COMMUN. MATH. SCI.
Vol. 22, No. 6, pp. 1635–1656

THE CONDITIONAL BARYCENTER PROBLEM, ITS DATA-DRIVEN FORMULATION AND ITS SOLUTION THROUGH NORMALIZING FLOWS*

ESTEBAN G. TABAK†, GIULIO TRIGILA‡, AND WENJUN ZHAO§

Abstract. A family of normalizing flows is introduced for selectively removing from a data set the variability attributable to a specific set of cofactors, while preserving the dependence on others. This is achieved by extending the barycenter problem of optimal transport theory to the newly introduced conditional barycenter problem. Rather than summarizing the data with a single probability distribution, as in the classical barycenter problem, the conditional barycenter is represented by a family of distributions labeled by the cofactors kept. The use of the conditional barycenter and its differences with the classical barycenter are illustrated on synthetic and real data addressing treatment effect estimation, super-resolution, anomaly detection and lightness transfer in image analysis.

Keywords. Optimal transport; barycenter problem; normalizing flows; conditional distributions.

AMS subject classifications. 49Q22; 62G07.

1. Introduction

Given a set of distributions $\rho(x|z)$ indexed by a variable z, the optimal transport barycenter problem [1, 12, 14] seeks a family of invertible maps $y=T(x,z)$ so that the push forward distribution $\mu=T(\cdot,z)\#\rho(\cdot|z)$ is independent of z. Among all maps $T(\cdot,z)$ satisfying this constraint, the barycenter $\mu(y)$ of the $\rho(x|z)$ is defined by the map that minimizes a cost function $C(T)$. In an optimal transport setting, the cost $C(T)$ is the expected value over the joint distribution $\rho(x,z)=\rho(x|z)\gamma(z)$ of a pairwise cost function $c(x,T(x,z))$. More general costs can be used, for instance to quantify some measure of the deformation incurred by the map or to compute the barycenter of distributions over different spaces. This extension gives rise to the distributional barycenter problem [18].

Direct applications of the barycenter problem include:

(1) Finding a single distribution representative of all $\rho(\cdot|z)$. The barycenter $\mu(y)$ provides a much sharper descriptor than the marginal

$$\rho(x)=\int \rho(x|z)\ \gamma(z)\ dz,$$

as the latter preserves all the variability in x due to z, merely averaging over the factor z that could have helped explain it. By contrast, the map $y=T(x,z)$ removes from x all variability attributable to z. For example, the natural variability in the heart rate x is highly diminished if one accounts for the effect of the patient's age z. Averaging the heart rate over the age, on the other hand, is equivalent to disregarding the age factor altogether, i.e. considering in lieu of the pairs $\{x^i,z^i\}$, their first component alone $\{x^i\}$.

*Received: August 26, 2022; Accepted (in revised form): January 16, 2024. Communicated by Jianfeng Lu.

†Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA (tabak@cims.nyu.edu).

‡Weissman School of Arts and Sciences, Baruch College, City University of New York, NY 10010, USA (Giulio.Trigila@baruch.cuny.edu).

§Division of Applied Mathematics, Brown University, Providence, RI 02912, USA (wenjun_zhao@brown.edu).

For a simple example, consider a one-dimensional normal conditional distribution of the form

$$\rho(x|z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-z)^2}{2}}, \quad \gamma(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}},$$

with barycenter

$$\mu(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}},$$

which is sharper than the marginal

$$\rho(x) = \int \rho(x|z)\ \gamma(z)\ dz = \frac{1}{\sqrt{4\pi}} e^{-\frac{x^2}{4}},$$

as it has half its variance. Not only the variance but also the barycenter's shape reflects the conditional distributions better than the marginal. Consider for instance a discrete, $z \in \{z_1, \ldots, z_K\}$, where the $\rho(x|z_k)$ are all Gaussian. Here the marginal is a Gaussian mixture, while the barycenter is a Gaussian, just as each conditional distribution.

In addition to providing a representative μ of the $\rho(\cdot|z)$, eliminating from $\rho(x)$ the variability explainable by known factors z paves the way to uncover additional, hidden sources of variability. It is clear that the use of the barycenter or the marginal ought to be made based on the application at hand. For instance, if one wishes to compute the mean of x independently of the value of z, computing the empirical mean of the marginal $\rho(x)$ is the fastest and simplest thing to do.

(2) Simulating the conditional distribution $\rho(x|z^*)$ from n sample pairs $\{x_i, z_i\}$. One first draws n samples $\{y_i\}$ from the barycenter μ through

$$y_i = T(x_i, z_i),$$

and then uses these to produce n samples $x_i^* \sim \rho(\cdot|z^*)$ through the inversion

$$x_i^* = T^{-1}(y_i, z^*).$$

Thus we are using samples from a population with heterogeneous factors z to simulate the distribution for one particular value z^*.

Factor discovery provides a further application [23]. Here a new factor z is determined by the condition that is should explain as much variability as possible, thus minimizing the unexplained variability remaining in the barycenter $\mu(y)$.

This article further extends the notion of distributional barycenter to that of conditional barycenter, where the individual, z-indexed objects to push forward are not distributions $\rho(x)$ but conditional distributions $\rho(x|r)$, further conditioned to z. The use of the distributional barycenter μ as a representative becomes richer in the conditional setting, where the conditional barycenter $\mu(y|r)$ summarizes the common information contained in all the $\rho(x|r,z)$. Consider for example the case where x is a measure of health state, such as cholesterol level, r quantifies treatment, such as the choice of a drug and its dosage, and z is a qualifier of the patient under consideration, such as age. Two main tasks are of particular relevance in this setting: assessing the overall effect of a treatment on a heterogeneous population of individuals, and estimating the treatment effect on a specific individual.

After eliminating the effect of z on x, $\mu(y|r)$ summarizes the effect of the treatment r over all values of z. As in the non-conditional case, using instead the marginal

$$\rho(x|r) = \int \rho(x|r,z)\ \gamma(z)\ dz$$

yields much higher variability as, for each treatment r, the distribution of x covers its range over all values of z. (In data-driven scenarios, this is the marginal estimated from pairs $\{x_i, r_i\}$, disregarding the corresponding value of z_i.) Moreover, the marginal may display Simpson's paradox [8,13]: consider the synthetic data in Figure 4.2 where, for each value of z (represented by color), the distribution of x (outcome) moves toward lower values as r (the treatment) increases, but x increases as a function of z, and therefore treatments with lower r are typically provided to patients with lower z. Then in the marginal $\rho(x|r)$, x may increase with r (since higher r correlates with higher z and therefore higher x). By contrast, a summary based not on marginalization but on the conditional barycenter $\mu(y|r)$, does not display Simpson's paradox, as shown in Figure 4.3.

A complementary use of the conditional barycenter for the same problem reverses the roles of the treatment r and the factors z, finding the family of maps $T(x,r,z)$ that push forward $\rho(x|r,z)$ to their barycenter $\mu(y|z)$ with respect to r. Here $T(\cdot,r,\cdot)$ is conceptualized as representing the treatment itself, i.e. transporting the health state x of a patient with given covariates z between two treatments r_1 and r_2, mediated by the barycenter:

$$x_2 = T^{-1}\left(T(x_1, r_1, z), r_2, z\right).$$

In practice, r_1 may represent the current treatment (such as no treatment at all) and x_1 the corresponding currently measured health state.

We will discuss below further applications where the conditional distributional barycenter problem provides a natural framework for analysis. In particular, we illustrate the use of the conditional barycenter problem on four examples:

- Super-resolution volume reconstruction from slice acquisitions. One of the ways to reconstruct a volumetric image of parts of the human body is through axial, low resolution MRI images. The images are in low resolution as their acquisition needs to be fast due to movements of the subject or to moving organs. The low resolution of the slices and their non-perfect alignment result in inter-slice artifacts.
- Treatment effects. This can refer to an actual medical treatment, to a habit –smoking, eating an apple a day, getting a medical check-up every year–, to a policy –raise taxes on gas to reduce CO_2 emissions–, in short, to any scenario where an outcome x, such as cholesterol level, depends on covariates z, such as a patient's age, and decisions r to be made, such as a course of medical treatment.
- Characterization of anomalies. Climate studies often address anomalies: a summer warmer than regular, a longer rainy season. Typically, such anomalies are quantified by comparing actual values to long-time averages, the *climatology*. Yet such characterization is far from ideal. On the one hand, defining the time-window for comparison may prove elusive: a particular day or week, a season, a year? More importantly, it is not only mean values that characterize "normality": the variability around them is a fundamental component of any

regime. Spring time weather in the American northeast, for instance, is not characterized so much by a mean temperature or precipitation value, but by their permanent change. The conditional barycenter distribution $\mu(y|r)$ provides a much better suited descriptor of climatology. Here r may represent, for instance, the day of the year, and z the year. Then the difference $x - T(x,r,z)$ characterizes anomaly: how much the observed value of x needs to be changed to make it consistent with the climatological distribution.

