Optimal transport with cost-free transformations for image co-registration

Yating Wang[1], Esteban Tabak[1], Debra F. Laefer[2,3]

[1] Courant Institute of Mathematical Sciences, New York University (NYU) New York, NY
(yw3087@nyu.edu, tabak@cims.nyu.edu)
[2] Dept. of Civil and Urban Engineering, Tandon School of Engineering, NYU, Brooklyn, NY
[3] Center for Urban Science and Progress, Tandon School of Engineering, NYU, Brooklyn, NY
(debra.laefer@nyu.edu)

Keywords: optimal transport, data co-registration, image co-registration, estimation of treatment effects.

Abstract

An extension of the optimal transport problem is proposed, which includes a family of transformations incurring no transportation costs. This extension improves the co-registration among imagery datasets where transformations such as rotations, displacements and changes of perspective are a natural component of data acquisition. More generally, it provides a strategy for co-registration that blends the robustness of optimal transport with the interpretability of models. The extended optimal transport problem pairs two distributions with minimal additional distortion, while identifying a cost-free, explainable component of the map. A data-driven formulation is developed, as well as a methodology for its numerical solution. The latter complements gradient descent with a game-theory inspired approach, favoring collaborative moves between the cost-free and the unrestricted transformations. Sample validations are provided. The methodology is illustrated through its successful application to matching pairs of both synthetic and real images, which are conceptualized as weighted samples from underlying distributions, and through the determination of treatment effects by co-registering treated and untreated populations in a synthetic example.

1. Introduction

In image processing, a common task in data analysis is to establish a *co-registration* between a pair of data sets: the process of aligning and overlaying data sets to improve the accuracy of analysis, such as matching objects or locations in photographs taken at different times and angles. One of the key applications of co-registration is to detect changes driven by different sources of variability that warrant attention. Factors contributing to the noteworthy changes that make image co-registration challenging include

- **physical (unexplained) factors** contributing to changes in data that causes content deformations, such as tree growth, presence or absence of a vehicle, or an object broken over time,
- and **perceptual (modeled) factors** bringing changes in data that are natural or well-understood components of data acquisition and, therefore, independent of object/location correspondence, such as changes in lighting, spatial extent, perspective, and resolution.

Thus, there is a need for co-registration methodologies that are robust to such a wide multitude of variabilities.

In Mathematics there are methodologies that are by their very nature robust to changes of the physical kind. One group of these is referred to as optimal transport (OT) based problems (Villani et al., 2009) OT methods conceptualize change as two distributions of information. One is considered the reference or "source" (e.g. the earlier photograph of a scene). The other is considered the target (e.g. the latter photograph). Selecting one over the other as source versus target has no effect, as each are treated as probability distributions. OT methods seek a pairing among of the distributions with minimal distortion. Thus, actual changes between the data sets are intrinsic to the methodology's conceptual framework. More precisely, the optimal transport problem seeks a co-registration between two probability distributions $\rho(x)$ and $\mu(x)$, $x \in R^d$, in the form of an invertible map $y = T(x)$ satisfying the push forward condition

$$\rho_T \stackrel{\text{def}}{=} T_{\#}\rho = \mu, \tag{1}$$

that the pre-images of all measurable sets B preserve their measure, $\rho(T^{-1}(B)) = \mu(B)$. Among all maps T satisfying (1), the optimal map minimizes a total *transportation cost* C of deforming $\rho(x)$ into $\mu(y)$.

While well-suited to address physical changes in a scence, the OT methodology does not intrinsically handle variability of the perceptual kind. For example, if an image has been rotated, for instance, the optimal map T will deform it, in order to minimize the Euclidean transportation cost. Instead, the aim of this research is to enable rotations and other admissible transformations to be performed by default, thereby incurring no transportation costs. In a rough sense, the approach presented herein can be considered as a pre-processing step. With this is mind, an extension to the classical OT problem is proposed that includes a family of cost-free transformations.

The choice of transformations, Q, that should be cost free is application-dependent. Below are some realistic examples, although this paper considers the formulation and validation of only a subset of these:

1. For comparing images of handwritten digits, a sensible cost-free family of transformations Q includes translations, rotations and zoom-in/out of the digits.
2. For some photographs, Q should filter out the perspective and distance from which the pictures were taken.

3. For pictures of natural settings, the desire may want to exclude the effect of the time of the day, day of the year, and weather from the analysis, so as to establish a co-registration among pictures of the same place taken at different times. For that Q would be the group of time and weather-sensitive transformations related to variations of sunlight and cloud cover.

The proposed formulation for a modified optimal transport problem in this article detects both sources of variability simultaneously, where the cost-free map Q accounts for the perceptual/-modeled changes and the cost-minimizing map T resolves the complete co-registration including physical changes in content.

While there are literally hundreds of approaches to image matching, the use of "Common Feature Detectors (CFDs)" has retained prominence over several decades. CFDs are deployed predominantly for object detection, image recognition, and camera motion estimation and generally can address image matching when translation, rotation, and scaling are needed (Kuai et al., 2010). As an example, the OpenCV library (Bradski, 2000) offers several well-established ones. These include Binary robust Independent Elementary Features (BRIEF), Oriented Fast and rotated BRIEF (ORB) (Kulkarni et al., 2013), binary robust invariant scalable keypoints (BRISK) (Leutenegger et al., 2011), and Learned Arrangement of Three Patches (LATCH), which is built on local binary patterns using trios of mini-patches for descriptor generation (Levi and Hassner, 2016). All have different performance metrics and specific attributes. For example, LATCH tends to outperform both Scale Invariant Feature Transform (SIFT) (Lowe, 2004) and Accelerated Kaze (AKAZE) (Alcantarilla et al., 2013), in terms of affine capabilities, despite AKAZE's use of finite element diffusion to accelerate the Gaussian scale estimation in conjunction with either a floating-point descriptor or a modified local difference, binary descriptor for improved detection and enhanced matching.

