Information and Inference: A Journal of the IMA (2021) **00**, 1–41
https://doi.org/DOI HERE

The Monge optimal transport barycenter problem

ANDREW D. LIPNICK
Department of Mathematics, Lafayette College, 730 High Street, Easton, 18042, Pennsylvania, USA

ESTEBAN G.TABAK
Department of Mathematics, Courant Intstitute of Mathematical Sciences, New York University, 251 Mercer Street, 10012, New York, USA

GIULIO TRIGILA
Department of Mathematics, Baruch College, City University of New York, 55 Lexington Avenue, 10010, New York, USA

YATING WANG AND XUANCHENG YE
Department of Mathematics, Courant Intstitute of Mathematical Sciences, New York University, 251 Mercer Street, 10012, New York, USA

AND

WENJUN ZHAO*
Department of Mathematics, University of British Columbia, 1984 Mathematics Rd, V6T 1Z2, BC, Canada
*Corresponding author: zhaow@wfu.edu

[Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year]

A novel methodology is developed for the solution of the data-driven Monge optimal transport barycenter problem, where the pushforward condition is formulated in terms of the statistical independence between two sets of random variables: the factors z and a transformed outcome y. Relaxing independence to the uncorrelation between all functions of z and y within suitable finite-dimensional spaces leads to an adversarial formulation, for which the adversarial strategy can be found in closed form through the first principal components of a small-dimensional matrix. The resulting pure minimization problem can be solved very efficiently through gradient descent driven flows in phase space. The methodology extends beyond scenarios where only discrete factors affect the outcome, to multivariate sets of both discrete and continuous factors, for which the corresponding barycenter problems have infinitely many marginals. Corollaries include a new framework for the solution of the Monge optimal transport problem, a procedure for the data-based simulation and estimation of conditional probability densities, and a nonparametric methodology for Bayesian inference.

Keywords: Optimal transport; Wasserstein barycenter; gradient flow; conditional density estimation.

1. Introduction

A central problem in the analysis of data is to estimate how a set of variables $x \in \mathscr{X}$, the *outcome*, depends on a set of covariates $z \in \mathscr{Z}$, the *factors*, a dependence that can be fully characterized by the conditional distribution $\rho(x|z)$. One seeks to extract from n observed data pairs $\{x_i, z_i\}$, either an evaluation procedure for ρ itself or a procedure to draw samples $\{x_j^*\}$ from $\rho(x|z^*)$ for any target value z^*. This is particularly challenging when z includes continuous components, since any particular value

z^* has small probability of having appeared among the $\{z_i\}$, even less of having shown up in enough observational pairs to warrant a statistical analysis based on those pairs alone.

A data-driven methodology for the simulation of conditional distributions based on the [Monge] optimal transport barycenter problem (OTBP) seeks a map

$$y = T(x,z) \in \mathscr{X}$$

that removes from x the variability that z can explain, i.e. such that the random variables y and z are independent. In order not to remove any additional variability from x, we select the map T that minimizes the expected value of a total "transportation" cost

$$C = \mathbb{E}_\pi[c(x,T(x,z))],$$

where π is the joint distribution of x and z. The pairwise cost function $c(x,y)$ quantifies the deformation of the data incurred by moving x to y. This results in an OTBP of the form

$$\min_{y=T(x,z)} \mathbb{E}_\pi\left[c\left(x,y\right)\right] \quad \text{s.t.} \quad y \perp\!\!\!\perp z,$$

where the symbol $\perp\!\!\!\perp$ stands for independence. We can use the solution to this problem to simulate the conditional distribution $\rho(x|z)$ for a target value $z = z_*$, extracting n samples $\{x_i^*\} \sim \rho(x|z_*)$ through

$$x_i^* = T^{-1}\left(y_i, z_*\right), \quad y_i = T\left(x_i, z_i\right).$$

In words, we remove from x_i the variability attributable to z having adopted the value z_i, and then restore that variability but with $z = z_*$. The variable y represents the variability in x that z does not explain.

Other uses of the OTBP include he following:

1. In order to eliminate the effect of confounding variables z from the data x, we simply move the $\{x_i\}$ to their counterpart in the barycenter, $\{y_i = T(x_i,z_i)\}$. Examples include the removal of batch effects, the consolidation of different data bases, where z represents the data source and, more generally, the removal of the confounding effects of any set of variables z that are not considered in the study under way. The removal of the effects of known factors z helps identifying further sources of variability by investigating y, a version of x cleaned of the effects of z.
2. The explanatory power of the covariates z can be quantified by the total cost C. This ranges from the extreme scenario where z has no explanatory value, so x is already independent of z, $y = x$ and $C = 0$, to the opposite extreme where all variability in x can be explained by z, so the barycenter reduces to a single point $\bar{y}$, which maximizes C. Quantifying through C the explanatory power of z gives rise to a rich methodology for factor selection and discovery [1, 2].
3. The barycenter problem permits not only simulating but also estimating conditional densities, and therefore yields a mode-free, non-parametric data-based procedure for Bayesian inference: given a prior distribution $\gamma_{pr}(z)$, a set of sample pairs $\{x_i, z_i\}$ drawn from an unknown joint distribution $\pi(x,z)$ and the observed current value of x, estimate the posterior distribution $\gamma_{pos}(z|x)$.
4. The optimal transport problem, a particular case of the OTBP with only two marginals, yields a natural horizontal distance among distributions. It also serves as a powerful tool for density estimation and sampling.

This article develops a novel, efficient methodology to solve the optimal transport barycenter problem, providing the capability to both simulate and estimate $\rho(x|z)$. Along the way, it clarifies the relation between the Monge optimal transport barycenter problem as posed above and the Wasserstein barycenter problem [3]. It also provides as corollaries new methodologies for the solution of Monge's regular optimal transport problem and for model-free Bayesian inference.

The methodology's central component is an adversarial formulation of the pushforward condition, where the independence between the random variables z and y is posed in terms of test functions. When these test functions are restricted to finite-dimensional inner-product spaces, the optimal adversarial strategy can be expressed in terms of the first principal components of a matrix, reducing the problem to a pure minimization, which can be solved very efficiently through a flow-based gradient-descent procedure. The corresponding optimal map $y = T(x,z)$ can be inverted in closed-form, which facilitates conditional density estimation and simulation. The closed form inversion formula itself has intrinsic value, as it extracts from the data natural factors $\{f^k(z)\}$ that encode the dependence of x on z.

1.1. *Relation to prior work*

The Kantorovich –or Wasserstein– optimal transport barycenter problem was introduced in [4], defining the barycenter μ_* of a set of distributions $\{\mu_i\}$ as the minimizer of a weighted sum of the squared Wasserstein distances between the $\{\mu_i\}$ and μ_*. At first sight, this problem appears to differ substantially from the Monge OTBP that we address in this article, which given a joint distribution $\pi(x,z)$ between the outcome x and factors z that are not necessarily discrete, seeks a cost-minimizing map $y = T(x,z)$ such that the resulting random variable y is independent of z, a problem introduced in [1, 5]. Yet, as discussed in section 2, the Monge OTBP is equivalent to the Wasserstein barycenter problem extended to general factors z, when the solution of the latter is supported on z-dependent maps. The extension of the OTBP to a continuous covariate z was studied in [6] in the context of its connection to the multimarginal optimal transport in the limit of infinitely many marginals.

There is a rich literature regarding the numerical solution of the OTBP, typically in their Kantorovich formulation. Within the data-driven problem alone, there is more than one way to classify the most popular approaches. We can first distinguish between discrete methods, which assume that the marginals at hand consist of convex linear combinations of Dirac delta functions, and continuous methods, which work instead under the hypothesis that smooth probability density functions underly the data at hand. The first category includes methods that leverage the Sinkhorn algorithm [7, 8, 9] and linear programming-based methods [10, 11]. Among the first algorithms to treat the problem in a continuous setting are [12, 13], both based on the dual of Kantorovich formulation.

Those algorithms solving the continuous problem can be further divided into families. One criterion, functional to this work, regards the nature of the parametrization of the maps pushing forward each $\rho(x|z)$ to the barycenter. Most algorithms parametrize this map via a deep neural network with problem-dependent architecture. An example of this approach in [14, 15] uses Convex Neural Networks to parametrize a potential related to the optimal map. A different approach is the flow-based methodology adopted in [5] and inspired by [16, 17, 18]. Flow-based numerical solvers do not require neural networks or an a priori parametrization of the map, as they rely on the composition of infinitesimal elementary maps. They lead naturally to the adoption of gradient descent methods, more straightforward than saddle point optimizers, whose convergence is harder to characterize [19, 20].

While the time complexity of the cited solvers varies, a state-of-the-art solver does not really exist. Depending on the problem at hand, one approach should be favored over another. For instance, when dealing with images defined in high-dimensional spaces, a Sinkhorn-based approach with fast convergence may be more suitable than other approaches if one is not interested in resolving sharp details that are smoothed out by the entropic regularizer [21]. The complexity of the methodology that we proposed, which scales linearly with the number of sample points, is to our knowledge better than in all previous approaches.

Another connected work concerns not the OTBP but the related vector quantile regression [22]. This also transforms a random variable through a factor dependent map to another that is independent of the factor. Unlike the barycenter, however, the target distribution –the quantile– is fixed by the user, and the transformation proposed is linear and computed through linear programming. We believe that the connection between the two problems, the OTBP and quantile regression, is worth exploring, as it could provide new interpretations for both, as well as new applications, particularly in economics.

Our approach focuses on the statistical analysis of data with factors z that typically include continuous components, therefore requiring the solution of OTBP problems with infinitely many marginals. To the best of our knowledge, the work presented here and in [2, 5, 23] are the only ones dealing numerically with this aspect of the problem. Another major point distinguishing the work presented here from the existing literature is the interpretation of the pushforward condition that drives the samples underlying the marginals towards the barycenter. As in [5], we characterize the pushforward condition in terms of the statistical independence between two random variables: the cofactors z and $y = T(x,z)$. This statistical characterization of the barycenter is critical for a number of applications, such as removal of variability and factor discovery [1, 2], and treatment effect estimation [24, 25] (See [26] for a similar characterization of independence through reproducing kernel Hilbert spaces.) Directly related to the statistical interpretation of the push-forward condition is the ability to solve numerically the barycenter problem for continuous factors and under costs different from the canonical Euclidean distance. While there is some literature dealing with more general costs (see for instance [27, 28]), to the best of our knowledge, the only alternative work on the solution of this problem in the continuous setting –i.e. with infinitely many marginals– is our own previous work in [2, 5, 23]. The formulations in those articles differ substantially from the current one: the first solved a minimax problem for a potential $\psi(y,z)$, extending the attributable component methodology of [29] beyond nonlinear regression, the second developed BaryNet, a network-based algorithm, and the third formulated the push-forward condition in terms of a test function of the form $F(y,z) = \rho(y|z)$, whose estimation through kernels has a computational cost that grows quadratically with the number of samples. By contrast, the current proposal, which is based on flows in phase space and formulates the pushforward condition in terms of the first principal components of a matrix, has a complexity that scales linearly with the number of observations. An additional property of the new methodology is that it automatically identifies nonlinear features $\{f^l(z)\}$ that encode the dependence of x on z, effectively performing factor extraction.

1.2. *Plan of the article*

This article is structured as follows. Section 2 discusses the formulation of the Monge OTBP, relates it to an extension of the Wasserstein barycenter problem, and justifies its use to identify hidden sources of variability. Section 3 introduces an adversarial formulation of independence, relaxed to finite dimensional functional spaces provided with a suitable inner product. This gives rise to a rather compact formulation of the problem in terms of the singular values of a matrix, discussed in Section 4, and an efficient procedure for its minimization through gradient descent. Section 5 derives a closed-form expression for the inverse $x = X(y,z)$ of the map $y = T(x,z)$, mediated by extracted factors. Section 6 discusses various implementation aspects: the choice of functional spaces, the determination of the penalization parameter and of the learning rate, the termination criteria and the virtues of solving various barycenter problems in a row, each contributing to further explain the variability of the outcome x. It also discusses the algorithm's complexity. Section 7 solves the regular optimal transport problem (in its Monge formulation) through a suitable reduction of the more general OTBP. This is used in Section 8 to perform regular and conditional density estimation. Section 9 illustrates the methodology through numerical examples. We first use synthetic data sets to demonstrate various aspects of the OTBP, such as the simulation and estimation of conditional densities, model-free Bayesian inference and the uncovering of hidden explanatory factors. Then we apply the procedure to real data sets related to weather and climate. Finally, Section 10 summarizes the procedure and suggests avenues for further improvement.

2. A Monge formulation of the optimal transport barycenter problem

Given a joint distribution

$$\pi(x,z) = \rho(x|z)\ \gamma(z)$$

between two sets of variables: the *outcome* $x \in \mathscr{X}$ and the *covariates* $z \in \mathscr{Z}$, we seek a map

$$y = T(x,z) \in \mathscr{X}$$

that removes from x the variability that z can explain, i.e. such that the random variables y and z are independent. We require that the space $\mathscr{X}$ have the structure of a smooth manifold, while the space $\mathscr{Z}$ can include both continuous and discrete components. We will further assume that $\rho(x|z)$ is absolutely continuous uniformly over z, vanishing on small subsets of $\mathscr{X}$. In order to remove from x only the variability that z can account for, we select the map T that minimizes the expected value of the total transportation cost $C = \mathbb{E}_\pi[c(x,y)]$, where $c(x,y)$, an externally provided pairwise cost function, measures the deformation of the data incurred by moving x to y. The canonical choice for c is the squared distance

$$c(x,y) = \frac{1}{2}\|y-x\|^2. \tag{2.1}$$

More general costs C, not necessarily based on pairwise cost functions, give rise to the Distributional Barycenter Problem [5]. Even though the methodology developed in this article applies to the more general problem almost without changes, we restrict attention for concreteness to the pairwise canonical cost in (2.1).

The resulting problem reads

$$\min_{y=T(x,z)} \mathbb{E}_\pi[c(x,y)] \quad \text{s.t.} \quad y \perp\!\!\!\perp z, \tag{2.2}$$

where the symbol $\perp\!\!\!\perp$ denotes independence, a problem introduced in [5]. At first sight, (2.2) looks quite different from the Wasserstein barycenter problem, introduced in [4], which reads

$$\mu_* = \arg\inf_{\mu} \sum_{i=1}^{p} \lambda_i {W_2}^2(\mu_i,\mu), \quad {W_2}^2(\rho,\mu) = \inf_{\xi(x,y)\in\Pi(\rho,\mu)} \mathbb{E}_\xi\left[\|y-x\|^2\right], \tag{2.3}$$

where $\Pi(\rho,\mu)$ is the set of joint distributions having ρ and μ as marginals. The differences between (2.3) and (2.2) stem from their conceptual origin: while (2.3) extends the geometrical notion of barycenter to sets of distributions equipped with the Wasserstein distance, (2.2) uses a map T to remove from x any variability that z can explain. It appears almost coincidental that there should be any connection between the two!

Yet the two problems relate as follows. The $i \in [1, \dots, p]$ in (2.3) correspond to the $z \in \mathscr{Z}$ in (2.2), their weights $\{\lambda_i\}$ to the distribution $\gamma(z)$, and the $\{\mu_i\}$ to the $\rho(x|z)$. Thus the barycenter problem in (2.3) is restricted to discrete covariates z, which play the role of indexes for the distributions $\{\mu_i\}$. By contrast, the z in (2.2) is a random variable of general type linked to x through their joint distribution π. The well posedness of the barycenter problem with infinitely many marginals has been studied in [6] in connection to the multi-marginal optimal transport problem.

Extending Kantorovich's relaxation of the optimal transport problem [3], (2.3) considers general couplings ξ between the μ_i and μ, while (2.2) extends Monge's original formulation of optimal transport [30] to the barycenter problem, restricting attention to maps T that push forward the $\rho(x|z)$ to μ. These maps are central to the applications that motivate (2.2), as they are used both for conditional density simulation and estimation. Importantly, they turn $y = T(x,z)$ into a random variable that derives from x and z, which leads to another critical distinction: while the argument of the minimization in (2.3) is the barycenter μ_* of the μ_i, the formulation in (2.2) does not involve the barycenter at all! The fact that y in this formulation is a random variable gives meaning to the alternative requirement of independence between y and z.

