (CDP), allowing fairness assessment within a population subgroup with $s = 0$. Consider the empirical version of "counterfactual fairness" in Kusner et al. (2017)

$$\text{CDP} = \frac{1}{n_0} \sum_{i \in \mathcal{D}_0} m(1, \boldsymbol{x}_i^\star) - m(0, \boldsymbol{x}_i), \tag{3}$$

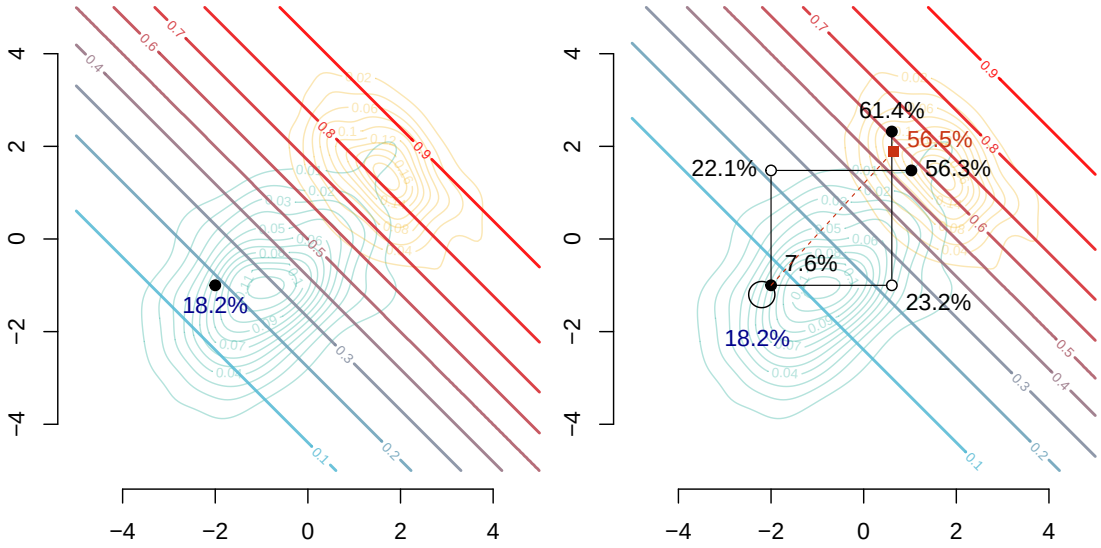


Figure 7: In the background, level curves for $(x_1, x_2) \mapsto m(0, x_1, x_2)$ and $m(1, x_1, x_2)$ respectively on the left and on the right. Then, on the left, individual $(s, x_1, x_2) = (s = 0, -2, -1)$ (predicted 18.2% by model m), and on the right, visualization of two counterfactuals $(s = 1, x_1^\star, x_2^\star)$ according to causal graphs 5a (bottom right path, predicted 61.4%) and 5b (top left path, predicted 56.3%). The red dot is the counterfactual obtained with multivariate OT (predicted 56.5%).

which corresponds to the "average treatment effect of the treated" in the classical causal literature. This can be computed more efficiently using Algorithm 2 in Appendix B, which offers a faster alternative compared to Algorithm 1. Other group fairness metrics, based on Equalized Odds, can be extended to aggregated counterfactual fairness measures (see Appendix C).

6 Application on Real Data

We analyze the Law School Admission Council dataset (Wightman, 1998), focusing on four variables: race $s \in \{\text{Black}, \text{White}\}$ (corresponding to 0 and 1), undergraduate GPA before law school (x_1, UGPA), Law School Admission Test (x_2, LSAT), and a binary response (y) indicating whether the first-year law school grade (FYA) is above the median, as described in Black et al. (2020). Unlike De Lara et al. (2024); Black et al. (2020); Kusner et al. (2017), we assume the causal graph in Figure 8, where UGPA influences LSAT. We aim to evaluate counterfactual fairness for Black individuals in logistic regression predictions ($\hat{y}|s = 0$), comparing an "aware" classifier, i.e., that includes s among the explanatory variables, with an "unaware" model that considers only $\boldsymbol{x} = (x_1, x_2)$. Fairness is measured using CDP (see Eq. 3). We apply the sequential transport method from Algorithm 2 to compute counterfactuals $\hat{y}^\star(s = 1)|s = 0$ following the network's topological order in Figure 8. These results are compared with those obtained from multivariate OT (De Lara et al., 2024) and quantile regressions (Plečko et al., 2024), namely Fairadapt.

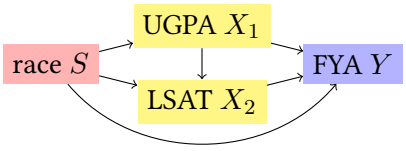


Figure 8: Causal graph of the Law School dataset.

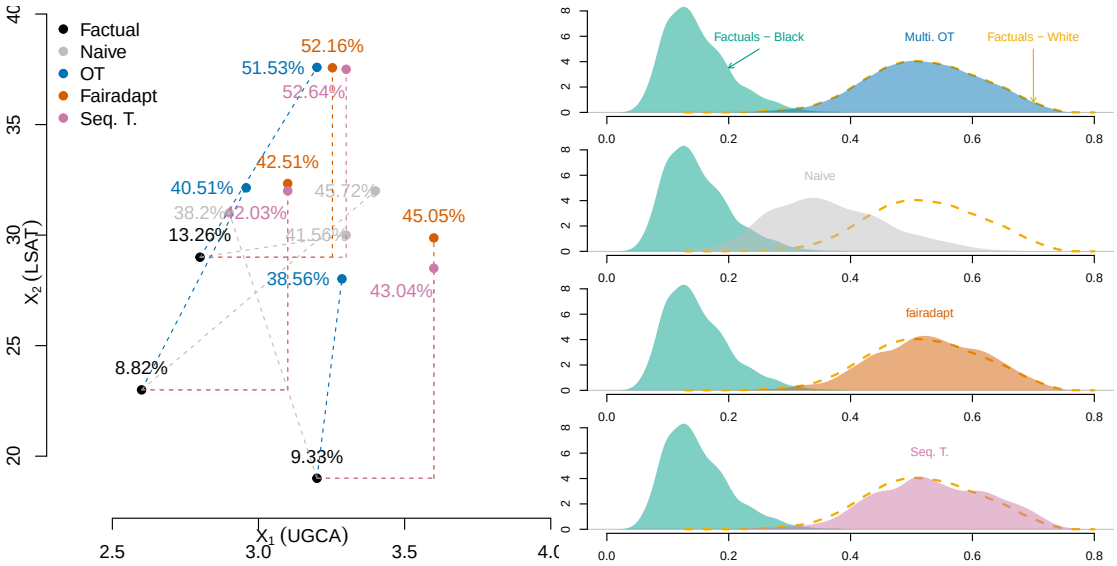


Figure 9: Counterfactual calculations for three Black individuals on the left (percentages indicate predicted scores), and densities of predicted scores (aware model) for all Black individuals with factuals and counterfactuals on the right. The dashed line represents the density of predicted scores for the observed White individuals.

	Fairadapt	multi. OT	seq. T
Aware model	0.3810	0.3727	0.3723
Unaware model	0.1918	0.1821	0.1817

Table 1: CDP *for Black individuals from Eq. 3 comparing classifier predictions over original features $\boldsymbol{x}$ (resp. $(s = 0, \boldsymbol{x})$) and their counterfactuals $\boldsymbol{x}^\star$ (resp. $(s = 1, \boldsymbol{x}^\star)$), using Fairadapt, multivariate OT, and sequential transport.*

Figure 9 illustrates the similarity between Fairadapt and sequential transport, both assuming a DAG, as shown by the counterfactual pathways for three Black individuals (left) and the alignment of counterfactual predicted score densities (right). The density of multivariate OT counterfactuals resembles factual White outcome distribution due to its matching process. Overall, the three methods yield similar results, as reflected in the aggregated counterfactual fairness metric in Table 1. Lastly, the "aware" model, which directly incorporates s into its covariates, is less counterfactually fair than the "unaware" model.

Conclusion

In this paper, we propose a sequential transport approach for constructing counterfactuals based on OT theory while respecting the underlying causal graph of the data. By using conditional univariate transport maps, we derive closed-form solutions for each coordinate of an individual's characteristics, which facilitates the interpretation of both individual counterfactual fairness of our predictive model, and global fairness through "counterfactual demographic parity." Future work could extend counterfactual fairness evaluation to mitigation by applying pre-processing or in-processing methods using sequential transport for counterfactual generation.

A Gaussian Case

The Gaussian case is the most simple one since mapping $T^\star$, corresponding to OT, can be expressed analytically (it will be a linear mapping). Furthermore, conditional distributions of a multivariate Gaussian distribution are Gaussian distributions, and that can be used to consider an iteration of simple conditional (univariate) transports, as a substitute to joint transport $T^\star$. Here Φ denotes the univariate cumulative distribution function of the standard Gaussian distribution $\mathcal{N}(0,1)$.

