The hierarchical barycenter: conditional probability simulation with structured and unobserved covariates

Esteban G. Tabak[1], Giulio Trigila[2†], Wenjun Zhao[3]

[1]Courant Insitute of Mathematical Sciences, New York University, 251 Mercer Street, New York, 10012, NY, USA.
[2]Mathematics Department, Baruch College, CUNY, 55 Lexington avenue, New York, 10010, NY, USA.
[3]Applied Mathematics Department, Brown University, 182 George Street, Providence, 02906, RI, USA.

Contributing authors: tabak@cims.nyu.edu; giulio.trigila@baruch.cuny.edu; zhaow@wfu.edu;
[†]Corresponding author

Abstract

A new methodology is presented for conditional probability density simulation, designed to work with unstructured data sets where not all data points have the same covariates, yet they share common information. Specific examples considered in this article fall into two main classes: homogeneous data with missing covariate values, and data sets divided into two or more groups with covariates that overlap only partially. The methodology, based on the mathematical theory of optimal transport, extends the optimal transport barycenter problem to a newly defined hierarchical barycenter problem. A data driven numerical procedure is developed for the solution of the hierarchical barycenter problem, which is used to illustrate the new problem's advantages over the classical barycenter problem on synthetic and real world data sets.

Keywords: Optimal transport, barycenter problem, data amalgamation, conditional estimation, multi-task learning.

1 Introduction

This article develops a data-based methodology for simulating conditional distributions, when some of the covariates are only observed or even defined in overlapping subsets of the data. Consider as an example a recommendation system where users assign a rate x to objects of different classes, such as movies, books and bicycles. How could one estimate, based on data, the rate that a particular user will assign to a particular object?

Every individual rate x_i has a few qualifiers attached. Some may relate to the user (e.g. age, national origin), some to the object (a book's number of pages, a bicycle's color), some to environmental factors (day of the week, weather, location). We will group these qualifiers into a factor z_i with entries $\{z_i^l\}$. Our goal is, given new factor values z_*, estimate the corresponding conditional density $\rho(x|z_*)$ or simulate it by drawing samples from it.

If the set of factors $\{z^l\}$ were common to all observations x_i, we could use existing tools, such as kernel conditional density estimation [1, 2], or conditional density simulation through the distributional barycenter problem, a methodology developed by the authors in prior work [3, 4]. The z-variable in our problem, however, is not a regular vector, as it is not globally defined in a uniform way. The z^l corresponding to number of pages, for instance, is defined for books but not for bicycles.

We could reduce this problem to the regular one by dividing it into sub-problems: one for movies, one for toys, and so on. Yet each of these sub-problems has a much smaller number of observations available than their aggregation, yielding less samples to base our estimation on. This sample-size problem is exponentiated by the fact that the tree of factors z may branch further, with only science books, for instance, classifiable into physics, zoology and so on. Yet each sub-problem could inform the others: the user's age, for instance, may have a consistent effect across the rates, and choices made on objects of one type could well inform those to be made among objects of other types.

Even though we have used a recommendation system for illustration, the setting described applies much more broadly. The variables x can quantify, for instance, the results of a medical treatment, where the condition being treated brings in its own qualifiers, such as body temperature for a cold or blood glucose content for diabetes, and each drug administered may be further qualified by the corresponding specification and dosage. A similar problem appears whenever one would like to bring together different data bases: the positive effect of their aggregation on the sample size comes at the price of having different sets of covariates available for each. Prediction with missing data, where the unavailability of different components of the vector z of predictors for each observation is due to lack of knowledge, not of definition, also fits in a similar framework.

This article proposes a methodology, the hierarchical barycenter, to address this broad category of data problems, extending the optimal transport-based methodology in [3] to covariates with a much more general structure.

For the sake of clarity, we will focus on two main categories of problems for which the hierarchical barycenter is particularly well suited:

1. Structured cofactors: the data in the training set have no missing values, they are however divided into groups characterized by different cofactors (e.g. data relative to books and to movies).
2. Missing data: the training set shares a common set of cofactors, yet only some of these are available for each data point.

The paper is structured as follows. Section 2 describes the hierarchical barycenter problem within the frame of the theory of optimal transport [5]. Section 4 describes the details of the implementation of the hierarchical barycenter that we then use in Section 5 to analyze synthetic and real world data relative to mineral bone density [6, 7].