Notably, these approaches rely on the essence of labeling and pairing and, hence, cannot be extended to higher dimensionality. The methodology proposed in this article, by contrast, extends naturally to any number of dimensions by discovering a distribution-wise correspondence. The reader is also directed to a different body of work of relevance titled Shuffled Regression [(Li et al., 2021) and (Pananjady et al., 2018)], which seeks the linear map between sets that best matches their underlying distributions.

The optimal-transport and model-based co-registration proposed, herein, can be extended to highly disparate problems, some physically very different from image processing. For instance, the same framework addresses the estimation of the effect of a treatment, which could refer to an actual long-term medical treatment, a habit such as smoking or exercising with potential effects on a person's health condition, indicated by metrics such as blood pressure or glucose level, or any other intervention whose effect one seeks to estimate. There the treatment's effect is represented by the map T between an untreated and a treated population, and Q consists of a parametric model of this effect based on prior knowledge. Notice that the data in this case consists of samples drawn from two populations, in contrast with those treatments where each patient's state can be observed before and after treatment, so the pairing between populations is known before hand. The same notion of treatment and toolbox for analysis applies well beyond medical applications, quantifying for instance the effect of economic policies on a nation's wealth.

2. Formulation

Traditionally, the regular optimal transport problem adopts the form

$$\min_T C(T) = \mathbb{E}_\rho \left[c\left(x, T(x)\right)\right] \quad \text{subject to} \quad T_{\#}\rho = \mu. \tag{2}$$

Here $c(x, y)$ is an externally provided pairwise transportation cost satisfying $c \geq 0$ and $c(x, y) = 0 \Leftrightarrow x = y$. A canonical cost in normed spaces is given by

$$c(x, y) = \frac{1}{2}\|y - x\|^2, \tag{3}$$

which will be adopted for this paper's numerical examples.

In order to include "free" displacements Q, this paper proposes to replace the formulation in (2) by

$$\min_{T,Q} \mathbb{E}_\rho \left[c\left(Q(x), T(x)\right)\right] \quad \text{subject to} \quad T_{\#}\rho = \mu, \quad Q \in \mathcal{F}, \tag{4}$$

where $\mathcal{F}$ is the family of allowed cost-free transformations. In this formulation, T is still required to push forward ρ to μ, but the pairwise cost c no longer measures the transportation cost between the original x and $T(x)$. Instead, it measures a reduced cost where x has been optimally displaced to $Q(x)$, using the additional freedom provided by the family $\mathcal{F}$. Thus the joint optimization problem (4) on T and Q pushes forward $\rho(x)$ to two different distributions: to $\mu = T_{\#}\rho$, and to $\tilde{\mu} = Q_{\#}\rho$, so that the corresponding expected value of the pairwise transportation cost between the two is minimal.

If the family $\mathcal{F}$ is sufficiently rich to push forward ρ to μ on its own through a map $Q \in \mathcal{F}$, the optimal solution has $T = Q$ and zero total cost. When such completely cost-free transportation is not available, $Q(x)$ can be thought of as a solution to the relaxed problem

$$Q_{\#}\rho \approx \mu, \quad Q \in \mathcal{F},$$

with minimal transportation cost between $\tilde{\mu} = Q_{\#}\rho$ and μ. This yields the alternative formulation

$$\min_{Q \in \mathcal{F}} W_c\left(Q_{\#}\rho, \mu\right), \text{ where } W_c\left(\tilde{\mu}, \mu\right) = \min_{\pi(z,y) \in \Pi_{\tilde{\mu},\mu}} \mathbb{E}_\pi \left[c\left(z, y\right)\right] \tag{5}$$

is the c-Wasserstein distance and $\Pi_{a,b}$ is the set of distributions coupling a and b. When $\mathcal{F}$ is the family of linear maps, this formulates Shuffled Regression in terms of optimal transport.

The formulation in (4) is well-suited for a double-flow procedure, with two flows that trace the maps $z = T(x)$ and $w = Q(x)$, respectively. In applications, the distributions ρ and μ are not provided in closed form but, instead, through a finite number of samples. Replacing (4) by a data-driven formulation requires writing in terms of the sample sets $X = \{x_i\}_{i=1}^N \sim \rho$ and $Y = \{y_j\}_{j=1}^M \sim \mu$:

1. The maps themselves $z = T(x)$ and $w = Q(x)$, which we represent through their values z_i, w_i on the available

samples x_i. Then

$$z_i(t) = T(x_i, t), \quad w_i(t) = Q(x_i, t),$$

where t represents an algorithmic time associated with the flows.

