Optimal Transport on Categorical Data for Counterfactuals Using Compositional Data and Dirichlet Transport

Agathe Fernandes Machado[1] , **Ewen Gallic**[2,3] and **Arthur Charpentier**[1]
[1]Département de Mathématiques, Université du Québec à Montréal, Montréal, QC Canada,
[2]CNRS - Université de Montréal CRM – CNRS
[3]Aix Marseille Univ, CNRS, AMSE, Marseille, France
fernandes_machado.agathe@courrier.uqam.ca, ewen.gallic@univ-amu.fr, charpentier.arthur@uqam.ca

Abstract

Recently, optimal transport-based approaches have gained attention for deriving counterfactuals, e.g., to quantify algorithmic discrimination. However, in the general multivariate setting, these methods are often opaque and difficult to interpret. To address this, alternative methodologies have been proposed, using causal graphs combined with iterative quantile regressions or sequential transport to examine fairness at the individual level, often referred to as "counterfactual fairness." Despite these advancements, transporting categorical variables remains a significant challenge in practical applications with real datasets. In this paper, we propose a novel approach to address this issue. Our method involves (1) converting categorical variables into compositional data and (2) transporting these compositions within the probabilistic simplex of the Euclidean space. We demonstrate the applicability and effectiveness of this approach through an illustration on real-world data, and discuss limitations.

1 Introduction

1.1 Counterfactuals

Counterfactual analysis, the third level in [Pearl, 2009]'s causal hierarchy, is widely used in machine learning, policy evaluation, economics and causal inference. It involves reasoning about "what could have happened" under alternative scenarios, providing insights into causality and decision-making effectiveness. An example could be the concept of counterfactual fairness, as introduced by [Kusner *et al.*, 2017], that ensures fairness by evaluating how decisions would change under alternative, counterfactual conditions. Counterfactual fairness focuses on mitigating bias by ensuring that sensitive attributes, such as race, gender, or socioeconomic status, do not unfairly influence outcomes.[1]

In the counterfactual problem, we consider data $\{(s_i, \mathbf{x}_i), i = 1, \cdots, n\}$, where s denotes a binary "treatment" (taking values in $\{0, 1\}$). With generic notations, the counterfactual version of $(0, \mathbf{x})$ can be constructed as $(1, T^\star(\mathbf{x}))$, where $T^\star$ is the optimal transport (OT) mapping from $\mathbf{X}|S = 0$ to $\mathbf{X}|S = 1$, as discussed in [Black *et al.*, 2020], [Charpentier *et al.*, 2023] and [De Lara *et al.*, 2024]. Unfortunately, this multivariate mapping is usually both complicated to estimate, and hard to interpret. If $\mathbf{x}$ is univariate, it is simply a quantile interpretation: if x is associated to rank probability u within group $s = 0$, then its counterfactual version should be associated with the same rank probability in group $s = 1$ (mathematically, $T^\star = F_1^{-1} \circ F_0$, where $F_j : \mathbb{R} \to [0, 1], j = \{0, 1\}$ denotes the cumulative distribution in group j, and F_j^{-1} is the generalized inverse, i.e., the quantile function). In higher dimensions, one could consider multivariate quantiles, as in [Hallin *et al.*, 2021] or [Hallin and Konen, 2024], but the heuristics is still hard to interpret. While OT-based counterfactual methods have been proposed to assess counterfactual fairness [Black *et al.*, 2020; De Lara *et al.*, 2024], an alternative approach introduced by [Plečko and Meinshausen, 2020] is grounded in causal graphs (DAGs). In this framework, the outcome y depends on variables $(s, \mathbf{x})$, where the sensitive attribute s "is a source" (a vertex without parents) and y is a "sink" (a vertex without outgoing edges). Recently, [Fernandes Machado *et al.*, 2025] unified these approaches by introducing sequential transport aligned with the "topological ordering" of a DAG.

For example, to test whether a predictor $\widehat{m}(\mathbf{x})$ is gender-neutral; let the sensitive attribute s be gender (binary genders for simplicity); compare its output on a woman's features $\mathbf{x}$ with that on her *mutatis mutandis* male counterpart. Unlike a *ceteris paribus* change, which flips s while holding all other features fixed, a *mutatis mutandis* intervention also adjusts any x_j causally influenced by s. Thus, if x_1 is height, the counterfactual of a 5'4" woman would not be a 5'4" man but, say, a 5'10" man, via an OT map. While OT handles continuous attributes naturally, categorical features (e.g. occupation or neighbourhood) lack a canonical distance. As a result, generating counterfactuals (e.g. the male counterpart of a female *nurse*, or where a Black *resident of X* would live if they were White) becomes particularly challenging.

1.2 The Case of Categorical Variables

For absolutely continuous variables, the approaches of [Plečko and Meinshausen, 2020; Plečko *et al.*, 2024] on the one hand (based on quantile regressions) and [Black *et al.*, 2020; Charpentier *et al.*, 2023; De Lara *et al.*, 2024;

[1]Extended Paper: https://arxiv.org/abs/2501.15549. Replication codes: https://github.com/fer-agathe/transport-simplex

Fernandes Machado *et al.*, 2025] (based on OT) are quite similar.

If [Plečko and Meinshausen, 2020] considered quantile regressions for absolutely continuous variables, the case of ordered categorical variables is considered (at least with some sort of meaningful ordering) in the section related to "Practical aspects and extensions." Discrete optimal transport between two marginal multinomial distributions is considered, but as discussed, it suffers multiple limitations. Here, we will consider an alternative approach, based on the idea of transforming categorical variables into continuous ones, coined "compositional variables" in [Chayes, 1971], and then, using "Dirichlet optimal transport," on those compositions.

While motivated by counterfactual fairness, the primary aim of this study is to present the core of a method for deriving counterfactuals for categorical data, applicable to any context requiring counterfactual analysis. Here, for simplicity, we have set aside considerations related to the assumption of a known Structural Causal Model (SCM).[2]

1.3 Agenda

After recalling notations on OT in Section 2, we discuss how to transform categorical variables with d categories into variables taking values in the simplex $\mathcal{S}_d$ in $\mathbb{R}^d$, i.e., compositional variables, in Section 3. In Section 4, we review the topological and geometrical properties of the probability simplex $\mathcal{S}_d \subset \mathbb{R}^d$. In Section 5, we introduce the first methodology, which transports distributions within $\mathcal{S}_d$ via Gaussian OT. This approach relies on an alternative representation of probability vectors in the Euclidean space $\mathbb{R}^{d-1}$ and assumes approximate normality in the transformed space. In Section 6, we present a second methodology, which operates directly on $\mathcal{S}_d$ using a tailored cost function instead of the standard quadratic cost. Theoretical aspects of this "Dirichlet transport" framework are discussed in Section 6.1, while empirical strategies for matching categorical observations are developed in Section 6.2. Section 7 provides two empirical illustrations using the `German Credit` and `Adult` datasets.