A.1 Univariate Optimal Gaussian Transport

One can easily prove that the optimal mapping, from a $\mathcal{N}(\mu_0,\sigma_0^2)$ to a $\mathcal{N}(\mu_1,\sigma_1^2)$ distribution is (see Figure 4):

$$x_1 = T^\star(x_0) = \mu_1 + \frac{\sigma_1}{\sigma_0}(x_0-\mu_0),$$

which is a nondecreasing linear transformation.

A.2 Multivariate Optimal Gaussian Transport

Recall that $\boldsymbol{X}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\boldsymbol{B}=\boldsymbol{\Sigma}^{-1}$, if its density, with respect to Lebesgue measure is

$$f(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^\top \boldsymbol{B}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right). \tag{4}$$

If $\boldsymbol{X}_0\sim\mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$ and $\boldsymbol{X}_1\sim\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$, the optimal mapping is also linear,

$$\boldsymbol{x}_1 = T^\star(\boldsymbol{x}_0) = \boldsymbol{\mu}_1 + \boldsymbol{A}(\boldsymbol{x}_0-\boldsymbol{\mu}_0),$$

where $\boldsymbol{A}$ is a symmetric positive matrix that satisfies $\boldsymbol{A}\boldsymbol{\Sigma}_0\boldsymbol{A}=\boldsymbol{\Sigma}_1$, which has a unique solution given by $\boldsymbol{A}=\boldsymbol{\Sigma}_0^{-1/2}\big(\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{1/2}\big)^{1/2}\boldsymbol{\Sigma}_0^{-1/2}$, where $\boldsymbol{M}^{1/2}$ is the square root of the square (symmetric) positive matrix $\boldsymbol{M}$ based on the Schur decomposition ($\boldsymbol{M}^{1/2}$ is a positive symmetric matrix), as described in Higham (2008). If $\boldsymbol{\Sigma}=\begin{pmatrix}1 & r\\ r & 1\end{pmatrix}$, and if $a=\sqrt{(1-\sqrt{1-r^2})/2}$, then:

$$\boldsymbol{\Sigma}^{1/2} = \begin{pmatrix}\sqrt{1-a^2} & a\\ a & \sqrt{1-a^2}\end{pmatrix}.$$

Observe further this mapping is the gradient of the convex function

$$\psi(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^\top\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{\mu}_0)+\boldsymbol{x}-\boldsymbol{\mu}_1^\top\boldsymbol{x}$$

and $\nabla T^\star=\boldsymbol{A}$ (see Takatsu (2011) for more properties of Gaussian transport). And if $\boldsymbol{\mu}_0=\boldsymbol{\mu}_1=\boldsymbol{0}$, and if $\boldsymbol{\Sigma}_0=\mathbb{I}$ and $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}$, $\boldsymbol{x}_1=T^\star(\boldsymbol{x}_0)=\boldsymbol{\Sigma}^{1/2}\boldsymbol{x}_0$. Hence, $\boldsymbol{\Sigma}^{1/2}$ is a linear operator that maps from $\boldsymbol{X}_0\sim\mathcal{N}(\boldsymbol{0},\mathbb{I})$ (the reference density) to $\boldsymbol{X}_1\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})$ (the target density).

A.3 Conditional Gaussian Transport

Alternatively, since $\boldsymbol{\Sigma}$ is a positive definite matrix, from the Cholesky decomposition, it can be written as the product of a lower (or upper) triangular matrix and its conjugate transpose,

$$\boldsymbol{\Sigma}=\boldsymbol{L}\boldsymbol{L}^\top=\boldsymbol{U}^\top\boldsymbol{U}.$$

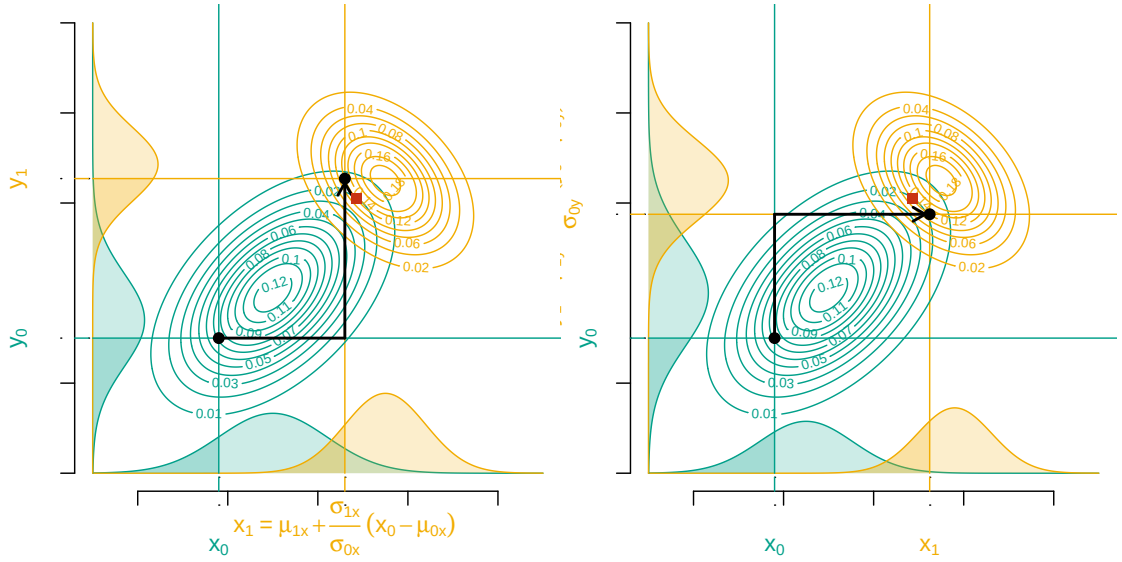


Figure 10: Two Gaussian conditional OTs. On the left-hand side, the process begins with a univariate transport along the x axis (using $T_x^\star$), followed by a transport along the y axis on the conditional distributions (using $T_{y|x}^\star$), corresponding to the "lower triangular affine mapping." On the right-hand side, the sequence is reversed: it starts with a univariate transport along the y axis (using $T_y^\star$) followed by transport along the x axis on the conditional distributions (using $T_{x|y}^\star$). The red square is the multivariate OT of the point in the bottom left, corresponding to the "upper triangular affine mapping."

Remark. *If* $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$, *then* $\boldsymbol{L} = \boldsymbol{\Sigma}_{2|1}^{1/2} = \begin{pmatrix} 1 & 0 \\ r & \sqrt{1-r^2} \end{pmatrix}$ *while* $\boldsymbol{U} = \boldsymbol{\Sigma}_{1|2}^{1/2} = \boldsymbol{\Sigma}_{2|1}^{1/2\top} = \boldsymbol{L}^\top$. *Then* $\boldsymbol{L}\boldsymbol{L}^\top = \boldsymbol{\Sigma} = \boldsymbol{U}^\top \boldsymbol{U}$.

Both $\boldsymbol{L}$ and $\boldsymbol{U}$ are linear operators that map from $\boldsymbol{X}_0 \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I})$ (the reference density) to $\boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ (the target density). $\boldsymbol{x}_0 \mapsto \boldsymbol{L}\boldsymbol{x}_0$ and $\boldsymbol{x}_0 \mapsto \boldsymbol{U}\boldsymbol{x}_0$ are respectively linear lower and upper triangular transport maps.

More generally, in dimension 2, consider the following (lower triangular) mapping $T(x_0, y_0) = (T_x(x_0), T_{y|x}(y_0|x_0))$,

$$\begin{aligned} &\mathcal{N}\left(\begin{pmatrix}\mu_{0x}\\ \mu_{0y}\end{pmatrix}, \begin{pmatrix}\sigma_{0x}^2 & r_0\sigma_{0x}\sigma_{0y}\\ r_0\sigma_{0x}\sigma_{0y} & \sigma_{0y}^2\end{pmatrix}\right) \\ &\xrightarrow{T} \mathcal{N}\left(\begin{pmatrix}\mu_{1x}\\ \mu_{1y}\end{pmatrix}, \begin{pmatrix}\sigma_{1x}^2 & r_1\sigma_{1x}\sigma_{1y}\\ r_1\sigma_{1x}\sigma_{1y} & \sigma_{1y}^2\end{pmatrix}\right), \end{aligned}$$

where

$$\begin{cases} T_x(x_0) = \mu_{1x} + \dfrac{\sigma_{1x}}{\sigma_{0x}}(x_0 - \mu_{0x}) \\ T_{y|x}(y_0) = \mu_{1y} + \dfrac{r_1\sigma_{1y}}{\sigma_{1x}}(T_x(x_0) - \mu_{1x}) \\ + \sqrt{\dfrac{\sigma_{0x}^2(\sigma_{1y}^2\sigma_{1x}^2 - r_1^2\sigma_{1y}^2)}{(\sigma_{0y}^2\sigma_{0x}^2 - r_0^2\sigma_{0y}^2)\sigma_{1x}^2}}(y_0 - \mu_{0y} - \dfrac{r_0\sigma_{0y}}{\sigma_{0x}}(x_0 - \mu_{0x})) \end{cases}$$

that are both linear mappings. This can be visualized on the left side of Figure 10.