- Image analysis. In data analysis, an image is often conceptualized as a big vector or matrix, such as a triad of color intensities for each pixel. However, a characterization better suited for many purposes is in terms of a conditional probability $\rho(x|r)$, where r is a two dimensional vector encoding position, and x represents again a triad of color intensities (physically, ρ represents the local photon density in three frequency windows). Thus, when images are qualified by factors z, such as the time or location at which a photograph was taken, we face a conditional barycenter problem. The example we describe concerns conditional lightness transfer, where the perceptual effect of an evolving luminosity due, for instance, to weather, season or time of the day, may depend on the color of the object considered.

Most applications are data-based, i.e. the family of conditional distributions $\rho(x|r,z)$ and the marginal distributions $\gamma(z)$ and $\eta(r|z)$ are not provided in closed form, but only through n observations, i.e. triplets of samples $\{x_i,r_i,z_i\}$ drawn from them. Thus the core of this article is devoted to formulating the problem in terms of samples and developing a methodology for its numerical solution.

To fully appreciate the usefulness of the conditional barycenter problem, one must distinguish it from two other related procedures that remove the factor z from a distribution $\rho(x|r,z)$:

- Marginalization, already discussed, simply averages over z or, in the data-driven case, truncates the samples, leaving just the pairs $\{x_i,r_i\}$. While simplifying the problem by reducing the number of variables considered, this throws away potentially useful information encoded in the factors z, enabling for instance occurrences of the Simpson paradox.
- Regular barycenter over z, which temporarily forgets r and eliminates the effect of z on x through a map $y = T(x,z)$. Unlike the conditional barycenter problem, this removes from x variability that r may account for, thus confounding the dependence of x on r.

This article is structured as follows: after this introduction, Section 2 formulates the problem, first as a cost minimization with the infinitely many constraints of the push-forward condition, then as a minimax problem over the transport T and a test function F that enforces the constraints, and finally as a penalized minimization where F adopts a prescribed adaptive form. Section 3 formulates the last problem in terms of samples while proposing an algorithm for its numerical solution, which involves a non-parametric smooth normalizing flow driven by gradient descent. Section 4 applies the algorithm to the four examples mentioned above, which simultaneously illustrate the numerical methodology and the breadth of applicability of the conditional barycenter problem. Finally, Section 5 concludes with some brief remarks.

2. Formulation

Consider the joint distribution

$$\rho(x,r,z) = \rho(x|r,z)\ \eta(r|z)\ \gamma(z)$$

of three variables x, r and z, each of which possibly has more than one component. Here x are the *variables of interest* or *outcome*, r may specify the *treatment* when describing an action intended to control the outcome x, or simply general coordinates that x depends on, and z are the *covariates*, *confounding factors* or simply *factors*. We are using ρ to denote all distributions with x as first argument, with its format specifying which variables are conditioned to which. Similarly, we will use γ and η for all distributions on z and r respectively.

We seek to transform the random variable x through a map T with parameters r and z, so that the resulting random variable $y = T(x,r,z)$ is independent of z:

$$\rho_T(y|r,z) = T_{\#}\rho(x|r,z) = \mu(y|r).$$

Among all maps T satisfying this constraint, we select the minimizer of a cost functional $C(T)$ and define the resulting distribution $\mu(y|r)$ as the z-conditional barycenter of the $\rho(x|r,z)$. In optimal transport settings, the cost function may adopt one of the two forms

$$C_1(T) = \int c(x,T(x,r,z))\,\rho(x,r|z)\ dx\ dr\ \gamma(z)dz \tag{2.1}$$

and

$$C_2(T) = \int c(x,T(x,r,z))\,\rho(x|r,z)\ dx\ \eta(r)dr\ \gamma(z)dz, \tag{2.2}$$

where the pairwise cost function $c(x,y)$ is externally provided. While we use C_1 in most applications, the use of C_2 is appropriate when searching for $\mu(y|r)$ as a representative of the $\rho(x|r,z)$. In that case one must weigh all values of z equally for each r, else we may observe bias in $\mu(y|r)$ due to the non-uniformity of $\gamma(z|r)$ analogous to the one responsible for the Simpson Paradox (see Section 4.2.1 for an example in more detail).

More general costs C not based on a pairwise function $c(x,y)$ are also possible, extending the distributional barycenter problem to its conditional counterpart. For any cost $C(T)$, the conditional barycenter problem can be formulated as follows:

$$\min_T C(T), \quad \rho_T(y|r,z) = T_{\#}\rho(x|r,z) = \mu(y|r).$$

Since the push forward of the map only affects the variable x, an equivalent formulation of the conditional independence constraint can be obtained by multiplying both sides of the constraint above by $\nu(z|r)$ leading to

$$\rho_T(y,z|r) = \rho_T(y|r,z)\nu(z|r) = \mu(y|r)\nu(z|r).$$

This is itself equivalent to the condition that every measurable test function F satisfying

$$\forall y \forall r \int F(y,r,z)\ \nu(z|r)\ dz = 0$$

must also satisfy

$$\forall r \int F(y,r,z)\rho_T(y,z|r)\ dy\ dz = 0, \tag{2.3}$$

as the following calculation shows: in one direction, if $\rho_T(y,z|r)=\mu(y|r)\nu(z|r)$, then

$$\int F(y,r,z)\rho_T(y,z|r)\ dy\ dz=\int dy\int F(y,r,z)\ \nu(z|r)\ dz=0.$$

In the other direction, it is enough to specialize $F(y,r,z)$ to $\rho(y|r,z)$ and apply Proposition 2.1, stated and proved below.

This observation allows us to re-write the constrained optimization problem defining the conditional barycenter as

$$\begin{cases}\min_T\max_F\ \left[C(T)+\int F(y,r,z)\ \rho_T(y,r,z)\ dy\ dr\ dz\right]\\ \forall y\forall r\ \int F(y,r,z)\ \nu(z|r)\ dz=0,\end{cases}\tag{2.4}$$

which can be re-stated as an unconstrained minimax problem by removing the z-mean from the test functions F:

$$\min_T\max_F\ C(T)+L_F,$$

$$\begin{aligned}L_F&=\int\left[F(y,r,z)-\int F(y,r,w)\nu(w|r)\ dw\right]\rho_T(y,r,z)\ dy\ dr\ dz\\&=\int F(y,r,z)[\rho_T(y|r,z)-\rho_T(y|r)]\ \eta(r|z)\ \gamma(z)\ dy\ dr\ dz.\end{aligned}$$

A simpler formulation, not requiring maximizing over the test function F, is enabled by the following proposition:

PROPOSITION 2.1. *Adopting $F(y,r,z)=\rho_T(y|r,z)$, L_F is non-negative, vanishing only when y and z are conditionally independent.*

Proof. For this choice of F, the first term in L_F becomes

$$\int\rho_T(y|r,z)^2\ \gamma(z|r)\ dz\ \eta(r)dr\ dy=\int\mathbb{E}^z[\rho_T(y|r,z)^2|r]\ \eta(r)dr\ dy,$$

while the second term equals

$$\int\rho_T(y|r,z)\gamma(z|r)dz\ \rho_T(y|r)\ dy\ \eta(r)dr=\int\mathbb{E}^z[\rho_T(y|r,z)|r]^2\eta(r)dr\ dy.$$

From Jensen's inequality, their difference is non-negative, vanishing only if $\rho_T(y|r,z)$ does not depend on z. ∎

This result permits restricting the test functions to the form $F(y,r,z)=\lambda\ \rho_T(y|r,z)$, since for this particular F, the test function L_F only vanishes when $\rho(y|r)$ is independent of z. Hence we have the alternative formulation:

PROBLEM 2.1.

$$\min_T\max_{\lambda>0}\ C(T)+\lambda\int[\rho_T(y|z,r)-\rho_T(y|r)]\rho_T(y,z,r)\ dy\ dz\ dr,\tag{2.5}$$

where a large penalty coefficient λ enforces the satisfaction of the push forward constraint. Section 3 discusses in detail how to pose this problem in data-driven scenarios and solve it numerically.