Our main contributions can be summarized as follows:

- We propose a novel method to handle categorical variables in counterfactual modeling by using optimal transport directly on the simplex. This approach transforms categorical variables into compositional data, enabling the use of probabilistic representations that preserve the geometric structure of the simplex.
- By integrating optimal transport techniques on this domain, the method ensures consistency with the properties of compositional data and offers a robust framework for counterfactual analysis in real-world scenarios.
- Our approach does not require imposing an arbitrary order on the labels of categorical variables.

2 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 $\varphi : \mathcal{X}_1 \to \mathbb{R}$,

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

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]. 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

$$\inf_{T_{\#}\mu_0 = \mu_1} \int_{\mathcal{X}_0} c\big(\mathbf{x}_0, T(\mathbf{x}_0)\big)\mu_0(\mathrm{d}\mathbf{x}_0), \tag{1}$$

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 [Kantorovich, 1942],

$$\inf_{\pi \in \Pi(\mu_0, \mu_1)} \int_{\mathcal{X}_0 \times \mathcal{X}_1} c(\mathbf{x}_0, \mathbf{x}_1)\pi(\mathrm{d}\mathbf{x}_0, \mathrm{d}\mathbf{x}_1), \tag{2}$$

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. Observe that $T^\star$ is an "increasing mapping," in the sense of being the gradient of a convex function, from [Brenier, 1991]). Finally, one should have in mind the the cost function c is related to the geometry of sets $\mathcal{X}$.

3 From Categorical to Compositional Data

Using the notations of the introduction, consider a dataset $\{s, \mathbf{x}\}$ where features $\mathbf{x}$ are either numerical (assumed to be "continuous"), or categorical. In the latter case, suppose that $\mathbf{x}_j$ takes values in $\{x_{j,1}, \cdots, x_{j,d_j}\}$, or more conveniently, $[\![d_j]\!] = \{1, \cdots, d_j\}$, corresponding to the d_j categories (as in the standard "One Hot" encoding).

The aim is to transform a categorical variable x, which takes values in $[\![d]\!]$, into a numerical one in the simplex $\mathcal{S}_d$. To achieve this, we suggest using a probabilistic classifier. This classifier is based on the other features in $\mathbf{x}$, denoted by $\mathcal{X}_{-x}$.Mathematically, we consider a mapping from $\mathcal{X}_{-x}$ to $\mathcal{S}_d$ (and not to $[\![d]\!]$ as in a standard multiclass classifier). The most natural model for this transformation is the Multinomial Logistic Regression (MLR), which is based on the "softmax" loss function. To normalize the output of the classifier into the simplex, we define the closure operator $\mathcal{C} : \mathbb{R}_+^d \to \mathcal{S}_d$ as

$$\mathcal{C}[x_1, x_2, \ldots, x_d] = \left[\frac{x_1}{\sum_{i=1}^d x_i}, \frac{x_2}{\sum_{i=1}^d x_i}, \cdots, \frac{x_d}{\sum_{i=1}^d x_i}\right],$$

or shortly

$$\mathcal{C}(\mathbf{x}) = \frac{\mathbf{x}}{\mathbf{x}^\top 1},$$

[2] Details on how the method can be integrated within an SCM are discussed in Appendix C of the extended version of the paper.

Algorithm 1 From categorical variables into compositions.

Input: training dataset $\mathcal{D} = \{(s_i, \mathbf{x}_i)\}$
Input: new observation $(s, \mathbf{x})$, with $\mathbf{x}_j$'s either in $\mathbb{R}$ or $[\![d_j]\!]$
Output: $(s, \tilde{\mathbf{x}})$, with $\tilde{\mathbf{x}}_j$'s either in $\mathbb{R}$ or $\mathcal{S}_{d_j}$

```
for j ∈ {1, ··· , k} do
  if x_j ∈ [[d_j]] then
    estimate a MLR to predict categorical x_j using D
    get estimates β̂_2, ··· , β̂_{d_j}
    x̃_j ← C(1, e^{x_j^⊤ β̂_2}, ··· , e^{x_j^⊤ β̂_d})
  else
    x̃_j ← x_j
  end if
end for
```

GAM-MLR (1)

x	$\tilde{x}_{\mathrm{C}}$	$\tilde{x}_{\mathrm{E}}$	$\tilde{x}_{\mathrm{O}}$
E	18.38%	**61.56%**	20.06%
C	40.86%	**42.38%**	16.76%
E	19.41%	**70.82%**	9.77%
C	**47.04%**	26.83%	26.13%

random forest

$\tilde{x}_{\mathrm{C}}$	$\tilde{x}_{\mathrm{E}}$	$\tilde{x}_{\mathrm{O}}$
23.68%	**46.32%**	30.00%
34.68%	**36.42%**	28.90%
16.87%	**76.51%**	6.63%
53.16%	26.84%	20.00%

GAM-MLR (2)

x	$\tilde{x}_{\mathrm{C}}$	$\tilde{x}_{\mathrm{E}}$	$\tilde{x}_{\mathrm{O}}$
E	9.22%	**75.92%**	14.86%
C	**46.80%**	24.06%	29.14%
E	11.23%	**79.07%**	9.71%
C	**50.74%**	26.98%	22.28%

gradient boosting model

$\tilde{x}_{\mathrm{C}}$	$\tilde{x}_{\mathrm{E}}$	$\tilde{x}_{\mathrm{O}}$
11.25%	**68.51%**	20.24%
61.14%	13.10%	25.76%
12.48%	**75.58%**	11.94%
51.12%	25.17%	23.71%

Table 1: Mappings from the `purpose` categorical variable x to the compositional one $\tilde{\mathbf{x}}$, (in the `german credit` dataset), for the first four individuals of the dataset. The first two models are GAM-MLR (multinomial model with splines for continuous variables), then, a random forest, and a boosting algorithm.

where 1 is a vector of ones in $\mathbb{R}^d$. Then, in the MLR model, the transformation $\widehat{T} : \mathcal{X}_{-x} \to \mathcal{S}_d$ is given by

$$\widehat{T}(\mathbf{x}) = \mathcal{C}(1, e^{\mathbf{x}^\top \widehat{\boldsymbol{\beta}}_2}, \cdots, e^{\mathbf{x}^\top \widehat{\boldsymbol{\beta}}_d}) \in \mathcal{S}_d,$$

where $\widehat{\boldsymbol{\beta}}_2, \ldots, \widehat{\boldsymbol{\beta}}_d$ are the estimated coefficients for each category, and the first category is taken as the reference. This procedure is described in Algorithm 1.