That the two problems are much closer than they appear at first sight follows from the fact that classical work [31] has shown that, for smooth distributions and under quite general assumptions, the solution to Kantorovich's formulation of the optimal transport problem also solves Monge's, a results that has been extended to the OTBP in [4, 6]. We put together the connection between the two problems in the framework of this article through the three theorems that follow. First, a partial equivalence between the two formulations is given by the following theorem:

Theorem 1 *Given a joint distribution $\pi(x,z)$, $x \in \mathscr{X}$, $z \in \mathscr{Z}$, define the marginal $\gamma(z) = \pi(\mathscr{X}, z)$ and the conditional distribution $\rho(x|z) = \frac{\pi(x,z)}{\gamma(z)}$, and consider the following two problems:*

1. *Extended Wasserstein barycenter:*
$$\mu^*, \xi_z^* = \arg\min_{\mu,\xi_z} C_K = \mathbb{E}_\gamma\left[\mathbb{E}_{\xi_z}[c(x,y)]\right], \quad \xi_z \in \Pi\left(\rho(x|z), \mu(y)\right)$$
(We call this problem "extended" because the covariates z are not necessarily discrete),
2. *Monge barycenter:*
$$T^* = \arg\min_T C_M = \mathbb{E}_\pi[c(x,y)], \quad y = T(x,z), \quad y \perp\!\!\!\perp z.$$

If the minimizing couplings ξ_z^* for the first problem are supported on maps Q_z,
$$\xi_z^*(x,y) = \pi(x,\mathscr{Z})\ \delta\left(y - Q_z(x)\right),$$
then
$$T^*(x,z) = Q_z(x), \quad \textit{and consequently } \ C_M = C_K \ \textit{ and } \forall z\ \mu^* = T^*(:,z)\#\rho(:|z).$$

Proof Since ξ_z^* solves problem 1, $Q_z\#\rho(:|z) = \mu^*$, so the joint distribution $\Theta(y,z)$ satisfies
$$\Theta(y,z) = \mu^*(y)\gamma(z),$$
which implies that $y \perp\!\!\!\perp z$. If there existed another $T \neq Q_z$ pushing forward all $\rho(x|z)$ to a single distribution $\mu(y)$ at a cost $C_M < C_K$, then $\xi_z(x,y) = \pi(x,\mathscr{Z})\ \delta\left(y - T(x,z)\right)$ would solve problem 1 with a smaller cost than the optimal ξ_z^*, a contradiction.

□

If $\rho(x|z)$ vanishes on small sets uniformly over z, the optimal ξ_z^* are indeed supported on maps, which also proves the existence of solutions to the Monge OTBP:

Theorem 2 *If the joint distribution $\pi(x,z)$ is absolutely continuous in x uniformly over z, then the barycenter $\mu(y)$ of the corresponding $\rho(x|z)$ under the canonical cost in (2.1)) is also absolutely continuous and the optimal couplings ξ_z^* are supported on maps.*

This theorem extends to general covariates Theorem 5.1 in [4], where it is proved for discrete z's. A similar result for continuous z was proved in [6], see Corollary 3.3.3 and Theorem 4.2.5. An alternative proof, not included here for conciseness, builds on the fact that any small set $A \in \mathscr{X}$ on which a distribution μ does not vanish can be mollified so that the optimal transport between any smooth distribution ρ and the mollified μ has a smaller total transportation cost (The bandwidth of the mollification relates to the degree of smoothness of ρ, hence the requirement of uniformity of the latter over z.)

We might conclude from theorems 1 and 2 that the Monge and [extended] Wasserstein barycenter problems are equivalent when applied to smooth distributions: after all, their unique solutions map to each other. Yet this equivalence applies to the problems' solutions, not to their formulations. All the applications described in this article, as well as the very methodology proposed for solving the problem numerically, are strongly based on the map $y = T(x,z)$ and its inverse $X(y,z)$, both parameterized by z and defining one of the random variables (x,y) in terms of the other.

One such application of the Monge OTBP is to uncover hidden sources of variability by removing the effects of known factors. The relation between a random variable $x \in \mathscr{X}$ and known covariates $z \in \mathscr{Z}$ can be specified alternatively through the conditional distribution $\rho(x|z)$ and through a functional relation

$$x = \phi(w,z), \quad w \in \mathscr{W}, \quad w \sim \nu(:|z),$$

where the random variable w represents all additional causes of variability in x, which we either do not currently consider, cannot measure or are simply not aware of. The function ϕ and the distribution ν underlying w determine $\rho(x|z)$ uniquely, but more than one pair (ν,ϕ) can give rise to the same ρ. One special such pair is provided by the solution to the OTBP:

Theorem 3 *Given any joint distribution $\pi(x,z)$ that vanishes on small sets in $\mathscr{X}$ uniformly over z, the random variable x can be written as*

$$x = X(y,z),$$

where $y = T(x,z) \sim \mu$ is the solution to the barycenter problem for $\rho(x|z)$, so $X(:,z) = T^{-1}(:,z)$.

Proof Neither $\rho(x|z)$ (for any z) nor $\mu(y)$ assign finite measure to small sets. Then the fact that $T(:,z)$, the least costly map that pushes forward $\rho(:|z)$ to μ, is invertible is a central result in optimal transport theory [3].

□

This theorem provides the ground for various applications.

1. In order to generate samples $\{x_i^*\}$ from $\rho(x|z_*)$ for any target value z_*, it is enough to generate samples $\{y_i\}$ from the barycenter $\mu(y)$ and write $x_i^* = X(y_i, z_*)$. A number n of such samples is already available through the barycentric map acting on the available data pairs, $y_i = T(x_i, z_i)$. More can be obtained if needed, performing optimal transport between μ and a given distribution ν and pushing forward samples of the latter to μ through the inverse of the corresponding transportation map.
2. Since our algorithm provides both $X(y,z)$ and its y-derivatives (see section 5), one can estimate $\rho(x|z_*)$ by the change-of-variable formula applied to an estimate for $\mu(y)$ (which can itself be obtained by optimal transporting μ to a known ν and applying again the change-of-variable formula.)

A third application addresses the following question: assuming that there exists a "true" additional source w of variability in x, such that that $x = \phi(w,z)$, how much does our $y = T(x,z)$ teach us about the true w? Here the notion of a true source is field dependent. For our purposes, we just assume that such true w exists.

An identifiability issue arises. In the absence of additional information, the conditional distribution $\rho(x|z)$ does not suffice to determine w. For instance, if for some value of z, $\phi(w_1,z) = \phi(w_2,z)$, then there is no way that using x and z alone we could distinguish between w_1 and w_2. More generally, any two z-dependent random variables W_1^z and W_2^z such that the distributions of both $\phi_1(z,W_1^z)$ and $\phi_2(z,W_2^z)$ agree with $\rho(x|z)$ explain the data equally well. In particular, y provides one such explanatory variable, with the additional property that it is necessarily a function of w and z:

$$y = T(x,z) = T(\phi(w,z),z) = Y(w,z).$$

Moreover, $Y(:,z)$ is invertible for all values of z for which w is identifiable, i.e. such that $\phi(w_1,z) = \phi(w_2,z) \Rightarrow w_1 = w_2$, since

$$Y(w_1,z) = Y(w_2,z) \;\Rightarrow\; T(\phi(w_1,z),z) = T(\phi(w_2,z),z) \;\Rightarrow\; \phi(w_1,z) = \phi(w_2,z) \;\Rightarrow\; w_1 = w_2.$$

Then, in order to uncover the "true" hidden explanatory w, one can use the fact that its identifiable component must have a –possibly z-dependent– one-to-one relation to y, together with any other information available on w, such as other variables that it may depend upon or correlate with. In the absence of any such additional information, y is the most natural explanatory variable among all w, since by construction it is independent of the known factors z, it is the closest to x itself, and it is the most "economical", since it is identifiable for all z. We illustrate these concepts through examples in Section 9.

3. Adversarial characterization of independence

Posing the Monge optimal transport barycenter problem (2.2) in data-driven scenarios, where the joint distribution $\pi(x,z)$ is only known through n sample pairs $\{x_i, z_i\}$, requires a sample-friendly formulation of the independence condition between the random variables y and z. We will use a weak characterization based on test functions [5, 32]: two variables $y \in \mathscr{Y}$ and $z \in \mathscr{Z}$ with joint distribution $\pi(y,z)$ are independent if and only if any two bounded measurable functions $g(y)$ and $f(z)$ satisfy

$$\mathbb{E}_\pi[g(y)f(z)] = \mathbb{E}_\rho[g(y)]\,\mathbb{E}_\gamma[f(z)], \quad \rho(y) \stackrel{\text{def}}{=} \pi(y,\mathscr{Z}),\ \gamma(z) \stackrel{\text{def}}{=} \pi(\mathscr{Y},z).$$

We can restrict the functions f and g to smaller spaces $\mathscr{F}$ and $\mathscr{G}$, provided that they contain suitable approximations to the delta function consistent with the smoothness of $\pi(y,z)$. To prove this, notice that if $\pi(y,z) > \rho(y)\,\gamma(z)$ in a neighborhood U_* of the pair (y_*, z_*), then $g(y) = \delta_{y_*}(y)$, $f(z) = \delta_{z_*}(z)$ satisfy

$$\mathbb{E}_\pi[g(y)f(z)] > \mathbb{E}_\rho[g(y)]\ \mathbb{E}_\gamma[f(z)],$$

where δ_{y_*} and δ_{z_*} are non-negative functions whose product is positive at (y_*, z_*) and vanishes outside of U_*.

We can further subtract from $f(z)$ its mean, yielding the following equivalence statement: two variables y and z are independent if and only if, for all functions $g(y) \in \mathscr{G}$ and $f(z) \in \mathscr{F}$ with $\mathbb{E}_\gamma[f(z)] = 0$,

$$\mathbb{E}_\pi[g(y)f(z)] = 0.$$

This equivalence gives rise to the following adversarial formulation of the barycenter problem (2.2):

$$\min_{y=T(x,z)} \max_{g,f,\lambda} L = \mathbb{E}_\pi\left[c(x,y) + \lambda\ \mathbb{E}_\pi[g(y)f(z)\right], \quad \mathbb{E}_\gamma[f] = 0,\ \|f\| = \|g\| = 1, \tag{3.1}$$

where we have decoupled the amplitude of f and g from their shape, absorbing their amplitude in the factor λ. Moreover, we can replace the maximization over λ by the external provision of a a penalization parameter $\lambda \gg 1$ for non-compliance of the independence condition, a relaxation that converges to (3.1) as $\lambda \to \infty$.

If we defined the norms of f and g in (3.1) through the canonical inner products

$$(f_1, f_2) = \int f_1(z)\ f_2(z)\ d\gamma(z), \quad (g_1, g_2) = \int g_1(y)\ g_2(y)\ d\rho(y), \tag{3.2}$$

these norms would represent the standard deviation of f and g, since not only $\mathbb{E}_\gamma[f] = 0$ by construction, but also $\mathbb{E}_\rho[g] = 0$ holds at the optimal solution: a constant added to g does not affect the value of L, and the norm of $g - a$ is smallest when $a = \bar{g}$. It follows that we could read the problem as the minimization of the transportation cost subject to the condition that the correlation between any two functions $f(z)$ and $g(y)$ vanishes [32]. Yet adopting this choice for a norm is neither required nor convenient for $g(y)$, since it depends on the unknown $\rho(y)$, which evolves through the optimization procedure. We will propose an alternative norm below.

A data-driven formulation of (3.1) replaces expected values by empirical means,

$$\min_{y_i=T(x_i,z_i)} \max_{g,f} L = \frac{1}{n}\sum_{i=1}^{n}\left[c(x_i,y_i) + \lambda\ g(y_i)\, f(z_i)\right], \quad \sum_{i=1}^{n} f(z_i) = 0,\ \|f\| = \|g\| = 1. \tag{3.3}$$

Since we cannot enforce infinitely many constraints on the finite set $\{y_i\}$ without trivializing the solution, we supplement (3.3) with the specification of two finite dimensional inner-product spaces of functions $\mathscr{F}$ and $\mathscr{G}$ over which to perform the maximization, writing

$$f(z) = F(z)a, \quad g(y) = G(y)b, \quad a \in R^{m_z}, \quad b \in R^{m_y},$$

where the m_z columns of F and the m_y columns of G are functions respectively of z and y (both have n rows when evaluated at the sample points) and the functions acting as columns of F have zero mean. Choices for the functions defining $\mathscr{F}$ and $\mathscr{G}$ will be discussed in Section 6.5. Independently of their choice, some further processing is required, which we describe here in terms of $\mathscr{G}$, since the same process applies to $\mathscr{F}$.

A first requirement is to eliminate redundancy: since the columns of the matrix G typically consist of smooth functions that we only evaluate at a finite set of points, the dimension of the effective range of G can be much smaller than m_y, particularly when the latter is chosen large so as to accommodate for a rich set of candidate test functions. Eliminating such redundancy makes the maximization problem over b smaller and better posed. The second requirement also relates to the ease of optimization over b: enforcing the requirement that $\|g(y)\| = 1$ would be much easier if G –an operator from R^{m_y} to $\mathscr{G}$– were orthogonal, as it would directly translate into the condition that $\|b\| = 1$.

Enforcing the orthogonality of G requires that we fix an inner product in $\mathscr{G}$. The canonical one

$$\langle \phi, \psi \rangle = \sum_i \phi\left(y_i\right) \psi\left(y_i\right), \tag{3.4}$$

the empirical version of (3.2), has the problem that its very definition depends on the unknown $\{y_i\}$. We can stick to the canonical inner product in z-space, since the $\{z_i\}$ are fixed, but we should use a different one for functions of y. We need a functional norm such that a function $g(y)$ of norm 1 cannot be very large on the data. As a counterexample, consider an inner product of the form

$$(\phi, \psi) = \int \phi(y)\psi(y)\, w(y) dy,$$

where the weighting function $w(y)$, though everywhere positive, is very small in at least one area where the true distribution $\rho(y)$ is not. Since the requirement that $\|g\| = 1$ will effectively only constrain $g(y)$ in areas where $w(y)$ is comparatively large, the algorithm's variables b can make L large by choosing test functions $g(y)$ not so much based on their correlation with $f(z)$ but just on their effective amplitude in areas where ρ is large but the norm of g does not truly act as a constraint. It follows that we must choose a weight $w(y)$ so that the Radon–Nikodym derivative $\frac{d\rho}{dw}$ is bounded. We adopt an inner product that is fixed through stages of the procedure, updating it only occasionally to reflect the evolving distribution of the $\{y_i\}$, replacing (3.4) by

$$\langle \phi, \psi \rangle = \sum_i \phi\left(y_i^0\right) \psi\left(y_i^0\right), \tag{3.5}$$

where y_i^0 is initially set to x_i and then updated every so often, to reflect more accurately the different stages of the $\{y_i\} \sim \rho(y)$.

In order to replace $G(y)$ by a smaller dimensional, orthogonal operator $Q_y(y)$ that spans the same effective range, we perform the reduced singular-value decomposition

$$G_i^j \overset{\text{def}}{=} G^j\left(y_i^0\right) \approx \sum_{k=1}^{n_y} \sigma_k\, u_i^k\, v_j^k,$$

where $n_y \le m_y$ is determined from the criterion that the sum $\sum_{k=1}^{n_y} (\sigma_k)^2$ be larger that a fraction of the squared norm of $\|G\|^2 = \sum_{i,j} \left(G_i^j\right)^2 = \sum_{k=1}^{m_y} (\sigma_k)^2$, and we adopt

$$Q_y(y) = G(y)B^y, \quad B^y_{jk} = \frac{1}{\sigma_k} v_j^k$$

(Notice that, in particular, $Q_y^k\left(y_i^0\right) = u_i^k$.) With a fixed inner product, the matrix B^y needs to be computed only once per stage, when the y^0 are updated. The sense in which this Q^y is orthogonal is not the conventional one: its columns represent the evaluation at arbitrary positions $\{y_i\}$ of a set of functions that are orthonormal under a chosen, fixed inner product based on the $\{y_i^0\}$.

The same procedure applied to F produces an orthogonal matrix Q_z of rank n_z and corresponding matrix B^z. For z, the distinction between matrices and operators is immaterial, since $F(z)$ and $Q_z(z)$ are only applied at the fixed set of points $\{z_i\}$.

4. A flow-based methodology

Replacing in (3.3) $f(z_i)$ by $Q_z(z_i)a$ and $g(y_i)$ by $Q_y(y_i)b$ yields

$$\min_{\{y_i\}} \left\{ \max_{a,b} \sum_{i=1}^{n} c(x_i, y_i) + \lambda \sum_{h=1}^{n_z} \sum_{l=1}^{n_y} \left(\sum_{i=1}^{n} Q_z^h(z_i) Q_y^l(y_i) \right) a_h b_l, \quad \|a\| = \|b\| = 1 \right\}.$$

The maximization over a and b can be carried out explicitly: they must align with the left and right first principal components of the $n_z \times n_y$ matrix

$$A^{hl}(y) \overset{\text{def}}{=} \sum_i Q_z^h(z_i) Q_y^l(y_i)$$

(A depends only on the $\{y_i\}$, since the $\{z_i\}$ are fixed), and the penalty term is given by the first singular value $\sigma_1(y)$ of A. It follows that we can write the problem as a minimization over y alone:

$$\min_{\{y_i\}} L = \sum_{i=1}^{n} c(x_i, y_i) + \lambda \|A(y)\|, \quad \|A\| \overset{\text{def}}{=} \sigma_1 = \max_{\|a\|=\|b\|=1} a' A b. \tag{4.1}$$

We can interpret the corresponding functions $f(z) = Q_z(z)a$, $g(y) = Q_y(y)b$ as the features whose correlation most strongly displays the current dependence between z and y.