3. Data-driven formulation

This section discusses the data-driven formulation of the continuous optimization problem (2.5) and its numerical solution. Replacing in (2.5) all expected values by their empirical counterpart yields

$$\min_y \max_\lambda \sum_i c(x_i,y_i)+\lambda\left[\rho_T(y_i|z_i,r_i)-\rho_T(y_i|r_i)\right], \tag{3.1}$$

where $y_i = T(x_i,z_i)$ and for concreteness we have adopted the cost C_1, with a pairwise cost function c that we will specialize to the canonical cost

$$c(x,y)=\frac{1}{2}\|y-x\|^2$$

in all numerical examples. (The algorithm extends with minor modifications to much more general costs; see [18] for examples of non-pairwise cost functions in a non-conditional setting.)

We can approximate both conditional densities by their Nadaraya-Waston estimates [6, 22]:

$$\rho_T(y|r_k,z_k)\approx\sum_i \mathcal{K}_a(y,y_i)\frac{\mathcal{K}_{b_r,b_z}([r_k,z_k],[r_i,z_i])}{\sum_j \mathcal{K}_{b_r,b_z}([r_k,z_k],[r_n,z_n])}, \tag{3.2}$$

$$\rho_T(y|r_k)\approx\sum_i \mathcal{K}_a(y,y_i)\frac{\mathcal{K}_{b_r}(r_k,r_i)}{\sum_n \mathcal{K}_{b_r}(r_k,r_n)}, \tag{3.3}$$

where $\mathcal{K}_h(:,c)$ is a kernel function, nonnegative and normalized so as to integrate to one, with center c and bandwidth h. We further factorize the kernels in $[r,z]$-space into the product of kernels in each individual space:

$$\mathcal{K}_{b_r,b_z}([r_k,z_k],[r_i,z_i])=\mathcal{K}_{b_r}(r_k,r_i)\ \mathcal{K}_{b_z}(z_k,z_i).$$

We have adopted isotropic Gaussian kernels in all three spaces, Y, R and Z for the numerical examples in this article, with bandwidths to be discussed below. Then we have

$$\rho_T(y|r_k,z_k)-\rho_T(y|r_k)\approx\sum_i \mathcal{K}_a(y,y_i)C_{ik}, \tag{3.4}$$

where the matrix

$$C_{ik}=\frac{\mathcal{K}_{b_r,b_z}([r_k,z_k],[r_i,z_i])}{\sum_n \mathcal{K}_{b_r,b_z}([r_k,z_k],[r_n,z_n])}-\frac{\mathcal{K}_{b_r}(r_k,r_i)}{\sum_n \mathcal{K}_{b_r}(r_k,r_n)}, \tag{3.5}$$

representing the difference between two affinity matrices, $\mathcal{Z}^{rz}$ in $[r,z]$ and $\mathcal{Z}^r$ in r space, can be pre-computed at the onset of the procedure. Then the problem (3.1) adopts the simple form

$$\min_y \max_\lambda L=\sum_i c(x_i,y_i)+\lambda\sum_{i,l}\mathcal{K}_a(y_l,y_i)C_{il}. \tag{3.6}$$

In the continuous version (2.5) of (3.6), Proposition 2.1 proves that L_F –the sum weighted by λ– is non-negative definite, vanishing only when the push forward condition

is satisfied. We need to guarantee that this is also the case for the discretization in (3.6), in order to apply a penalization method where λ is increased gradually until the push forward condition is satisfied to the level of accuracy sought. Otherwise a map T may be found that makes L negative –and large for large values of λ– in an uncontrolled way, bypassing the satisfaction of the push forward condition and the minimization of the transportation cost. Another key property that the L_F in (2.5) has and that we must enforce on its discrete version is that it vanishes when y is conditionally independent of z, i.e. when $\rho_T(y|r,z)$ is a function of r alone.

Being a discrete approximation to the continuous L_F, the discrete one in (3.6) satisfies these properties approximately too. Yet we can make them hold exactly by slightly modifying the matrix C as follows:

(1) Symmetrize $C \to \frac{1}{2}\left(C + C^T\right)$, a step of little consequence, as C only appears in L acting on the symmetric kernel $\mathcal{K}_a$.

(2) Perform the eigen-decomposition of C, keeping only the eigenvalues λ_k larger than a threshold $\epsilon > 0$ and the corresponding eigenvectors U_k (we used $\epsilon = 10^{-10}$), in order to guarantee the non-negative nature of L_F. *This step only produces changes in C of $O(\epsilon)$.*

(3) In order for C to vanish when applied to functions of r alone, we need to make each U_k orthogonal to those vectors that represent smooth functions of r. Since any such vector can be well-approximated through kernel regression in r, it must belong to the range of the affinity matrix $\mathcal{Z}^r$. Therefore we project U onto the subspace orthogonal to this range, spanned by the left singular vectors Q of $\mathcal{Z}^r$ with corresponding singular values above a threshold ϵ:

$$U \to (I - QQ^T)U.$$

This step also changes C only slightly. To see this, notice that $C(z,r)$ is orthogonal to all smooth functions of r in the limit of infinitely many sample points and vanishingly small bandwidths, since then we converge to the measure-based problem and Proposition 2.1 holds. This orthogonality must still hold approximately in the sample-based case under a robust choice of bandwidths for the kernels –discussed below–, though proving it rigorously would entail a long technical exercise.

(4) The final $\tilde{C}$ adopts the form

$$\tilde{C} = \sum_k \lambda_k U_k U_k^T.$$

To verify in an example that the resulting modification of C is small, see for instance Figure 3.1 for the matrices before and after modification in the volume reconstruction example from Section 4.1.

3.1. A normalizing flow. We build the map $y = T(x,r,z)$ through a time dependent flow T^t discretized into the composition of near-identity maps, as in normalizing flows [5, 7]. Unlike regular normalizing flows, the target distribution is generally not a Gaussian but the unknown barycenter $\mu(y)$. The corresponding current conditional distribution of x, evolving from $\rho(x|r,z)$ to $\mu(y|r)$, is represented by its samples $y_i^t = T^t(x_i, z_i, r_i)$. In most previous works in normalizing flows and flow-based optimal transport [16, 20], each near-identity map is a simple function with a small number of parameters to optimize over, as first proposed in [15]. We adopt here instead the free-flow approach of [18], where the y_i are individual degrees of freedom, updated

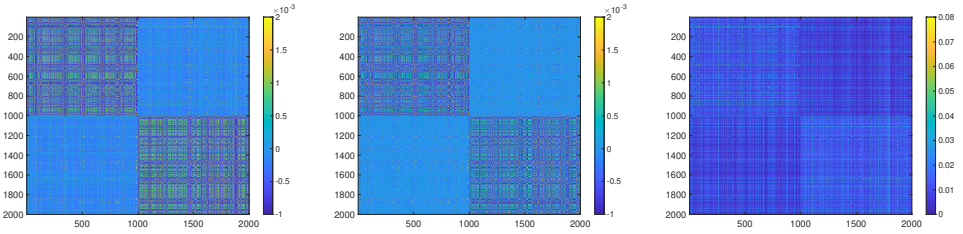

FIG. 3.1. *The C matrix before (left) and after (middle) modification in the volume reconstruction example, and their relative difference (right).*

through gradient descent of L. The smoothness of the underlying map T follows from the smoothness in y, r and z of the kernels defining L. There is no need to propose a closed form for T in an optimal transport setting since, unlike in Kullback–Leibler divergence driven normalizing flows, the procedure does not require the explicit knowledge of the Jacobian of the map. Treating the y_i as independent degrees of freedom makes the algorithm essentially non-parametric, except for the choice of bandwidths for the three kernel functions.

The algorithm starts with $y_i = x_i$ and small initial values for the learning rate η and for the penalty coefficient λ. Then at every step n, we perform the following updates:

(1) Tentatively increase η through

$$\eta \to \alpha\eta, \quad \alpha > 1,$$

where α has been determined via cross validation in the examples below: once near optimality, we want η to increase, since the implicit gradient descent procedure converges to Newton's as $\eta \to \infty$.

(2) Writing

$$L = C + \lambda L_F, \quad C = \frac{1}{n}\sum_i c(x_i, y_i), \quad L_F = \frac{1}{n}\sum_{i,l} \mathcal{K}_a(y_i, w_l) C_{il},$$

where the $w_l = y_l$ are considered as fixed parameters, compute $\nabla_y C$ and $\nabla_y L_F$, their derivatives with respect to the y_i. The reason not to name y_l the second argument of $\mathcal{K}_a$ and differentiate with respect to it too, is that the two arguments of $\mathcal{K}_a$ play different roles: the first represents the value of y at which F is applied, the second a parameter defining the test function F itself. The objective function L must be minimized over y only in its first role: recall that, in the original minimax formulation, L was not minimized but maximized over F.