As an illustration, consider the `purpose` variable from the `German` dataset. For simplicity, this variable has been reduced to three categories: C, E, O (representing cars, equipment, and other, respectively). More details on the dataset are provided in Section 7.1. The `purpose` variable is converted into a continuous variable using four models: (i) a GAM-MLR with splines for three continuous variables, (ii) a GAM-MLR incorporating these variables and seven categorical ones, (iii) a random forest, and (iv) a gradient boosting model. Table 1 presents the observed values in the first column for each model, along with the estimated scores for each category in the three remaining columns, corresponding to the transformed values $T^\star(\mathbf{x})$.

Note that if we want to go back from compositions to categories, the standard approach is based on the majority (or argmax) rule.

In the rest of the paper, given a dataset $\{s, \mathbf{x}\}$, all categorical variables are transformed into compositions, so that $\mathcal{X}$ is a product space of sets that are either $\mathbb{R}$ for numerical variables or $\mathcal{S}_d$ (type) for compositions (d will change according to the number of categories).

In fact, for privacy issues, a classical strategy is to consider aggregated data on small groups (usually on a geographic level, per block, or per zip code), even if there is an ecological fallacy issue (that occurs when conclusions about individual behaviour or characteristics are incorrectly drawn based on aggregate data for a group, see [King *et al.*, 2004]). Hence, using "compositional data" is quite natural in many cases, as unobserved categorical variables can often be represented as compositions predicted from observed variables serving as proxies. For example, in U.S. datasets, racial information about individuals may not always be available. However, the proportions of groups such as "White and European," "Asian," "Hispanic and Latino," "Black or African American," etc., within a neighbourhood may be observed instead (see, e.g., [Cheng *et al.*, 2010], [Naeini *et al.*, 2015] and [Zadrozny and Elkan, 2001] for more general discussions, or [Imai *et al.*, 2022] about the use of predicted probabilities when categories are not observed).

4 Topology and Geometry of the Simplex

The standard simplex of $\mathbb{R}^d$ is the regular polytope $\mathcal{S}_d = \left\{\mathbf{x} \in \mathbb{R}^d_+ \,\middle|\, \mathbf{x}^\top \mathbf{1} = 1\right\}$, but for convenience, consider the open version of that set,

$$\mathcal{S}_d = \left\{\mathbf{x} \in (0,1)^d \,\middle|\, \mathbf{x}^\top \mathbf{1} = 1\right\}.$$

Following [Aitchison, 1982], define the inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{d} \sum_{i<j} \log \frac{x_i}{x_j} \log \frac{y_i}{y_j} \qquad \forall x, y \in \mathcal{S}_d, \tag{3}$$

and the simplex becomes a metric vector space if we consider the associated "Aitchison distance," as coined in [Pawlowsky-Glahn and Egozcue, 2001]. Figure 1 shows $n = 61$ points in $\mathcal{S}_3$. Each point $\mathbf{x}$ can be seen as a probability vector over $\{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$, drown either from a distribution $\mathbb{P}_0$ for red points or $\mathbb{P}_1$ for blue points.

If we define the binary operator $\diamond$ on $\mathcal{S}_d$,

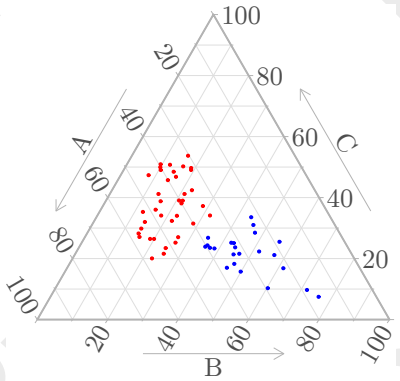
$$\mathbf{x} \diamond \mathbf{y} = \left[\frac{x_1 y_1}{\sum_{i=1}^d x_i y_i}, \cdots, \frac{x_d y_d}{\sum_{i=1}^d x_i y_i}\right],$$

then $(\mathcal{S}_d, \diamond)$ is a commutative group, with identity element $d^{-1}\mathbf{1}$, and the inverse of $\mathbf{x}$ is

$$\mathbf{x}^{-1} = \left[\frac{1/x_1}{\sum_{i=1}^d 1/x_i}, \cdots, \frac{1/x_d}{\sum_{i=1}^d 1/x_i}\right] = \mathcal{C}(1/\mathbf{x}).$$

5 Using an Alternative Representation of Simplex Data

A first strategy to define a transport mapping could be to use some isomorphism, $h : \mathcal{S}_d \to \mathcal{E}$ and then define the inverse mapping $h^{-1} : \mathcal{E} \to \mathcal{S}_d$, where $\mathcal{E}$ is some Euclidean space, classically $\mathbb{R}^{d-1}$, where the standard quadratic cost can be considered. This idea corresponds to the dual transport problem in [Pal and Wong, 2018].



	A	B	C
•	29.83%	48.60%	21.56%
•	43.71%	18.57%	37.72%

Figure 1: $n = 61$ points in $\mathcal{S}_3$, with a toy dataset.

5.1 Classical Transformations

The additive log ratio (alr) transform is an isomorphism where $\mathrm{alr} : \mathcal{S}_d \to \mathbb{R}^{d-1}$, given by

$$\mathrm{alr}(\mathbf{x}) = \left[\log \frac{x_1}{x_d}, \cdots, \log \frac{x_{d-1}}{x_d}\right].$$

Its inverse is, for any $\mathbf{z} \in \mathbb{R}^{d-1}$,

$$\mathrm{alr}^{-1}(\mathbf{z}) = \mathcal{C}(\exp(z_1), \cdots, \exp(z_{d-1}), 1) = \mathcal{C}\big(\exp([\mathbf{z}, 0])\big).$$

Such a map, from $\mathcal{S}_d$ to $\mathbb{R}^{d-1}$ is related to the so-called "exponential coordinate system" of the unit simplex, in [Pal, 2024]. The center log ratio (clr) transform is both an isomorphism and an isometry where $\mathrm{clr} : S^d \to \mathbb{R}^d$,

$$\mathrm{clr}(\mathbf{x}) = \left[\log \frac{x_1}{\overline{\mathbf{x}}_g}, \cdots, \log \frac{x_D}{\overline{\mathbf{x}}_g}\right],$$

where $\overline{\mathbf{x}}_g$ denotes the geometric mean of $\mathbf{x}$. Observe that the inverse of this function is the softmax function, i.e.,

$$\mathrm{clr}^{-1}(\mathbf{z}) = \mathcal{C}(\exp(z_1), \cdots, \exp(z_d)) = \mathcal{C}\big(\exp(\mathbf{z})\big),\ \mathbf{z} \in \mathbb{R}^d.$$

Finally, the isometric log ratio (ilr) transform, defined in [Egozcue *et al.*, 2003], is both an isomorphism and an isometry where $\mathrm{ilr} : \mathcal{S}_d \to \mathbb{R}^{d-1}$,