This suggests a flow-based procedure, whereby y, initially set equal to x, follows gradient descent of (4.1),

$$y_i^{n+1} = y_i^n - \eta^n \left[\frac{1}{n} \nabla_y c(x_i, y)|_{y_i^n} + \lambda\, a'\, \nabla_y A|_{y_i^n} b \right] \tag{4.2}$$

(for which all $\{y_i\}$ decouple), where

$$\nabla_y A^{hl}\Big|_{y_i^n} = Q_z^h(z_i) \sum_j \nabla G^j(y)\big|_{y_i} B_{jl}^y$$

and a and b are updated in an alternate step.

It might appear that we are taking an uncontrolled approximation to the y-gradient of L in (4.1) by differentiating only A in (4.2) at fixed a and b. The principal components of A do of course depend on A, so they too change when A varies. Yet this way of computing derivatives is exact:

Theorem 4 *The derivative of the k-th principal value σ_k of a matrix A with respect to any parameter s on which A may depend, is given by*

$$\frac{\partial}{\partial s} \sigma_k = a' \left(\frac{\partial}{\partial s} A \right) b,$$

where a and b are the left and right kth principal components of A.

Proof By definition, $\sigma_k = a'Ab$, so

$$\frac{\partial}{\partial s}\sigma_k = a'\left(\frac{\partial}{\partial s}A\right)b + \left(\frac{\partial}{\partial s}a\right)' Ab + a'A\left(\frac{\partial}{\partial s}b\right).$$

But the principal components satisfy $Ab = \sigma_k a$, $A'a = \sigma_k b$ and $\|a\| = \|b\| = 1$, so

$$\left(\frac{\partial}{\partial s}a\right)' Ab = \sigma_k \left(\frac{\partial}{\partial s}a\right)' a = \sigma_k \frac{\partial}{\partial s}\frac{\|a\|^2}{2} = 0 \quad \text{and} \quad a'A\left(\frac{\partial}{\partial s}b\right) = \sigma_k b'\left(\frac{\partial}{\partial s}b\right) = \sigma_k \frac{\partial}{\partial s}\frac{\|b\|^2}{2} = 0.$$

□

The penalty term $\sigma_1(y)$ is not smooth at its arg-min $y = y^*$: for y and z to be independent, the first singular value of A must vanish, so $A(y^*)$ itself must equal zero, and the first singular value $\sigma_1(y)$ of a matrix that depends smoothly on y typically has corners where $A(y)$ vanishes (The simplest example is the 1×1 matrix $A = y \in R$, whose only singular value $\sigma = |y|$ has a corner at $y = 0$.) To address this, we square the penalty term:

$$\min_{\{y_i\}} L = \sum_{i=1}^{n} c(x_i, y_i) + \lambda\ \sigma_1^2(y), \quad \sigma_1(y) \overset{\text{def}}{=} \max_{\|a\|=\|b\|=1} a'A(y)b.$$

There still remains one issue to address to make the methodology fully functional. Because every step of the algorithm brings down the largest singular value of $A(y)$, the first few of those singular values will tend to coalesce at convergence at a common value $\sigma_* \ll 1$. Then the derivatives of the penalty term with respect to the $\{y_i\}$ are not well defined, as they depend on the arbitrary choice of one pair among the various singular components (a,b) associated to the singular value σ_*. In terms of test functions, more that one pair of functions (f,g) have reached the threshold correlation σ_*.

To address this, we modify the algorithm so that it tracks the first K pairs (a_k, b_k) of principal component of A, where $K = \min(\text{rank}(A), K_{max})$, with K_{max} fixed by the user. Then we descend over y

$$\min_{\{y_i\}} L = \sum_{i=1}^{n} c(x_i, y_i) + \lambda \sum_{k=1}^{K} \sigma_k{}^2(y), \quad \sigma_k(y) \overset{\text{def}}{=} a_k{}'A(y)b_k. \tag{4.3}$$

Notice that this extension carries little computational cost, since the $A(y)$ to differentiate is common to all the $\{\sigma_k\}$. If performed using reproducing Kernel Hilbert spaces, this extension could be thought as interpolating between COCO [26] and HSIC [33].

5. Map inversion

The procedure described so far finds n samples $y_i = T(x_i, z_i)$ of the barycenter $\mu(y)$. In order to simulate $\rho(x|z_*)$ for a target z_*, we need to invert T to obtain n samples $\{x_i^*\}$ from $\rho(:|z_*)$ through

$$x_i^* = X(y_i, z_*) \overset{\text{def}}{=} T^{-1}(y_i, z_*).$$

Since we do not know $T(x,z)$ in closed form, it could appear that we can only invert it by learning $X(y,z)$ from its n available samples $\{x_i, y_i, z_i\}$, for instance through kernel regression, nearest neighbor or neural networks. Yet we can do much better than that and obtain a closed form for T^{-1}, exploiting the fact that the penalization parameter λ is large but finite. Since at convergence the gradient $\nabla_y L$ is zero, or at least sufficiently small to satisfy a termination criterion, we have

$$\nabla_{y_i} c(x_i, y_i) + 2\lambda \sum_{k=1}^{K} \sigma_k \nabla_y \sigma_k \Big|_{y_i, z_i} = 0,$$

which for the canonical quadratic cost in (2.1) yields $x_i = y_i + 2\lambda \sum_{k=1}^{K} \sigma_k \nabla_y \sigma_k \Big|_{y_i, z_i}$. In order to invert the map for arbitrary values of y and z, we extend the validity of this expression and write

$$X(y,z) = y + 2\lambda \sum_{k=1}^{K} \sigma_k \nabla_y \sigma_k \Big|_{y,z}, \quad \text{where} \quad \nabla_y \sigma_k \Big|_{y,z} = f_k(z) \nabla g_k(y), \tag{5.1}$$

an expression that is smooth in (y,z) and yields $X(y^i, z^i) = x^i$ on all the available samples.

Finding $x_i^* = X(y_i, z_*)$ requires the values of $f_k(z_*)$ and $\nabla g_k\big|_{y_i}$. Out of these, all the

$$\nabla g_k\big|_{y_i} = \nabla_y G(y)\big|_{y_i} B^y b_k$$

are already known, since they have been used at the last descent step. Hence we only need the K numbers $\{f_k(z_*)\}$, which we calculate by introducing a new row $F(z_*)$ of F and setting

$$f_k(z_*) = F(z_*) B^z a_k.$$

The inversion formula in (5.1) provides us with a valuable bonus: it shows that the dependence of x on z that our algorithm has uncovered is mediated by the K functions $\{f_k(z)\}$, so we have inadvertently performed *factor extraction*. The extracted factors bring in insights about the mechanisms of the dependence of x on z, while the gradient of these functions inform us of the sensitivity of x with respect to changes in z.

When we consider density estimation in Section 8, it will be useful to notice that we have access not only to $X(y,z)$ but also to its derivatives,

$$\frac{\partial X^p(y,z)}{\partial y^q} = \delta_p^q + 2\lambda \sum_{k=1}^{K} \sigma_k \frac{\partial^2}{\partial y^p \partial y^q} \sigma_k \Big|_{y,z}, \quad \frac{\partial^2}{\partial y^p \partial y^q} \sigma_k \Big|_{y,z} = f_k(z) \left(\frac{\partial^2 G(y)}{\partial y^p \partial y^q} \right) B^y b_k. \tag{5.2}$$

6. Implementation

6.1. *Choices for the functional spaces $\mathscr{F}$ and $\mathscr{G}$*

The methodology is not fully specified until we select the functional spaces $\mathscr{F}$ and $\mathscr{G}$ to use. This section describes some practical choices and their consequences, including the choices that we have made for the simulations in Section 9. A more complete exploration of data-adapted functional spaces is beyond the scope of this article, it will be pursued elsewhere.

Even though the notion of independence is symmetric in y and z, the consequences of restricting the two functional spaces $\mathscr{G}$ and $\mathscr{F}$ enforcing independence through (3.1) are conceptually quite different. While the richness of $\mathscr{F}$ relates to the level of resolution of the dependence of x on z, the richness of $\mathscr{G}$ specifies the level of detail with which the distributions $\rho(x|z)$ are captured for each value of z. Imagine temporarily that $\mathscr{F}$ is rich enough to capture all functions $f(z)$ that may be required, and concentrate on the effect of making different choices for $\mathscr{G}$. Since for any $g \in \mathscr{G}$, the condition

$$\int g(y) f(z)\; d\rho(y|z)\; d\gamma(z) = 0$$

must hold for any $f \in \mathscr{F}$ with zero mean –and $\mathscr{F}$ is by hypothesis rich enough– it follows that

$$\int g(y)\; d\rho(y|z) \perp\!\!\!\perp z,$$

i.e. the conditional expectation $\bar{g}(z)$ is in fact a constant $\bar{g}$ independent of z.

When $X = R^d$, the simplest choice for $\mathscr{G}$ is the space of linear functions, spanning the columns of a matrix G with the d independent functions y^l, $l \in [1,\ldots,d]$. It follows from the argument above that the conditional mean of y is independent of z. On the other hand, from (5.1), $X(y,z) = y + h(z)$, where $h(z) = 2\lambda\ \sum_{k=1}^{K} \sigma_k\ f_k(z)\ v_k$ and $v_k \overset{\text{def}}{=} \nabla_y g_k(y)$ is a constant, since $g_k(y)$ is linear. It follows from taking the conditional expectation of both sides that

$$h(z) = \bar{x}(z) - \bar{y},$$

so the procedure captures –and removes from x– the conditional mean $\bar{x}(z)$, i.e. it performs [nonlinear] regression, as in the "poorest man solution" of [1]. When the columns of G span a general quadratic function of y, not only the conditional mean but also the conditional covariance matrix of y is independent of z, and (5.1) implies that the relation between x and $y = T(x,z)$ is linear (with z-dependent coefficients), as in the "poor man solution" of [1]. We typically start our experiments with a first run where the columns of G span all quadratic functions of y, to capture the conditional mean and covariance matrix of x, leaving a more detailed characterization of $\rho(x|z)$ to subsequent runs.

One can of course extend the choices above and fill all columns of G with externally provided functions, such as Hermite polynomials of a given degree. Similarly, we can use as columns of F polynomials when $z \in R^d$, trigonometric functions when z is a periodic variable, such as the time of the day, and indicator functions when z can only adopt a finite set of categorical values. Yet it is generally preferable to use a less parametric approach and let the data dictate the form of the functions to use. For our experiments, we have used a simple class of data-adapted spaces described in the appendix, where the columns of F and G are given by asymmetric kernel-like functions with column dependent center and bandwidths, a flexible and economic variation of reproducing kernel Hilbert spaces:

$$F^j(z) = K^z\left(z, z_j^c\right), \quad G^j(z) = K^y\left(y, y_j^c\right).$$

Even though the $\{y_i\}$ evolve, the centers $\{y_j^c\}$ are fixed throughout stages of the procedure, so as to have a fixed functional space $\mathscr{G}$. Their cardinality does not need to match that of the $\{y_i\}$, in practice it is typically much smaller.

6.2. *Choice of the penalization parameter*

The penalization parameter λ establishes a balance at the final $y = y_*$ between the gradients of the transportation cost function and the penalization term. It follows from (4.3) that if one would like the

$\sigma_k\left(y_*\right)$ to have values smaller than $\sigma_* \ll 1$, one must adopt a parameter λ satisfying

$$\sqrt{\overline{\left\|\nabla_y c\left(x,y\right)\big|_{y^*}\right\|^2}} = 2\lambda\sqrt{\overline{\left\|\sum_k \tilde{\sigma}_k \nabla_y \sigma_k\left(y\right)\big|_{y^*}\right\|^2}},$$

where $\tilde{\sigma}_k \overset{\text{def}}{=} \min(\sigma_k, \sigma_*)$ and $\overline{\|s\|^2} \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \|s_i\|^2$ for any $s = \{s_i\}_{i=1,\ldots,n}$, $s_i \in R^d$. Based on this characterization of λ at convergence, we adopt a state-dependent penalization parameter that evolves over algorithmic time:

$$\lambda = \frac{1}{2}\frac{\sqrt{\overline{\left\|\nabla_y c\left(x,y\right)\right\|^2 + 0.1\operatorname{var}(x)}}}{\sqrt{\overline{\left\|\sum_{k,l} \tilde{\sigma}_k \nabla_y \sigma_k^l(y)\right\|^2}}}.$$

The addition of a small fraction of the variance to the numerator addresses the fact that $\nabla_y c = 0$ at the onset of the algorithm, when $x = y$ (The variance is a natural reference value, since $\overline{\left\|\nabla_y c\left(x,y\right)\right\|^2} = \frac{1}{n}\sum_i \|x_i - y_i\|^2 \leq \frac{1}{n}\sum_i \|x_i - \bar{x}\|^2 = \operatorname{var}(x)$.)

6.3. *Learning rate*

We minimize the objective function L in (4.3) through gradient descent, with steps of the form

$$y_i^{n+1} = y_i^n - \eta^n \, \nabla_{y_i} L\big|_{y^n},$$

determining the learning rates η^n through back-tracking limited to a small interval:

1. At time t_n, pick an initial candidate η through $\eta = \theta\ \eta^{n-1}$, where θ is only slightly larger than one, so as to explore an interval for η^n that is only marginally larger than the accepted learning rate from the previous step.
2. Back-track from this η through the Armijo-Goldstein algorithm: defining $G = \nabla_y L\big|_{y^n}$, set $\eta^n = \eta$ if $L\left(y^n - \eta G\right) \leq L\left(y^n\right) - \kappa\eta\|G\|^2$, with $0 < \kappa < 1$. Otherwise, set $\eta \to \tau\eta$, $0 < \tau < 1$ and repeat this step.

We have adopted for our numerical examples $\theta = 1.1$ and $\kappa = \tau = \frac{1}{2}$.

6.4. *Termination criterion*

At convergence, $y = y^*$ must satisfy two natural criteria for termination:

1. Any remaining dependence of y on z must be within an acceptable range:

$$\forall k\ \sigma_k\left(y^*\right) < \sigma_* \ll 1. \tag{6.1}$$

   In order to assign a value to σ_*, notice that σ_k represents the empirical correlation between $f_k(z)$ and $g_k(y)$, which should be uncorrelated. It follows that a reference value for σ_* is the standard deviation of the empirical correlation between two independent variables, which equals $\frac{1}{\sqrt{n}}$. We have adopted in our experiments

$$\sigma_* = \frac{\nu}{\sqrt{n}}, \quad \nu = 0.2. \tag{6.2}$$

   Once the criterion in (6.1) is satisfied, we freeze λ at its current value until criterion 2 (below) is satisfied.

2. The gradient $\nabla_y L$ of the objective function must be sufficiently small for the inversion formula (5.1) to be valid. Since the error in the determination of x_i from this formula is given by

$$\|x_i - X(y_i, z_i)\| = \left\| \nabla_{y_i} L \big|_{y^*} \right\|$$

and a natural reference scale for the square of this error is the variance of x, we use as termination criterion

$$\frac{1}{n} \sum_i \left\| \nabla_{y_i} L \big|_{y^*} \right\|^2 < \alpha \ \mathrm{var}(x), \tag{6.3}$$

with $\alpha \ll 1$. We end the run when both (6.1) and (6.3) are satisfied.

Another, internal termination criterion starts a new stage once the current y differ significantly from their values y^0 at the outset of the current stage, i.e. when

$$\sum_{i=1}^{n} \left\| y_i - y_i^0 \right\|^2 > \delta \sum_{i=1}^{n} \left\| y_i^0 - \bar{y}^0 \right\|^2, \quad 0 < \delta < 1.$$

In our experiments, we have adopted $\alpha = 0.0025$ and $\delta = 0.1$,

6.5. *Successive barycenter problems*

The barycenter problem removes from x any z-dependence detectable through the functional spaces $\mathscr{F}$ and $\mathscr{G}$, so the resulting $y = T(x,z)$ can still depend on z in ways that $\mathscr{F}$ and $\mathscr{G}$ do not capture. For instance, if $\mathscr{G}$ consists only of quadratic function of y, just the conditional mean and covariance of x are removed, leaving in y any other z-dependent property of $\rho(x|z)$, such as higher moments or the distribution's modality. Similarly, if $\mathscr{F}$ includes only functions of a subset of the $\{z^l\}$, y may still depend on the remaining ones, if $\mathscr{F}$ includes only functions of the individual $\{z^l\}$, any non-additive dependence of x on the $\{z_l\}$ will remain in y, and if the bandwidths of the functions in $\mathscr{F}$ are large, only long-scale trends are removed, leaving small-scale signals behind.