Of course, this is highly dependent on the axis parametrization. Instead of considering projections on the axis, one could consider transport in the direction $\vec{u}$, followed by transport in the direction $\vec{u}^\perp$ (on conditional distributions). This can be visualized in Figure 11.

A.4 Gaussian Probabilistic Graphical Models

An interesting feature of the Gaussian multivariate distribution is that any marginal and any conditional distribution (given other components) is still Gaussian. More precisely, if

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix},\ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

then $\boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$, while, with notations of Eq. 4, we can also write $\boldsymbol{B}_1 = \boldsymbol{B}_{11} - \boldsymbol{B}_{12}\boldsymbol{B}_{22}^{-1}\boldsymbol{B}_{21}$ (based on properties of inverses of block matrices, also called the Schur complement of a block matrix). Furthermore, conditional distributions are also Gaussian, $\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2 \sim \mathcal{N}(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})$,

$$\begin{cases} \boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\boldsymbol{x}_2 - \boldsymbol{\mu}_2\right) \\ \boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \end{cases}$$

and the inverse of the conditional variance is simply $\boldsymbol{B}_{11}$.

It is well known that if $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $X_i \perp\!\!\!\perp X_j$ if and only if $\Sigma_{i,j} = 0$. More interestingly, we also have the following result, initiated by Dempster (1972):

Proposition A.1. *If $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with notations of Eq. 4, $\boldsymbol{B} = \boldsymbol{\Sigma}^{-1}$, $\boldsymbol{X}$ is Markov with respect to $\mathcal{G} = (E, V)$ if and only if $B_{i,j} = 0$ whenever $(i, j), (j, i) \notin E$.*

Proof. This is a direct consequence of the following property : if $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $X_i \perp\!\!\!\perp X_j|\boldsymbol{X}_{-i,j}$ if and only if $B_{i,j} = 0$ (since the log-density has separate terms in x_i and x_j). □

A.5 Sequential Transport

In the Gaussian case we obviously recover the results of Section A.3, if we plug Gaussian distributions in the expressions of Section 4

$$\begin{aligned} X_{0:1} &\sim \mathcal{N}(\mu_{0:1}, \sigma_{0:1}^2), \text{ hence } F_{0:1}(x) = \Phi\big(\sigma_{0:1}^{-1}(x - \mu_{0:1})\big) \\ X_{1:1} &\sim \mathcal{N}(\mu_{1:1}, \sigma_{1:1}^2), \text{ hence } F_{1:1}^{-1}(u) = \mu_{1:1} + \sigma_{1:1}\Phi^{-1}(u) \end{aligned}$$

thus

$$T_1^\star(x) = F_{1:1}^{-1}\big(F_{0:1}(x)\big) = \mu_{1:1} + \frac{\sigma_{1:1}}{\sigma_{0:1}}(x - \mu_{0:1}),$$

while

$$\begin{cases} X_{0:2}|x_{0:1} \sim \mathcal{N}(\mu_{0:2|1}, \sigma_{0:2|1}^2), \\ X_{1:2}|x_{0:1} \sim \mathcal{N}(\mu_{1:2|1}, \sigma_{1:2|1}^2), \end{cases}$$

i.e.,

$$\begin{cases} F_{0:2|1}(x) = \Phi\big(\sigma_{0:2|1}^{-1}(x - \mu_{0:2|1})\big), \\ F_{1:2|1}^{-1}(u) = \mu_{1:2|1} + \sigma_{1:2|1}\Phi^{-1}(u), \end{cases}$$

where we consider $X_{0:2}$ conditional to $X_{0:1} = x_{0:1}$ in the first place,

$$\begin{cases} \mu_{0:2|1} = \mu_{0:2} + \dfrac{\sigma_{0:2}}{\sigma_{0:1}}(x_{0:1} - \mu_{0:1}), \\ \sigma_{0:2|1}^2 = \sigma_{0:2}^2 - \dfrac{r_0^2\sigma_{0:2}^2}{\sigma_{0:1}^2}, \end{cases}$$

and $X_{1:2}$ conditional to $X_{1:1} = T_1^\star(x_{0:1})$ in the second place,

$$\begin{cases} \mu_{1:2|1} = \mu_{1:2} + \dfrac{\sigma_{1:2}}{\sigma_{1:1}}\big(T_1^\star(x_{0:1}) - \mu_{1:1}\big), \\ \sigma^2_{1:2|1} = \sigma^2_{1:2} - \dfrac{r_1^2\sigma^2_{1:2}}{\sigma^2_{1:1}}, \end{cases}$$

thus

$$T_{2|1}(x) = F^{-1}_{1:2|1}\big(F_{0:2|1}(x)\big) = \mu_{1:2|1} + \frac{\sigma_{1:2|1}}{\sigma_{0:2|1}}(x - \mu_{0:2|1}),$$

which is

$$\begin{aligned} &\mu_{1:2} + \frac{r_1\sigma_{1:2}}{\sigma_{1:1}}\Big(\mu_{1:1} + \frac{\sigma_{1:1}}{\sigma_{0:1}}(x_{0:1} - \mu_{0:1}) - \mu_{1:1}\Big) \\ &+\sqrt{\frac{\sigma^2_{0:1}(\sigma^2_{1:2}\sigma^2_{1:1} - r_1^2\sigma^2_{1:2})}{(\sigma^2_{0:2}\sigma^2_{0:1} - r_0^2\sigma^2_{0:2})\sigma^2_{1:1}}} \\ &\times\Big(x - \mu_{0:2} - \frac{r_0\sigma_{0:2}}{\sigma_{0:1}}(x_{0:1} - \mu_{0:1})\Big). \end{aligned}$$

A.6 General Conditional Transport

An interesting property of Gaussian vectors is the stability under rotations. In dimension 2, instead of a sequential transport of $\boldsymbol{x}$ on $\vec{e}_x$ and then (conditionally) on $\vec{e}_y$, one could consider a projection on any unit vector $\vec{u}$ (with angle θ), and then (conditionally) along the orthogonal direction $\vec{u}^\perp$. In Figure 11, we can visualize the set of all counterfactuals $\boldsymbol{x}^\star$ when $\theta \in [0, 2\pi]$. The (global) OT is also considered.

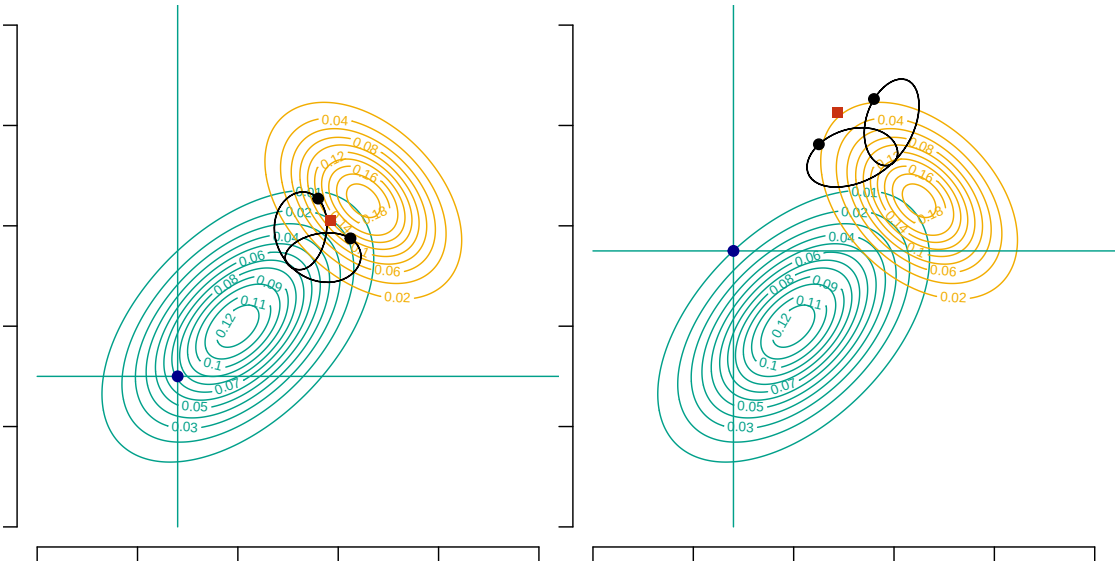


Figure 11: Gaussian conditional OTs. Each graph illustrates the transport starting from a different point (black point in the bottom left corner). The process begins with a univariate transport along the direction $\vec{u}$ (using $T^\star_{\vec{u}}$) followed by a transport along the orthogonal direction $\vec{u}^\perp$, on conditional distributions (using $T_{\vec{u}^\perp|\vec{u}}$). The curves in the upper right corner of each panel represent the set of all transport maps of the same point (bottom left corner) for all possible directions $\vec{u}$, the black points correspond to classical x (horizontal) and y (vertical) directions. The red point corresponds to the global OT.

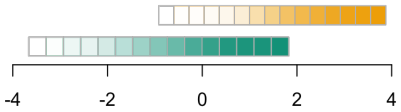


Figure 12: Visualization of vectors $F_{1|0}$ (below) and $F_{1|1}$ (on top) for the example of Figure 6, with $k = 15$.