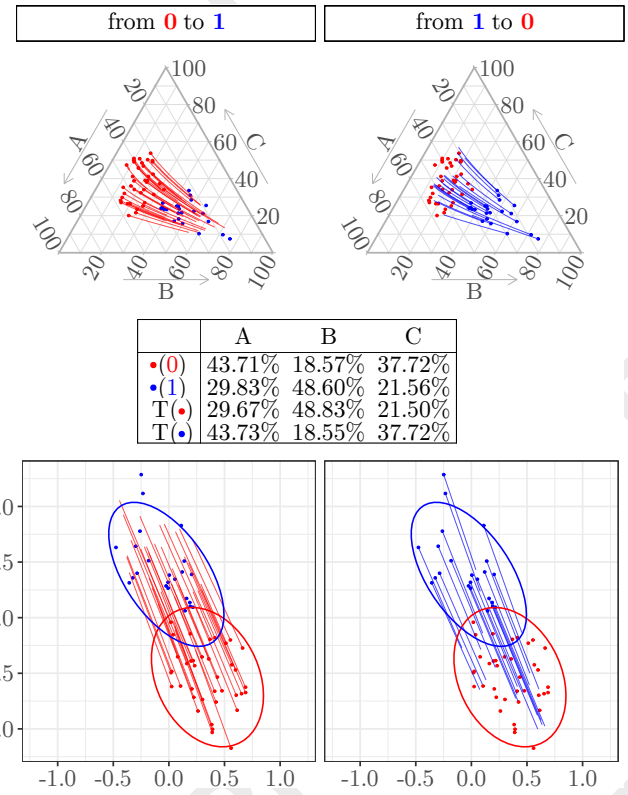
$$\mathrm{ilr}(\mathbf{x}) = \big[\langle \mathbf{x}, \vec{e}_1\rangle, \cdots, \langle \mathbf{x}, \vec{e}_{d-1}\rangle\big]$$

for some orthonormal base $\{\vec{e}_1, \cdots, \vec{e}_{d-1}, \vec{e}_d\}$ of $\mathbb{R}^d$. One can consider some matrix $\mathbf{M}$, $d \times (d-1)$ such that $\mathbf{M}\mathbf{M}^\top = \mathbb{I}_{d-1}$ and $\mathbf{M}^\top\mathbf{M} = \mathbb{I}_d + \mathbf{1}_{d\times d}$. Then

$$\mathrm{ilr}(\mathbf{x}) = \mathrm{clr}(\mathbf{x})\mathbf{M} = \log(\mathbf{x})\mathbf{M},$$

and

$$\mathrm{ilr}^{-1}(\mathbf{z}) = \mathcal{C}\big((\exp(\mathbf{z}\mathbf{M}^\top)\big),\ \mathbf{z} \in \mathbb{R}^{d-1}.$$



	A	B	C
•(0)	43.71%	18.57%	37.72%
•(1)	29.83%	48.60%	21.56%
T(•)	29.67%	48.83%	21.50%
T(•)	43.73%	18.55%	37.72%

Figure 2: Counterfactuals using the ilr transformation, and Gaussian optimal transports, $\mu_0 \mapsto \mu_1$ on the left, and $\mu_1 \mapsto \mu_0$ on the right. Below are the averages of $\mathbf{x}_{0,i}$'s and $\mathbf{x}_{1,i}$'s, and of the transported points. The lines are geodesics in the dual spaces, mapped in the simplex. Optimal transport in $\mathbb{R}^2$, on $\mathbf{z}_{0,i}$'s and $\mathbf{z}_{1,i}$'s, can be visualized at the bottom (with linear mapping since Gaussian assumptions are made).

5.2 Gaussian Mapping in the Euclidean Representation

Given a random vector $\mathbf{X}$ in $\mathcal{S}_d$, we say that $\mathbf{x}$ follows a "normal distribution on the simplex" if, for some isomorphism h, the vector of orthonormal coordinates, $\mathbf{Z} = h(\mathbf{X})$ follows a multivariate normal distribution on $\mathbb{R}^{d-1}$. If we suppose that both $\mathbf{X}_0$ and $\mathbf{X}_1$, taking values in $\mathcal{S}_d$, follow "normal distributions on the simplex," then we can use standard Gaussian optimal transport, between $\mathbf{Z}_0$ and $\mathbf{Z}_1$. For convenience, suppose that the same isomorphism is used for both distributions (but that assumption can easily be relaxed). Hence, if $\mathbf{Z}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\mathbf{Z}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, the optimal mapping is linear,

$$\mathbf{z}_1 = T^\star(\mathbf{z}_0) = \boldsymbol{\mu}_1 + \boldsymbol{A}(\mathbf{z}_0 - \boldsymbol{\mu}_0), \tag{4}$$

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]. Interestingly, it is possible to derive McCann's displacement interpolation, from [McCann, 1997], to have some sort of continuous mapping

Algorithm 2 Gaussian Based Transport of $\mathbf{x}_0$ on $\mathcal{S}_d$

Input: $\mathbf{x}_0$ $(\in \mathcal{S}_d)$
Parameter: $\{\mathbf{x}_{0,1}, \cdots, \mathbf{x}_{0,n_0}\}$ and $\{\mathbf{x}_{1,1}, \cdots, \mathbf{x}_{1,n_1}\}$ in $\mathcal{S}_d$; isomorphic transformation $h : \mathcal{S}_d \to \mathbb{R}^{d-1}$
Output: $\mathbf{x}_1$

for $i \in \{1, \cdots, n_0\}$ **do**
  $\mathbf{z}_{0,i} \leftarrow h(\mathbf{x}_{0,i})$
end for
for $i \in \{1, \cdots, n_1\}$ **do**
  $\mathbf{z}_{1,i} \leftarrow h(\mathbf{x}_{1,i})$
end for
$\mathbf{m}_0 \leftarrow$ average of $\{\mathbf{z}_{0,1}, \cdots, \mathbf{z}_{0,n_0}\}$
$\mathbf{m}_1 \leftarrow$ average of $\{\mathbf{z}_{1,1}, \cdots, \mathbf{z}_{1,n_1}\}$
$\mathbf{S}_0 \leftarrow$ empirical variance matrix of $\{\mathbf{z}_{0,1}, \cdots, \mathbf{z}_{0,n_0}\}$
$\mathbf{S}_1 \leftarrow$ empirical variance matrix of $\{\mathbf{z}_{1,1}, \cdots, \mathbf{z}_{1,n_1}\}$
$\boldsymbol{A} \leftarrow \boldsymbol{S}_0^{-1/2}\big(\boldsymbol{S}_0^{1/2}\boldsymbol{S}_1\boldsymbol{S}_0^{1/2}\big)^{1/2}\boldsymbol{S}_0^{-1/2}$
$\mathbf{x}_1 \leftarrow h^{-1}\big(\mathbf{m}_1 + \boldsymbol{A}(h(\mathbf{x}_0) - \mathbf{m}_0)\big)$