(3) Update the penalty coefficient λ through

$$\lambda^{n+1} = \min\left\{\lambda_{max}, \max\left\{\lambda^n, \gamma - \frac{\langle \nabla_y C, \nabla_y L_F\rangle}{\langle \nabla_y L_F, \nabla_y L_F\rangle}\right\}\right\} \quad \gamma > 0.$$

The rationale behind this choice is to have λ (a) never decrease, (b) not become so large as to force the procedure to overfit the push forward condition, and (c) be large enough to make the total gradient ∇L point toward the satisfaction of the push forward constraint –as opposed to the direction of decreasing transportation cost, which would have the y_i return toward their corresponding original x_i. In the runs below, we have adopted $\lambda_{max} = 1000$ and $\gamma = 0.1\lambda_{max}$.

(4) Compute

$$\nabla_y L = \nabla_y C + \lambda^{n+1} \nabla_y L_F$$

and the Hessian

$$H_i^k = \frac{\partial^2}{\partial y_i \partial y_k} L,$$

with special care regarding the second derivatives of $\mathcal{K}_a$, which must be read as

$$\frac{\partial^2}{\partial y^2} \mathcal{K}_a(y,w)\Big|_{y=y_i, w_k=y_k} + \frac{\partial^2}{\partial y \partial w} \mathcal{K}_a(y,w)\Big|_{y=y_i, w_k=y_k},$$

since the implicit anticipation of the future value of ∇_y requires a prediction not only for y but also for $w=y$.

(5) Perform a tentative implicit gradient descent step [3] with the current value of the learning rate η:

$$\tilde{y} = y - \eta (I + \eta H)^{-1} \nabla_y L,$$

and verify that

$$L\left(\tilde{y}, \lambda^{n+1}\right) \leq L\left(y, \lambda^{n+1}\right).$$

If not, reduce $\eta \to \frac{1}{2}\eta$ and repeat until L does decrease. Then set $y = \tilde{y}$.

(6) Stop when a termination criterion is met, $\|\nabla_y L\| \leq \epsilon$, and verify that $L_F \leq \epsilon_F$.

3.2. Choice of bandwidths. The penalty term L_F involves three different bandwidth parameters: b_r for r, b_z for z and a for y. While the first two are fixed throughout the procedure, a must evolve from a large initial value associated with the coarse features of $\rho(x)$ to a smaller one resolving the fine structure of the barycenter $\mu(y)$.

If either r or z are categorical variables, the corresponding bandwidths must vanish, with the kernels replaced by binary indicator functions, so the description that follows applies to continuous r and z. Yet keeping the discrete case in mind helps conceptualize the choice of a bandwidth. If r were categorical, we would –at least to some extent– perform an independent regular barycenter problem for the distribution $\rho(x|z)$ for each value of r. This suggests that, for continuous r, the bandwidth adopted should not be very small, or else each "individual barycenter problem" would contain too few samples. In the examples below, we have adopted $b_r = \sigma_r$, the standard deviation of $\eta(r)$. Given b_r, we define the corresponding 'effective sample size' as

$$n(r) = \sum_i \frac{K^r(r_i, r)}{K^r(r_i, r_i)},$$

a number that will be used in the selection of the remaining bandwidths.

For continuous $z \in \mathbb{R}^{L_z}$, the kernel can be chosen via Silverstein's rule-of-thumb, with appropriate modification to account for the dependence on r as follows:

$$b_z = \left(\frac{4}{L_z + 2}\right)^{1/(L_z+4)} SD(z) \min_i \{n(r_i)\}^{-1/(L_z+4)}.$$

Given b_r and b_z, the effective number of samples for jointly given values or r and z is

$$n(r,z)=\sum_i \frac{K^{r,z}([r_i,z_i],[r,z])}{K^{r,z}([r_i,z_i],[r_i,z_i])}.$$

The initial value h_y^0 for the bandwidth in y can be conceptualized in terms of preconditioning, addressing only the coarser features of $\rho(x|r,z)$, such as its conditional mean and variance. The examples below use $h_y^0=SD(x)$. For the final value, instead, we use

$$h_y^{end}=\left(\frac{4}{L_y+2}\right)^{1/(L_y+4)} SD(x-xC)\min_i\{n(r_i,z_i)\}^{-1/(L_y+4)}.$$

Note that the standard deviation is evaluated on a surrogate for y: the residual of x after applying kernel regression on z conditioned on r, which reflects the rough features of the conditional barycenter. Then, at step n, we adopt

$$a=max\left\{h_y^0+\frac{n}{n_{end}}\left(h_y^{end}-h_y^0\right),h_y^{end}\right\},$$

where n_{end} is an estimate of the number of steps required by the algorithm (the inversion procedure described below requires that, after $n=n_{end}$ we still perform additional steps with a and λ kept constant until reaching convergence.)

3.3. Map inversion. Simulating the conditional distribution $\rho(x|r,z)$ for specific values (r^*,z^*) of r and z requires mapping back the samples y_i in the barycenter to

$$x_i^*=T^{-1}(y_i,r^*,z^*).$$

Yet we only have access to the map $y=T(x,r,z)$ through its action on the original samples, $y_i=T(x_i,r_i,z_i)$. Thus we need a procedure to invert this sample-given T.

One possibility is to perform a nonlinear regression based on the available points. A kernel-based regression is a natural candidate, since we already have kernels available. Yet there exists a faster and more accurate way to perform this inversion. It assumes that we have run the algorithm to convergence so that, for all practical purposes, $\frac{\partial L}{\partial y_i}=0$. Using (3.6), this reads

$$\frac{\partial c(x_i,y_i)}{\partial y_i}+\lambda\sum_l C_{il}\frac{\partial}{\partial y_i}\mathcal{K}_a(y_i,y_l)=0,$$

with care taken to differentiate the kernels $\mathcal{K}_a$ only with respect to their first argument. This is a relation between x_i and y_i that can immediately be extended to a smooth regression for arbitrary values of y. In particular, for the canonical cost $c(x,y)=\frac{1}{2}\|y-x\|^2$, it reads

$$x=T^{-1}(y,r_i,z_i)=y+\lambda\sum_l C_{il}\frac{\partial}{\partial y}\mathcal{K}_a(y,y_l).$$

(The only requirement for extending this procedure to more general cost functions is that the expression $\frac{\partial c(x,y)}{\partial y}=\alpha$ be invertible for x.) For each pair (r_i,z_i) in the data set, this provides n samples of $\rho(x|r_i,z_i)$, including x_i, the only one originally available in the data. If required to simulate $\rho(x|r,z)$ for pairs (r,z) not in the dataset, one needs to generate a new row for the matrix C. For this, one can simply choose $C(i,:)$, where (r_i,z_i) is the closest neighbor to (r,z), or more accurately interpolate among the rows of C corresponding to K near neighbors of (r,z).

4. Examples

4.1. Super-resolution volume reconstruction from slice acquisitions. This subsection illustrates the use of the conditional barycenter problem on simple synthetic data designed to mimic volume reconstruction from slice acquisition in the analysis of medical images.

The data points on the left panel of Figure 4.1 are triplets (x_1, x_2, z) drawn from a distribution $\rho(x_1, x_2, r, z) = \rho(x_1, x_2|r, z)\ \eta(r|z)\ \gamma(z)$, where r shows through the coloring of the markers. The distribution $\gamma(z)$ is uniform in $[0,1]$. The distributions $\rho(x_1, x_2|r, z)$ and $\eta(r|z)$ are constructed in the following way: the observations are assigned to two different branches, colored in yellow and blue in Figure 4.1, according to a Bernoulli random variable with probability $p = 0.5 + 0.4\tanh(3(z-1/2))$; such assignment is reflected in the sign of r, positive and negative in the yellow and blue branches respectively. The absolute value of r encodes the distance d to the origin of an auxiliary variable $\tilde{x}$ distributed uniformly in the two-dimensional disk or radius 0.3, through $|r| = \exp(-d^2)$. Thus, for the yellow and blue branches, r takes values in $[e^{-0.09}, 1]$ and $[-1, -e^{-0.09}]$ respectively. In Figure 4.1, the absolute value of r is represented through color intensity.

In order to generate x, we apply to $\tilde{x}$ an affine map depending on z and the sign of r:

$$x = A(z)\tilde{x} + B(r, z), \quad \tilde{x} \sim U_{D_{0.3} = \{\bar{x} \in \mathbb{R}^2 : ||\bar{x}||_2 \leq 0.3\}}, \tag{4.1}$$

with

$$A(z) = \begin{bmatrix} -\frac{9}{5}\cos(2\pi z) & \frac{9}{5}\sin(2\pi z) \\ \frac{5}{4}\sin(2\pi z) & \frac{5}{4}\cos(2\pi z) \end{bmatrix}, \tag{4.2}$$

$$\begin{aligned} B(r, z) = \mathbb{1}_{blue} \left(\frac{1}{2} \begin{bmatrix} \cos(2\pi z) - \sin(2\pi z) \\ -\sin(2\pi z) - \cos(2\pi z) \end{bmatrix} + \frac{2}{5} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right) \\ + \mathbb{1}_{yellow} \left(\frac{1}{2} \begin{bmatrix} -\cos(2\pi z) + \sin(2\pi z) \\ \sin(2\pi z) + \cos(2\pi z) \end{bmatrix} + \frac{2}{5} \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right), \end{aligned} \tag{4.3}$$

whereby the discs are separated by color, rescaled into ellipses and rotated.