B Algorithms

While Algorithm 1 is intuitive, it becomes inefficient when computing counterfactuals for thousands of individuals, as conditional densities, c.d.f.'s and quantile functions should be computed for each individual. An alternative is to compute these quantities on a grid, store them, and then retrieve them as needed. These objects are generated using Algorithm 2 (see also Figures 12 and 13, which visualize the vectors $F_{1|0}$, $F_{1|1}$, $F_{2|0}[\cdot, i]$ and $F_{2|1}[\cdot, j]$—with light colors corresponding to small probabilities and darker ones to large probabilities). Algorithm 3 computes the counterfactual for a given individual with $s = 0$ using the stored functions. In Algorithm 2 if j has no parents, $\text{parents}(j) = \varnothing$, $d_j = 0$ and then $F_{j|s}$ and $Q_{j|s}$ are vectors (of length k). In Algorithm 3, in that case, $\boldsymbol{i}_0 = \boldsymbol{i}_1 = \varnothing$.

Algorithm 2 Faster sequential transport on grids (1)

Require: graph on $(s, \boldsymbol{x})$, with adjacency matrix $\boldsymbol{A}$
Require: dataset $(s_i, \boldsymbol{x}_i)$ and $k \in \mathbb{N}$ some grid size,
Require: grids $\boldsymbol{g}_{j|s} = (g_{j,1|s}, \cdots, g_{j,k|s})$, for all variable j
Require: grid $\boldsymbol{u} = (1, \cdots, k)/(k+1)$, for all j
  $(s, \boldsymbol{v}) \leftarrow \boldsymbol{A}$ the topological ordering of vertices (DFS)
  for $j \in \boldsymbol{v}$ **do**
    $\boldsymbol{p}(j) \leftarrow$ parents(j), dimension d_j
    $\mathcal{G}_{j|s} \leftarrow$ grid $\boldsymbol{g}_{\boldsymbol{p}(j)_1|s} \times \cdots \times \boldsymbol{g}_{\boldsymbol{p}(j)_{d_j}|s}$
    $F_{j|s} \leftarrow$ tensors $k \times k^{d_j}$, taking values in $\boldsymbol{u}$
    $Q_{j|s} \leftarrow$ tensors $k \times k^{d_j}$, taking values in $\boldsymbol{g}_{j|s}$
    for $\boldsymbol{i} = (i_1, \cdots, i_{d_j}) \in \{1, \cdots, k\}^{d_j}$ **do**
      $F_{j|s}[\cdot, \boldsymbol{i}] \leftarrow$ c.d.f. of $X_j | \boldsymbol{X}_{\boldsymbol{p}(j)} = \boldsymbol{g}_{\boldsymbol{i}|s}, S = s$
      $Q_{j|s}[\cdot, \boldsymbol{i}] \leftarrow$ quantile of $X_j | \boldsymbol{X}_{\boldsymbol{p}(j)} = \boldsymbol{g}_{\boldsymbol{i}|s}, S = s$
    end for
  end for

Even though Algorithm 2 allows for calculating counterfactuals for a new observation without recalculating distribution quantities, it becomes complex as the number of parents of variables increases, resulting in a grid of dimension k^{d+1} for a node with d parents. To address this, we propose a novel algorithm, Algorithm 5, to compute counterfactuals using sequential transport, while still leveraging the quantities calculated during the training step for a new observation. It is based on Algorithm 4, which provides simple codes to estimate the empirical CDF and the empirical quantile function when observations are weighted. More precise estimates are obtained using Harrell (2024)'s `HMisc` R package and functions `wtd.stats` (inspired by Harrell and Davis (1982)). Algorithm 5 is similar to Algorithm 1, but weights are now calculated using distance metrics rather than Gaussian kernels.

Following Section 4, the existence and uniqueness of sequential transport maps are guaranteed when the source and target conditional distributions are atomless. However, in many practical scenarios, causal graphs include categorical variables (e.g., Figures 14 and 15), which

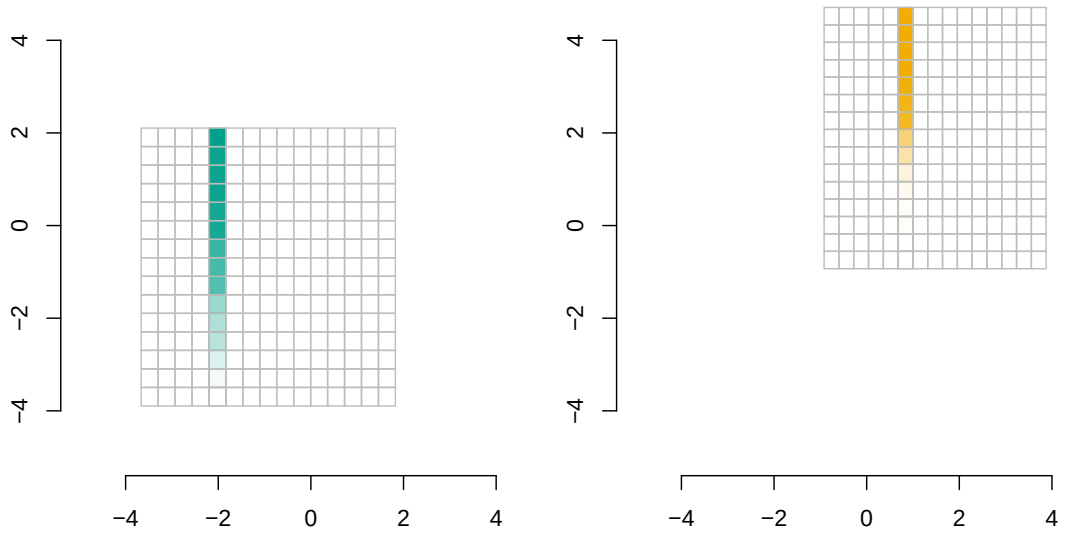


Figure 13: Visualization of matrices $F_{2|0}$ (on the left) and $F_{2|1}$ (on the right) for the example of Figure 6, with $k = 15$. Vertical vectors are $F_{2|0}[\cdot, i]$ and $F_{2|1}[\cdot, j]$.

Algorithm 3 Counterfactual calculation on causal graph (2)

Require: $F_{1|s}, \cdots, F_{d|s}$ and $Q_{d|s}, \cdots, Q_{d|s}$ (algorithm 2)
Require: grids $\boldsymbol{g}_{1|s}, \cdots, \boldsymbol{g}_{d|s}, \mathcal{G}_{1|s}, \cdots, \mathcal{G}_{d|s}$ and $\boldsymbol{u}$
Require: features $\boldsymbol{a} \in \mathbb{R}^d$ (group 0)
  $\boldsymbol{b} \leftarrow \boldsymbol{a}$
  for $j \in \boldsymbol{v}$ **do**
    $\boldsymbol{i}_0 \leftarrow \boldsymbol{a}_{\boldsymbol{p}(j)}$ on grid $\mathcal{G}_{j|0}$
    $k_0 \leftarrow a_j$ on grid $\boldsymbol{g}_{j|0}$
    $p \leftarrow F_{j|0}[k_0, \boldsymbol{i}_0]$
    $\boldsymbol{i}_1 \leftarrow \boldsymbol{b}_{\boldsymbol{p}(j)}$ on grid $\mathcal{G}_{j|1}$
    $k_1 \leftarrow p$ on grid $\boldsymbol{u}$
    $b_j \leftarrow Q_{j|1}[k_1, \boldsymbol{i}_1]$
  end for
  return $\boldsymbol{b}$ (counterfactual in group 1)

Algorithm 4 Weigthed ecdf and eqf

Require: n observations $\boldsymbol{x}$ and weights $\boldsymbol{w}$
  $\boldsymbol{x} \leftarrow$ sorted $\boldsymbol{x}$, and $\boldsymbol{w}$ accordingly (and set $x_0 = -\infty$)
Require: points x or probability level u
  $\overline{\boldsymbol{w}} \leftarrow$ cumulated sum of $\boldsymbol{w}$ ($\overline{w}_0 = 0$ and $\overline{w}_n = 1$)
  $\hat{F}[x; \boldsymbol{x}; \boldsymbol{w}] \leftarrow \overline{w}_{j-1}$ where j satisfies $x_{j-1} \leq x < x_j$
  $\hat{Q}[u; \boldsymbol{x}; \boldsymbol{w}] \leftarrow x_j$ where j satisfies $\overline{w}_{j-1} \leq u < \overline{w}_j$
  return $\hat{F}[x; \boldsymbol{x}; \boldsymbol{w}]$ (ecdf) and $\hat{Q}[u; \boldsymbol{x}; \boldsymbol{w}]$ (eqf)

leads to the loss of these theoretical guarantees. To address this issue in practice, if x_i is a categorical variable, we propose an alternative approach. First, we fit a multinomial model to predict the probabilities of category membership for x_i in the group $s = 1$. Next, using this model, we predict category probabilities for x_i based on the transported parent characteristics from group $s = 0$. For each prediction, we obtain a probability vector representing the likelihood of belonging to each category. Finally, we randomly draw a category using these probabilities as weights, thereby determining the transported category for the node x_i.