$T_t^\star$ such that $T_1^\star = T^\star$ and $T_0 = Id$, and so that $\mathbf{Z}_t = T_t^\star(\mathbf{Z}_0)$ has distribution $\mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ where $\boldsymbol{\mu}_t = (1-t)\boldsymbol{\mu}_0 + t\boldsymbol{\mu}_1$ and

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_0^{-1/2}\left((1-t)\boldsymbol{\Sigma}_0 + t\left(\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{1/2}\right)^{1/2}\right)^2\boldsymbol{\Sigma}_0^{-1/2}.$$

Empirically, this can be performed using Algorithm 2, and a simulation can be visualized in Figure 2, where $h = \mathrm{clr}$. On the left, we can visualize the mapping of red points to the blue distribution, and on the right, the "inverse mapping" of blue points to the red distribution. Transformed points $\mathbf{z} = h(\mathbf{x})$, that are plotted at the bottom, are supposed to be normally distributed, and a multivariable Gaussian Optmial Transport mapping is used. Hence, $T_t^\star$ is linear in $\mathbb{R}^{d-1}$, as given by expression 4, as well as displacement interpolation, corresponding to red and blue segments. But, as we can see on top of Figure 2, in the original space, $t \mapsto \mathbf{x}_t := h^{-1}(\mathbf{z}_t)$ will be nonlinear. Tables are average values of the three components of $\mathbf{x}$'s and $T^\star(\mathbf{x})$'s.

6 Optimal Transport for Measures on $\mathcal{S}_d$

6.1 Theoretical Properties

A function $\psi : \mathcal{S}_d \to \mathbb{R}$ is exponentially concave if $\exp[\psi] : \mathcal{S}_d \to \mathbb{R}_+$ is concave. As a consequence, such a function ψ is differentiable almost everywhere. Let $\nabla\psi$ and $\nabla_{\vec{u}}\psi$ denote, respectively, its gradient, and its directional derivative. Following [Pal and Wong, 2016; Pal and Wong, 2018; Pal and Wong, 2020], define an allocation map generated by ψ, $\pi_\psi : \mathcal{S}_d \to \mathcal{S}_d$ defined as

$$\pi_\psi(\mathbf{x}) = \left[x_1\big(1 + \nabla_{\vec{e}_1 - \mathbf{x}}\psi(\mathbf{x})\big), \cdots, x_d\big(1 + \nabla_{\vec{e}_d - \mathbf{x}}\psi(\mathbf{x})\big)\right],$$

where $\{\vec{e}_1, \cdots, \vec{e}_d\}$ is the standard orthonormal basis of $\mathbb{R}^d$. Consider the optimal transport problem with the following cost function, on $\mathcal{S}_d \times \mathcal{S}_d$, i.e., the L-divergence corresponding to the cross-entropy,

$$c(\mathbf{x}, \mathbf{y}) = \log\left(\frac{1}{d}\sum_{i=1}^d \frac{y_i}{x_i}\right) - \frac{1}{d}\sum_{i=1}^d \log\left(\frac{y_i}{x_i}\right), \quad (5)$$

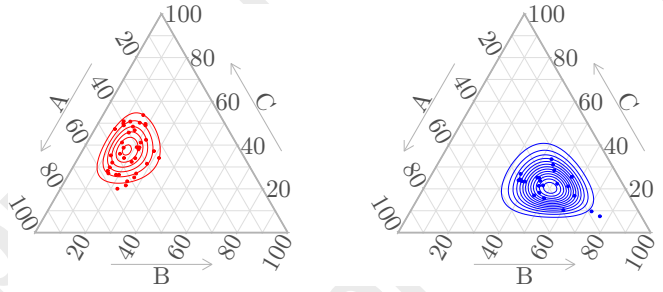


Figure 3: Densities of Dirchlet distributions in $\mathcal{S}_3$ fitted on observations of the toydataset of Figure 1.

called "Dirichlet transport" in [Baxendale and Wong, 2022]. See [Pistone and Shoaib, 2024] for a discussion about the connections with the distance induced by Aitchsion's inner product of Equation (3). From Theorem 1 in [Pal and Wong, 2020], for this cost function, there exists an exponentially concave function $\psi^\star : \mathcal{S}_d \to \mathbb{R}$ such that

$$T^\star(\mathbf{x}) = \mathbf{x} \diamond \pi_{\psi^\star}\left(\mathbf{x}^{-1}\right)$$

defines a push-forward from $\mathbb{P}_0$ to $\mathbb{P}_1$, and the coupling $(\mathbf{x}, T^\star(\mathbf{x}))$ is optimal for problem (1), and is unique if $\mathbb{P}_0$ is absolutely continuous. Observe that if $\mathbf{y} = T^\star(\mathbf{x})$,

$$\mathbf{y} = \mathcal{C}\big(\pi_{\psi^\star}(\mathbf{z})_1/z_1, \cdots, \pi_{\psi^\star}(\mathbf{z})_d/z_d\big),$$

where $\mathbf{z} = \mathbf{x}^{-1}$.

One can also consider an interpolation,

$$T_t^\star(\mathbf{x}) = \mathbf{x} \diamond \pi_t\left(\mathbf{x}^{-1}\right)$$

where $\pi_t = (1-t)d^{-1}\mathbf{1} + t\pi_{\psi^\star}$ (even if this approach differs from McCann's displacement interpolation).

Note that a classical distribution on $\mathcal{S}_d$ is Dirichlet distribution, with density

$$f(x_1, \ldots, x_d; \boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{i=1}^d x_i^{\alpha_i - 1}$$

for some $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}_+^d$, and a normalizing constant denoted $\mathrm{B}(\boldsymbol{\alpha})$. Level curves of the density of Dirichlet distributions fitted on our toy dataset can be visualized in Figure 3. Unfortunately, unlike the multivariate Gaussian distribution, there is no explicit expression for the optimal mapping between Dirichlet distribution (regardless of the cost). Therefore, to remain within $\mathcal{S}_d$ and avoid the $\mathbb{R}^{d-1}$ representation, numerical techniques should be considered.