Our tasks are to compute the z-barycenter $\mu(y|r)$ of the $\rho(x|r, z)$ and to simulate $\rho(x|r, z^*)$ for given values of $z = z^*$. Since for each value of z there is at most one sample available from the probability distribution, the latter task mimics the reconstruction of high-resolution images from sets of very sparse observations. Examples include tomography, where z represents longitudinal distance and the subject is moving rapidly, hence the rotations and the short exposure time of each image. In the visualization of embryo development, z represents time and r labels the organs, whose evolution is mimicked in our example by the changing mass in the yellow and blue systems.

The top right panel of Figure 4.1 shows the points in the barycenter $\mu(y|r)$ as a function of their original label z, showing that the dependence on the parameter z has indeed been removed. The bottom left panel, by contrast, displays all available samples from $\mu(y|r)$ without their original label z. This result could not have been obtained from the classical barycenter problem, in which the map $T(x, r, z)$ pushes forward $\rho(x|r, z)$ to the single probability density function $\tilde{\mu}(y)$. Samples from the classical barycenter would, in this example, fall on a closed line, since only the angle in $\tilde{x}$ remains unexplained when both r and z are taken into account. Moreover, the blue and yellow points would be mixed, as would their color intensities, since y could not depend on r. This would not

serve the purpose, for instance, of reconstructing the shape of evolving organs, when z represents time, the blue and yellow signal different organs and their intensity indicates a radial coordinate along them. Finally, the bottom right panel of Figure 4.1 shows the high resolution reconstruction of slices for different values of z, reminiscent of volume reconstruction, computed through the inversion procedure described in Section 3.3. Ellipses with varying semi-axis r and center, are reconstructed robustly, even though only one sample point is available in the data for each value of z.

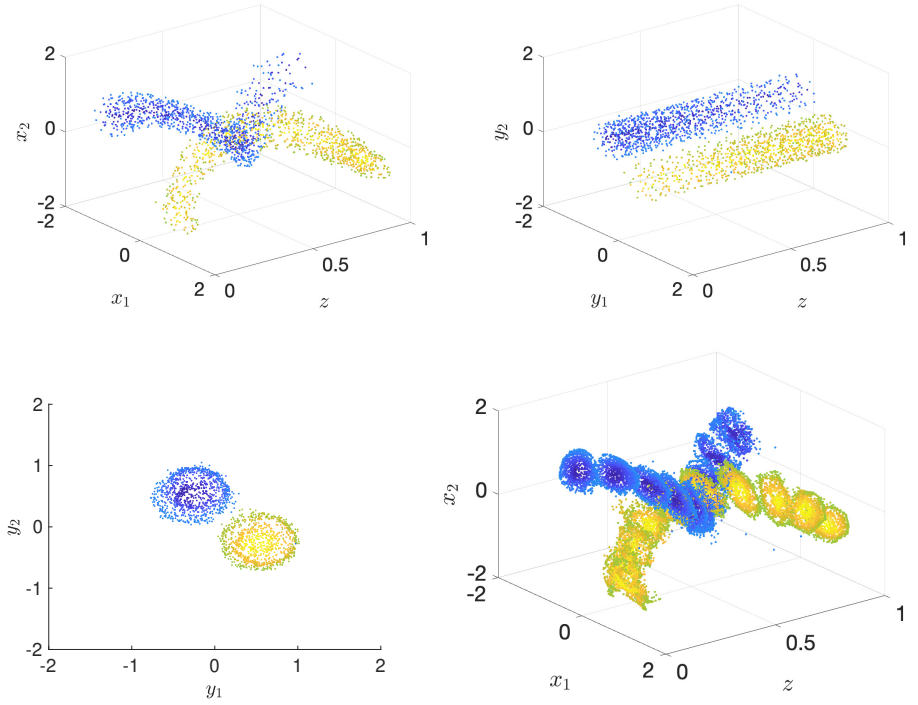

FIG. 4.1. *Top left: original input data sampled from the $\rho_1(x|r,z)$ (blue) and $\rho_2(x|r,z)$ (yellow). Top right: removing the dependence from z results in two steady images for the blue and the yellow density. Bottom left: barycenter $\mu(y|r)$. Bottom right: high resolution reconstruction of the input data in the top left panel for various values of z.*

4.2. Treatment effects: two complementary approaches. A significant application of the conditional barycenter problem is to the estimation of treatment effects. Let us introduce some general terminology:

> **Treatment r:** the action taken. It can adopt binary values –to treat or not–, continuous –a dosage, the reduction of emitted CO_2–, or far more general structures –to treat or not and under which protocol, a diet, an economic package.
>
> **State x:** variables possibly affected by the treatment, such as a patient's blood pressure, the average sea-surface temperature and global inequality indexes.
>
> **Factors z:** the age of a patient, the state x before treatment, the distribution of income in a population.

We describe below two alternative uses of the conditional barycenter problem for the assessment of treatment effects. The map $y = T(x,r,z)$ pushes forward the $\rho(x|r,z)$

to their r-barycenter $\mu_1(y|z)$ in the first use, and to their z-barycenter $\mu_2(y|r)$ in the second. Figure 4.2 shows an example of synthetic data illustrating the discussion of the Simpson paradox in Section 1. Here x represents arterial pressure, r a dosage of a pressure-reducing drug, and z the patient's age. Because young people are less likely to develop high pressure, their prescribed values of r will typically be lower than for older people. As a consequence, even though for each fixed value of the age z, x decreases with r, a plot of x against r for the whole population may display x growing with r, contrary to the actual effect of the drug considered, a typical instance of the Simpson paradox.

In order to quantify the paradox, we fitted a line to the data by minimizing the L_1 residual. Ignoring the age z results in a positive slope for the best fit line for the state variable x as function of the treatment r, the opposite of the actual effect of the treatment for any fixed age.

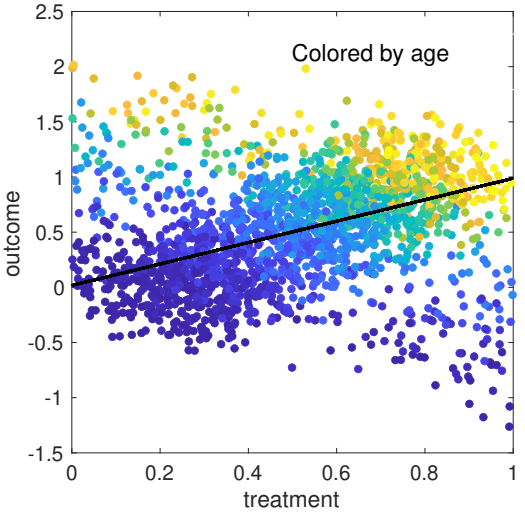

FIG. 4.2. *Data displaying the Simpson paradox and the regression line of outcome given treatment that minimizes the L_1 residual. For each patient's age, represented through color, larger values of the treatment yield lower values of the outcome. Yet when all ages are combined, the treatment–outcome relationship is inverted, as showed by the regression line: outcome* $= 0.9701 \times$ *treatment* $+ 0.0178$.

The literature on treatment effect estimation is vast [2,4,11,21], most of it focusing on the computation of the average treatment effect $\tau = E[Y(1) - Y(0)]$ (where $y(1) = y$ represents the value of the outcome in presence of the treatment and $y(0) = x$ in the non-treated case) or the average treatment effect conditional on the some covariates $\tau(z) = E[Y(1) - Y(0)|Z = z]$. This section presents a different approach, which further extends the idea introduced in [17], by leveraging the ability of the conditional barycenter problem of selectively removing from a data set the variability attributable to a specific set of cofactors while preserving the dependence on others.