2. The push-forward constraint that $\rho_T \equiv T\#\rho = \mu$, which is rewritten in terms of the Kullback-Leibler divergence

$$KL(\rho_T, \mu) = \int \log\left(\frac{\rho_T(z)}{\mu(z)}\right) \rho_T(z)\, dz = 0. \quad (6)$$

Since the relative entropy between any two distributions is non-negative, the condition in (6) minimizes $KL(\rho_T, \mu)$ over the map T, a formulation that will be instrumental in the construction of a flow-based descent algorithm. In terms of the samples $\{z_i\} \sim \rho_T$, thus

$$KL(\rho_T, \mu) \to \hat{KL}(Z, Y) = \frac{1}{N}\sum_i \log\left(\frac{\hat{\rho}_T(z_i)}{\hat{\mu}(z_i)}\right), \quad (7)$$

which is complemented with kernel density estimation,

$$\begin{aligned} \hat{\rho}_T(y; \{z_i\}) &= \frac{1}{N}\sum_{i=1}^{N} \kappa_a(y, z_i), \\ \hat{\mu}_a(z; \{y_j\}) &= \frac{1}{M}\sum_{j=1}^{M} \kappa_a(z, y_j). \end{aligned} \quad (8)$$

The demonstrated numerical examples use Gaussian kernels of the form $\kappa_a(\cdot, \gamma) = N(\gamma, A^2)$, where $A^2 \in \mathbb{R}^{d\times d}$ is a diagonal covariance matrix, with entries $A_{kk} = a_k$ determined by the rule of thumb along each independent dimension.

3. The constraint that $Q \in \mathcal{F}$. Families $\mathcal{F}$ are considered with an explicit parametric form, so any $Q \in \mathcal{F}$ is specified by its defining parameters α.
4. The cost function $\mathbb{E}_\rho[c(Q(x), T(x))]$, for which a sample-based formulation replaces the expected value by the empirical mean:

$$\begin{aligned} C(Q(X), T(X)) &= \int c(Q(x), T(x))\, \rho(x)\, dx \\ \to \hat{C}(W, Z) &= \frac{1}{N}\sum_i c(w_i, z_i). \end{aligned}$$

3. Description of the algorithm

A sample-based driven formulation of the problem in (4) reads

$$\min_{z,\alpha} L \equiv \frac{1}{N}\sum_i c(w_i, z_i) + \frac{\lambda}{N}\sum_i \log\left(\frac{\rho_T(z_i)}{\mu(z_i)}\right), \quad (9)$$

$$w_i = Q(x_i, \alpha), \quad Q \in \mathcal{F},$$

where μ and ρ_T are estimated from their samples $\{y_i\}$ and $\{z_j\}$ through (8), and the parameter $\lambda > 0$ penalizing non-compliance with the push forward condition evolves over time as described in Appendix A.1. Adopting parametric maps that form a group over α,

$$Q(Q(x, \alpha), \beta) = Q(x, \gamma(\alpha, \beta)), \quad Q(x, 0) = x,$$

allows discretization of the flow $w_i(t)$ into near-identity maps:

$$w_i\left(t^{n+1}\right) = Q\left(w_i\left(t^n\right), \alpha^n\right), \quad \|\alpha^n\| \ll 1.$$

Then the minimization problem in (9) can be solved by tracing discrete flows (z^n, w^n), determined through gradient descent of L:

$$\begin{aligned} z_i^0 = w_i^0 = x_i, \quad z_i^{n+1} &= z_i^n - \eta\frac{\partial L}{\partial z_i}, \\ w_i^{n+1} = Q(w_i^n, \alpha^n), \quad \alpha^n &= -\eta\frac{\partial L}{\partial \alpha}, \end{aligned}$$

where η is an adaptive learning rate (described in Appendix A.2) and all partial derivatives of L are evaluated at $(z = z^n, w = w^n, \alpha = 0)$.

Yet such a straightforward descent procedure has limitations. On the one hand, the initial values adopted, $z_i = w_i = x_i$ may be far from the optimal $\{z_i^*, w_i^*\}$, possibly beyond their basin of attraction under gradient descent. On the other, it is more efficient for z and w not to descend L independently but to "act collaboratively". To see this, consider a deformation-free scenario, where a map $Q \in \mathcal{F}$ could satisfy the push-forward condition on its own, so the optimal answer should have $z_i = w_i$. Yet whenever this condition holds, $\frac{\partial L}{\partial \alpha} = 0$, so only z evolves, leaving w behind. If instead, the minimization process is considered as a two-player game in which w and z could "anticipate" how the other would react to their displacement, they could move together, conspiring to make L decrease in ways that neither alone could.

To address these limitations, the initialization of z and w is improved through a preconditioning procedure, thereby moving from gradient descent to a game-theory-based approach, whereby two players, with strategies z and α respectively, seek to optimize the objective function both directly and through actions that lead the other player to act in a convenient way.

3.1 Preconditioning with the free transformation

In the alternative problem formulation in (5), the family of free transformations approximates the full map as much as possible, so as to minimize the transportation cost between $\tilde{\mu} = Q_\#\rho$ and μ. Hence a natural preconditioning procedure considers Q alone, bringing the $\{w_i\}$ as close as possible to the $\{y_j\}$ distribution-wise. This step is now formulated in terms of the relative entropy between $\rho_Q := Q_\#\rho$ and μ:

$$\min_{Q\in\mathcal{F}} \hat{KL}(W = Q(X), Y) = \frac{1}{N}\sum_i \log\left(\frac{\rho_Q(Q(x_i))}{\mu(Q(x_i))}\right), \quad (10)$$

where $\rho_Q(\cdot)$ and $\mu(\cdot)$ are replaced by their Kernel density estimations based on the $\{w_i\}$ and $\{y_j\}$, respectively. Next, equation (10) is minimized through gradient descent over the parameters α in $Q(w, \alpha)$, and set at the end of the preconditioning procedure $z_i = w_i$ so that the initial values of Z and W for the second phase of the algorithm are paired and close to the Y distribution-wise.

This preconditioning procedure is similar in nature to a maximal likelihood-driven, normalizing flow (Tabak and Vanden-Eijnden, 2010, Tabak and Turner, 2013) with three main differences: a) the target distribution is not necessarily Gaussian, b) the target is specified not by a closed form expression but

by a finite set of samples, and c) the parametric maps at each step are constrained to adopt the form of the specified cost-free transformations.