2 Formulation

2.1 The distributional barycenter problem

Before addressing the structure of the factors z, this section summarizes the distributional barycenter problem [4], an extension of the Wasserstein barycenter described in [8], and its use to perform conditional density simulation. This problem can be posed as follows: given a set of N sample pairs $\{x_i, z_i\}$ –the observations– drawn from an unknown joint distribution

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

simulate

$$\rho\left(x|z_*\right)$$

for any target value z_* of z, i.e. extract samples $\{x_j^*\}$ thereof.

The distributional barycenter removes from the $\{x_i\}$ all the variability that the $\{z_i\}$ can explain –and only that variability– transforming them into new variables $\{y_i\}$ that are independent of z. If one then brings back to the $\{y_i\}$ the variability that $z = z_*$ would entail, the resulting $\{x_i^*\}$ are the samples of $\rho\left(x|z_*\right)$ sought.

Describing this procedure in more detail, one first seeks a map $y = T(x, z)$ with $x, y \in \mathbb{R}^{d_x}$ and the factors $z \in \mathbb{R}^{d_z}$ acting as parameters. Removing from x all variability that z can explain translates into the condition that the resulting y must be independent of z. Not removing any additional variability from x can be implemented through the requirement that the map T deform x as little as possible. Then the distributional barycenter problem reads

$$\min_{y=T(x,z)} C(x, y) \quad \text{s.t.} \quad y \perp\!\!\!\perp z, \tag{1}$$

where the symbol "$\perp\!\!\!\perp$" stands for independence, and $C(x, y)$ is a measure of the deformation incurred by transforming the x into y's. In an optimal transport framework, C adopts the form of the expected value of a pairwise deformation cost c,

$$C(x, y) = \mathbb{E}[c(x, y)], \tag{2}$$

where a typical choice for c in normed spaces is the squared norm

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

Once T has been found, we draw samples $\{x_i^*\} \sim \rho(x|z_*)$ for any target z_* through

$$x_i^* = T^{-1}(y_i, z_*), \quad \text{where} \quad y_i = T(x_i, z_i). \tag{3}$$

The formulas in (3) carry out the program described above, removing first from the $\{x_i\}$ the variability attributable to the corresponding $\{z_i\}$, and then replacing it by the variability entailed by the target z_*.

2.2 Extension to hierarchical covariates

Applying the procedure just described to our problem requires understanding what it means for y to be independent of z when the latter has components that are defined only for a subset of the data. The simplest setting has a single variable z^1 that is known for some of the $\{x_i\}$ and not for the others. Let I_k be the set of observations $\{i\}$ where z^1 is known and I_u its complement. If one simply removes the variability in x attributable to z^1 from the $\{x_i,\ i \in I_k\}$ while leaving the $\{x_i,\ i \in I_u\}$ untouched, the resulting $\{y_i\}$ will be divided into two groups, one with reduced variability and the other not. This spurious source of variability in y can be explained away by adding a new, binary factor $z^0 \in \{k, u\}$, discriminating observations in I_k from those in I_u.

It follows that we should use in this case a factor z structured as a tree, with z^0 partitioning the root into two branches, one with and one without the covariate z^1. Then the independence between y and z acquires a clear meaning: y must be independent of z^0, and those y with $z^0 = k$ must be independent of z^1. Notice that, if the availability of z^1 was caused by a hidden binary confounder that could also affect the distribution of the x, this additional source of variability is also taken into consideration by the inclusion of z^0 as a factor. On the other hand, if this confounder was not hidden but interpretable –we can count the number of pages z^1 of a book but not those of a bicycle– then z^0 would have been already included as a factor among the z^l –in this example, through the type of object being rated.

In the general case, the structure of z may be not that of a tree but of a more general graph, for instance when different subgroups of objects have partially overlapping factors, such as color, size and age. The general rule is that, for any observation x_i, the subset of the factors $\{z_i^l\}$ available should be fully determined by some of those factors themselves, either in an interpretable fashion –as with weight being an available factor for luggage– or through a factor added explicitly to account for possibly missing observations. With such *covariate extension*, the problem formulation reduces to a regular barycenter problem.

The factors z_i^l associated to different data sets need not be of the same type or dimensionality, as for instance, books and movies may have a different number of covariates. To keep the notation simple, we will omit to specify the type and dimensionality of each factor z_i^l .