4.2.1. The conditional barycenter $\mu(y|r)$ as a summary of the treatment effects. One may wish to visualize the effect of a treatment r on a variable x without providing a detailed description on how this effect depends on the covariates z. Yet, as seen in Figure 4.2, forgetting the factor z altogether in the description may convey a treatment effect with the wrong sign, a manifestation of Simpson's paradox. To sum-

marize the true treatment effect over all ages, one needs to average over z all conditional probabilities $\rho(x|r,z)$. The barycenter $\mu(y|r)$ provides a natural, *horizontal* way of performing this averaging. In order to give equal weight to all ages z for each value of r, we use as transportation cost C_2 from (2.2). This requires weighting the samples in the cost function by a factor

$$w_i = \frac{\eta(r_i)}{\eta(r_i|z_i)} \approx \frac{\left(\sum_j \mathcal{K}_{b_r}(r_i,r_j)\right)\left(\sum_j \mathcal{K}_{b_z}(z_i,z_j)\right)}{N\sum_j \mathcal{K}_{b_r,b_z}([r_i,z_i],[r_j,z_j])}.$$

The resulting barycenter is displayed in Figure 4.3. Contrasting it to Figure 4.2, one can see a much decreased variability, as the effect of the age z has been explained away, and a resolution of the Simpson paradox, as now y decreases with the treatment r, since x decreases with r for any fixed value of z.

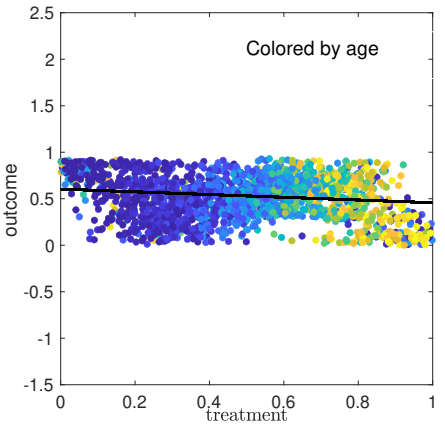

FIG. 4.3. *Conditional barycenter $\mu(y|r)$ and the regression line of outcome given treatment that minimizes the L_1 residual. The regression line is outcome $= -0.1467 \times$ treatment $+ 0.6036$. The barycenter summarizes the treatment effect regardless of the age z of the patients, displayed through color. A larger dosage of the treatment results in a lower value of the outcome, even though the barycenter ignores the age of the patient.*

4.2.2. The map T as a representation of the treatment. How can one quantify the effect of a treatment? For a binary treatment $r \in \{0,1\}$, one should compare the values of x with and without treatment, x_1 and x_0. Since both are random variables –seldom does one have full control over all factors affecting x– we are really comparing the conditional distributions $\rho(x|z,r=0)$ and $\rho(x|z,r=1)$. A natural way to perform this assessment is through the estimation of a map $T(x,z)$ that pushes forward $\rho(x|z,r=0)$ to $\rho(x|z,r=1)$. Such map has a natural interpretation as the effect of the treatment on x, conditioned on z.

Moving next to more general, non-binary treatments, the goal broadens to estimating how r affects the state x, i.e. how the distribution $\rho(x|z,r)$ depends on r. Again, this can be achieved through a family of maps T_r that transform the conditional distributions $\rho_{r_{1,2}}(x|z)$ corresponding to pairs of values of r into each other. More economically, these pairwise maps can be described through invertible, r-dependent maps that, for each value of r, push forward $\rho(x|z,r)$ to a common conditional distribution $\mu(x|z)$. When these maps minimize a transportation cost, $\mu(x|z)$ becomes the conditional barycenter of the $\rho(x|z,r)$.

The characterization of treatment effects above assumes that we have access to the distributions $\rho(x|z,r)$. A problem arises though when the treatment r depends on some factor w that may also affect x, i.e. a confounder. Consider, for instance, a binary treatment $r \in \{0,1\}$ for high blood pressure x. If this treatment (i.e. $r=1$) is provided only to patients above a certain age, and high blood pressure is more prevalent for older patients, we could observe that the distribution of blood pressure for treated patients concentrates at higher values than for untreated ones, even for effective treatments, an instance of the Simpson paradox. In terms of maps, T should not push forward $\rho(x|r=0)$ to $\rho(x|r=1)$, since a share of the population at $r=0$, the younger patients, have never been treated, so they do not show up for $r=1$: mass conservation does not hold between the two distributions. If we could consider the two distributions but only for patients above the threshold age for treatment, this problem would disappear.

But this is exactly the case when w, the age, is included among the factors z in a conditional transport problem. Then the push forward condition holds for each value of z, and the procedure retains its validity. Similarly, the state x before treatment could play the role of w, when only patients with high pressure are subject to treatment. Again, including this prior condition among the z solves the problem.

This makes a strong case for the use of conditional optimal transport, as well as the conditional barycenter problem for non-binary treatments. Expanding the set of factors z so as to more likely include most confounders makes the estimation of treatment effects more robust to instances of the Simpson paradox.

We illustrate the characterization and prediction of treatment effects using the same synthetic dataset as in the prior subsection. For each individual patient with age z, treatment level r and outcome x, in order to predict the outcome after switching to treatment r^*, we may adopt the identical two-step procedure as in the super-resolution example, though exchanging the roles of z and r, as follows: (1) first we find the map from the original dataset which removes the conditional dependence on the treatment. The resulting barycenter $\mu(y|z)$, shown in Figure 4.4 (upper left), is free of treatment effects, given the age, and will be used for simulating outcome at different treatment levels. (2) For each treatment level r^*, we apply the inverse map for their corresponding age z to obtain the predicted outcome. Figure 4.4 compares the true model prediction for each treatment level (upper right) and the data-driven prediction (lower left). Each vertical bar contains information for all patients present in the dataset at their corresponding age z, but is now associated with the treatment as indicated on the horizontal axis. The difference between truth and prediction is displayed through the root-mean-square-error (RMSE), averaged over the whole population, for each treatment level (Figure 4.4, lower right, green line labeled 'CBary'). Three other methodologies for the same purpose are compared: (1) CBary+Bary: first remove the conditional effect on age, then effect on treatment, then apply two inverse maps in the procedures for prediction, (2) Bary: remove effects on both age and treatment in one step and use the inverse map for prediction, and (3) 2Bary: remove first the effect of age, then effect of treatment, through two barycenter problems, and use inverse maps for prediction. For all treatment levels, the proposed procedure reaches the minimal error among all four procedures. The reason of this superior performance may be explained by minimal accumulated numerical error: procedures (1), (2) and (3) all first explain away the effect of z and then bring it back at the same value of z, while the procedure proposed only removes and restores (at a different value of r) the variability explained by r at constant z.

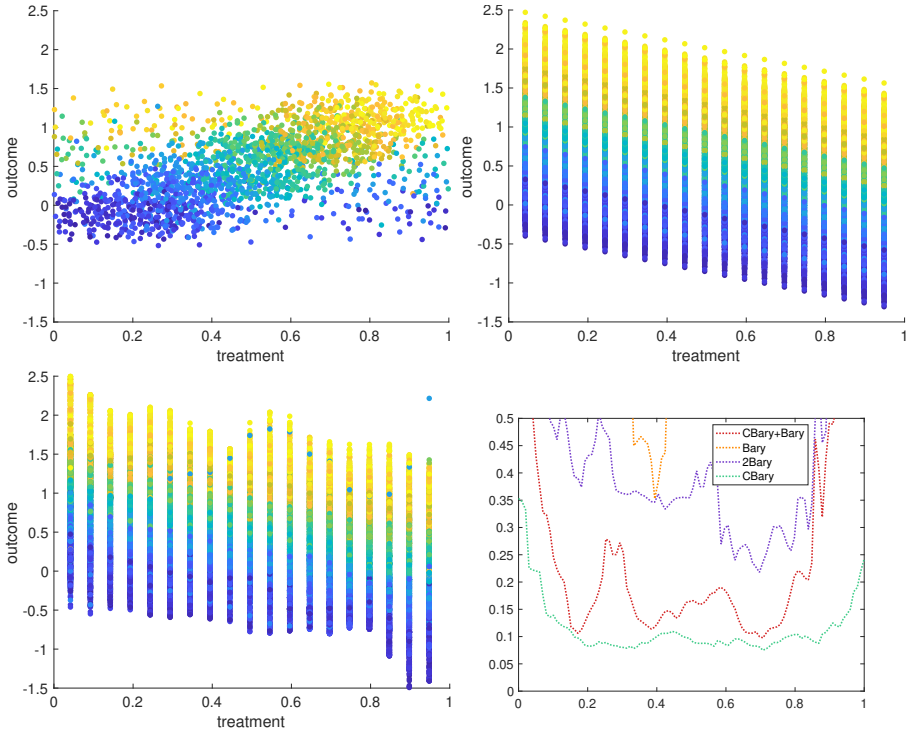

FIG. 4.4. *Upper left: conditional barycenter after removing the treatment effect (nearly horizontal for each age). Upper right: true treatment effect for the whole population at different treatment levels. Lower left: prediction of the previous subplot, by applying the inverse of the map. Lower right: RMSE as a function of treatment level for the procedure proposed and three alternatives.*

4.3. Ground temperature anomalies. This section presents a meteorological example to further illustrate the applicability of the conditional barycenter with real data. Here we use daily averaged measurements of the ground-level temperature in Ithaca, NY, publicly available from NOAA (https://www1.ncdc.noaa.gov/pub/data/uscrn/products/daily01/). The overall goal is to characterize temperature anomalies on the yearly time scale. The variable of interest, $x \in \mathbb{R}$, is the ground temperature, measured in degrees Celsius, available from January 2011 to June 2022 (4116 data points). A natural set of covariates is the following:

(1) The day of the year $r \in [0, 365.25]$, modeled as continuous and periodic, required to capture the seasonal cycle.