3.2 Multiplayer games

This minimization problem is now rewritten as a game between two players with strategies W and Z, each trying to minimize the objective function L not only through the direct effect of their move on L, but also through the indirect, second order effect produced by the expected response of the other player to their move.

Next is the application of a general methodology for multiplayer games restricted to two players with the same objective function L in (9), which yields following system for evolving w and z:

$$\begin{aligned} z^{n+1} &= z^n - \eta_z \left[\nabla_z L - \eta_z [\nabla^2_{z\alpha} L] \cdot \nabla_\alpha L\right], \\ \alpha^{n+1} &= -\eta_\alpha \left[\nabla_\alpha L - \eta_\alpha [\nabla^2_{z\alpha} L] \cdot \nabla_z L\right], \\ w^{n+1} &= Q\left(w^n, \alpha^{n+1}\right). \end{aligned}$$

Here z, w and α stand for the vectors with components $\{z_i\}$, $\{w_i\}$ and $\{\alpha_j\}$ respectively, ∇_a for the gradient operator in the a-variables, and $\nabla^2_{z\alpha} L$ for the off-diagonal block in the Hessian of $L(z, \alpha)$. The logic underlying this procedure is that each player anticipates how the other player will react to their action and how the reaction will affect their own objective function.

4. Numerical examples

This section illustrates the effectiveness of the methodology proposed through a series of pairs of two-dimensional datasets corresponding to both synthetic and real images and through the determination of treatment effects in a synthetic example. These were selected to illustrate highly specific aspects of the cost-free transformations.

4.1 Synthetic images

Rotated rectangle In this first example, both the source and the target distributions ρ and μ are uniform on 20×4 rectangles centered at the origin, but the two rectangles are rotated by an angle $\theta = \frac{\pi}{3}$ with respect to each other. In order to formulate the data-based optimal transport problem, 500 particles are drawn from each distribution, as shown on the leftmost and rightmost panels on the top row in Figure 1. The source and target are sampled independently, so there is no one-to-one correspondence between their samples themselves.

The natural cost-free maps for this problem are rigid rotations about the origin, parameterized by the rotation angle θ. To appreciate the need for cost-free transformations, consider the second panel on the top row of Figure 1, with respect to the solution to the regular optimal transport problem between source and target. For better visualization of the map $Z = T(X)$, the particles are aligned with a color scheme based on their initial positions in the source distribution. As the image shows, pure optimal transport fails to identify the rotation, producing instead significant deformation that destroys the color-alignment, resulting in a very counter-intuitive co-registration. The optimal map could not possibly have consisted of a pure rotation, since under the canonical cost (3), the optimal maps are necessarily curl free (McCann, 1995).

By contrast, the third panel of the top row in Figure 1 shows that when cost-free rotations $\{Q(\mathbf{x})\} = \{(R_\theta)\mathbf{x} \big| \theta \in [0, 2\pi)\}$ are allowed, then the transported particles align correctly under a close-to-zero cost, since the rotation is fully recovered by the cost-free map Q alone.

Modified rectangle In the previous example, the initial and final distributions were rotated versions of each other, so the optimal solution had $T = Q$. In order to include physical deformations that Q alone cannot capture, this example is modified by perforating a circle of radius 1 from the target distribution, as shown on the right panel of the 2nd row of Figure 1. In this example, preconditioning rotates the source rectangle with zero cost, while another 300 steps of multiplayer game successfully complete the job, emptying out the hole, as shown on the 2nd row, resulting in an overall transportation with non-zero cost.

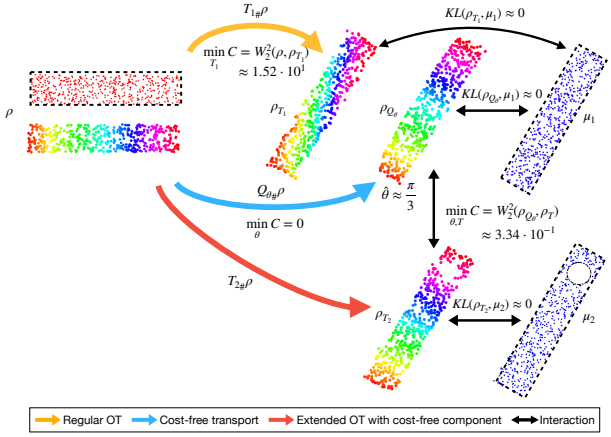

Figure 1. Initial samples from the source (leftmost) and samples from the target (rightmost), their position after transportation, displaying $Z = T(X)$ with color assignment, under pure optimal transport (2nd panel, top row) and with cost-free rotations (3rd panel, top row) for a rigidly rotated target and, on the second row, after 100 steps of preconditioning and 200 steps of the multiplayer-game with cost-free rotations for a perforated and rotated target.

4.2 Real images

Next the algorithm is applied to real world images drawn from the photograph of Mount Fuji shown in Figure 2 and the MNIST data (Deng, 2012). Co-registration among images, matching objects and locations in noisy backgrounds are common in data applications. In order to apply the proposed methodology to images, the images must first be interpreted as distributions with pixels serving as samples thereof. The following subsection develops such interpretation for black-and-white pictures.

4.2.1 Images as distributions: a formulation in terms of weighted pixels

So far the problem has been posed in terms of two distributions ρ and μ, known through samples $\{x_i\}$ and $\{y_j\}$. The situation with images is somewhat different: the data consists of two images known through two sets of values associated to a uniform grid of pixels. This raises two questions: (1) can an image be described as a probability distribution, and (2) are pixel values in any sense equivalent to random samples thereof? For simplicity, these questions are addressed for black-and-white pictures, though similar considerations apply to color pictures and higher spectral datasets.