6.2 Matching

Consider two samples in the $\mathcal{S}_d$ simplex, $\{\mathbf{x}_{0,1}, \cdots, \mathbf{x}_{0,n_0}\}$ and $\{\mathbf{x}_{1,1}, \cdots, \mathbf{x}_{1,n_1}\}$. The discrete version of the Kantorovich problem (corresponding to Equation 2) is

$$\min_{\mathrm{P} \in U(n_0, n_1)} \left\{\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} \mathrm{P}_{i,j}\mathrm{C}_{i,j}\right\} \quad (6)$$

where, as in [Brualdi, 2006], $U(n_0, n_1)$ is the set of $n_0 \times n_1$ matrices corresponding to the convex transportation polytope

$$U(n_0, n_1) = \left\{\mathrm{P} : \mathrm{P}\mathbf{1}_{n_1} = \mathbf{1}_{n_0} \text{ and } \mathrm{P}^\top\mathbf{1}_{n_0} = \frac{n_0}{n_1}\mathbf{1}_{n_1}\right\},$$

Algorithm 3 Coupling samples on $\mathcal{S}_d$

Input: $\{\mathbf{x}_{0,1}, \cdots, \mathbf{x}_{0,n_0}\}$ and $\{\mathbf{x}_{1,1}, \cdots, \mathbf{x}_{1,n_1}\}$ in $\mathcal{S}_d$;
Output: weight matching matrix $n_0 \times n_1$ $\mathbf{P}^*$

$\mathbf{C} \leftarrow$ matrix $n_0 \times n_1$, $\mathbf{C}_{i,j} = c(\mathbf{x}_i, \mathbf{x}_j)$ using (5)
$\mathbf{P}^\star \leftarrow$ solution of Equation (6), using LP libraries

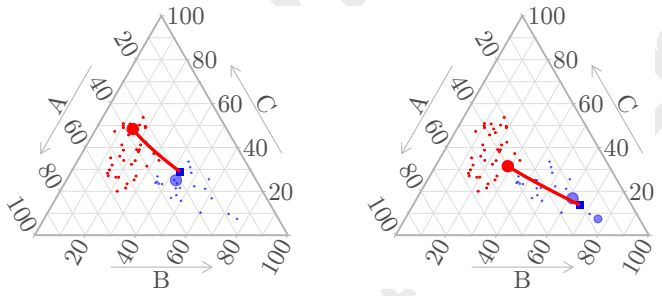


Figure 4: Getting empirical counterfactuals using matching techniques, with $\mathbf{x}_{0,i}$ in red (on the top-left hand-side), and counterfactuals $\mathbf{x}_{1,j}$'s in blue (bottom-right hand-side), with size proportional to $\mathbf{P}_i^\star = [\mathbf{P}_{i,1}^\star, \cdots, \mathbf{P}_{i,n_1}^\star] \in \mathcal{S}_{n_1}$.

and where C denotes the $n_0 \times n_1$ cost matrix, $\mathrm{C}_{i,j} = c(\mathbf{x}_i, \mathbf{x}_j)$, associated with cost from Equation (5).

In Algorithm 3, we recall how this procedure works, which is the one explained in [Peyré *et al.*, 2019], with a specific cost function (from Equation (5)). In the toy dataset, this can be visualized for two specific observations $\mathbf{x}_{0,i}$ in Figure 4 (big red dots). If $n_0 \neq n_1$, it is not a one-to-one coupling, and "the counterfactual" (blue square) is in fact a weighted average of $\mathbf{x}_{1,j}$'s (blue dots), where weights are given in row $\mathbf{P}_i^\star = [\mathbf{P}_{i,1}^\star, \cdots, \mathbf{P}_{i,n_1}^\star] \in \mathcal{S}_{n_1}$.

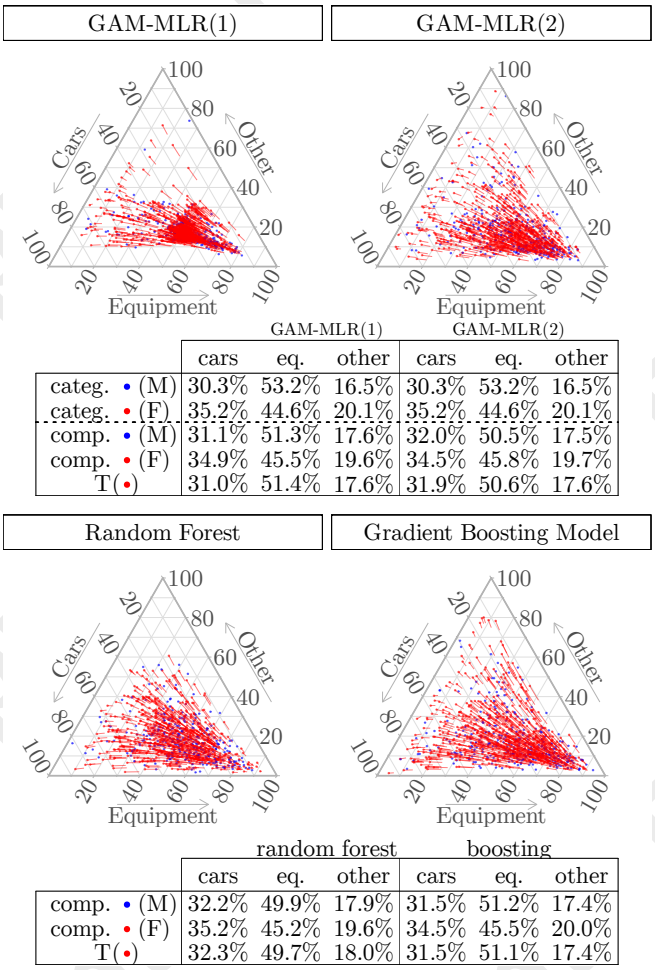
7 Application on Sequential Transport for Counterfactuals

Variables $\mathbf{x}_j$ in tabular data are either continuous or categorical. If $\mathbf{x}_j$ is continuous, since $\mathbf{x}_j \in \mathbb{R}$, transporting from observed $\mathbf{x}_j|s=0$ to counterfactual $\mathbf{x}_j|s=1$ is performed using standard (conditional) monotonic mapping, as discussed in [Fernandes Machado *et al.*, 2025], using classical $F_1^{-1} \circ F_0$. If $\mathbf{x}_j$ is categorical, with d categories, consider some fitted model $\widehat{m}(\mathbf{x}_j|\mathbf{x}_{-j})$, using some multinomial loss, and let $\widehat{\mathbf{x}}_j = \widehat{m}(\mathbf{x}_j|\mathbf{x}_{-j})$ denote the predicted scores, so that $\widehat{\mathbf{x}}_j \in \mathcal{S}_d$. Then use Algorithm 2, with a Gaussian mapping in an Euclidean representation space, to transport from observed $\widehat{\mathbf{x}}_j|s=0$ to counterfactual $\widehat{\mathbf{x}}_j|s=1$, in $\mathcal{S}_d$.

7.1 German Credit: Purpose

In the popular `German Credit` dataset, from [Hofmann, 1994], the variable `Purpose` described the reason an individual took out a loan. This variable is an important predictor for explaining potential defaults. The original variable is based on ten categories, that are merged here into three main classes, cars, equipment and other, in order to visualize the transport in a ternary plot (or Gibbs triangle). The sensitive variable s is here `Sex`.