(2) The year $z \in [2011, 2012, \cdots, 2021, 2022]$, modeled as a categorical variable, capturing inter-annual variability, which may reflect relevant global phenomena, such as El Niño or global warming.

The original data is shown in the top panels of Figures 4.5 and 4.6, which display in different ways the dependence of the ground temperature on both covariates. To investigate variation over one time scale without being confused by the other, one needs to decompose the variability into two parts, explained by r and z respectively.

The removed variability itself is also informative as a new notion of temperature anomaly, the departure from the year-independent temperature distribution character-

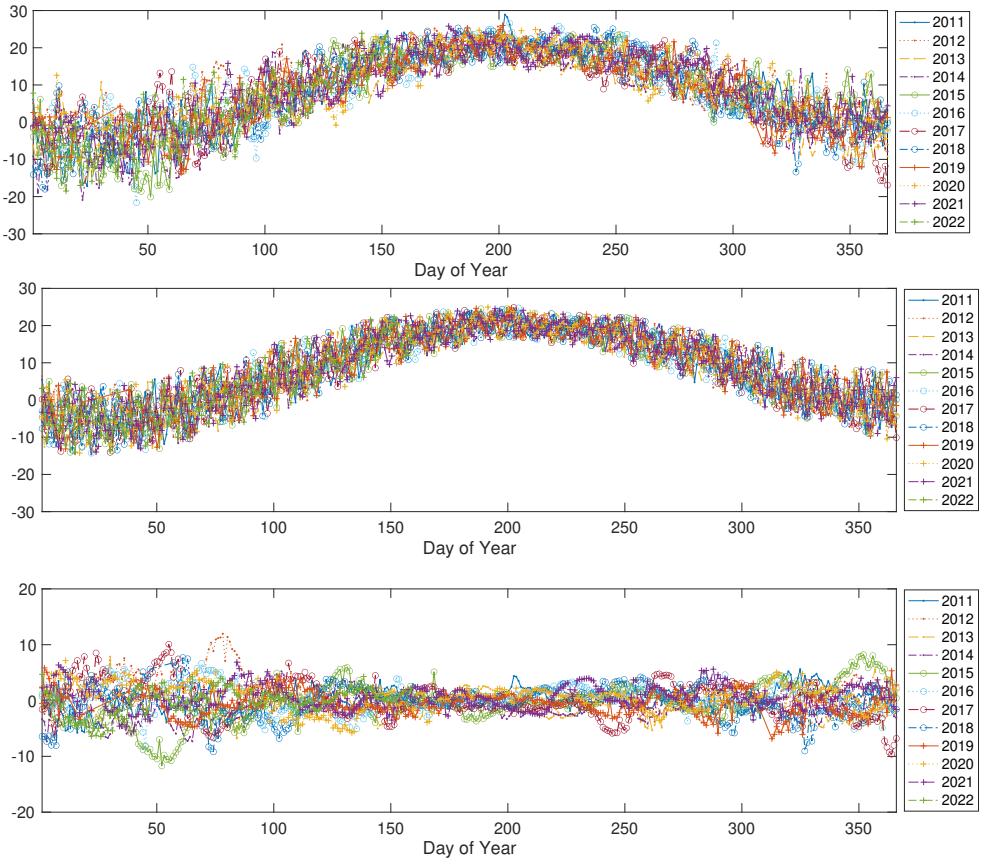

FIG. 4.5. *Daily averaged ground-level temperature as functions of day of year r. The year z in which the data are measured is indicated via colors. From top to bottom: original data x, points $y=T(x,r,z)$ of conditional barycenter $\mu(y|r)$, and anomalies measured as $x-T(x,r,z)$. Notice that, in displaying the points y in the barycenter, we still color them according to the color z relative to the year they were originally from. This allowed us to find the year in which a temperature anomaly occurred as a function of the day.*

ized by the conditional barycenter. The difference between each data point x and its image $y=T(x,r,z)$ in the conditional barycenter serves this purpose, as it represents the anomalous behavior attributable to the year. This is a much more appropriate description of anomalies than the standard difference from the mean (the *climatology* in climate studies), as it takes into account the full distribution, not just its mean value. The differences $x-T(x,r,z)$ are shown in the bottom panels of Figures 4.5 and 4.6. These anomalies display, for instance, how extreme were the beginning and end of the the year 2015, in which one of the strongest El Niño event occurred.

4.4. Lightness transfer. This section applies the conditional barycenter to lightness transfer problem in image processing. Consider the four images of different flowers photographed under different light conditions in the top panel of Figure 4.7. Lightness is distributed unequally between the purple flower and the green background across the different images. We seek to transform all the images to some notion of

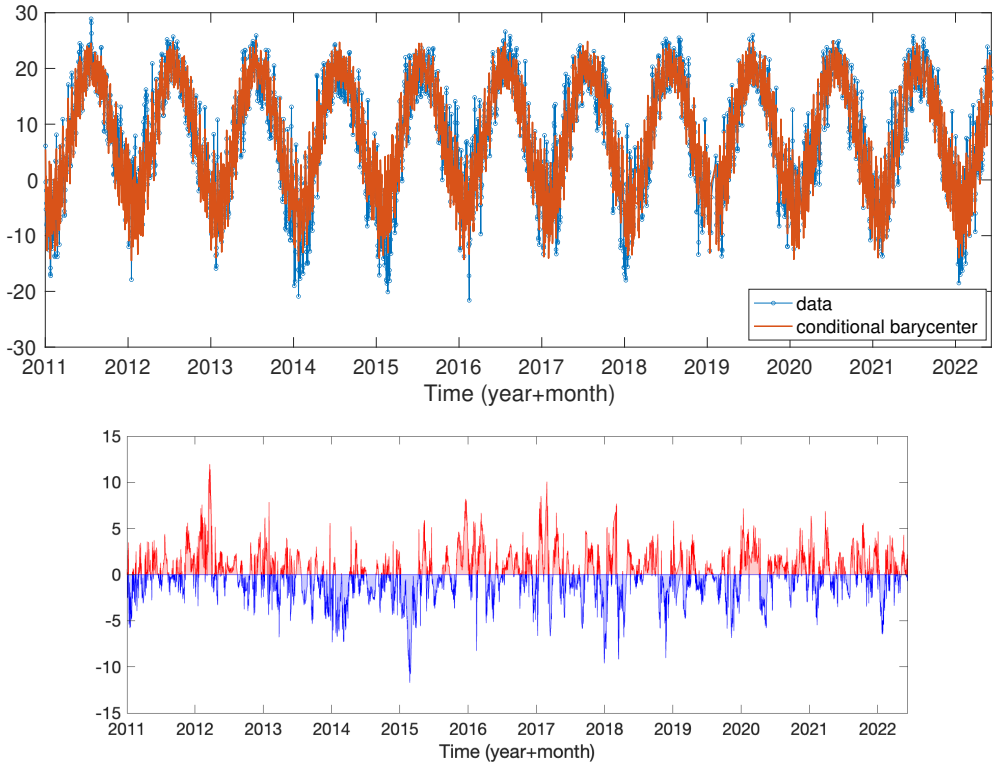

FIG. 4.6. *Same as Figure 4.5, but with the years z displayed in succession, not through colors. Top: original data x and data $y=T(x,r,z)$ in the conditional barycenter, which removes the conditional effect of z. Bottom, anomaly $x-T(x,r,z)$.*

"average" light condition and, using the transportation map and its inverse, to the light condition in a specific image. The bottom panels show the images represented in the three dimensional CIELAB space, with coordinates L, A and B for lightness, red/green contrast and blue/yellow contrast respectively.

It is in this space that we perform conditional optimal transport following 3 main steps:

(1) We first reduce the number of sample points in each image by defining superpixels following the work in [10,19], which roughly cluster each image into 200 superpixels.

(2) We apply the conditional barycenter procedure to the reduced CIELAB representation, using the lightness L for x, the color contrasts A and B for r, and the image identity (a categorical variable with four values) for z. The rationale for conditioning lightness on color is that a changing light setting affects the various colors differently.