Figure 2. Full image of Mount Fuji with 664*1601 pixels, each containing one color channel in a gray scale between 0 and 1.

Physically, images are indeed probability distributions, as they represent the local density of photons hitting the screen. Even though understanding the physical underpinning of images is not required for their analysis, it is comforting to know that treating images as distributions is not just a convenient mathematical construct. Thus an image can be described as a distribution $\rho(x)$, where $x = \left(x^1, x^2\right)$ is supported on a rectangular frame. Larger values of ρ correspond to a higher density of photons per unit area weighted by their energy (i.e. to more whiteness).

In this conceptualization, the ρ and μ are still distributions. Yet instead of random samples from them, values in a gray scale are readily available $\rho_i, \mu_j \in [0,1]$ attached to equally-sized pixels, as demonstrated in Figure 3. Even though the location of the pixels is fixed, it is important to think of them as movable, since moving them is what optimal transport procedure does. A natural way to reconcile these two perspectives is to think of the pixel locations as particularly regular samples from a uniform prior distribution U and of $\{\rho_i\}$, $\{\mu_j\}$ as weights that correct for the fact that the actual distributions are not uniform:

$$\rho_i \propto \frac{\rho\left(x_i\right)}{U\left(x_i\right)}, \quad \mu_j \propto \frac{\mu\left(y_j\right)}{U\left(y_j\right)},$$

where the constant U is provided with arguments just to emphasize the sample re-weighting concept. Then the modified data-driven problem setup seeks a map that pushes forward a source to a target distribution, respectively known through two sets of weighted samples $\{(x_i, \rho_i)\}_{i=1}^N$ and $(\{(y_j, \mu_j)\}_{j=1}^M$. The corresponding flow moves the samples $z_i(t) = T(x_i, t)$, while keeping their weights ρ_i fixed.

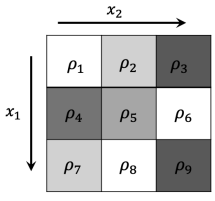

Figure 3. Visualization of a 3*3 image as a set of weighted pixels.

Since both the relative entropy and the transportation cost are defined through expected values under ρ, their weighted empirical versions are

$$\hat{KL}(\rho, \mu) = \frac{1}{\sum_i \rho_i} \sum_i \rho_i \log\left(\frac{\rho_T\left(z_i\right)}{\mu\left(z_i\right)}\right), \qquad (11)$$

and

$$\hat{C}\left(Q(X), T(X)\right) = \frac{1}{\sum_i \rho_i} \sum_i \rho_i \, c\left(Q(x_i), T(x_i)\right). \qquad (12)$$

Kernel density estimation also derives from an expected value. Thus, μ and ρ can be estimated from their sampled pairs $\{(y_j, \mu_j)\}_{j=1}^M$, $\{(x_i, \rho_i)\}_{i=1}^N$, through a weighted sum of kernel functions:

$$\hat{\mu}_a(y; \{y_j\}) = \frac{1}{\sum_j \mu_j} \sum_{j=1}^M \mu_j \kappa_a\left(y, y_j\right),$$

$$\hat{\rho}_b(x; \{x_i\}) = \frac{1}{\sum_i \rho_i} \sum_{i=1}^N \rho_i \kappa_b\left(x, x_i\right).$$

where the bandwidths a and b are proportional to the pixel size along each axis.

To visualize the weighted samples throughout the transportation, the samples are reconstructed into pixels drawn from the same distribution via kernels as elaborated in A.3.

Rotated, blurred and distorted image of Mount Fuji To demonstrate the processes, first only small fractions of the picture of Mount Fuji in Figure 2 are modified by rotating, blurring and deforming, so as to test both the free transformations (rotations in this case) and the additional changes.

Starting with a simple co-registration case, the source image is selected as a 80*80 section of the top of Mount Fuji. For the target, the same image rotated by an angle of $\pi/6$ is employed. To ensure that co-registration is purely based on the picture's content and not on recognizing the rectangular contour of the images, the corner edges of both source and target images were removed. This left only the two inscribed circles visualized in Figure 4.

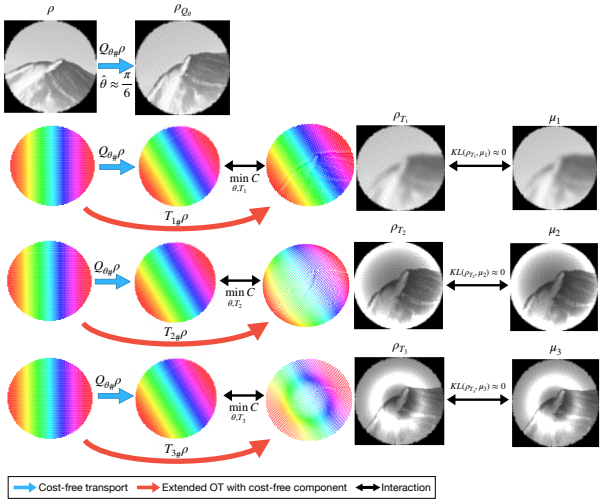

Figure 4. Images (80*80 pixels) of the top of Mount Fuji as the source, a rotated version as a first target in the 1st row, a further blurred version in the 2nd row, and two other version with extra local deformations of the pixels in the last two rows as alternative targets.