Algorithm 5 Sequential transport with weights

Require: graph $\mathcal{G}$ on $(s, \boldsymbol{x})$, with adjacency matrix $\boldsymbol{A}$
Require: dataset $(s_i, \boldsymbol{x}_i)$ and one individual $(s = 0, \boldsymbol{a})$
  $(s, \boldsymbol{v}) \leftarrow \boldsymbol{A}$ the topological ordering of vertices (DFS)
  $\boldsymbol{a}^\star \leftarrow \boldsymbol{a}$
  for $j \in \boldsymbol{v}$ **do**
    $\boldsymbol{p}(j) \leftarrow$ parents(j)
    $(x_{i,j|s}, \boldsymbol{x}_{i,\boldsymbol{p}(j)|s}) \leftarrow$ subsets when $s \in \{0, 1\}$
    $\boldsymbol{w}_{j|0} \leftarrow 1/\text{dist}(\boldsymbol{x}_{\boldsymbol{p}(j)|0}; \boldsymbol{a}_{\boldsymbol{p}(j)})$
    $\boldsymbol{w}_{j|1} \leftarrow 1/\text{dist}(\boldsymbol{x}_{\boldsymbol{p}(j)|1}; \boldsymbol{a}^\star_{\boldsymbol{p}(j)})$
    $\boldsymbol{a}^\star_j \leftarrow \hat{Q}[\hat{F}[\boldsymbol{a}_j; \boldsymbol{x}_{j|0}; \boldsymbol{w}_{j|0}]; \boldsymbol{x}_{j|1}; \boldsymbol{w}_{j|1}]$ (from Alg. 4)
  end for
  return $(s = 1, \boldsymbol{a}^\star)$, counterfactual of $(s = 0, \boldsymbol{a})$

C Fairness Metrics

In this section, we describe the calculation of aggregated counterfactual fairness measures, extending commonly used group fairness metrics such as Equality of Opportunity (EqOp) (Hardt et al., 2016), Class Balance (CB) or False Negative Rate (FNR) (Kleinberg, 2018), and Equal Treatment (EqTr) (Berk et al., 2021). To achieve this, the scoring classifier $m(\cdot)$ is transformed into a threshold-based classifier $m_t(\cdot)$, defined as $m_t(\cdot) = 1$ if $m(\cdot) > t$, and $m_t(\cdot) = 0$ otherwise, with the threshold set at $t = 0.5$.

The Counterfactual Equality of Opportinities (CEqOp) is defined as

$$\text{CEqOp} := \text{TPR}_0^\star - \text{TPR}_0,$$

where $\text{TPR}_0^\star$ is the True Positive Rate (TPR) in the sample $\mathcal{D}_0$ (group with $s = 0$) when predictions are made using the counterfactuals $m_t(1, \boldsymbol{x}^\star)$, and TPR_0 is the TPR in $\mathcal{D}_0$ when predictions are based on the individuals' original values in $\mathcal{D}_0$. A positive value of CEqOp indicates that the initial model is unfair towards the protected class.

The Counterfactual Class Balance (CCB or FNR)

$$\text{CCB(F)} := \frac{\text{TNR}_0^\star}{\text{TNR}_0}$$

where $\text{TNR}_0^\star$ is the True Negative Rate (TNR) of individuals in $\mathcal{D}_0$ calculated based on $m_t(1, \boldsymbol{x}^\star)$, and TNR_0 is the TNR of these individuals computed using $m_t(1, \boldsymbol{x})$.

Finally, Counterfactual Equal Treatment (CEqTr) corresponds to

$$\text{CEqTr} := \frac{\text{FPR}_0^\star}{\text{FNR}_0^\star} - \frac{\text{FPR}_0}{\text{FNR}_0}$$

where $\text{FPR}_0^\star$ and $\text{FNR}_0^\star$ are the False Positive Rate (FPR) and FNR computed based on the counterfactuals in the protected group, and FPR_0 and FNR_0 are their counterparts computed using the factual values for the same individuals.

D Additional Applications on Real Data

The application exercise from Section 6 is replicated here using two more complex datasets: `adult income` and COMPAS. These datasets include a greater number of variables as well as a mix of numerical and categorical variables. We use cleaned version of these datasets available in the `fairadapt` R package Plečko and Meinshausen (2020).

The first dataset is the `adult income` dataset, available from the UCI Machine Learning Repository Becker and Kohavi (1996). We follow the causal graph proposed by Plečko et al. (2024), reproduced in Figure 14. The binary target variable y indicates whether annual income exceeds \$50k, while the sensitive attribute s represents the gender of employees (protected group: $s =$ "female"; other group: $s =$ "male"). The other variables, $\boldsymbol{x}$, provide information on age (numerical), country of birth (categorical), marital status (categorical), education level (numerical), work class (categorical), weekly working hours (numerical), and occupation (categorical).

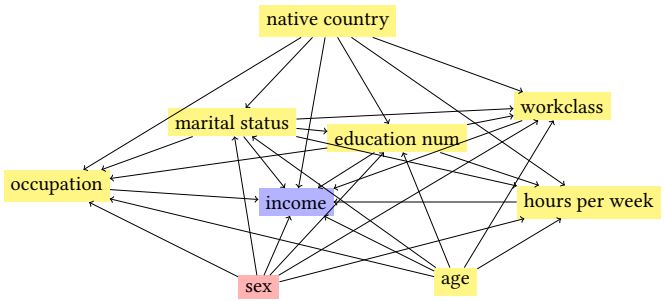


Figure 14: Causal graph of the `adult` dataset.

The second dataset, COMPAS Larson et al. (2016), contains individual-level information used to predict whether a criminal defendant is likely to reoffend within two years (y). Again, we follow the causal graph proposed by Plečko et al. (2024), reproduced in Figure 15. The sensitive attribute is race (protected value: $s =$ "Non-White"; other value: $s =$ "White"). The individual features $\boldsymbol{x}$ include age (numerical), gender (binary), the number of juvenile felonies (numerical), juvenile misdemeanors (numerical), other juvenile offenses (numerical), prior offenses (numerical), and the degree of charge (binary).

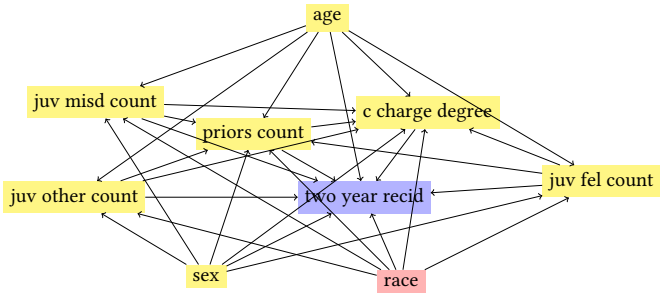


Figure 15: Causal graph of the COMPAS dataset.

For each dataset, we train a logistic regression model to predict the binary target variable y

using the features $\boldsymbol{x}$. Two versions of the model are considered: an "aware" model, where the sensitive attribute s is included as a feature, and an "unaware" model, where the model is blind to the sensitive attribute. These models are then employed to compare predictions based on the original features (factuals) and the counterfactual features. As in Section 6, we consider four types of counterfactuals: (i) naive (or *ceteris paribus*), where only the sensitive attribute for the protected group is changed ($s \leftarrow 1$); (ii) multivariate OT, where the features of the protected group are transformed using multivariate OT; (iii) fairadapt; and (iv) sequential transport based on the causal graphs shown in Figures 14 and 15.

The score distributions are plotted in Figure 16 for the `adult income` dataset (left) and the COMPAS dataset (right), for the "aware" model only (the results of the "unaware" model based on different counterfactuals for the protected group are provided in the Online Replication Ebook). The green curves represent the estimated density of scores predicted by the "aware" model using the factual features of individuals from the protected group. The dashed gold curves show the score distributions of the same model for the reference group. All other curves represent the estimated densities of scores predicted by the "aware" model based on the different counterfactuals for the protected group.

As observed with the `law school` dataset in the main part of the paper, the score distributions obtained from counterfactuals of the protected group using our sequential transport approach closely align with those obtained using fairadapt for the `adult income` dataset. The results are slightly more nuanced for the COMPAS dataset. Nevertheless, for both additional examples, the score distributions predicted by the "aware" model for counterfactual individuals approach the score distribution of the reference group. This stands in contrast to the distinct distribution observed for the protected group when counterfactuals are not considered.