(3) Finally, we apply a TMR filter [9] to recover the sharp details in the original images under the new, average palette.

Figure 4.8 presents results of the conditional and the regular barycenter problems, where the light-transfer in the latter simply ignores the color contrasts A and B. The main luminosity difference among the images lies in the green background. Consequently, a conditional barycenter is required to determine how to resolve the discrepancy in lightness. Ignoring the foreground-background contrast, the regular barycenter

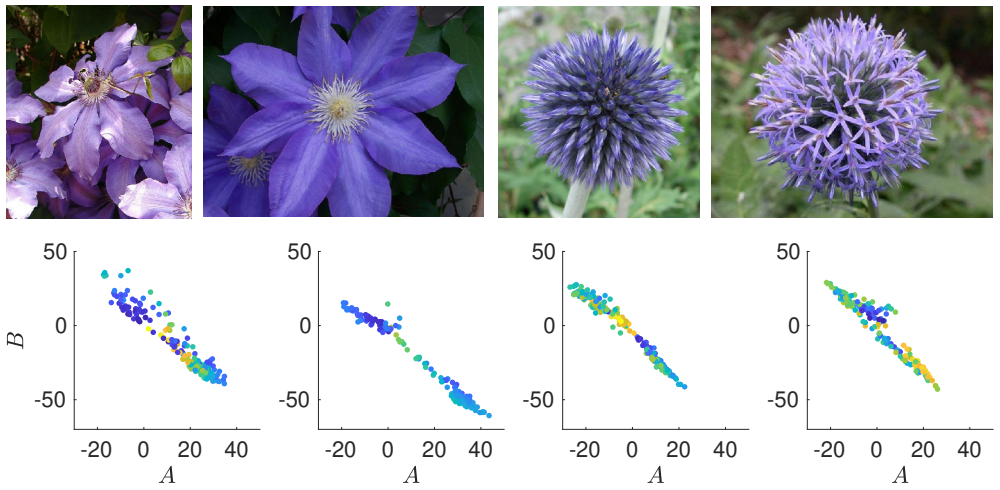

FIG. 4.7. *Source images (top) and their representation in CIELAB coordinates (bottom). The lightness coordinate L is represented through the color on the scatter plot.*

(a) Source images

(b) Light transfer results using the conditional barycenter

(c) Light transfer results using the regular barycenter

FIG. 4.8. *Source images (top), conditional barycenter (middle), and regular barycenter (bottom).*

FIG. 4.9. *Source images adapted to light condition in the first image (bright flower, dark background, top) and third image (dark flower and brighter background, bottom).*

adjusts the light homogeneously, while the conditional barycenter is able to adjust the background independently, with minimal alteration in the flowers. Consider for instance the first image, with a dark background, and the third image, with a much brighter one. Unlike the regular barycenter, which could not alter this contrast, the light transfer to the conditional barycenter managed to make the two backgrounds about equally bright without substantially altering the lightness of the flower that appears to be much brighter in the regular barycenter.

The conditional barycenter procedure also allows one to transform back the conditional barycenter to any of the four $\rho_{i^*}(L|A,B)$, using the map inversion described in Section 3.3. This enables us to adapt each source image to any of the four light conditions in the source images. Two light conditions are used as targets for illustration in Figure 4.9. The first image has a brighter flower and a dark background; it was likely taken in the shade on a not very bright day. The third image, with a well-lit background, was probably taken in the morning. After the two corresponding inverse maps are applied to the barycenter, all four resulting images have the luminosity features of the corresponding target.

5. Conclusions

This work extends the optimal transport barycenter to the conditional barycenter problem, which selectively removes from data the variability attributable to some cofactors while preserving the dependence on others. This allows, for instance, to characterize the effect of a medical treatment on a heterogeneous population as a function of selected cofactors, such as dosage, irrespective of others, such as the patient's age.

In order to pose the data-driven conditional barycenter problem and solve it numerically, we introduce a new class of normalizing flows that extend the work in [18]. This procedure is essentially non-parametric, having as only tunable parameters the bandwidths of three kernel functions.

The conditional barycenter problem provides a new conceptual and a computational framework for data analysis. Numerical examples illustrate various uses of this tool: to mitigate the Simpson paradox, to provide a new characterization of anomalous variation in climate data, to volume reconstruction from slice acquisitions and to lightness transfer

in image processing. Despite their broad scope, these applications may only scratch the surface of the conditional barycenter's rich field of applicability.

Acknowledgments. Tabak's work was partially supported by ONR grant N00014-15-1-2355.

REFERENCES

[1] M. Agueh and G. Carlier, *Barycenter in the Wasserstain space*, SIAM J. Math. Anal., 43(2):904–924, 2011. 1

[2] R.K. Crump, V. Joseph Hotz, G.W. Imbens, and O.A. Mitnik, *Nonparametric tests for treatment effect heterogeneity*, Rev. Econ. Stat., 90(3):389–405, 2008. 4.2

[3] M. Essid, E. Tabak, and G. Trigila, *An implicit gradient-descent procedure for minimax problems*, Math. Meth. Oper. Res., 97:57–89, 2023. 5

[4] K. Hirano, G.W. Imbens, and G. Ridder, *Efficient estimation of average treatment effects using the estimated propensity score*, Econometrica, 71(4):1161–1189, 2003. 4.2

[5] I. Kobyzev, S.J.D. Prince, and M.A. Brubaker, *Normalizing flows: An introduction and review of current methods*, IEEE Trans. Pattern Anal. Mach. Intell., 43(11):3964–3979, 2020. 3.1

[6] E.A. Nadaraya, *On estimating regression*, Theory Probab. Appl., 9(1):141–142, 1964. 3

[7] G. Papamakarios, E.T. Nalisnick, D.J. Rezende, S. Mohamed, and B. Lakshminarayanan, *Normalizing flows for probabilistic modeling and inference*, J. Mach. Learn. Res., 22(57):1–64, 2021. 3.1

[8] K. Pearson, A. Lee, and L. Bramley-Moore, *Reproductive or genetic selection*, Science, 9(217): 283–286, 1899. 1

[9] J. Rabin, J. Delon, and Y. Gousseau, *Removing artefacts from color and contrast modifications*, IEEE Trans. Image Process., 20(11):3073–3085, 2011. 3

[10] J. Rabin, S. Ferradans, and N. Papadakis, *Adaptive color transfer with relaxed optimal transport*, 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 4852–4856, 2014. 1

[11] P.R. Rosenbaum and D.B. Rubin, *Constructing a control group using multivariate matched sampling methods that incorporate the propensity score*, Amer. Stat., 39(1):33–38, 1985. 4.2

[12] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, Springer, 2015. 1

[13] E.H. Simpson, *The interpretation of interaction in contingency tables*, J. Royal Stat. Soc. Ser. B, 13(2):238–241, 1951. 1

[14] S. Srivastava, V. Cevher, Q. Dinh, and D. Dunson, *WASP: Scalable Bayes via barycenters of subset posteriors*, Proc. Mach. Learn. Res., 912–920, 2015. 1

[15] E.G. Tabak and C.V. Turner, *A family of nonparametric density estimation algorithms*, Commun. Pure Appl. Math., 66(2):145–164, 2013. 3.1

[16] E.G. Tabak and G. Trigila, *Explanation of variability and removal of confounding factors from data through optimal transport*, Commun. Pure Appl. Math., 71(1):163–199, 2018. 3.1

[17] E.G. Tabak, G. Trigila, and W. Zhao, *Data driven conditional optimal transport*, Mach. Learn., 110(11):3135–3155, 2021. 4.2

[18] E.G. Tabak, G. Trigila, and W. Zhao, *Distributional barycenter problem through data-driven flows*, Pattern Recognit., 130:108795, 2022. 1, 3, 3.1, 5

[19] Y.-W. Tai, J. Jia, and C.-K. Tang, *Local color transfer via probabilistic segmentation by expectation-maximization*, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), IEEE, 1:747–754, 2005. 1

[20] G. Trigila and E.G. Tabak, *Data-driven optimal transport*, Commun. Pure Appl. Math., 66:613–648, 2016. 3.1

[21] S. Wager and S. Athey, *Estimation and inference of heterogeneous treatment effects using random forests*, J. Amer. Stat. Assoc., 113(523):1228–1242, 2018. 4.2

[22] G.S. Watson, *Smooth regression analysis*, Ind. J. Stat. Ser. A, 359–372, 1964. 3

[23] H. Yang and E.G. Tabak, *Conditional density estimation, latent variable discovery, and optimal transport*, Commun. Pure Appl. Math., 75(3):610–663, 2022. 1