In the rotation-only example with target represented in the right panel of the 1st row of Figure 4, the preconditioning alone can solve the problem, i.e. $T = Q$, as shown in the middle panel.

When the target image is not just rotated but also blurred, deformations are required. The blurring is introduced as a change in resolution. The rightmost panel of the 2nd row in Figure 4 depicts the target, which is both rotated and artificially blurred, using as high-to-low pass filter the convolution of the matrix of pixel intensities with a Gaussian kernel. The results of the co-registration procedure, displayed in the 2nd row, show how the rigid rotation of the image as displayed in the 3rd panel –nearly completed during pre-conditioning– is complemented by small, local deformations that perform the task of blurring the image as shown in the 3rd panel. The resulting transported image distribution is almost identical to the actual target with essentially zero relative entropy.

In addition to artificial blurring, local stretching and compressing of the pixels can also bring in deformations. The 3rd row in Figure 4 depicts the results of applying the procedure to target –rightmost panel– that is both rotated and locally stretched out using a radial-based quadratic kernel

$$f_1\left(\begin{pmatrix} r \\ \theta \end{pmatrix}\right) = \begin{pmatrix} r + \epsilon \cdot r(R-r) \\ \theta \end{pmatrix}, \quad \epsilon = 1,$$

where $(r, \theta)^T$ is the polar coordinates of each pixel in 2D, R is the largest radius to the image center, and $\epsilon > 0$ is some small perturbation factor.

The last row of figure 4's displays similar results when the target is obtained through an angle dependent, radial-based stretching and compressing of the image. After rotation, the pixels are re-located through

$$f_2\left(\begin{pmatrix} r \\ \theta \end{pmatrix}\right) = \begin{pmatrix} r + \epsilon \cdot \sin(\frac{2\pi r}{R}) \\ \theta \end{pmatrix}, \quad \epsilon = 0.03.$$

MNIST data To further illustrate the procedure, co-registration is applied to different instances of hand-written digits imported from the MNIST dataset (which contains several images of each of the 0-to-9 digits, differing in size and hand-writing style). Each image of a digit has 28×28 pixels with gray-scaled weight, so the same set up of weighted samples can be used as that employed for the examples on Mount Fuji. Co-registering two instances of the same digit provides a more interesting challenge than the simple rotations and blurring we considered before, since the variety of hand-writing styles can bring in non-standard deformation. Consider the two handwritten digits "3" displayed in the first row of Figure 5.

As the target digit 3 seems to be slightly counter-clockwise-rotated and vertically-compressed compared to the source digit 3, the family of cost-free transformations are extended – adding to the rigid rotation the possibility of global translation and stretching factors:

$$\{Q(\mathbf{x})\} = \{R_\theta\, \mathbf{x} + s\, \mathbf{x} + \Delta, \quad \text{with } \theta \in [0, 2\pi), s \in \mathbb{R}, \Delta \in \mathbb{R}^2\}.$$

The results, displayed in the middle row of Figure 5, show a consistent co-registration. Precondition rotates, translates, and compresses the source digit 3, as shown in the 2nd panel. The ensuing multiplayer-game algorithmic steps make small adjustments to the rotation angle, displacement distance, and the stretching factor, while further deforming the digit to complete the map to reach the final status in the rightmost panel. The extended explainability derived from the more comprehensive family of cost-free transformations, results in a small transportation cost.

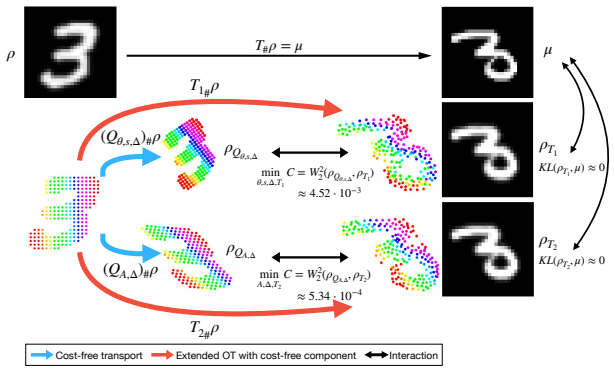

Figure 5. Two gray-scaled, hand-written digits 3 in different styles as the source and the target. The middle column displays cost-free components from the family of linear combinations of rotation, scaling, and translation, and the corresponding completion of the map. The right column shows the cost-free affine transformation and the corresponding unrestricted push forward map.

Affine transformation Instead of considering rotation, displacement, and stretching separately, a more general family of affine transformations can be introduced

$$\{Q(\mathbf{x})\} = \{(I + A)\mathbf{x} + \Delta = \mathbf{x} + A\mathbf{x} + \Delta, \\ \text{with } A \in \mathbb{R}^{d \times d}, \Delta \in \mathbb{R}^d\}$$

where $I \in \mathbb{R}^{d \times d}$ is the identity matrix and d represents the dimension. This indicates any linear mapping A with some bias Δ persists substantial integrity and incurs zero cost.

The results, displayed in the last row of Figure 5, show a different, more robust correspondence between the source and the target. Precondition performs affine transformation of the source digit 3, which obviously include rotation, translation, and compression, as shown in the 2nd panel. The following multiplayer-game algorithmic steps make small adjustments to the linear mapping and the free displacement, while bringing extra deformation of the digit to reach the final status in the rightmost panel. The even stronger explainability derived from the cost-free family of affine transformations, results in an even small transportation cost. In fact, the enhanced cost-free components result in a different transportation destination with a better alignment with the target digit 3 as shown in the rightmost two panels in the last row.