We aim to construct a counterfactual value for the loan purpose, assuming the individuals were of a different sex. To



	GAM-MLR(1)			GAM-MLR(2)		
	cars	eq.	other	cars	eq.	other
categ. • (M)	30.3%	53.2%	16.5%	30.3%	53.2%	16.5%
categ. • (F)	35.2%	44.6%	20.1%	35.2%	44.6%	20.1%
comp. • (M)	31.1%	51.3%	17.6%	32.0%	50.5%	17.5%
comp. • (F)	34.9%	45.5%	19.6%	34.5%	45.8%	19.7%
T(•)	31.0%	51.4%	17.6%	31.9%	50.6%	17.6%

	random forest			boosting		
	cars	eq.	other	cars	eq.	other
comp. • (M)	32.2%	49.9%	17.9%	31.5%	51.2%	17.4%
comp. • (F)	35.2%	45.2%	19.6%	34.5%	45.5%	20.0%
T(•)	32.3%	49.7%	18.0%	31.5%	51.1%	17.4%

Figure 5: Optimal transport using the clr transformation, and Gaussian optimal transports, on the `purpose` scores in the `German Credit` dataset, with two logistic GAM models to predict scores, on top, and below a random forest (left) and a boosting model (right). Points in red are compositions for women, while points in blue are for men. Lines indicate the displacement interpolation when generating counterfactuals.

achieve this, we apply our suggested procedure from to represent the `purpose` categorical variable as a compositional variable, using the same four models outlined in Section 3 and then apply Gaussian mapping from Section 5.2. The results provided by all of the models, shown in Figure 5, suggest that, had the individuals been of a different sex, the purpose of the loan would have changed. Specifically, if the average scores in each group (cars, equipment, and other) were approximately $[35\%, 45\%, 20\%]$ in the female population, after transporting to obtain the counterfactuals, the average scores become $[31\%, 51\%, 18\%]$, which closely resemble the actual frequencies of each category in the original male population.

One can also consider our second approach, using matching in $\mathcal{S}_3$. Consider individual i among women, e.g., the left of Figure 6, $x_{0,i}$ = "other." Using a MLR model, we obtain composition $\mathbf{x}_{0,i}$, here $[36.98\%, 23.81\%, 39.2\%]$. Using Algorithm 3, three points $\mathbf{x}_{1,j}$'s are matched, respectively

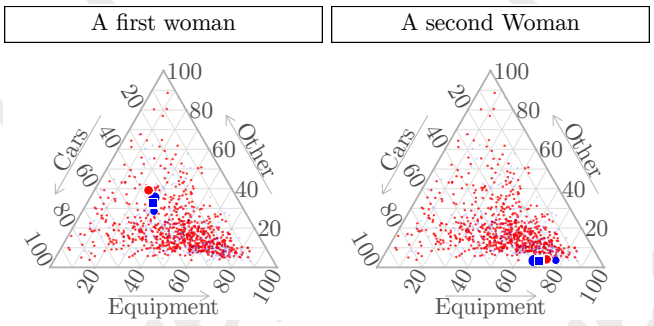


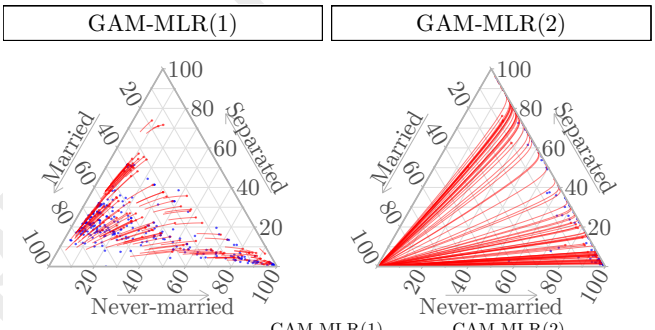
Figure 6: Empirical matching of two women, in red, from the `German Credit` dataset, with 2 or 3 men, in blue. Size of blue dots are proportional to the weights $\mathbf{P}_i^\star$.

with weights $[0.226, 0.548, 0.226]$. For those three points, the propensity for the purpose to be "cars' is the highest ($[39.9\%, 36.3\%, 42.3\%]$). Therefore, the counterfactual version of woman i with an "other" credit is a man with the purpose "cars". In fact, using Gaussian transport (blue square), we obtain $T^\star(\mathbf{x}_{0,i}) = [38.47\%, 29.05\%, 32.49\%]$.

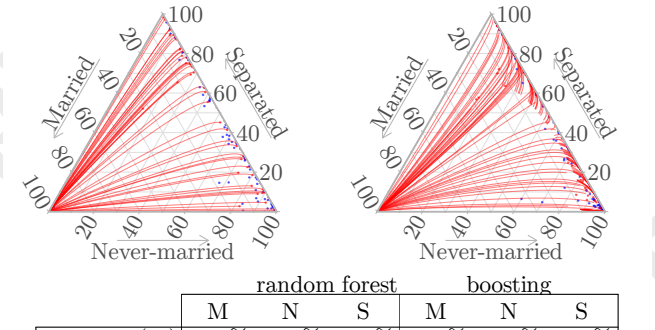
7.2 Adult: Marital Status

Following the numerical applications in [Plečko *et al.*, 2024] and [Fernandes Machado *et al.*, 2025], we consider here the `Adult` dataset, from [Becker and Kohavi, 1996]. We regrouped categories of the `Marital Status` variable to create three generic ones (that can be visualized in a ternary plot, as in Figure 7), namely Married (M), Never-married (N) and Separated (S). This example is interesting because if we compare status with respect to the `Sex` variable, proportions are quite different. In the dataset, proportions for married, never married, and separated are (roughly) $[62\%, 27\%, 12\%]$ for men, $[14\%, 44\%, 41\%]$ for women (more precise values are at the top of the table in Figure 7). Thus, the counterfactual of a "separated" woman is more likely to be a "married" man than a "separated" man. Four models are used to convert the categorical variable `Marital Status` into a composition, as previously. The first MLR is based on three variables: a categorical variable, `occupation`, and two continuous ones, `age`, and `hours_per_week`, modeled nonlinearly using b-splines (hence, it is referred to as a logistic GAM). This model is clearly underfitted. Therefore, observations $\mathbf{x}_{0,i}$'s for women and $\mathbf{x}_{1,i}$'s for men clearly are in the interior of $\mathcal{S}_d$. In contrast, the more complex MLR (which uses additional features), as well as the random forest and boosting models, can produce predictions near the simplex boundary, $\partial\mathcal{S}_d$.