This suggests proceeding in N_s stages: calling $y^0 = x$, we compute in stage l the barycenter of the y^{l-1} over z, as captured through the functional spaces $\mathscr{F}^l$ and $\mathscr{G}^l$. The final y^{N_s} are not samples of the barycenter of the original x, since the composition of optimal maps is not necessarily optimal. However, since we know how to invert each of the maps, we can still simulate and estimate by composition $\rho(x|z^*)$ for any target z^*. This procedure resembles boosting [34], in which multiple models are trained sequentially so that each new problem removes further variability from the barycenter of the prior one.

6.6. *Complexity*

One major advantage of the new methodology is its efficiency, which makes it applicable to large data sets. This efficiently derives from formulating the independence conditions between y and z in terms of the uncorrelation between test functions and subsequently relaxing it to the vanishing of the first singular value of a matrix $A(y)$ whose rank does not depend on the sample size. Previous methods used kernel density estimation to formulate independence [5], requiring kernels where every data point acted as a center, yielding at least $O(n^2)$ time complexity. By contrast, the new algorithm's time complexity scales bilinearly with the number of samples (n) and the dimension of the data (d_x). Because of this, most of the numerical examples in Section 9 had running times on a laptop raging between a fraction of a second and a few seconds.

The algorithm's operations can be broken into three different categories: those which are performed only once per run, those performed once per stage, and those performed at each descent step. Even though it is only the third category that determines the time complexity of the algorithm in practice, we analyze all three parts for completeness.

1. Operations performed only once. Since the factors z do not evolve through a run, the orthogonal matrix Q_z is computed only once, making the algorithm's complexity insensitive to the dimensionality of z. When using kernels, calculating Q_z requires k-means clustering to determine the centers for z, which with a fixed maximum number of iterations requires $\mathrm{O}(n * d_z * m_z)$ operations. Evaluating the kernel function also requires $\mathrm{O}(n * d_z * m_z)$ operations. A standard singular value decomposition of the matrix $F \in \mathbb{R}^{n \times n_z}$ requires $\mathrm{O}(n * n_z^2)$ steps where n_z is a user's provided input. Therefore, the number of operations performed only once scales as $\mathrm{O}(n * d_z * m_z)$. For very large data-sets with high-dimensional factors z, this number can be further reduced by adopting state-of-the-art methodologies for finding the first few principal components of large matrices [35, 36], but we found no need for this in our experiments so far. Even for time-like factors z, which as discussed in the appendix require a number of centers that grows linearly with the extent of the time series analyzed, the corresponding matrix $F(z)$ is sparse, which keeps the complexity of its principal component analysis at $O(n)$, since the complexity scales not with the number of columns of F (i.e. the number of centers) but with the number of non-zero elements in each row, which can be kept fixed using kernel functions with compact support.
2. Operations performed once per stage. To recall, a new stage is started when the average squared distance between the current values of the $\{y_i\}$ and their values $\{y_i^0\}$ at the start of the current stage is larger than a prescribed fraction of the variance of the latter. At the beginning of each stage, the orthogonal matrix Q_y needs to be computed. The same scaling arguments apply here as for the calculations of Q_z, yielding a total of $\mathrm{O}(n * d_y * m_y)$ operations per stage.
3. Operations performed at each descent step. The main loop iteration requires calculating derivatives of the cost function and of the penalty function. The complexity of the former is $\mathrm{O}(n * d_y * m_y))$. The latter requires calculating the gradient of the matrix $G(y)$ which, when using kernels, involves calculating the kernel in y and its derivatives, with complexity $\mathrm{O}(n * d_y * m_y)$ if the number of kernel centers is fixed. This is followed by a matrix multiplication which is $\mathrm{O}(n)$. So overall each iteration performs $\mathrm{O}(n * d_y * m_y)$ operations. Moreover, these operations decouple among the sample points, making them trivially parallelizable.

One thing not captured by the complexity analysis above is the number of iterations required for convergence. Additionally, in practice one may adopt larger values of n_y and n_z for problems with more complex dependence between x and z. Yet, for a fixed problem, the number of iterations should not depend on the number of data points n, an observation confirmed in our numerical experiments. We verify the algorithm's complexity by plotting the time of the pre-calculations, stage calculations, average descent iteration and total time as n, d_x and d_z vary. In each case, each factor z^l is a normal random variable with mean 0 and variance 0.25 and x is drawn from the z-dependent isotropic gaussian

$$x \sim \mathscr{N}\left(\cos\left(2\pi \sum_{l=1}^{d_z} z^l\right)\vec{1}_{d_x}, 0.05\left[\left[\sin\left(0.1 * \left(\sum_{l=1}^{d_z} z^l + 0.2\right)^2\right) + 0.25\right]^{-1}\right] I_{d_x} \right).$$

Figure 1 displays the data and barycenter for $d_x = d_z = 1$ and figure 2 shows the running times for various values of n, d_x, and d_z. Each data point displayed is the median across 10 trials of the mean time spent in each portion of the algorithm. When the dependence on n is being considered, d_x and d_z are both kept at 1. When either d_x or d_z are being varied, the other is kept at 1 and n is kept at 2000. The first two rows use kernel based functions for both F and G while the third row uses only linear and quadratic functions of y for G, which has the same complexity but with lower constants and thus is orders of magnitude faster. These experiments confirm that the complexity of all stages grow linearly with the number of samples and that the descent steps contribute most heavily to the total run time. Similarly we see that, as predicted, the dimension of x increases linearly the time complexity of the main descent steps and the calculations per stage, and the dimension of z increases linearly the complexity of the pre-calculations.

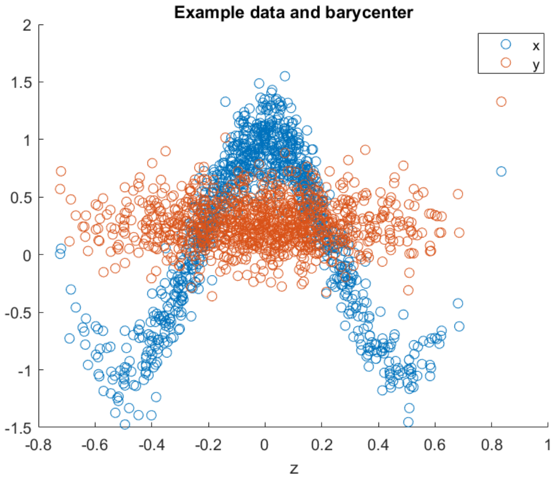


FIG. 1. Data points $\{x_i\}$ and corresponding barycenter samples $\{y_i\}$ with $d_x = d_z = 1$.

7. The optimal transport problem

The methodology developed for the optimal transport barycenter problem can be easily adapted to solve regular [Monge] optimal transport problems. In addition to the importance of the optimal transport problem per se, we will use it for conditional density estimation, as detailed in the next section. Optimal transport problems are simpler than the OTBP, as they involve only two distributions, a source ρ_0 and a target ρ_1:

$$\min_{w=Q(x)} E_{\rho_0}\left[c\left(x,w\right)\right], \quad Q\#\rho_0 = \rho_1. \tag{7.1}$$

For the purpose of relating them to the barycenter problem, we introduce a binary covariate $z \in \{0,1\}$, and think of the source and target distributions as instances of the single conditional distribution $\rho(x|z)$:

$$\rho_0(x) = \rho(x|z=0), \quad \rho_1(x) = \rho(x|z=1).$$

Then the map $y = T(x,z)$ that solves the barycenter problem for $\rho(x|z)$, automatically provides the solution to (7.1) through

$$Q(x) = T^{-1}\left(T(x,0),1\right),$$

a standard result in interpolation displacement [3].

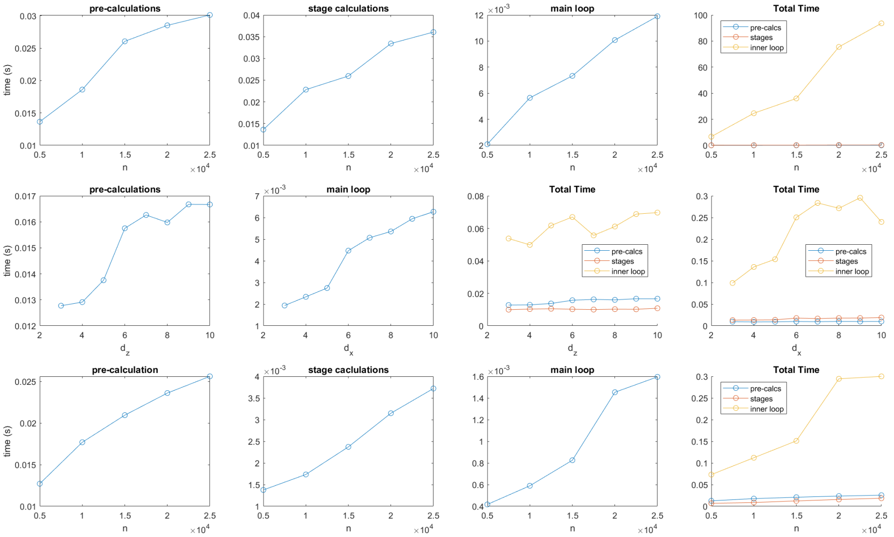


FIG. 2. Median time dependency on number and dimension of data. The top and bottom rows display the run-time dependence of the pre-calculations, calculations per stage, calculations per time-step and total run-time, on the number of samples n, with $d_x = d_z = 1$, the top one using kernels in y, the bottom only linear and quadratic functions. The middle row displays the relevant dependences on d_x and d_z, with the other fixed at 1 and n at 2000.

In the data-driven case, we have n_0 samples $\{x_i^0\}$ from ρ_0 and n_1 samples $\{x_j^1\}$ from ρ_1, for a total of $n = n_0 + n_1$ pairs $\{x_i, z_i\}$ from $\pi(x, z)$. Since our methodology provides all the values $y_i = T(x_i^0, 0)$ and an explicit formula for $T^{-1}(y_i, 1)$, we have direct access to all $Q(x_i^0)$ and, mutatis mutandis, we have also access to its inverse, $Q^{-1}(x_j^1)$. Under the canonical cost, the corresponding formula simplifies to

$$Q(x_i^0) = T\left(x_i^0, 0\right) - \frac{n_0}{n_1}\left(x_i^0 - T\left(x_i^0, 0\right)\right),$$

as follows from the fact that every point in the barycenter is the weighted geometrical c-barycenter of its pre-images [12]. Then, with only two distributions, a point at the barycenter and one of its pre-images suffice to find the other.

The fact that there is only one, binary covariate z simplifies our methodology considerably, since except for an arbitrary sign, there is only one function $f(z)$ with zero mean and norm one:

$$f(z) \propto \begin{cases} \frac{1}{n_0} & \text{for } z = 0 \\ -\frac{1}{n_1} & \text{for } z = 1 \end{cases}.$$

Then the barycenter problem reduces to

$$\min_{y_i} L = \sum_{i=1}^{n} c\left(x_i, y_i\right) + \lambda \left\| f' Q_y(y) \right\|^2, \quad f_i = f(z_i), \tag{7.2}$$

where we have used the fact that, since f is fixed, the maximizing vector b replacing the right principal component of A is proportional to $Q_y{}'f$:

$$\arg\max_{\|b\|=1} f'Q_y b = \frac{Q_y' f}{\|Q_y' f\|} \Rightarrow \sigma \stackrel{\text{def}}{=} \max_{\|b\|=1} f'Q_y b = \|f'Q_y\|.$$

Other than the simplifying facts that we do not need to update a and b and that the matrix Q_z consists of a single column, the procedure to solve (7.2) follows the same steps as the one for the full barycenter problem (4.3).

In order to bypass the barycenter μ in the procedure above, finding a map $Q(x) = T(x,0)$ that pushes forward ρ_0 to ρ_1 directly, use the same objective function L in (7.2), but minimize it only over the y_i with corresponding $z_i = 0$, i.e. over $T(x,0)$, leaving the remaining y_i fixed at x, i.e. setting $T(x,1) = x$. This enforces the condition that $T(x,0)\#\rho_0 = \rho_1$, since the final

$$y_i^* = \begin{cases} T(x_i,0) & \text{for } z_i = 0 \\ x_i & \text{for } z_i = 1 \end{cases}$$

must be independent of z. This procedure, while lacking the symmetry of the prior one with respect to ρ_0 and ρ_1, is far more straightforward. In particular, it is very well-suited for density estimation.

8. Conditional density estimation

Our methodology simulates $\rho(x|z)$, by producing n samples $\{x_i^*\}$ from $\rho(x|z^*)$ for any target z^*. Simulation is at the core of many applications, but others, such as Bayesian inference, require the evaluation of $\rho(x|z)$ for arbitrary values of x and z. The fact that typically there is none or at most one observation available for any target value z makes estimating $\rho(x|z)$ directly from the data $\{x_i, z_i\}$ challenging. A slight extension of our procedure produces such conditional density estimation.

Regular –as opposed to conditional– density estimation can be obtained through optimal transport as follows. Given n samples $\{x_i\}$ drawn from the unknown distribution $\rho(x)$ that we seek to estimate, select a target distribution $\mu(y)$ that one can easily both evaluate and sample, such as a Gaussian, and find the optimal map $Q(x)$ pushing forward ρ to μ. Then

$$\rho(x) = |\det(\nabla_x Q)|\ \mu(Q(x)), \quad \rho(X(y)) = \frac{1}{|\det(\nabla_y X)|}\ \mu(y), \quad X = Q^{-1},$$

so for any x,

$$\rho(x) = \frac{1}{\left|\det\left(\nabla_y X(y)\big|_{y=Q(x)}\right)\right|}\ \mu(Q(x)).$$

In our procedure, $Q(x) = T(x,0)$, and $\nabla_y X(y)$ is known from (5.2).

If the density $\rho(x)$ is sought for values of x different from the $\{x_i\}$, one can carry these values through the procedure to their final $y = Q(x)$ as passive tracers that do not affect L in (7.2). Alternatively, one can solve the reciprocal optimal transport problem from μ to ρ, and then write

$$\rho(x) = |\det(\nabla_x Y(x))|\ \mu(Y(x))$$

for any value of x sought.

This density estimation procedure requires selecting a target measure $\mu = \rho_1$. We can either adopt a fixed target, such as a standard Gaussian, or to adapt it to the data, using for instance a Gaussian with the same mean and covariance matrix as the data or a Gaussian mixture fitted to the data through Expectation Maximization. The advantage of such more tailored approaches is that the corresponding optimal transport problem becomes easier, since even the trivial map $Q(x) = x$ provides a regular parametric density estimation.

We can apply this procedure to conditional density estimation, i.e. estimate $\rho(x|z)$ from n samples $\{x_i, z_i\}$ in at least two distinct ways:

1. Obtain n samples $\{x_i^*\}$ from $\rho(x|z^*)$ and apply density estimation to these directly.
2. Estimate the density $\mu(y)$ of the barycenter, and then compute

$$\rho\left(X(y,z)|z\right) = \frac{1}{|\nabla_y X(y,z)|}\, \mu(y),$$

with $X(y,z)$ given by (5.1) and $\nabla_y X(y,z)$ by (5.2).

One would choose the first approach when seeking $\rho(x|z)$ for many values of x and only a handful of values of z, and the second approach when exploring the dependence of the conditional density on z, as beholds for instance Bayesian inference.

9. Numerical examples

We illustrate the methodology through numerical examples, using synthetic data first and then two real data applications: uncovering hidden patterns in the atmospheric temperature at ground level and forecasting ocean states.