4.3 Treatment effect

The additional cost-free transformations can also account for prior knowledge on the potential effect of a treatment. When studying the effect of a medical treatment, the control group and the treatment group often consist of different populations, i.e. one group of patients receive the treatment and the other does not. Hence, unlike scenarios where the state of the same patient before and after treatment are known, there is no point-wise correspondence between the control group and the treated group, corresponding to the source and target in the transportation set up. Then, co-registration serves to figure out the effect of the treatment based on a distribution-wise map, attributed to either modeled or data-driven sources of variability.

Lacking more detailed knowledge of the effect of a treatment, a common and most straightforward characterization of the potential effect is through a linear model. In this case, the family of cost-free transformations consist again of affine maps, which is the example that will be used for illustration,

$$\{Q(\mathbf{x})\} = \{(I + A)\mathbf{x} + \Delta = \mathbf{x} + A\mathbf{x} + \Delta, \\ \text{with } A \in \mathbb{R}^{d\times d},\ \Delta \in \mathbb{R}^d\}.$$

The affine map pushing forward one distribution to another is unique up to a group of affine-invariant surfaces. For example, an isotropic Gaussian distribution is mapped to itself by any rotation. Therefore, the discovery of the map through transportation is only unique and well-conditioned when the source data (control group) exhibits sufficient complexity and heterogeneity, which is typically the case in real data.

In the artificial experimental setup displayed in Figure 6, the source data (control group) $\{x_i\}$ is distributed as a mixture (ρ) of Gaussians distributions. The target (or experiment group) μ is built from another set of patients $\{\tilde{x}_j\}$ generated from the same Gaussian mixture and then "treated" through an affine map (defined by A and Δ), with added noise that mimics the uncertainty of treatment effects and a nonlinear correction that mimics model error.

$$\{x_i\},\ \{\tilde{x}_j\} \sim \rho, \\ y_j = A\tilde{x}_j + \Delta + \epsilon_1\sigma_1(\tilde{x}_i) + \epsilon_2\sigma_{2i} \sim \mu,$$

where ϵ_1, ϵ_2 are small positive numbers, $\sigma_1(\mathbf{x}) = x_1^3$ represents model error and $\sigma_{2i} \sim N(0, 1)$ represents random noise.

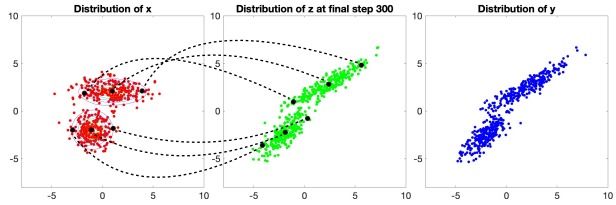

Figure 6. Source data (control group) generated from a Gaussian mixture, the final screenshot after the optimal transport, and target data (experimental group), created through an affine treatment with added noise and nonlinear corrections applied to independent samples or ρ. The black circles and connecting dashed lines show the predicted treatment effect on six patients.

Provided with no prior pairwise correspondence among the two sets of patients in a noisy environment, the proposed algorithm discovers the effect of the treatment, as shown in Figure 6. The cost-free transformation successfully detects the modeled affine transformation with little error. In this example, with zero bias $\Delta = 0$, it yields

$$A = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \quad \hat{A} = \begin{pmatrix} 0.9565 & 0.6432 \\ 0.4539 & 1.1102 \end{pmatrix},$$

where A is the true linear effect, and $\hat{A}$ is the discovered affine component of the map, with $\|A - \hat{A}\|_2 \approx 0.1912$.

5. Conclusions

This article proposes an extension of the optimal transport problem, accounting for natural transformations of the source distribution that should be free of cost. Allowing such transformations is key to a proper co-registration of data sets. The article develops a data-driven formulation of the extended problem and a numerical methodology for its solution. Applying the new methodology to images requires a novel conceptualization of images as probability distributions, with pixel values as weighted samples thereof.

Four elements stand out among those requiring further development:

1. In the examples provided herein, the free transformations adopted a simple parametric form. Yet there are instances, such as changes in luminosity or vegetation growth between the source and target distributions, that would require to learn the corresponding family of transformations from the data.
2. Implementing the push-forward condition with less computational cost than through the relative entropy and in a way that extends to high-dimensional settings, to allow for an effective co-registration of large data sets.
3. Extending the characterization as probability distributions of images to more complex and irregular data sets requiring co-registration, such as point clouds.
4. Historically, the class of OT methods has been computationally expensive, thus there is significant room for acceleration. One possible approach is further optimization of the selected items for distribution where a set of guidance could be developed that considers both the size of the data sets and their complexity.

6. Acknowledgments

The work of Esteban Tabak and Debra Laefer was partially supported by ONR #13456472.

References

Alcantarilla, P., Nuevo, J., Bartoli, A., 2013. Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces. *Proceedings of the Proceedings of the British Machine Vision Conference*.

Bradski, G., 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 430.

Deng, L., 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6), 141–142.

Kuai, X., Yang, K., Fu, S., Zheng, R., Yang, G., 2010. Simultaneous localization and mapping (SLAM) for indoor autonomous mobile robot navigation in wireless sensor networks. *Proceedings of the 2010 International Conference on Networking, Sensing and Control*, 128–132.

Kulkarni, A., Jagtap, J., Harpale, V., 2013. Object recognition with ORB and its Implementation on FPGA. *International Journal of Advanced Computer Research*, 164.

Leutenegger, S., Chli, M., Siegwart, R., 2011. BRISK: Binary robust invariant scalable keypoints. *2011 International conference on computer vision*, 2548-2555.