For the underfitted model (top left), transported scores have a distribution very close to the ones in the population of men. For the more accurate MLR model (top right), proportions are very close to the actual proportions (which is not surprising since GLMs are usually well calibrated), but the transported scores are slightly different than the proportions of categories (proportions were $[62\%, 27\%, 12\%]$ while average transported scores are $[67\%, 25\%, 8\%]$). At least, we are different from the original ones, but the mapping is not as accurate as it should be. This might come from the fact that when the points $\mathbf{x}_i$ are close to the border $\partial\mathcal{S}_d$, it is quite unlikely



	GAM-MLR(1)			GAM-MLR(2)		
	M	N	S	M	N	S
categ. • (M)	61.8%	26.6%	11.6%	61.8%	26.6%	11.6%
categ. • (F)	15.3%	44.1%	40.7%	15.3%	44.1%	40.7%
comp. • (M)	50.4%	31.0%	18.6%	59.6%	27.3%	13.1%
comp. • (F)	35.2%	39.7%	25.1%	13.3%	46.1%	40.6%
T(•)	49.1%	32.1%	18.8%	77.6%	7.8%	14.6%



	random forest			boosting		
	M	N	S	M	N	S
comp. • (M)	59.3%	27.7%	13.0%	59.5%	27.5%	13.0%
comp. • (F)	13.2%	48.7%	38.1%	13.5%	46.0%	40.4%
T(•)	48.7%	33.4%	18.0%	51.7%	31.4%	17.0%

Figure 7: Optimal transport using the clr transformation, and Gaussian optimal transports, on the `Marital Status` scores in the `Adult` dataset, with two logistic GAM-MLR models to predict scores, on top, and below a random forest (left) and a boosting model (right). Points in red are compositions for women, while points in blue are for men. Lines indicate the displacement interpolation when generating counterfactuals.

that the sample $\mathbf{z}_i$ is Gaussian.

8 Conclusion

In this article, we introduce a novel approach for constructing counterfactuals for categorical data by transforming them into compositional data using a probabilistic classifier. Our approach avoids imposing arbitrary assumptions about label ordering. However, our methodology is not without limitations. OT computations, particularly on the simplex, can be computationally intensive for large-scale datasets, posing challenges in high-dimensional settings. Additionally, the reliance on a probabilistic classifier in the initial step introduces potential vulnerabilities. Biases may arise from a poorly calibrated or inaccurate classifier, impacting the quality of the subsequent analysis–especially with scarce categories that may need grouping to apply the proposed method.

Ethical Statement

There are no ethical issues.

Acknowledgments

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 acknowledges 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.

References

[Aitchison, 1982] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

[Baxendale and Wong, 2022] Peter Baxendale and Ting-Kam Leonard Wong. Random concave functions. *The Annals of Applied Probability*, 32(2):812–852, 2022.

[Becker and Kohavi, 1996] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996.

[Black *et al.*, 2020] Emily Black, Samuel Yeom, and Matt Fredrikson. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121, 2020.

[Brenier, 1991] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[Brualdi, 2006] Richard A Brualdi. *Combinatorial matrix classes*, volume 13. Cambridge University Press, 2006.

[Charpentier *et al.*, 2023] Arthur Charpentier, Emmanuel Flachaire, and Ewen Gallic. Optimal transport for counterfactual estimation: A method for causal inference. In *Optimal Transport Statistics for Economics and Related Topics*, pages 45–89. Springer, 2023.

[Chayes, 1971] Felix Chayes. *Ratio correlation*. University of Chicago Press, 1971.

[Cheng *et al.*, 2010] Weiwei Cheng, Eyke Hüllermeier, and Krzysztof J Dembczynski. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 279–286, 2010.

[De Lara *et al.*, 2024] Lucas De Lara, Alberto González-Sanz, Nicholas Asher, Laurent Risser, and Jean-Michel Loubes. Transport-based counterfactual models. *Journal of Machine Learning Research*, 25(136):1–59, 2024.

[Egozcue *et al.*, 2003] Juan José Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barcelo-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical geology*, 35(3):279–300, 2003.

[Fernandes Machado *et al.*, 2025] Agathe Fernandes Machado, Arthur Charpentier, and Ewen Gallic. Sequential conditional transport on probabilistic graphs for interpretable counterfactual fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):19358–19366, Apr. 2025.

[Hallin and Konen, 2024] Marc Hallin and Dimitri Konen. Multivariate quantiles: Geometric and measure-transportation-based contours. In *Applications of Optimal Transport to Economics and Related Topics*, pages 61–78. Springer, 2024.

[Hallin *et al.*, 2021] Marc Hallin, Eustasio Del Barrio, Juan Cuesta-Albertos, and Carlos Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139–1165, 2021.

[Higham, 2008] Nicholas J Higham. *Functions of matrices: theory and computation*. SIAM, 2008.

[Hofmann, 1994] H. Hofmann. German Credit Data. UCI Machine Learning Repository, 1994.

[Imai *et al.*, 2022] Kosuke Imai, Santiago Olivella, and Evan T.R. Rosenman. Addressing census data problems in race imputation via fully bayesian improved surname geocoding and name supplements. *Science Advances*, 8(49):eadc9824, 2022.

[Kantorovich, 1942] Leonid V Kantorovich. On the translocation of masses. In *Doklady Akademii Nauk USSR*, volume 37, pages 199–201, 1942.

[King *et al.*, 2004] Gary King, Martin A Tanner, and Ori Rosen. *Ecological inference: New methodological strategies*. Cambridge University Press, 2004.

[Kusner *et al.*, 2017] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. NIPS, 2017.

[McCann, 1997] Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

[Monge, 1781] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[Naeini *et al.*, 2015] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[Pal and Wong, 2016] Soumik Pal and Ting-Kam Leonard Wong. The geometry of relative arbitrage. *Mathematics and Financial Economics*, 10:263–293, 2016.

[Pal and Wong, 2018] Soumik Pal and Ting-Kam Leonard Wong. Exponentially concave functions and a new information geometry. *The Annals of probability*, 46(2):1070–1113, 2018.

[Pal and Wong, 2020] Soumik Pal and Ting-Kam Leonard Wong. Multiplicative schrödinger problem and the dirichlet transport. *Probability Theory and Related Fields*, 178(1):613–654, 2020.

[Pal, 2024] Soumik Pal. On the difference between entropic cost and the optimal transport cost. *The Annals of Applied Probability*, 34(1B):1003–1028, 2024.

[Pawlowsky-Glahn and Egozcue, 2001] Vera Pawlowsky-Glahn and Juan José Egozcue. Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15:384–398, 2001.

[Pearl, 2009] Judea Pearl. *Causality*. Cambridge university press, 2009.

[Peyré *et al.*, 2019] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[Pistone and Shoaib, 2024] Giovanni Pistone and Muhammad Shoaib. A unified approach to aitchison's, dually affine, and transport geometries of the probability simplex. *Axioms*, 13(12):823, 2024.

[Plečko and Meinshausen, 2020] Drago Plečko and Nicolai Meinshausen. Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44, 2020.

[Plečko *et al.*, 2024] Drago Plečko, Nicolas Bennett, and Nicolai Meinshausen. Fairadapt: Causal reasoning for fair data preprocessing. *Journal of Statistical Software*, 110(4):1–35, 2024.

[Santambrogio, 2015] Filippo Santambrogio. *Optimal transport for applied mathematicians*. Springer, 2015.

[Villani, 2003] Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Society, 2003.

[Zadrozny and Elkan, 2001] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 609–616, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.