9.1. *A Gaussian distribution with z-dependent mean and variance*

As a first example, we draw 1500 independent samples from the distribution

$$x \sim N\left(\mu(z), \sigma^2(z)\right), \quad \mu(z) = \cos(2\pi z_1) + \sin(\pi z_2), \quad \sigma(z) = 0.2\sqrt{(1-2z_1)(1-2z_2)},$$

with $z = (z_1, z_2)$ uniformly distributed in the square $-\frac{1}{2} \leq z_{1,2} \leq \frac{1}{2}$. The data is displayed on the top left panel of figure 3. Since for each value of z the distribution for x is Gaussian, it can be fully captured using the two-dimensional test function space $\mathscr{G}(y)$ spanned by the functions y and y^2, while keeping for $\mathscr{F}(z)$ a general adaptive space based on kernels. We display the results of the run in figure 3 through the corresponding $\{y_i\}$, whose z-independent distribution is a Gaussian, as beholds the barycenter of a set of Gaussians, and the simulation and estimation of $\rho(x|z_*)$ for two selected values of z_*, with the true underlying distribution also drawn for comparison.

9.2. *Two z-dependent Gaussian mixtures*

In order to consider non-Gaussian examples –more precisely, examples where the dependence of x on z cannot be reduced to a z-dependent translation and scaling– we perform first a one-dimensional

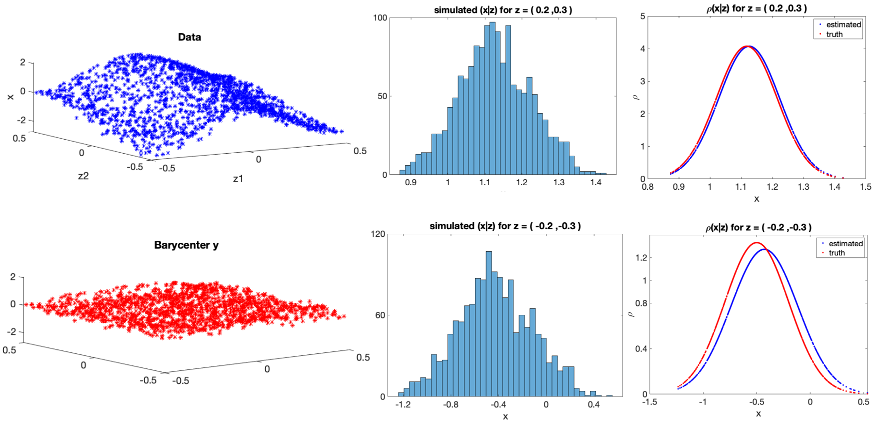


FIG. 3. First Example. The leftmost column displays the samples $\{x_i\}$ in the top row and the corresponding $\{y_i\}$ in the barycenter in the bottom row. The middle and rightmost columns show the simulated samples $\{x_i^*\}$ and the estimated versus the true density $\rho(x|z_*)$ for $z_* = (0.2, 0.3)$ and $z_* = (-0.2, -0.3)$.

experiment, drawing 1500 samples from the z-dependent Gaussian mixture

$$x \sim \sum_{k=1}^{2} \gamma_k \, N\left(\mu_k(z), \sigma_k^2(z)\right), \quad \gamma_1 = \gamma_2 = \frac{1}{2},$$

$$\mu_1(z) = 3 + 2z, \quad \sigma_1^2(z) = \frac{1}{2}e^z, \quad \mu_2(z) = \frac{z}{2} - z^2, \quad \sigma_2^2(z) = 0.25 - 0.1z, \quad z \sim \mathrm{U}([-2.5, 2.5]).$$

The datapoints are displayed on the top left panel of figure 4.

This example requires a test function space $\mathscr{G}(y)$ that goes beyond linear and quadratic functions. We performed four successive barycenter problems, the first with just linear and quadratic $\mathscr{G}(y)$, the rest with adaptive kernels, with the bandwidths of the kernels for both $\mathscr{G}(y)$ and $\mathscr{F}(z)$ adopted smaller for each successive run. The results, displayed in figure 4, show how the simulated samples and conditional density estimation recover the original z-dependent Gaussian mixture.

We extend this example to the two-dimensional Gaussian mixture

$$x \sim \sum_{k=1}^{2} \gamma_k \, N\left(\mu_k(z), \Sigma_k(z)\right), \quad \gamma_1 = \gamma_2 = \frac{1}{2}, \quad z \sim \mathrm{U}([-2.5, 2.5]),$$

$$\mu_1(z) = \begin{pmatrix} 3+2z \\ 2+z \end{pmatrix}, \quad \Sigma_1(z) = \begin{pmatrix} \frac{1}{2}e^z & 0 \\ 0 & 0.5 \end{pmatrix}, \quad \mu_2(z) = \begin{pmatrix} \frac{z}{2} - z^2 \\ -3 \end{pmatrix}, \quad \Sigma_2(z) = \begin{pmatrix} 0.25 - 0.1z & 0 \\ 0 & 0.25 + 0.1z \end{pmatrix}.$$

The data points and the barycenter are visualized in figure 5 and the results of the procedure are displayed in figure 6.

9.3. *An example of Bayesian inference*

This section illustrates model-free Bayesian inference using the OTBP. A different use of push-forward maps for Bayesian inference [37] pushes forward the prior to the posterior measure. To demonstrate our

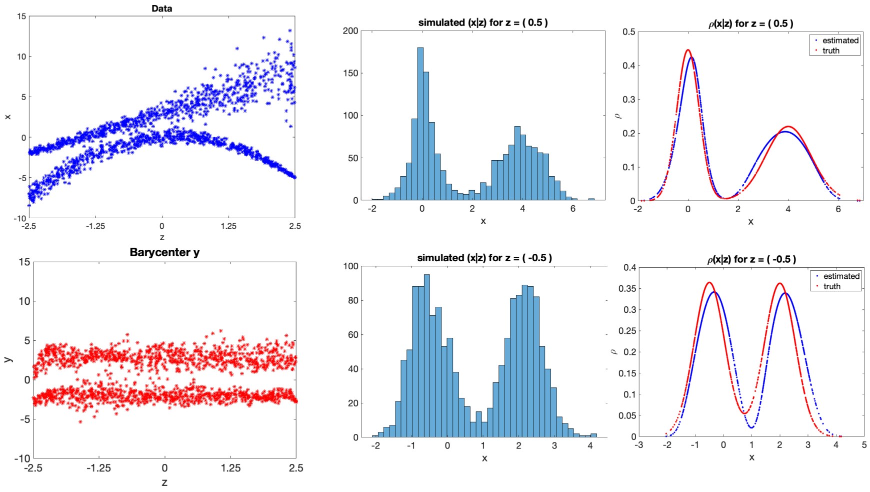


FIG. 4. One-dimensional Gaussian mixture. The leftmost column displays the data samples in the top row and the barycenter in the bottom row. The middle and rightmost columns show the simulated samples and the estimated versus the true density for $z_* = 0.5$ and $z_* = -0.5$.

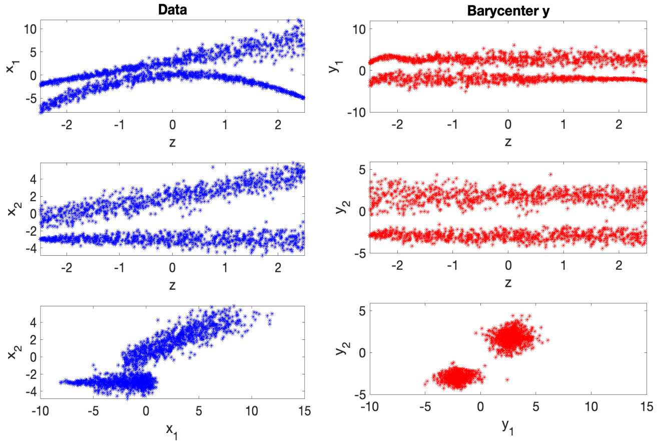


FIG. 5. A Two-dimensional Gaussian mixture. The left panel displays the data points, while the right panel shows the equivalent plots for the barycenter.

approach, we draw samples from the distribution

$$\rho(x|z) = N(z^2 - 2, \sigma^2), \quad \sigma = 0.5, \quad z \sim \mathrm{U}([-2,2]).$$

The left panel of figure 7 displays the data $\{x_i\}$ and the discovered $\{y_i\}$ as functions of the corresponding $\{z_i\}$. From these, we can directly infer the distribution of z given an observation x:

$$\gamma(z|x) = \frac{\rho(x|z)}{\rho(x)} \gamma_{pr}(z) \propto \rho(x|z) \cdot \gamma_{pr}(z),$$

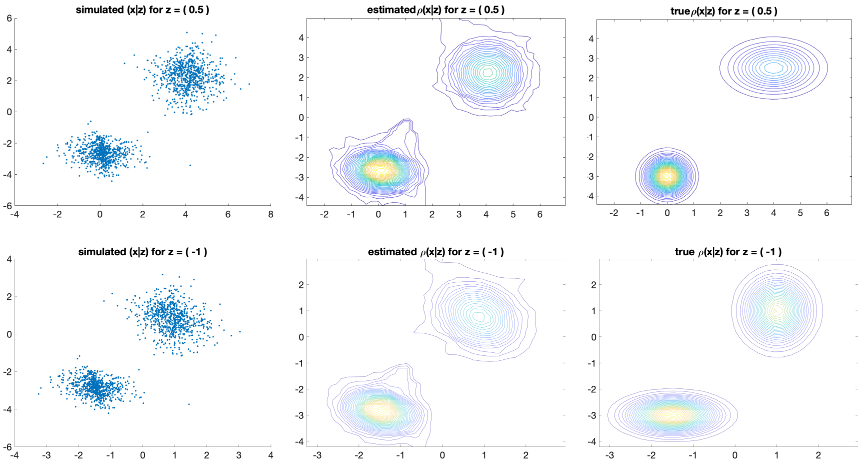


FIG. 6. Two-dimensional Gaussian mixture. The top and bottom rows show the simulated samples, the estimated conditional density, and the true density for $z_* = 0.5$ and $z_* = -1$ respectively.

where $\rho(x|z)$ is the not the distribution we know in closed form but the one inferred from the data through the OTBP. We have adopted as a natural default prior γ_{pr} the distribution underlying the observed $\{z_i\}$. The results for two values of x are displayed on the middle and right panels of Figure 7, overlapped with the exact answer. They succeed in capturing the transition from unimodal to bimodal distributions corresponding to the parabolic dependence of the conditional mean of x on z.

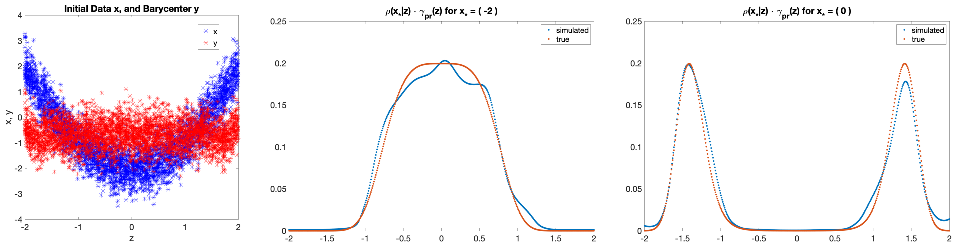


FIG. 7. Original data, barycenter and simulated versus true posterior density $\gamma(z|x^*)$ evaluated at $x^* = -2$ and $x^* = 0$.

9.4. *Online model estimation in Ornstein–Uhlenbeck processes and Lotka-Volterra pray-predator models with observational noise*

We consider next the online estimation of parameters, a key component of data assimilation. Given successive samples from a time series X^n drawn from some transitional distribution $\rho(X^{n+1}|X^n, Z^n, \alpha)$ depending on known and unknown parameters Z^n and α respectively, and assuming a prior distribution $\gamma^0(\alpha)$ for the latter, we seek to successively improve on these priors as new observations arrive, using Bayes rule:

$$\gamma^{n+1}(\alpha) \propto \rho(X^{n+1}|X^n, Z^n, \alpha) \cdot \gamma^n(\alpha),$$

with the proportionality constant determined by the normalizing condition that $\int d\gamma(\alpha) = 1$. In the conventional setting, the conditional distribution ρ is known except for the parameters α. We can extend this framework to situations where $\rho(X^{n+1}|X^n, Z^n, \alpha)$ itself is only known from partial observations of time series under different values of α and Z. In a medical setting for instance, X may represent glucose concentration in the bloodstream, Z^n the caloric intake at time t_n, and α a patient's parameter that may only be determined after treatment. Having observed in the past a number of patients under different diets and having determined their corresponding parameters α, we can use this for the online estimation of α for a patient currently under treatment.

For a first simple example, consider the time-discretized 1D Ornstein–Uhlenbeck process

$$X^{n+1} = (1-\alpha)X^n + \beta + \sigma W^n, \quad W^n \sim N(0,1),$$

where $\alpha \in (0,1)$ is an unknown model parameter and $\beta = \sigma = 0.5$ are fixed drift and noise levels. Our goal is to learn the model from a set of training data pairs $(X^{n+1}_{\text{train}}; X^n_{\text{train}}, \alpha^n_{\text{train}})$ and use the model learned to estimate α online from a testing series, while making increasingly more accurate forecasts. For the training data, we draw α^n_{train} from a beta distribution $B(2,2)$ over $(0,1)$, which we also adopt as prior $\gamma^0(\alpha)$, and X^n_{train} from the uniform distribution $\mathrm{U}([a,b])$.

We carry out experiments with two different parameter values, $\alpha = 0.2, 0.8$, with the corresponding test data displayed in Figure 8. The results in Figure 9 demonstrate that the posterior densities converge to delta functions centered around the corresponding true parameters.

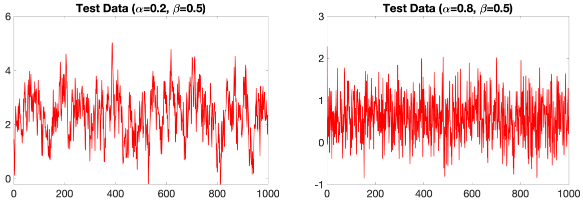


FIG. 8. Time-discretized Ornstein–Uhlenbeck process. Testing time series for $\alpha = 0.2, 0.8$.

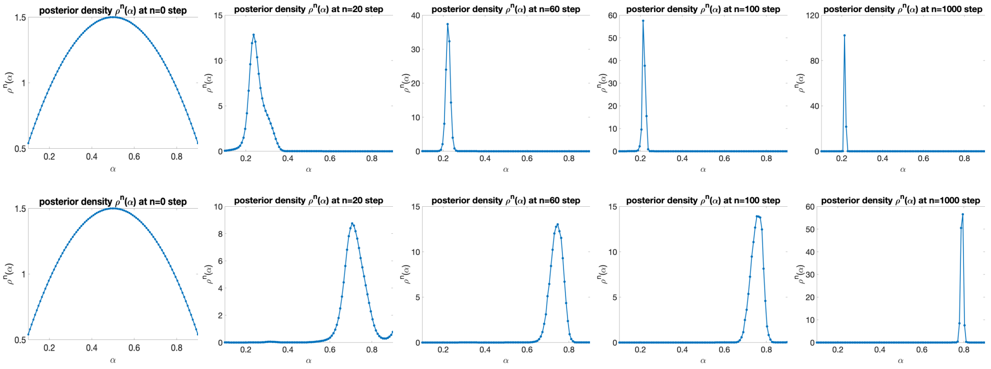


FIG. 9. Estimated posterior densities as the number of time steps grows for $\alpha = 0.2$ in the top row and $\alpha = 0.8$ in the bottom row.

In order to apply the procedure to a more complex scenario, we consider next the Lotka-Volterra predator-prey model

$$\begin{aligned}\frac{dx_1}{dt} &= \alpha x_1 - \beta x_1 x_2,\\ \frac{dx_2}{dt} &= -\gamma x_2 + \delta x_1 x_2,\end{aligned}$$

with data observed at irregular discrete times with non-uniform time intervals $\Delta t^n \sim \mathrm{U}[0.5, 1.5]$ and corrupted by noise. The time Δt^n here plays the role of the known covariate Z^n, so the Barycenter problem needs to resolve the data dependence on this additional factor. The data are generated through the explicit trapezoidal numerical scheme but with a much smaller $\tilde{\Delta t}$, to accurately solve the system of ODEs. Gaussian noise with amplitude ε is added after simulating the time series to represent noisy observations.

We adopt as test data a simulation with parameters $\alpha = 0.3, \beta = 0.9, \gamma = 0.5, \delta = 0.4$ and $\varepsilon = 0.1$, yielding the periodic results displayed in Figure 10. For a first experiment, we take α as the only unknown parameter. Again, random training data pairs $(X^{n+1}_{\text{train}}; X^n_{\text{train}}, \alpha^n_{\text{train}}, \Delta t^n)$ are generated, drawing X^n_{train} from a uniform distribution, α^n_{train} from the beta distribution $B(2,2)$, and deriving the corresponding X^{n+1}_{train} from the model with additive noise of level ε.