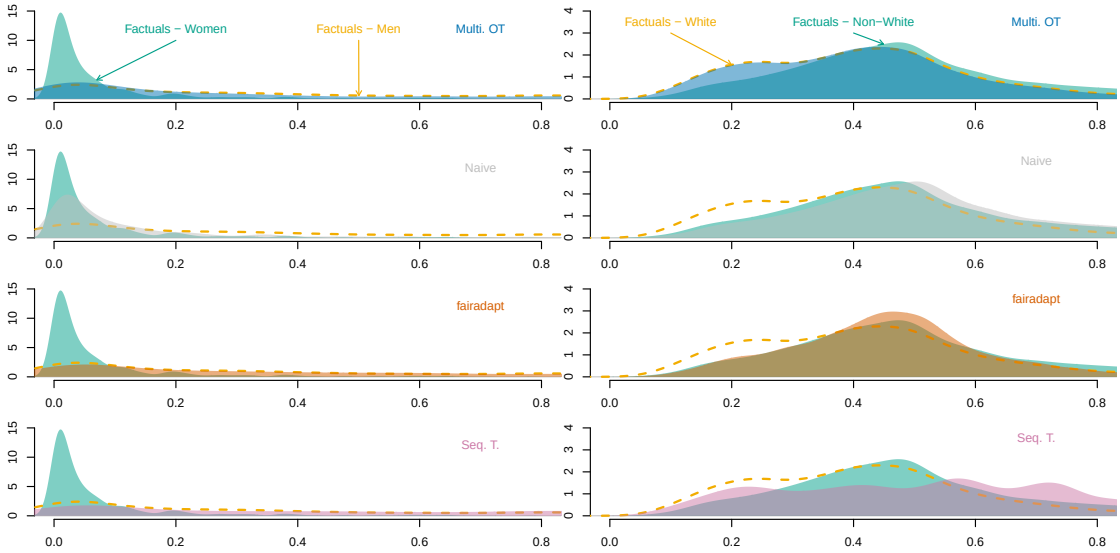


Figure 16: Densities of predicted scores (aware model) for all individuals from the minority class (Women on the left, Black individuals on the right) with factuals and counterfactuals, for the `adult` dataset (on the left) and the COMPAS datset (on the right). The dashed line represents the density of predicted scores for the observed individuals from the majority group.

Table 2 presents various metrics computed from the scores of both "aware" and "unaware" models, using either the initial observations or the counterfactuals, for these datasets. For each model, the first two rows report the true positive rate (TPR) and false positive rate (FPR) within the groups: protected group (either Women or Non-White, depending on the dataset), using

different counterfactual constructions, and other group (Men or White) in the final column. The scoring classifier $m(\cdot)$ is transformed into a threshold-based classifier $m_t(\cdot)$, where $m_t(\cdot) = 1$ if $m(\cdot) > t$, and $m_t(\cdot) = 0$ otherwise. The remaining rows provide the counterfactual fairness metrics presented in Appendix C, computed exclusively for individuals in group $\mathcal{D}_0$.

Table 3 provides the same metrics for the `law school` dataset, to complement Table 1.

E Wrong Causal Assumptions

In this section, we illustrate the impact of incorrect causal assumptions between two independent variables through an example. Similar to the approach in Section 5, we simulate a dataset $\{(y_i, s_i, \boldsymbol{x}_i)\}_{i=1}^n$ composed of two legitimate variables X_1 and X_2, an outcome Y, and a sensitive attribute S. The variables X_1 and X_2 are uniformly distributed and independent of each other. However, they depend on the sensitive attribute, as the distribution parameters vary according to S. Specifically:

$$\begin{cases} X_1 \sim \mathcal{U}(0,1), & X_2 \sim \mathcal{U}(0,1) \quad \text{if } S = 0, \\ X_1 \sim \mathcal{U}(1,2), & X_2 \sim \mathcal{U}(1,2) \quad \text{if } S = 1. \end{cases}$$

The outcome values are drawn from a Bernoulli distribution $Y_i \sim \mathcal{B}(p_i)$, where the success parameter $p_i \in [0,1]$ is individual-specific:

$$p_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)},$$

with

$$\eta_i = \begin{cases} 0.6x_1 + 0.2x_2, & \text{if } s_i = 0, \\ 0.4x_1 + 0.3x_2, & \text{if } s_i = 1. \end{cases}$$

The causal graph corresponding to this simulated data is shown in Figure 17(a). We investigate the impact of incorrect causal assumptions on sequential transport estimates by assuming that X_1 causes X_2 (Figure 17(b)) or, alternatively, that X_2 causes X_1 (Figure 17(c)).

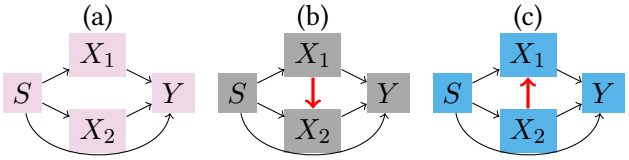


Figure 17: Causal assumptions on simulated data, with a correct assumption on the left and two wrong assumptions in the middle (where X_1 is assumed to cause X_2) and on the right (where X_2 is assumed to cause X_1).

We consider a hypothetical scoring model $m(\cdot)$, a logistic regression, which estimates the outcome based on the two covariates and the sensitive. Specifically:

$$m(x_1, x_2, s) = \big(1 + \exp\big[-\big((x_1 + x_2)/2 + \mathbf{1}(s = 1)\big)\big]\big)^{-1}.$$

The iso-curves of this scoring classifier are shown in Figure 18 for $s = 0$ (left) and $s = 1$ (right). This figure depicts the individual $(s = 0, x_1 = 0.5, x_2 = 0.5)$, predicted at 62.2% by the model $m(\cdot)$. When only the sensitive attribute is changed to $s = 1$, the naive model predicts a value of 81.8%. Under the correct causal assumption (purple point), the counterfactual values $(s = 1, x_1^\star, x_2^\star)$ are close to those obtained using multivariate OT. The model then predicts a

	Adult dataset					
	Women $s=0$	Naive $s\leftarrow 1$	OT $s\leftarrow 1$	Fairadapt $s\leftarrow 1$	Seq $s\leftarrow 1$	Men $s=1$
No. Obs.	662	662	662	662	662	1,338
Aware model						
TPR	0.26	0.44	0.67	0.74	0.61	0.54
FPR	0.01	0.02	0.15	0.23	0.32	0.10
		Individuals in $\mathcal{D}_0$				
CDP		0.05	0.17	0.24	0.29	
CEqOp		0.19	0.41	0.48	0.35	
CCB(FNR)		0.75	0.45	0.35	0.52	
CEqTr		0.04	0.44	0.86	0.81	
Unaware model						
TPR	0.41	0.41	0.63	0.74	0.56	0.54
FPR	0.02	0.02	0.14	0.21	0.31	0.10
		Individuals in $\mathcal{D}_0$				
CDP		0.00	0.12	0.20	0.25	
CEqOp		0.00	0.22	0.33	0.15	
CCB(FNR)		1.00	0.62	0.44	0.75	
CEqTr		0.00	0.35	0.80	0.67	
	COMPAS dataset					
	Non-White $s=0$	Naive $\leftarrow 1$	OT $\leftarrow 1$	Fairadapt $s\leftarrow 1$	Seq $s\leftarrow 1$	White $s=1$
No. Obs.	4,760	4,760	4,760	4,760	4,760	24,2454
Aware model						
TPR	0.57	0.67	0.39	0.53	0.65	0.45
FPR	0.23	0.31	0.18	0.22	0.41	0.19
		Individuals in $\mathcal{D}_0$				
CDP		0.02	-0.07	-0.02	0.02	
CEqOp		0.10	-0.18	-0.04	0.08	
CCB(FNR)		0.76	1.42	1.10	0.82	
CEqTr		0.41	-0.23	-0.07	0.62	
Unaware model						
TPR	0.61	0.61	0.35	0.47	0.62	0.40
FPR	0.25	0.25	0.14	0.17	0.38	0.15
		Individuals in $\mathcal{D}_0$				
CDP		0.00	-0.09	-0.05	0.00	
CEqOp		0.00	-0.26	-0.14	0.02	
CCB(FNR)		1.00	1.66	1.36	0.96	
CEqTr		0.00	-0.42	-0.33	0.36	

Table 2: Fairness metrics for the `adult` *dataset (top) and the COMPAS dataset (bottom), comparing classifier predictions based on original features $(s, \boldsymbol{x})$ and counterfactuals $(s = 1, \boldsymbol{x})$, constructed using different techniques: naive, OT, Fairadapt, and sequential transport. For metrics computed exclusively on individuals in $\mathcal{D}_0$ (Women or Non-White), the values obtained using counterfactuals are compared to those obtained using factuals.*

	Black $s=0$	Naive $s \leftarrow 1$	OT $s \leftarrow 1$	Fairadapt $s \leftarrow 1$	Seq $s \leftarrow 1$	White $s=1$
No. Obs.	1,282	1,282	1,282	1,282	1,282	18,285
Aware model						
TPR	0.00	0.15	0.64	0.66	0.68	0.65
FPR	0.00	0.08	0.57	0.63	0.64	0.51
			Individuals in $\mathcal{D}_0$			
CDP		0.22	0.37	0.38	0.39	
CEqOp		0.15	0.64	0.66	0.68	
CCB(FNR)		0.85	0.36	0.34	0.32	
CEqTr		0.10	1.59	1.86	2.04	
Unaware model						
TPR	0.11	0.11	0.60	0.62	0.62	0.60
FPR	0.07	0.07	0.52	0.56	0.56	0.45
			Individuals in $\mathcal{D}_0$			
CDP		0.00	0.18	0.19	0.20	
CEqOp		0.00	0.49	0.51	0.51	
CCB(FNR)		1.00	0.45	0.42	0.43	
CEqTr		0.00	1.22	1.41	1.40	

Table 3: Fairness metrics for the `law school` dataset, comparing classifier predictions based on original features $(s, \boldsymbol{x})$ and counterfactuals $(s=1, \boldsymbol{x})$, constructed using different techniques: naive, OT, Fairadapt, and sequential transport. For metrics computed exclusively on individuals in $\mathcal{D}_0$ (Black individuals), the values obtained using counterfactuals are compared to those obtained using factuals.

value of 92.5% with sequential transport under the correct causal assumption, which is close to the 90.4% predicted using the counterfactual constructed via OT.