2.3 Enforcing independence between y and z

The objective function of the barycenter problem (1) has two components: the cost $C(x, y)$ to minimize and the constraint that y and z should be independent. Translating the cost C to a sample-based formulation is straightforward, particularly in an optimal transport setting, where when the joint distribution $\pi(x, z)$ is only known through N sample pairs $\{x_i, z_i\}$, we can replace expectation with empirical mean:

$$C(x, y) = \mathbb{E}[c(x, y)] \rightarrow \frac{1}{N} \sum_i c\left(x_i, y_i\right), \quad y_i = T\left(x_i, z_i\right). \tag{4}$$

In order to complete the problem's formulation, we also need to implement the independence condition in a sample-friendly way.

There are a number of ways to write down the condition that two variables y and z be independent, i.e. that their joint distribution factorizes:

$$\pi(y, z) = \rho(y)\gamma(z). \tag{5}$$

Some choices are:

1. A weak formulation of (5) in terms of measurable test functions:

$$\forall F(y, z), \int \left[F(y, z) - \int F(y, w)\ d\gamma(w)\right] d\pi(y, z) = 0,$$

   which is implementable in terms of samples (see [4] for more details):

$$\forall F(y, z) \in \mathcal{F}, \sum_i \left[F\left(y_i, z_i\right) - \frac{1}{N} \sum_k F\left(y_i, w_k\right)\right] = 0,$$

   where $\mathcal{F}$ is a class of functions adapted to the number and distribution of the samples available.
2. Same as above but with just one special test function F_0:

$$F_0(y, z) = \rho(y|z)$$

   (See in [4] how the condition that the vanishing of the non-negative quantity

$$\int \left[F_0(y, z) - \int F_0(y, w)\ d\gamma(w)\right] d\pi(y, z)$$

   suffices to guarantee independence.)
3. The vanishing of the mutual information between Y and Z:

$$MI(Y, Z) = \int \log\left(\frac{\pi(y, z)}{\rho(y)\gamma(z)}\right) d\pi(y, z) = 0.$$

The second and third options pose the independence condition in terms of the vanishing of a non-negative functional $\Theta(\pi)$, with two immediate advantages: on the one hand, it permits adopting a standard penalty optimization procedure [9], and solve

$$\min_T C(x, T(x, z)) + \lambda\Theta(\pi_T),$$

where

$$\pi_T(y, z) = T\#\pi(x, z)$$

is the push forward by the map T of the original joint distribution π, and $\lambda > 0$ is a penalization parameter. On the other hand, it allows us to enforce a detailed notion of independence. When z has more than one component, the independence between y and z implies that y should be independent of any subset of the $\{z^l\}$, including each individual factor alone. Because of the non-negative nature of the functional Θ, we can enforce each of these requirement separately, writing

$$\min_T C(x, T(x, z)) + \sum_k \lambda_k \Theta_k(\pi_T).$$

In option 3, for instance, Θ_k measures the mutual information between y and the kth subset of the $\{z^l\}$ considered.

This is important for the barycenter problem in general, and more so for the hierarchical one. When the number of factors $\{z^l\}$ is large and the number of observations is comparatively small, most $\{z_i\}$ will be far from each other. This makes a global characterization of independence necessarily inaccurate, while assessing the independence between y and lower dimensional subsets of the $\{z^l\}$ may be within reach. This statement applies even more strongly to the hierarchical barycenter problem, where some of the $\{z^l\}$ are available for only a fraction of the data.

Option 1 addresses the same issue automatically, since the set of all measurable functions $F(y, z)$ includes those that depend only on any given subset of the $\{z^l\}$. Hence all the options above are suitable candidates, each with its own advantages and challenges. For concreteness, we restrict attention to option 3, since this article's main objective is to develop the hierarchical barycenter's conceptual framework, rather than perfecting one technical approach or another.

Thus we will pose the problem in the form

$$\min_T L[T] = C(X, Y) + \sum_k \lambda_k MI\left(Y, Z_k\right), \quad Y = T(X, Z). \tag{6}$$

Here the $\{Z_k\}$ represent subsets of the full set of cofactors $\{Z^l\}$. Their choice, as well as the values of the penalization parameters $\{\lambda_k\}$ and the data-based formulation of the problem are discussed below.

3 Related work

The methodology presented here has partial overlap with two well established procedures: multi-task learning (MTL) [10, 11] and transfer learning (TL) [12, 13]. Both

methodologies aim to exploit common information shared by different data sets and both are usually framed in terms of learning tasks. The main difference between the two is that while MTL aims to learn multiple tasks at the same time, TL leverages information available from an initial task to improve the performance of another, single target task.