As before, we learn the conditional density $\rho(X^{n+1}|X^n, \alpha, \Delta t^n)$ by solving the barycenter problem for the training data. Then we apply Bayes rule online to the testing data, updating at each step

$$\rho^{n+1}(\alpha) \propto \rho(X^{n+1}|X^n, \alpha, \Delta t^n) \cdot \rho^n(\alpha).$$

The results are shown in the top panel of Figure 11.

We consider next a situation where two parameters, α and γ, are unknown, so the training data consists of quintuples $(X^{n+1}_{\text{train}}; X^n_{\text{train}}, \alpha^n_{\text{train}}, \gamma^n_{\text{train}}, \Delta t^n)$, and the joint posterior density should be estimated through

$$\rho^{n+1}(\alpha, \gamma) \propto \rho(X^{n+1}|X^n, \alpha, \gamma, \Delta t^n) \cdot \rho^n(\alpha, \gamma),$$

with joint Gaussian prior $\rho^0(\alpha,\gamma) = N\left(\begin{pmatrix}0.5\\0.5\end{pmatrix}, \begin{pmatrix}0.2 & 0\\ 0 & 0.2\end{pmatrix}\right)$. The results are displayed in the bottom panel of Figure 11. We see how in both cases the estimation converges to a delta function centered at the right underlying value of the parameters.

9.5. *Uncovering a hidden signal*

The solution to the barycenter problem helps uncover a hidden signal w that, together with the known factors z, fully explain the outcome variable x. In order to demonstrate this through examples, rather than simulating a distribution $\rho(x|z)$, we propose a function

$$x = \phi(z, w), \quad z \sim \gamma(z), \quad w \sim \nu(w)$$

where w, playing the role of noise in the distribution, is a hidden cause of variability in x.

Recall from Theorem 3 that the solution $y = T(x,z)$ of the barycenter problem is a proxy for the variable w, with $\phi(z,w) \to X(y,z)$. Moreover, y is related to any "true" hidden variable w through a possibly z-dependent function

$$y = Y_z(w),$$

which is invertible if w is identifiable, i.e. if a single value of x cannot originate from a single z and two different values of w.

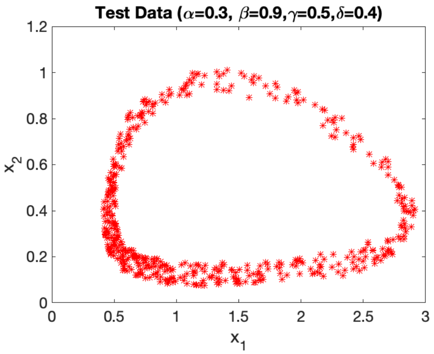


FIG. 10. Lotka-Volterra model: Testing time series for $\alpha = 0.3, \beta = 0.9, \gamma = 0.5, \delta = 0.4$.

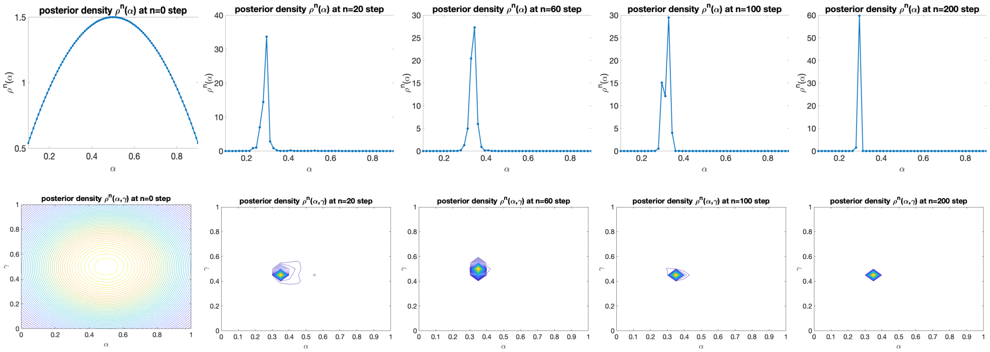


FIG. 11. Estimated posterior densities as the number of time steps grows for the single unknown $\alpha = 0.3$ in the top row, and for two unknowns $\alpha = 0.3, \gamma = 0.5$ in the bottom row.

Figure 12 presents three synthetic examples in order to illustrate the different kind of dependence between y and w typically observed in applications. The panels of each row are relative to the different synthetic examples described below. The panels on the left column display the $\{x_i\}$ and corresponding $\{y_i\}$ in terms of the $\{z_i\}$, the middle column displays $y(z)$ again, colored according to the corresponding value of w, and the right column displays $y(w)$, colored according to z. In the first row, $z \sim \mathrm{U}[0.25, 1]$, $w \sim \mathrm{U}[-1, 1]$ and $x = \phi(z, w) = zw^3$ (we exclude values of z near 0 because $\forall w\ \phi(0, w) = 0$, i.e. $\rho(x|0)$ does not vanish on small sets.) In this example $Y_z(w) = Y(w)$ does not depend on z and $Y(w)$ is invertible. In the second row $x = zw^2$ under the same distributions for z and w. We still have that $Y_z(w)$ does not depend on z but now $Y(w)$ is not globally invertible, a reflection of the fact that the sign of w is not identifiable, since $\forall z$ and $\forall w$ we have that $\phi(z, -w) = \phi(z, w)$. In the third row, $w \sim \mathrm{U}[0, 1]$, $z \sim \mathrm{U}[(-1, -0.25) \cup (0.25, 1)]$ and $x = zw$, for which $Y_z(w)$ depends on the sign of z (we leave it as a challenge to the interested reader to uncover why this is so.)

The analysis of the barycenter underlying the points $\{y_i\}$ may at first seem similar to residual analysis, whereby the difference between actual and predicted values is further analyzed to assess model adequacy and improve its predictive power [38, 39]. Both procedures aim to remove variability in the data x attributed to the cofactors z, yet while residual analysis only removes $\bar{x}(z)$, the conditional expected value of x, the barycenter does this at the level of the full probability distributions $\rho(x|z)$

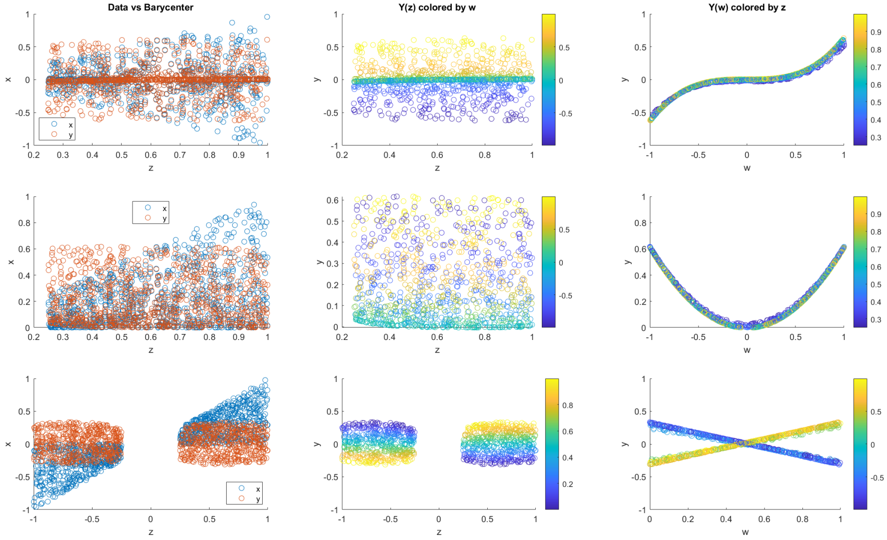


FIG. 12. Three examples with different relations $Y_z(w)$ between w and y: one-to-one on the first row, two-to-one on the second and z-dependent on the third.

underlying the data. Consider the example in the first row of Figure 12, where the data are generated according to $x = zw^3$ with z and w drawn uniformly. The distribution of the residuals (obtained by subtracting the regression of x vs z from the actual value of x) would, in this case, be identical to the original distribution underlaying the x, providing no new useful information. By contrast, the points y_i in the barycenter represent the full variability of x not explainable by z. If the y's can be related to known factors w, this can be used to improve the model for x, for instance by regressing y against these factors and then using both z and the reconstructed y to predict x. Better still, instead of regression, the barycenter problem can be used once again to simulate $\rho(y|w)$.

When the hidden signal w is lower dimensional than x, it follows that y must lie in a lower dimensional manifold of $\mathscr{X}$. Consider an example where $z \in \mathbb{R}^2$ with $z \sim \mathcal{N}(0,I)$, $w \in \mathbb{R}$ with $w \sim \mathcal{N}(0,1)$ and $x = [8z^1z^2 + 2w, 2z^1 + 8z^2 + 3w]$. As shown in Figure 13, after solving the barycenter problem, the resulting y lies on a 1-D manifold that is parametrized (and therefore completely explained) by w, which is not generally the case for residual analysis.

9.6. *Hidden patterns in ground-level atmospheric temperature*

Switching to applications to real data, we consider next the hourly ground-level temperature in Ithaca, NY from 2007 to 2023. The data, available from National Oceanic and Atmospheric Administration, is displayed in the top panel of Figure 16. We will use the OTBP to investigate the dependence of this temperature on the diurnal and seasonal cycles, and to uncover hidden signals at the synoptic weather and multi-year scales.

We first solve the OTBP for $\rho(x|z_1)$, where x is the hourly temperature in Ithaca and $z_1 \in [0, 365.25]$ is the day of the year, a continuous, periodic factor. Panel (a) of Figure 14 displays the corresponding

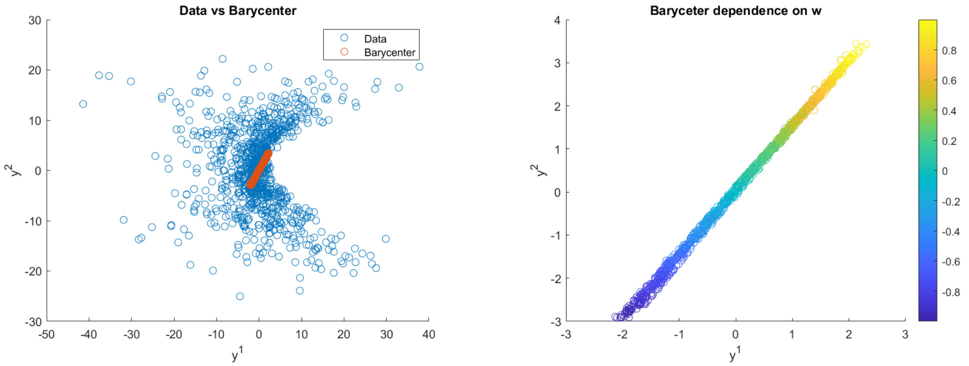
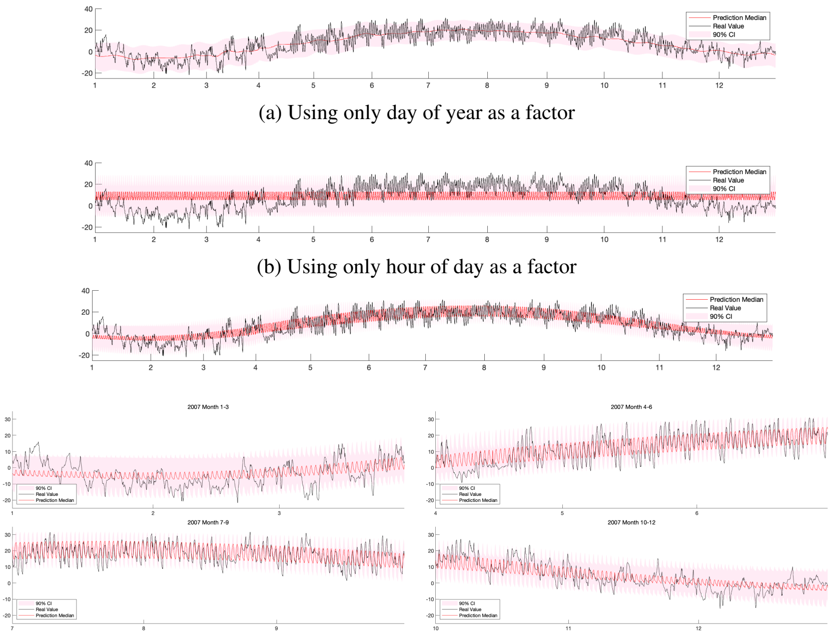


FIG. 13. Two-dimensional y dependence on a one-dimensional w (denoted with a colorbar)

day-dependent median value of the simulated temperature $X(:,z_1)$ and the corresponding conditional 90% confidence interval, capturing seasonal effects, superimposed for reference on the true observed temperatures for the year 2007. We use the conditional median and confidence intervals rather than the conditional mean and standard deviation of x because they are more robust statistics and they are also much cheaper to compute: while computing the mean involves averaging $X(y,z)$ over all $\{y_i\}$ for each value of z, and similarly for the variance, the monotonicity of X implies that the conditional median of x is just $X(\bar{y},z)$, where $\bar{y}$ is the median of the $\{y_i\}$), and similarly for confidence intervals, since conditional percentiles of $\rho(x|z)$ translate directly from the corresponding percentiles in y. Next we consider instead $\rho(x|z_2)$, where $z_2 \in [0,24]$ is the time of the day, another continuous and periodic factor, displaying in panel (b) the simulated median diurnal cycle and corresponding confidence interval together with the true x for 2007. Then we combine the two factors and consider $\rho(x|z_1,z_2)$, with results displayed on panel (c) both for the the full year 2007 at once and for zoomed-in versions for each season. One may notice in all panels how the 90% confidence interval depends on z, often adopting asymmetric shapes around the median and displaying for instance interesting contrasts between day and night. This is one manifestation of the power of having captured the full conditional distribution $\rho(x|z)$, as opposed to just a few statistics, such as the conditional mean value computed in regression.

We can see how the diurnal cycle changes over the year not only in mean but also in amplitude and shape. This is seen more clearly in the left panel of Figure 15, displaying the median diurnal cycle for four specific days of the year, corresponding to the solstices and equinoxes. We can see in detail, for instance, how the Winter Solstice day is colder, shorter and has smaller day/night contrast than its summer counterpart, and how the day at the Spring Equinox, despite having exactly the same duration as the one at the Fall Equinox, is much colder, has smaller amplitude and a slightly different shape. The right panel of Figure 15 similarly shows how the median seasonal cycle depends on the time of the day at which it is considered.

We switch next to consider the variability of x not explained by (z_1,z_2), as captured in y. In order to analyze both synoptic weather and short-term multi-year variability, we introduce two new time factors, z_3 and z_4, built by rescaling time using different scales. In other words, both z_3 and z_4 consist just of the time t (measured in hours), but the bandwidths used for the corresponding kernels are of the order of 30 and 3 days, respectively (we set these scales through the parameter γ_z defined in the appendix.) As we did for $z_{1,2}$, we remove the variability in y attributable to z_3 and z_4 separately, then together. Figures 16 and 17 show the results of introducing these new factors. Panels (a) and (b) of



(a) Using only day of year as a factor

(b) Using only hour of day as a factor

(c) Using the two periodic time factors together

FIG. 14. Median temperature and 90% confidence interval as a function of day of year, time of day and both, displayed over the true temperature for 2007.