When an incorrect causal assumption is made (gray point for assuming X_1 causes X_2, and light blue point for assuming X_2 causes X_1), the counterfactual values remain very close to those obtained with a correct assumption on the causal structure.

To gain a better understanding of this example beyond the analysis of a single point, Figure 19 presents the bivariate densities of the counterfactuals $(s=1, x_1^\star, x_2^\star)$ estimated using kernel density estimation. The densities are shown for counterfactuals obtained via OT (top left) and sequential transport, using the correct causal graph (top right) and incorrect causal assumptions (bottom). The estimated densities of the factual values are also displayed, in green for group $s=0$ and yellow for group $s=1$. The conclusions observed for the single point extend to the sample level. The density of the counterfactuals is very to the factual distribution either when the correct causal assumption is made or when this assumption is wrong.

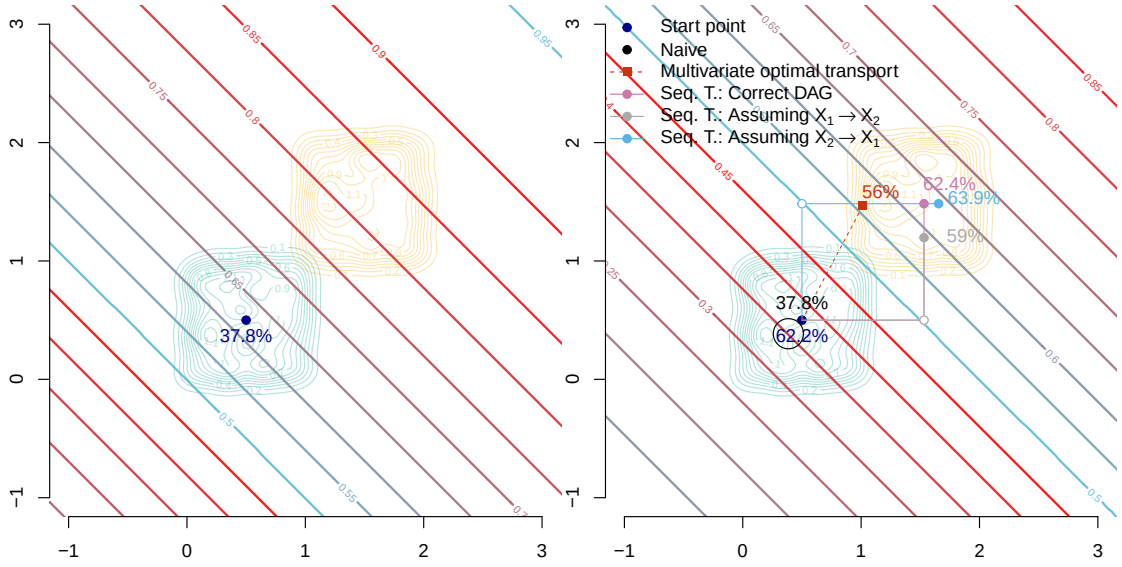


Figure 18: The iso-curves for $m(0, x_1, x_2)$ (left) and $m(1, x_1, x_2)$ (right) are shown in the background. The blue dot represents the individual $(s = 0, x_1 = 0.5, x_2 = 0.5)$, predicted at 62.2% by the model $m(\cdot)$. On the right, the purple dot corresponds to the counterfactual $(s = 1, x_1^\star, x_2^\star)$ obtained using sequential transport under the correct causal graph from Figure 17(a). Counterfactuals derived under incorrect assumptions (Figure 17(b) and Figure 17(c)) are depicted by the gray and blue dots, respectively. The red square represents the counterfactual obtained using multivariate OT.

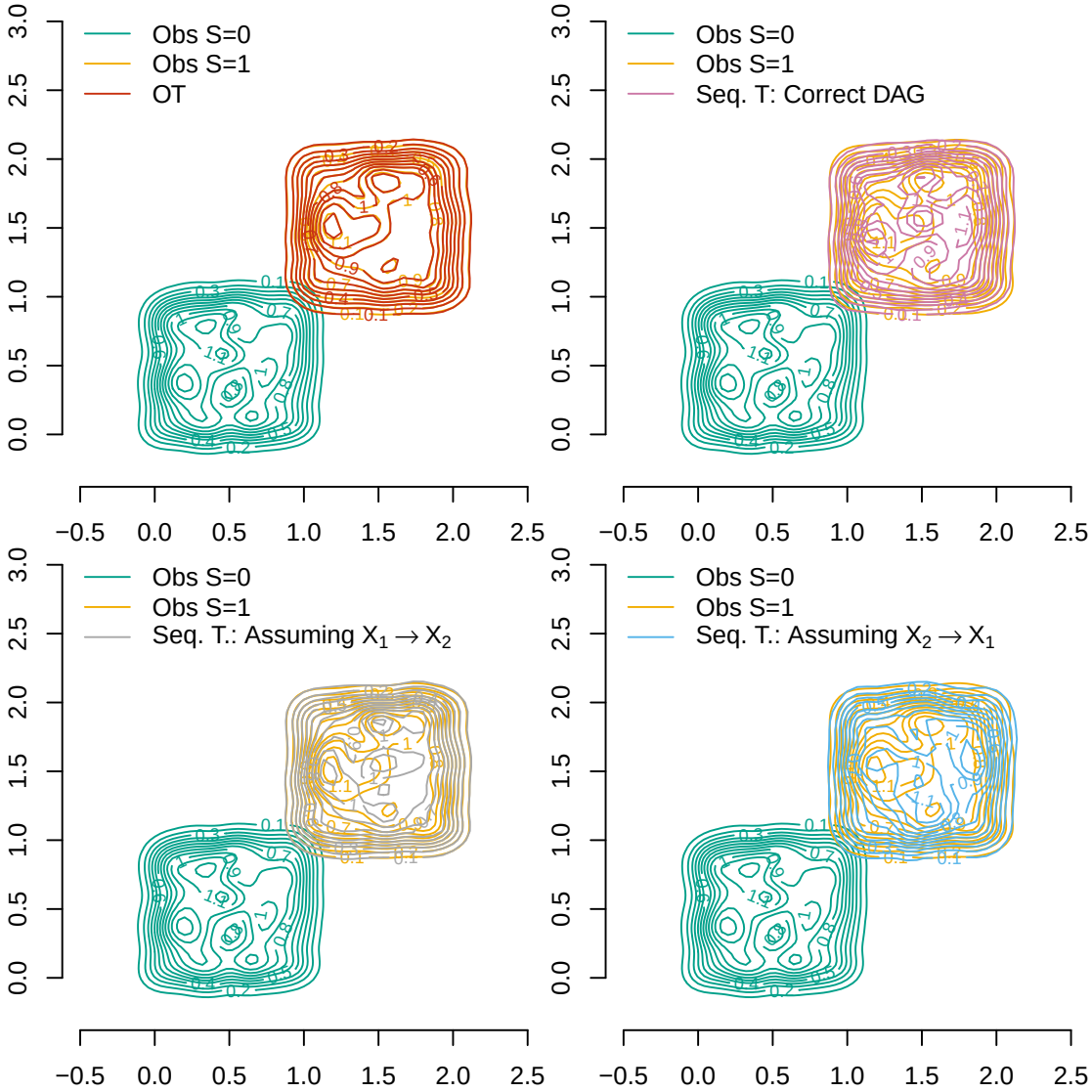


Figure 19: Estimated densities of the factuals in both groups and estimated densities using the counterfactuals either with optimal transport (top left), or sequential transport under a correct causal assumption (top right), a wrong assumption where X_1 causes X_2 (bottom left) and another wrong assumption where X_2 causes X_1 (bottom right).

References

Ahuja, R. K., Magnanti, T. L. and Orlin, J. B. (1993). *Network flows: Theory, algorithms, and applications*. Prentice Hall.

Ambrosio, L., Gigli, N. and Savaré, G. (2005). *Gradient flows: in metric spaces and in the space of probability measures*. Springer.

Backhoff, J., Beiglbock, M., Lin, Y. and Zalashko, A. (2017). Causal transport in discrete time and applications. *SIAM Journal on Optimization* 27: 2528–2562, doi:10.1137/16M1080197.

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Barocas, S., Hardt, M. and Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT press.

Bartl, D., Beiglböck, M. and Pammer, G. (2021). The wasserstein space of stochastic processes. doi:10.48550/arXiv.2104.14245.

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository, doi:10.24432/C5XW20, doi:10.24432/C5XW20.

Berk, R., Heidari, H., Jabbari, S., Kearns, M. and Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50: 3–44.