Levi, G., Hassner, T., 2016. LATCH: Learned arrangements of three patch codes. *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1-9.

Li, F., Fujiwara, K., Okura, F., Matsushita, Y., 2021. Generalized shuffled linear regression. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6454–6463.

Lowe, D., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 91–110.

McCann, R. J., 1995. Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.*, 80(2), 309–323.

Pananjady, A., Wainwright, M. J., Courtade, T. A., 2018. Linear Regression With Shuffled Data: Statistical and Computational Limits of Permutation Recovery. *IEEE Transactions on Information Theory*, 64(5), 3286-3300.

Tabak, E. G., Trigila, G., Zhao, W., 2022. Distributional barycenter problem through data-driven flows. *Pattern Recognition*, 130, 108795.

Tabak, E. G., Turner, C. V., 2013. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2), 145–164.

Tabak, E. G., Vanden-Eijnden, E., 2010. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1), 217–233.

Villani, C. et al., 2009. *Optimal transport: old and new*. 338, Springer.

A. Appendix / supplemental material

A.1 Choice of the penalization parameter λ

We allow the penalization parameter λ to depend on algorithmic time, determining its value at each step through the extension of a methodology first proposed in (Tabak et al., 2022). At each step, the evolution of Z is determined from the competition between two typically conflicting goals: to satisfy the push-forward condition and to remain as close as possible to W, through the respective minimization of $\hat{KL}(Z,Y)$ and $\hat{C}(W,Z)$. The parameter λ assigns relative weights to these two components. Since the push-forward condition is a hard constraint, we require that the full Z-gradient of the objective function with respect to Z projects positively onto the Z-gradient of $\hat{KL}(Z,Y)$. This provides a lower bound for λ.

In order to ensure that $\hat{KL}(Z,Y)$ is non-increasing,

- We first compute the gradient of both components of the objective function at the current position $Z(t)$:

$$\mathbf{u} := -\nabla_Z \hat{KL}(Z,Y;Z)\Big|_{Z(t)}, \quad \mathbf{v} := -\nabla_Z \hat{C}(W,Z)\Big|_{Z(t)}.$$

- We choose a direction $\mathbf{s}$ of descent as a linear combination of $\mathbf{u}$ and $\mathbf{v}$ that projects positively onto $\mathbf{u}$, so that $\hat{KL}(Z,Y)$ decreases to leading order in η:

$$Z(t+\eta) - Z(t) \propto \mathbf{s}, \quad \mathbf{s} := \mathbf{v} + \lambda \mathbf{u}, \quad \langle \mathbf{s}, \mathbf{u} \rangle \geq 0.$$

  To this end,

  - if $\langle \mathbf{u}, \mathbf{v} \rangle > 0$, set $\lambda = 0$,
  - else, set $\langle \mathbf{s}, \mathbf{u} \rangle = \epsilon \|\mathbf{u}\|^2$, with $0 < \epsilon < 1$:

$$\begin{aligned} &\langle \mathbf{s}, \mathbf{u} \rangle = \epsilon \|\mathbf{u}\|^2 \\ &\implies \langle \mathbf{v}, \mathbf{u} \rangle + \lambda \|\mathbf{u}\|^2 = \epsilon \|\mathbf{u}\|^2 \\ &\implies \lambda = \epsilon - \frac{\langle \mathbf{v}, \mathbf{u} \rangle}{\|\mathbf{u}\|^2}. \end{aligned}$$

A.2 Choice of the learning rate η

We use a simple adaptive procedure to evolve the learning rates for the transporting particles Z and W:

- Set an initial learning rate η_0.
- At time t,
  - double the learning rate $\eta_t = 2\eta_{t-1}$,
  - while necessary, repeatedly half the learning rate until the objective function at the new potential points $Z^* = Z(t) + \eta_t \dot{Z}$ is smaller than at at $Z(t)$, and set $Z(t+1) = Z^*$ (and similarly for W.)

Algorithm 1 Update learning rate

$\eta_t \leftarrow 2\eta_{t-1}$ ▷ double step size
$\mathbf{z}(t+1) \leftarrow \mathbf{z}(t) - \eta_t \cdot \nabla_{\mathbf{z}(\mathbf{t})} L$ ▷ move data with current step size
while $L(\mathbf{z}(t+1)) > L(\mathbf{z}(t))$ **do** ▷ terminate when the loss decreases
  $\eta_t \leftarrow \eta_t/2$ ▷ half step size
  $\mathbf{z}(t+1) \leftarrow \mathbf{z}(t) - \eta_t \cdot \nabla_{\mathbf{z}(\mathbf{t})} L$ ▷ move data with current step size
end while

A.3 Image reconstruction

Having conceptualized an image as a set of weighted samples drawn from a distribution, we need to visualize the results of the optimal transport procedure, which moves these samples $\{x_i\}$ to the corresponding $\{z_i\}$. This requires the inverse process, taking as input a set of weighted samples –no longer uniformly distributed– (z_l, ρ_l) and producing an image: a set of intensity values I_i on the original regular, rectangular grid $\{x_i\}$. For this, we can use kernels again:

$$I(x_i) = I_i = \sum_l \rho_l \frac{\kappa_h(x_i, z_l)}{\sum_h \kappa_a(x_k, z_l)}.$$

Because images are inherently local, we choose a kernel with strictly local support and linear growth:

$$\kappa_h(\cdot, c) = \max\left\{1 - \frac{\|\cdot - c\|_\infty}{h}, 0\right\}.$$

We pick as bandwidth h for image reconstruction the length of the pixels's sides, so that the effect of each z_i on the image does not extend beyond the boundaries of the pixel within which it falls.