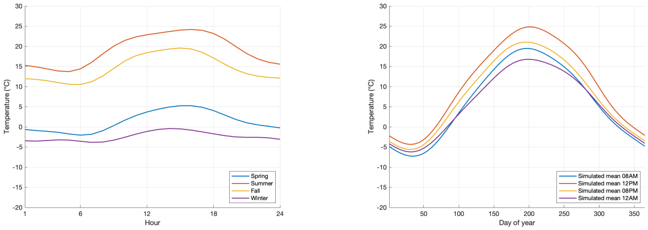
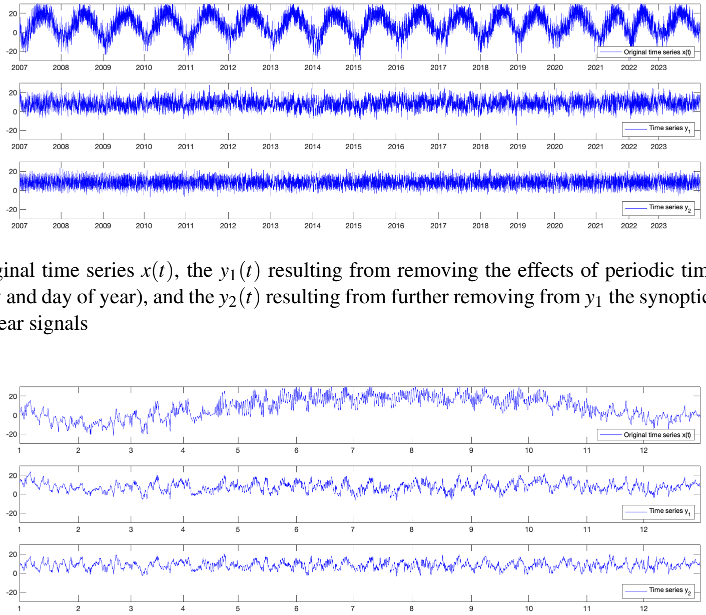


FIG. 15. Median diurnal cycle and seasonal cycle of the temperature in Ithaca, NY, displayed for four days of the year (one per season) and four hours of the day respectively.

figure 16 display the original time series $x(t)$, the signal $y_1(t)$ resulting from removing $z_{1,2}$ from x and the signal $y_2(t)$ resulting from removing z_3 and z_4 from y_1. The removal of variability is reflected in the signals' decreasing variance, from 113.71 for x, through 24.71 for y_1, to 17.10 for y_2. Beyond the decreasing variance, one can observe the further explanation of variability in the fact that y_2 is much more homogeneous in time than y_1, which has a clear inhomogeneity associated with the synoptic weather signal (The fact that similarly y_1 does not display the time dependence on the seasonal cycle present on x is far more obvious to the eye.) Panel (a) of figure 17 displays the median temperature dependence on z_3, a multi-year signal, and panel (b) the dependence on z_4, corresponding to synoptic weather, over the year 2007. The median temperature and 90% confidence interval determined by z_3 and z_4 together is displayed in panel (c) and zoomed-in over 2007 in panel (d). A climate scientist looking at these reconstructions may not only confirm that the method has captured the right scales (a roughly 2-4 year scale for the multiyear signal and around 15 days for the synoptic weather) but also detect individual signals, such as in panel (a) a signal resulting from the El Niño years 2007, 2010 and 2016, and in panel (b) a signal from the North American heat wave of 2007, which may have contributed to the elevated temperature of Ithaca during the late summer and early autumn. Finally, panel (e) displays the full $z_{1,2,3,4}$-dependent conditional median $\bar{x}(z(t))$ and 90% confidence interval over 2007. We can see in this plot not only how well the reconstruction has captured the dependence of temperature on time, season and the synoptic and multi-year time scales, but also how it has not captured (by construction) weather signals shorter than a week long. These could of course be captured by another factor z_5 with shorter bandwidths, but doing this would take us too far afield in the context of the current article. Notice that these shorter scales are nonetheless represented as noise in the 90 percentile, a general property of the OTBP methodology: as new factors z^l are introduced, these explain away part of the variability previously present in the conditional distribution $\rho(x|z)$.



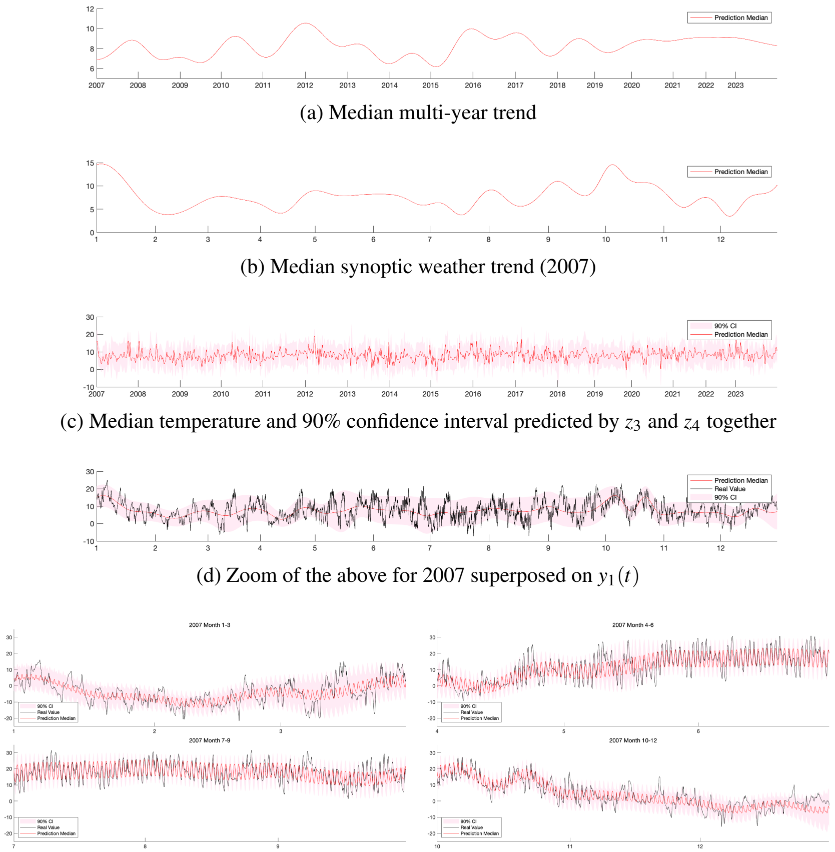
(a) The original time series $x(t)$, the $y_1(t)$ resulting from removing the effects of periodic time factors (time of day and day of year), and the $y_2(t)$ resulting from further removing from y_1 the synoptic weather and multi-year signals

(b) A zoom-in version of the previous figure, restricted to 2007

FIG. 16. Effect of removing from the temperature x, in succession, the effect of periodic time factors (time of day and day of year) and of time itself at the synoptic and multiyear scales.

9.7. *Forecasting of global ocean states*

We further illustrate the methodology, using it to forecast six months ahead the global sea surface temperature (SST). The data (available at Met Office Hadley Centre observations) consists of monthly values of the SST from 1870 to 2024, over a global 1×1 latitude-longitude grid. The resulting dataset $T^l_{i,j}$ has a dimension of $180 \times 360 \times 1860$, corresponding to latitude (indexed by i), longitude (j) and



(a) Median multi-year trend

(b) Median synoptic weather trend (2007)

(c) Median temperature and 90% confidence interval predicted by z_3 and z_4 together

(d) Zoom of the above for 2007 superposed on $y_1(t)$

(e) Median temperature and 90% confidence interval as a functions of day of year, time of day, multi-year trend and the synoptic weather trend, displayed over the true temperature for 2007

FIG. 17. Reconstruction of the conditional median temperature $\bar{x}(z)$ and 90% confidence interval as functions of different combinations of the factors $z_{1,2,3,4}$.

time (l in months). The goal is to use historical observations to predict the global SST 6 months into the future.

In order to extract a lower dimensional time signal from the data, we apply a standard pre-processing to the whole dataset:

- Filter out spatial grid points that either lie over land (where SST is undefined) or contain missing data, resulting in 31,094 valid spatial grid points; all subsequent analysis, including EOF computation, is restricted to this filtered spatial domain;
- De-trend by fitting a linear function of the temporal variable;
- Explain away the seasonal cycle by removing the mean value at each day of the year from each point on the spatial grid, reducing $T_{i,j}^l$ to the anomaly signal $A_{i,j}^l = A(x_{i,j}, t_l)$;

- Obtain through principal component analysis the first K empirical orthogonal functions of the data [40],
$$A^l_{i,j} \approx \sum_{k=1}^{K} \sigma_k \mathrm{C}^k_l \, \mathrm{EOF}^k_{i,j},$$
where $\mathrm{EOF}^k_{i,j} = \mathrm{EOF}^k(x_{ij})$ are static, geographically dependent components of the SST profiles, and the $\mathrm{C}^k_l = \mathrm{C}^k(t_l)$ capture their magnitude at time t_l. The components are sorted according to σ_k, proportional to the fraction of the variance of A that they explain.

(We could replace the standard pre-processing by a far more informative one based on the OTBP methodology, but this would take us too far afield in the current article, it is currently investigated as part of a general methodology for the analysis of high-dimensional time series.) It is known [41] that the first EOF component correlates strongly with ENSO events. Figure 18 depicts the first 3 $\mathrm{EOF}^k(x)$ as well as their temporal coefficients $\mathrm{C}^{1,2,3}(t)$. The prediction task then reduces to forecasting the coefficients C^k from their lagged observations $\mathrm{C}^l(t-\Delta t)$, with $\Delta t = 6$ months.

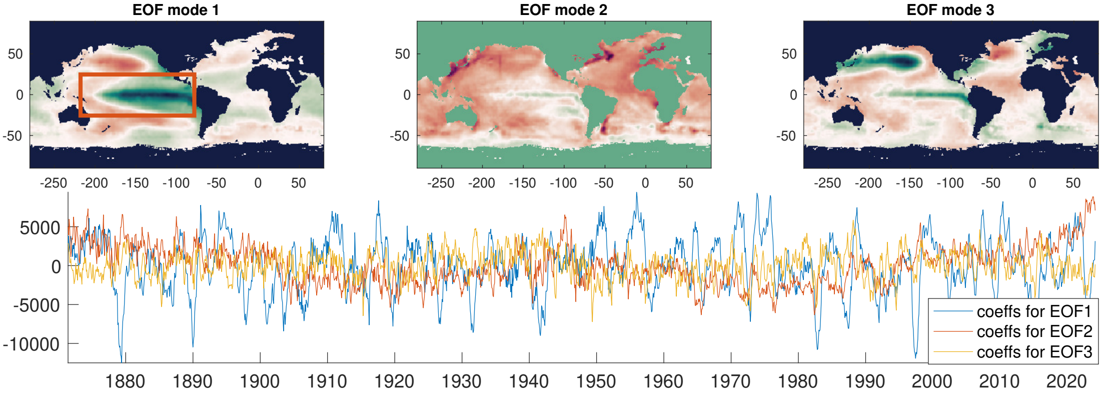
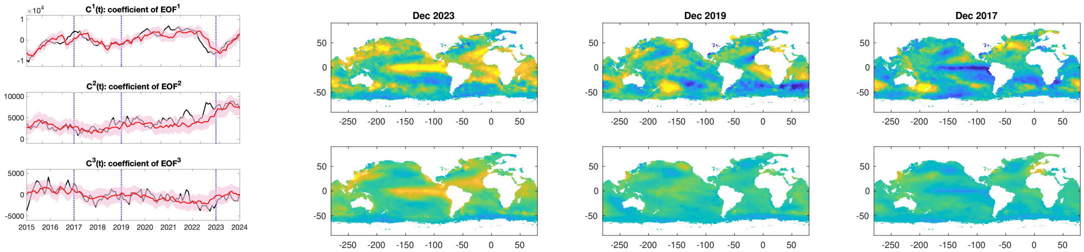


FIG. 18. Upper panels: $EOF^{1,2,3}$. The red box in EOF^1 (left most panel) indicates the region of El Niño events. Lower panel: $C^{1,2,3}_l$ as a function of time t_l.

After approximating the original time series by its first $K = 50$ EOFs, we split the data into in-sample (145 years, 1870-2014) and out-of-sample (10 years, 2015-2024) sets. We apply the procedure to each component C^k independently, so x is a one dimensional outcome. The covariate space z is multidimensional, consisting of two types of factors: (1) the $\mathrm{C}^l(t-\Delta t)$ with lagged correlation with C^k of absolute value greater than 0.1, and (2) time-lagged observations of the same component C^k, with lags of 6, 12, 24, and 36 months.

We restrict the family of functions $\mathscr{G}$ and $\mathscr{F}$ in Section 6 to include linear and quadratic terms in y and kernels in z space respectively. We cross-validate over the optimal z-space kernel bandwidth parameter γ_z, defined in the appendix. For each value of γ_z among 40 points uniformly distributed in $[0.2, 20]$, we find the barycenter and compute the ℓ_2 norm of the difference between true and predicted mean, evaluated over the out-of-sample data. The results shown in Figure 19 correspond to the optimal bandwidth that minimizes the norm of difference.

Figure 19 depicts the prediction of $A(x,t)$, focusing on SST anomalies for December of 2023, 2019, and 2017, corresponding to El Niño, neutral, and La Niña events. The predictions performance can be visually inspected in two different spaces: through the prediction of each EOF component C^k (panel (a)) and of $A(x,t)$ for specific times t (panel (b)). For the second option, we truncated the prediction to the first 50 components, which explain above 83% of the variability of the original SST anomaly. In both spaces, our method always recovers consistent anomalies both globally and locally within the El Niño region.



(a) The first 3 coefficients $C^k(t)$ (black) and our prediction (red for prediction mean, and pink for one standard deviation away from prediction mean).

(b) Anomaly at 3 dates (from left to right: strong El Niño, neutral, strong La Niña), color limit: $\pm 1.5°C$. The top row displays the ground truth, while the bottom row contains our prediction, truncated to the top 50 components.

FIG. 19. Forecast 6 months ahead of the global Sea Surface Temperature: (a) visualization of the signal's first 3 time components, (b) global anomaly.

10. Summary and discussion

We have developed in this article an efficient methodology for solving the sample-based Monge optimal transport barycenter problem, which takes as input n observed sample pairs $\{x_i, z_i\}$ drawn from an unknown joint distribution $\pi(x,z)$, and produces as output n associated samples $\{y_i\}$ from the barycenter μ of the conditional distributions $\rho(x|z)$ under $\gamma(z) = \pi(\mathscr{X}, z)$. In addition, it produces n samples $\{x_i^*\}$ drawn from the estimated $\rho(:|z_*)$ for any proposed target value z_* of the covariates z and it estimates $\rho(x|z)$, instrumental for model-free Bayesian inference. A corollary extends the procedure to solve the regular [Monge] optimal transport problem.

Central to the methodology and to most of its applications is a formulation of the OTBP not in terms of the barycenter μ itself but of the underlying random variable $y = T(x,z)$, which must be statistically independent of the factors z. A test-based formulation of independence through the uncorrelation between all functions $\{f(z), g(y)\}$ within suitable functional spaces $\{\mathscr{F}, \mathscr{G}\}$ provides an adversarial formulation of the OTBP. Since the best adversarial functions f and g can be found exactly in terms of the first principal components of a matrix $A(y)$, the problem reduces to a single minimization over the map T. Solving this problem through gradient descent over y yields a flow in phase space that transports each y_i from x_i to $T(x_i, z_i)$. Fortiutously, the resulting map T can be inverted in closed form, which facilitates much both the simulation and the estimation of $\rho(x|z)$. A byproduct of this closed-form inversion is the extraction of factors $\{f^k\}(z)$ that encode the dependence of x on z.

Numerical examples illustrate the applicability of the Monge OTBP and the effectiveness of the methodology proposed. These examples range from synthetic demonstrations of density estimation and simulation, model-free Bayesian inference and hidden signal discovery, to real data applications to weather and climate. Within this article, the latter two are intended only as illustrations of the algorithm at work. A more in-depth study, which requires further extensions of the methodology, should be pursued in field-specific contexts.

This article lays the methodology's general framework. Much more can be done regarding the adaptive choice of the functional spaces $\mathscr{F}$ and $\mathscr{G}$, for which we have proposed here just a handful of simple choices. Since the range of options to explore on adaptive functional spaces is too broad for a single article, exploring them further here would take us too far afield. We also choose not to dwell in this article on other extensions, such as going beyond gradient descent, as required for factor discovery, or building functional spaces $\mathscr{G}$ better-suited for high-dimensional outcome spaces $\mathscr{X}$. We believe that the proposed methodology can be extended in a number of meaningful directions, making it an effective, robust, versatile and conceptually sound approach to a broad set of tasks in data analysis.

A. Appendix: a data-adapted functional space

This appendix describes the choice of functional spaces $\mathscr{F}$ and $\mathscr{G}$ used in our numerical examples. Since the two constructions are entirely similar, we describe only the space $\mathscr{F}$. Exploring other, potentially much richer choices of adaptive functional spaces goes beyond this article's scope.

A.1. *Embedding z in an Euclidean space*

Since the components of z can be of arbitrary type, including real, periodic, categorical and more (no so those of y, which are typically real), we first embed z in an Euclidean space R^k as follows. For each component $z^l \notin R$ of z,

1. When z^l is periodic with period T, we embed it in R^2, mapping z^l to w on the unit circle,
$$w\left(z^l\right) = \left[\cos\left(\frac{2\pi}{T}z^l\right), \sin\left(\frac{2\pi}{T}z^l\right)\right].$$
2. When z^l is categorical with h discrete values v_k, we embed it in R^{h-1}, mapping the $\{v_k\}$ to h equidistant points, the vertices of a regular simplex.
3. For variables z of a more complex type, such as images, distributions or graphs, we introduce an application-specific distance among them and embed them into some R^h accordingly.

Having done this, we can restrict attention to $\mathscr{Z} = R^{d_z}$.