Black, E., Yeom, S. and Fredrikson, M. (2020). Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 111–121, doi:10.1145/3351095.3372845.

Bogachev, V. I., Kolesnikov, A. V. and Medvedev, K. V. (2005). Triangular transformations of measures. *Sbornik: Mathematics* 196: 309, doi:10.1070/SM2005v196n03ABEH000882.

Bongers, S., Forré, P., Peters, J. and Mooij, J. M. (2021). Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics* 49: 2885–2915, doi:10.1214/21-AOS2064.

Bonnotte, N. (2013). From Knothe's rearrangement to Brenier's optimal transport map. *SIAM Journal on Mathematical Analysis* 45: 64–87, doi:10.1137/120874850.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics* 44: 375–417, doi:10.1002/cpa.3160440402.

Bunne, C., Krause, A. and Cuturi, M. (2022). Supervised training of conditional Monge maps. *Advances in Neural Information Processing Systems* 35: 6859–6872.

Cai, H., Wang, Y., Jordan, M. and Song, R. (2023). On learning necessary and sufficient causal graphs. doi:10.48550/arXiv.2301.12389.

Cannon, A. J. (2018). Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic Environmental Research and Risk Assessment* 32: 3207–3225, doi:10.1007/s00477-018-1573-6.

Carlier, G., Galichon, A. and Santambrogio, F. (2010). From Knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis* 41: 2554–2576, doi:10.1137/080740647.

Charpentier, A., Flachaire, E. and Gallic, E. (2023). Optimal transport for counterfactual estimation: A method for causal inference. In *Optimal Transport Statistics for Economics and Related Topics*. Springer, 45–89, doi:10.1007/978-3-031-35763-3_3.

Cheridito, P. and Eckstein, S. (2023). Optimal transport and Wasserstein distances for causal models. doi:10.48550/arXiv.2303.14085.

Chernozhukov, V., Fernández-Val, I. and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica* 81: 2205–2268, doi:10.3982/ECTA10582.

Cordero-Erausquin, D. (2004). Non-smooth differential properties of optimal transport. *Contemporary Mathematics* 353: 61–72.

Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C. (2022). *Introduction to algorithms*. MIT

press.

De Lara, L., González-Sanz, A., Asher, N., Risser, L. and Loubes, J.-M. (2024). Transport-based counterfactual models. *Journal of Machine Learning Research* 25: 1–59.

Dempster, A. P. (1972). Covariance selection. *Biometrics* 28: 157–175, doi:10.2307/2528966.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226, doi:10.1145/2090236.2090255.

Hallin, M. and Konen, D. (2024). Multivariate quantiles: Geometric and measure-transportation-based contours. doi:10.48550/arXiv.2401.02499.

Hardt, M., Price, E. and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29: 3315–3323.

Harrell, F. E. (2024). Hmisc: Harrell Miscellaneous. R package version 5.2-1.

Harrell, F. E. and Davis, C. (1982). A new distribution-free quantile estimator. *Biometrika* 69: 635–640, doi:10.2307/2335999.

Higham, N. J. (2008). *Functions of matrices: theory and computation*. SIAM.

Hosseini, B., Hsu, A. W. and Taghvaei, A. (2023). Conditional optimal transport on function spaces. doi:10.48550/arXiv.2311.05672.

Kahn, A. B. (1962). Topological sorting of large networks. *Commun. ACM* 5: 558–562, doi:10.1145/368996.369025.

Kantorovich, L. V. (1942). On the translocation of masses. In *Doklady Akademii Nauk USSR*, *37*, 199–201.

Kearns, M. and Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D. and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in neural information processing systems* 30.

Kleinberg, J. (2018). Inherent trade-offs in algorithmic fairness. *SIGMETRICS Perform. Eval. Rev.* 46: 40, doi:10.1145/3292040.3219634.

Knothe, H. (1957). Contributions to the theory of convex bodies. *Michigan Mathematical Journal* 4: 39–52, doi:10.1307/mmj/1028990175.

Koenker, R. (2005). *Quantile regression*, *38*. Cambridge university press.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society* : 33–50doi:10.2307/1913643.

Koenker, R., Chernozhukov, V., He, X. and Peng, L. (2017). *Handbook of quantile regression*. CRC Press.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Kusner, M. J., Loftus, J., Russell, C. and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R. (eds), *Advances in Neural Information Processing Systems 30*. NIPS, 4066–4076.

Larson, S., Jeff ans Mattu, Kirchner, L. and Angwin, J. (2016). How we analyzed the compas recidivism algorithm.

Lauritzen, S. L. (2020). *Lectures on graphical models*. University of Copenhagen.

Ma, J., Guo, R., Zhang, A. and Li, J. (2023). Learning for Counterfactual Fairness from Observational Data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23. New York, NY, USA: Association for Computing Machinery, 1620–1630, doi:10.1145/3580305.3599408.

Ma, L. and Koenker, R. (2006). Quantile regression methods for recursive structural equation models. *Journal of Econometrics* 134: 471–506, doi:10.1016/j.jeconom.2005.07.003.

Meinshausen, N. and Ridgeway, G. (2006). Quantile regression forests. *Journal of machine learning research* 7.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris* .

Pearce, T., Jeong, J.-H., jia, y. and Zhu, J. (2022). Censored Quantile Regression Neural Networks for Distribution-Free Survival Analysis. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K. and Oh, A. (eds), *Advances in Neural Information Processing Systems*, *35*. Curran Associates, Inc., 7450–7461.

Pearl, J. (2000). Comment. *Journal of the American Statistical Association* 95: 428–431, doi:10.1080/01621459.2000.10474213.

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

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Plečko, D. and Meinshausen, N. (2020). Fair data adaptation with quantile preservation. *Journal of Machine Learning Research* 21: 1–44.

Plečko, D., Bennett, N. and Meinshausen, N. (2024). fairadapt: Causal reasoning for fair data preprocessing. *Journal of Statistical Software* 110: 1–35, doi:10.18637/jss.v110.i04.

Robertson, J., Hollmann, N., Awad, N. and Hutter, F. (2024). Fairpfn: Transformers can do counterfactual fairness. doi:10.48550/arXiv.2407.05732.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The annals of mathematical statistics* 23: 470–472.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* 100: 322–331, doi:10.1198/016214504000001880.

Russell, C., Kusner, M. J., Loftus, J. and Silva, R. (2017). When worlds collide: integrating different counterfactual assumptions in fairness. *Advances in neural information processing systems* 30.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians*. Springer, doi:10.1007/978-3-319-20828-2.

Shpitser, I., Richardson, T. S. and Robins, J. M. (2022). *Multivariate Counterfactual Systems and Causal Graphical Models*. New York, NY, USA: Association for Computing Machinery. 1st ed., 813–852, doi:10.1145/3501714.3501757.

Susser, M. (1991). What is a cause and how do we know one? A grammar for pragmatic epidemiology. *American Journal of Epidemiology* 133: 635–648, doi:10.1093/oxfordjournals.aje.a115939.

Takatsu, A. (2011). Wasserstein geometry of gaussian measures. *Osaka Journal of Mathematics* 48: 1005–1026.

Toth, C., Lorch, L., Knoll, C., Krause, A., Pernkopf, F., Peharz, R. and Kügelgen, J. von (2022). Active Bayesian Causal Inference. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K. and Oh, A. (eds), *Advances in Neural Information Processing Systems*, *35*. Curran Associates, Inc., 16261–16275, doi:10.48550/arXiv.2206.02063.

Villani, C. (2003). *Topics in optimal transportation*, *58*. American Mathematical Society.

Villani, C. (2009). *Optimal transport: old and new*, *338*. Springer, doi:10.1007/978-3-540-71050-9.

Watson, D. S., Gultchin, L., Taly, A. and Floridi, L. (2021). Local explanations via necessity and sufficiency: unifying theory and practice. In Campos, C. de and Maathuis, M. H. (eds), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, Proceedings of Machine Learning Research *161*. PMLR, 1382–1392.

Wightman, L. F. (1998). LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. Tech. rep., Law School Admission Council, Newtown, PA.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research* 20.

Wright, S. (1934). The method of path coefficients. *The annals of mathematical statistics* 5: 161–215.

Yu, Y., Chen, J., Gao, T. and Yu, M. (2019). DAG-GNN: DAG Structure Learning with Graph Neural Networks. In Chaudhuri, K. and Salakhutdinov, R. (eds), *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research *97*. PMLR, 7154–7163, doi:10.48550/arXiv.1904.10098.

Zech, J. and Marzouk, Y. (2022a). Sparse approximation of triangular transports, part I: The finite-dimensional case. *Constructive Approximation* 55: 919–986, doi:10.1007/s00365-022-09569-2.

Zech, J. and Marzouk, Y. (2022b). Sparse approximation of triangular transports, part II: The infinite-dimensional case. *Constructive Approximation* 55: 987–1036, doi:10.1007/s00365-022-09570-9.

Zheng, X., Aragam, B., Ravikumar, P. and Xing, E. P. (2018). DAGs with NO TEARS: continuous optimization for structure learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 9492–9503, doi:10.48550/arXiv.1803.01422.