A.2. *Approximation of functions through mollifiers of the delta function*

For any function $f(z)$ and any probability density $\gamma(z)$, we have

$$\gamma(z) = \int \gamma\left(z'\right) \delta\left(z - z'\right) dz', \quad f(z) = \int f\left(z'\right) \delta\left(z - z'\right) dz' = \int f\left(z'\right) \frac{\delta\left(z - z'\right)}{\int \gamma(z'')\, \delta\left(z' - z''\right) dz''}\, \gamma\left(z'\right) dz'$$

(Notice that the last expression is only valid for values of z within the support of the distribution γ.)

Mollifying $\delta(x-y)$ into a smooth non-negative function $K(x,y)$ that concentrates near $x=y$ and integrates to 1 over x, we have

$$f(z) \approx \int f\left(z'\right) \frac{K\left(z, z'\right)}{\int \gamma(z'')\, K\left(z', z''\right) dz''} \gamma\left(z'\right) dz',$$

which in terms of samples $z_j^c \sim \gamma(z)$ yields the empirical version

$$f(z) \approx \sum_j f\left(z_j^c\right) \frac{K\left(z, z_j^c\right)}{\sum_h K\left(z_j^c, z_h^c\right)}. \tag{A.1}$$

Notice that the normalization factor for the weights in this formula is different from the one in kernel regression, which is z-dependent:

$$w_{KR}^j = \frac{K\left(z, z_j^c\right)}{\sum_h K\left(z, z_h^c\right)}.$$

Unlike kernel regression, our approximation is based on the distribution underlying the $\{z_j^c\}$. Notice also that nothing in our argument requires $K(x,y)$ to be a symmetric function, so it is not a "kernel" in the conventional sense. This allows us to use center-dependent bandwidths, coarser where the data is sparser.

If K is a smooth function of z, so is the right-hand side of (A.1) for any choice of $f\left(z_j^c\right)$. We conclude that the expression

$$f(z) = \sum_j a_j \frac{K\left(z, z_j^c\right)}{\sum_h K\left(z_j^c, z_h^c\right)}$$

parameterizes arbitrary smooth functions of z within the support of $\gamma(z)$. Since the denominator does not depend on z, we can absorb it into the definition of a_j, which yields

$$f(z) = F(z)a, \quad F^j(z) = K\left(z, z_j^c\right), \tag{A.2}$$

where it is no longer required that the $\{F^j(z)\}$ integrate to one, since the corresponding normalizing constants can also be absorbed into the $\{a_j\}$. Then the space $\mathscr{F}$ of smooth functions $f(z)$ agrees with the column space of the operator F. In order to consider functions with zero mean, it is enough to subtract the mean of each column of F.

There is no need for the set of centers $\{z_j^c\}$ for K and the set of points $\{z_i\}$ where f is to be evaluated to agree; it is enough that the support of the distribution γ underlying the former contains the support of the latter. Thus, when z is restricted to the sample points $\{z_i\}$, $F(z)$ is a rectangular matrix. Using a number $m \ll n$ of centers reduces the computational cost associated to evaluating the kernels. Moreover, when applied to $g(y)$, one needs to decouple the centers $\{y_j^c\}$ from the samples $\{y_i\}$, as only the latter are arguments over which the objective function L is minimized.

The centers $\{z_i^c\}$ should be well-balanced and representative of the distribution underlying the $\{z_i\}$. A simple procedure for selecting $m \ll n$ centers satisfying these conditions is through k-means applied to the $\{z_i\}$, which has been the choice adopted for all examples in Section 9, where we have set $m = \min([\sqrt{n}], m_{max})$, with m_{max} set by the user. Notice that, when we embed periodic or discrete z^l in R^d, the corresponding l-component of the centers z_j^c need not lie on the unit circle or on the vertices of a simplex.

We adopted for $K(z, z^c)$ a Gaussian function with center-dependent inverse covariance matrix $S_j = S(z_j^c)$:

$$K\left(z, z_j^c\right) = e^{-\frac{1}{2}\left(z - z_j^c\right)' S_j \left(z - z_j^c\right)}.$$

A.3. *Data-adapted determination of bandwidths*

Tuning the $\{S_j\}$ is critical for extracting as much dependence of x on z as possible, by capturing the right functions $f(z)$ and $g(y)$. A simple example illustrates how the appropriateness of a functional space for $f(z)$ depends not just on the samples $\{z_i\}$ but also on the $\{x_i\}$. Consider a situation where x depends only on the first component z^1 of a multidimensional $z \in R^k$, i.e. $x \sim \rho(: |z^1)$. Capturing this dependence requires a family of functions $f(z)$ that depend only on z^1. The ideal S_j for this adopt the form

$$S_j \propto e_1 {e_1}', \quad e_1 = (1, 0, \dots, 0)',$$

since these yield functions $K(z, z^c)$ of z^1 alone, disregarding all other components of both z and z^c. One would not have been able to make these selection from only looking at the $\{z_i\}$.

Similarly, consider a situation where the distribution of x depends on z through some functions $f_k(z)$, with the strength of this dependence quantified by numbers σ_k, $0 \le \sigma_k \le 1$. An appropriate set $\{S_j\}$ could adopt the form

$$S_j \propto \sum_k \sigma_k v_k {v_k}', \quad v_k = \nabla_z f_k(z)\big|_{z_j^c},$$

so that $K(z, z^c)$ changes only in the span of the local $\{\nabla_z f_k\}$, weighted by their relevance. Thus, an appropriate choice of the $\{S_j\}$ yields mollifiers $K_j(z, z_j^c)$ of $\delta(z - z_j^c)$ that single out the sub-manifold of the tangent space to Z at z_j^c on which y may depend.

Yet we do not know the form of this dependence before hand, since determining it is precisely our algorithm's goal. Thus we first make a choice based not on the relation between x and z but on the data available for each. We can subsequently refine this choice by iteratively capturing the dependence between z and x, though we do not pursue such refinement within this article, other than through some straightforward cross-validation sketched below.

A.3.1. Initial determination of the $\{S_j^0\}$

The most natural function $f(z)$ to attempt to capture when looking only at the $\{z_i\}$ is their underlying probability density $\gamma(z)$. We apply the following adaptive procedure to determine the corresponding $\{S_j^0\}$.

1. Compute first a global empirical mollified covariance matrix Σ and its inverse S_g,

$$\Sigma^{kl} = \frac{1}{m} \sum_{i=1}^{m} \left((z_i^c)^k - \bar{z}^k \right) \left((z_i^c)^l - \bar{z}^l \right) + \varepsilon I, \quad \varepsilon = \frac{\operatorname{var}(z)}{m}, \qquad S_g = \Sigma^{-1}.$$

2. Introducing an adjustable parameter α, define

$$K_i^j = K_g\left(z_i^c, z_j^c\right) = e^{-\frac{1}{2\alpha^2}\left(z_i^c - z_j^c\right)' S_g \left(z_i^c - z_j^c\right)},$$

estimate $\gamma_\alpha^i\left(z_i^c\right)$ through leave-one-out kernel density estimation,

$$\gamma_\alpha^i\left(z_i^c\right) \propto \frac{1}{\alpha^d} \sum_{j \neq i} K_i^j,$$

determine α through leave-one-out maximal likelihood,

$$\alpha^* = \arg\max_\alpha L = \sum_{i=1}^{m} \log\left(\gamma_\alpha^i\left(z_i^c\right)\right),$$

and define accordingly $S_{\alpha_*} = \frac{S_g}{\alpha_*^2}$. A practical choice is to maximize L over a finite set of candidate α's centered around the rule-of-thumb value

$$\alpha_{r.o.th.} = \frac{\left(\frac{4}{d+2}\right)^{\frac{1}{d+4}}}{m^{\frac{1}{d+4}}}.$$

3. Next rescale the S_{α_*} locally using the estimated $\gamma_{\alpha_*}^j$,

$$S_j = \left(\gamma_{\alpha_*}\left(z_j^c\right)\right)^{\frac{2}{d}} S_{\alpha_*},$$

so that the the number of points $\{z_i^c\}$ within the effective support of the corresponding function $K(z, z_j^c)$ is roughly independent of j.

4. Finally, introducing a new global adjustable parameter β, write

$$K_j^{\beta}(z) = e^{-\frac{1}{2\beta^2}\left(z-z_j^c\right)' S_j \left(z-z_j^c\right)},$$

with β determined again through leave-one-out maximal likelihood:

$$\beta_* = \arg\max_{\beta} L = \sum_{i=1}^{m} \log\left(\gamma_\beta\left(z_i^c\right)\right), \quad \gamma_\beta\left(z_i^c\right) \propto \frac{1}{\beta^d} \sum_{j \neq i} \frac{K_i^j}{\gamma_{\alpha_*}\left(z_j^c\right)}, \quad K_i^j = K_j^{\beta}\left(z_i^c\right),$$

and define

$$S_j^0 = \frac{S_j}{\beta_*^2}.$$

The procedure so far is automatic and based exclusively on the datapoints $\{z_i\}$. Yet there are at least two reasons why we may want to add one or more free parameters to the determination of the bandwidths. One is that, as discussed above, the ideal S_j should depend not just on the $\{z_i\}$ but also on their relation to the $\{x_i\}$. One straightforward way to address this dependence is through cross-validation over such free parameters. The second reason is that often the bandwidths are determined not by the data alone but also from the scales that one would like to resolve, as in the multi time-scale analysis of ground temperature of Section 9.6. In view of this, we divide our S_j by an externally provided constant ${\gamma_z}^2$, which we either set based on the scales that we have chosen to resolve, as in Section 9.6, or cross-validate over, as in the prediction of global sea surface temperature in Section 9.7.

A.4. *The case of time-like variables z*

The discussion above applies to bounded variables z, all smooth functions of which can be approximated through kernels centered at a relatively small number of well-chosen points $\{z_i^c\}$. Yet in time series-analysis, some of the z's can consist of time itself under different scalings (such as $z_{3,4}$ in the ground-temperature example of Section 9.6.) The number of centers required for such time-like variables grows linearly with the time-extent of the data, which could be very large. From a practical perspective, the problem of such secular growth is that the computation of the orthogonal matrix Q_z from F requires finding the dominant singular components of a potentially vary large matrix F, a computationally costly task.

Yet this problem comes with its own solution. Even though F is large, it is also very sparse, since $K(t_1,t_2)$ can be made to vanish for time pairs such that $|t_2 - t_1|$ is much larger than the kernel's bandwidth. The computational cost of finding the principal components of large but sparse matrices grows only linearly with the number of a matrix rows.

Acknowledgments

The work by Y. Wang and E. G. Tabak was partially supported by a grant from the ONR, and W. Zhao's research was supported in part by the Pacific Institute for the Mathematical Sciences and the Simons Foundation.

Data availability statement

Ground-level temperature data are available from the NOAA U.S. Climate Reference Network. Global sea surface temperature data can be accessed from the Met Office Hadley Centre. All other data used in this study are simulated as described in the corresponding subsections.

REFERENCES

1. Esteban G Tabak and Giulio Trigila. Explanation of variability and removal of confounding factors from data through optimal transport. *Communications on Pure and Applied Mathematics*, 71(1):163–199, 2018.
2. Hongkang Yang and Esteban G Tabak. Conditional density estimation, latent variable discovery, and optimal transport. *Communications on Pure and Applied Mathematics*, 2020.
3. Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.
4. M Agueh and G Carlier. Barycenter in the Wasserstein space. *SIAM J. MATH. ANAL.*, 43(2):094–924, 2011.
5. Esteban G Tabak, Giulio Trigila, and Wenjun Zhao. Distributional barycenter problem through data-driven flows. *Pattern Recognition*, 130:108795, 2022.
6. Brendan Pass. Optimal transportation with infinitely many marginals. *Journal of Functional Analysis*, 264(4):947–963, 2013.
7. Montacer Essid and Michele Pavon. Traversing the schrödinger bridge strait: Robert fortet's marvelous proof redux. *Journal of Optimization Theory and Applications*, 181(1):23–60, 2019.
8. Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.
9. Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.
10. Gennaro Auricchio, Federico Bassetti, Stefano Gualandi, and Marco Veneroni. Computing wasserstein barycenters via linear programming. In *International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research*, pages 355–363. Springer, 2019.
11. Adam M Oberman and Yuanlong Ruan. An efficient linear programming method for optimal transportation. *arXiv preprint arXiv:1509.03668*, 2015.
12. Max Kuang and Esteban G Tabak. Sample-based optimal transport and barycenter problems. *Communications on Pure and Applied Mathematics*, 72(8):1581–1630, 2019.
13. Lingxiao Li, Aude Genevay, Mikhail Yurochkin, and Justin M Solomon. Continuous regularized wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33:17755–17765, 2020.
14. Alexander Kolesov, Petr Mokrov, Igor Udovichenko, Milena Gazdieva, Gudmund Pammer, Evgeny Burnaev, and Alexander Korotin. Estimating barycenters of distributions with neural optimal transport. *arXiv preprint arXiv:2402.03828*, 2024.

15. Amirhossein Taghvaei and Bamdad Hosseini. An optimal transport formulation of bayes' law for nonlinear filtering algorithms. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 6608–6613. IEEE, 2022.
16. Esteban G Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Comm. Math. Sci.*, 8, 2010.
17. Esteban G Tabak and Cristina V Turner. A family of non-parametric density estimation algorithms. *CPAM*, LXVI, 2013.
18. Giulio Trigila and Esteban G Tabak. Data-driven optimal transport. *Communications on Pure and Applied Mathematics*, 69(4):613–648, 2016.
19. Eric V Mazumdar, Michael I Jordan, and S Shankar Sastry. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838*, 2019.
20. Montacer Essid, Esteban G Tabak, and Giulio Trigila. An implicit gradient-descent procedure for minimax problems. *Mathematical Methods of Operations Research*, 97(1):57–89, 2023.
21. Sebastian Claici, Edward Chien, and Justin Solomon. Stochastic wasserstein barycenters. In *International Conference on Machine Learning*, pages 999–1008. PMLR, 2018.
22. Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. Vector quantile regression: An optimal transport approach. *Ann. Statist*, 44(3):1165–1192, 2016.
23. Esteban G Tabak, Giulio Trigila, and Wenjun Zhao. Conditional density estimation and simulation through optimal transport. *Machine Learning*, pages 1–24, 2020.
24. Esteban G Tabak, Giulio Trigila, and Wenjun Zhao. Data driven conditional optimal transport. *Machine Learning*, pages 1–21, 2021.
25. Esteban G Tabak, Giulio Trigila, and Wenjun Zhao. The conditional barycenter problem, its data-driven formulation and its solution through normalizing flows. *Communications in Mathematical Sciences*, 22(6):1635–1656, 2024.
26. Arthur Gretton, Alexander Smola, Olivier Bousquet, Ralf Herbrich, Andrei Belitski, Mark Augath, Yusuke Murayama, Jon Pauls, Bernhard Schölkopf, and Nikos Logothetis. Kernel constrained covariance for dependence measurement. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, volume R5 of *Proceedings of Machine Learning Research*, pages 112–119. PMLR, 06–08 Jan 2005. Reissued by PMLR on 30 March 2021.
27. Matthew Staib, Sebastian Claici, Justin M Solomon, and Stefanie Jegelka. Parallel streaming wasserstein barycenters. *Advances in Neural Information Processing Systems*, 30, 2017.
28. Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pages 2664–2672. PMLR, 2016.
29. Esteban G Tabak and Giulio Trigila. Conditional expectation estimation through attributable components. *Information and Inference: A Journal of the IMA*, 128(00), 2018.
30. Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.
31. Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications in Pure and Applied Mathematics*, 44:371–417, 1991.
32. Jean Jacod and Philip Protter. *Probability essentials*. Springer Science & Business Media, 2012.
33. Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, editors, *Algorithmic Learning Theory*, pages 63–77. Springer Berlin Heidelberg, 2005.
34. Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.
35. Vladimir Rokhlin, Arthur Szlam, and Mark Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2010.
36. Matej Artac, Matjaz Jogan, and Ales Leonardis. Incremental pca for on-line visual learning and recognition. In *2002 International Conference on Pattern Recognition*, volume 3, pages 781–784. IEEE, 2002.
37. Tarek A. El Moselhy and Youssef M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

38. Francis J Anscombe and John W Tukey. The examination and analysis of residuals. *Technometrics*, 5(2):141–160, 1963.
39. Julia Martin, David Daffos Ruiz De Adana, and Agustin G Asuero. Fitting models to data: Residual analysis, a primer. *Uncertainty quantification and model calibration*, 133, 2017.
40. A. Hannachi, I. T. Jolliffe, and D. B. Stephenson. Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology*, 27(9):1119–1152, May 2007.
41. K. Takahashi, A. Montecinos, K. Goubanova, and B. Dewitte. Enso regimes: Reinterpreting the canonical and modoki el niño: Reinterpreting enso modes. *Geophysical Research Letters*, 38(10), May 2011.