In this sense, the hierarchical barycenter (HB) has points in common with both procedures: it uses information from different data sets to enhance the estimate of the conditional density underlying one –any– data set. This data set does not have to be known a-priori as in TL: the information is "transferred" by the pull back of the barycenter to one of the specific data sets (see eq. 3).

There are many differences between HB and the MTL-TL procedures. Starting with MTL, the problem is usually formulated through an objective function

$$MTL(w) = \sum_{k=1}^{l} L_k(f_k(w_k)) + \lambda R(w), \tag{7}$$

where L_k is a loss function for the k-th task and the w_k are the corresponding parameters for label's estimation. The function $R(w)$ regularizes the parameters within each task and specifies how much the parameters referring to different tasks are related. Different forms of MTL are represented by different functions R. The use of (7) assumes that there are features that are related to all tasks and that these features are learned by optimizing (7) (see [10, 14]).

A first difference between this set up and the HB is that the latter does not impose a parametrization for each task. Since the pushforward is defined through the flow associated to the minimization in (1), HB is non-parametric (see Section 4 for more details). From this point of view, HB is more similar to [15], where task-specific estimators are considered as random variables and the task relationships are discovered by measuring the statistical dependence between each pair of random variables. The overall goal is to leverage information between random variables displaying higher degrees of dependence. As in our approach, the work in [15] uses kernel approximation of the mutual information [16, 17], yet the scope is different. In HB, the mutual information is used to enforce the pushforward condition relating the marginals to the barycenter rather than to measure the degree of dependence.

Another major difference between the HB and MTL is the barycenter itself. A byproduct of HB, absent in both MTL and TL, is the merging of multiple datasets into the barycenter. This distribution is characterized in a precise way, representing all the variability that cannot be explained by any of the cofactors across the different datasets. The barycenter represents a tool for at least two important tasks: the removal of variability [18] and factor discovery [19]. Section 5.4 shows examples in which we use HB to remove the variability explainable by known cofactors to reveal a hidden signal in the data. The work in [19] shows then how one can look for hidden factors that explain the hidden signal.

We do not claim that HB necessarily leads to better results than MTL in scenarios in which both procedures apply. We propose the hierarchical barycenter as a general,

first principled procedure, based on the mathematical theory of optimal transport, well-suited to the analysis of heterogeneous datasets.

4 The Algorithm

4.1 Problem formulation in terms of samples

In order to develop a data-driven methodology to solve the hierarchical barycenter, we need to transform the formulation in (6) into one that uses not the joint distribution $\pi(X, Z)$ itself but samples thereof (x_i, z_i), $i \in \{1, \ldots . N\}$ and their image, $y_i = T(x_i, z_i)$, under the unknown map. This transformation is straightforward for the transportation cost:

$$C(X, Y) = \int c\left(x, T(x, z)\right) \pi(x, z)\, dxdz \rightarrow \frac{1}{N} \sum_i c(x_i, y_i).$$

Similarly, for the mutual information $MI(Y, Z^k)$, where Z^k is a subset of the variables Z, we can write

$$MI(Y, Z^k) = KL\left[\pi_T^k\left(Y, Z^k\right) || \mu\left(Y\right) \nu\left(Z^k\right)\right] \approx \\ \approx \frac{1}{N_k} \sum_{i \in I_k} \left[\log\left[\pi_T^k\left(y_i, z_i^k\right)\right] - \log\left[\mu\left(y_i\right) \nu\left(z_i^k\right)\right]\right], \quad (8)$$

where I_k represents the subset of observations where all covariates $z^l \in Z_k$ are defined, and $N_k = |I_k|$, their cardinality. For this formulation to depend only on the data points, we replace the probability densities by their kernel-based estimation:

$$MI^{est}\left(Y, Z^k\right) = \frac{1}{N_k} \sum_{i \in I_k} R(y_i, z_i^k) \quad (9)$$

where

$$R(y_i, z_i^k) = \log\left(\frac{1}{N_k} \sum_{j \in I_k} K^y\left(y_i, y_j\right) K^z\left(z_i^k, z_j^k\right)\right) \\ - \log\left(\frac{1}{N_k} \sum_{j \in I_k} K^y\left(y_i, y_j\right)\right) - \log\left(\frac{1}{N_k} \sum_{j \in I_k} K^z\left(z_i^k, z_j^k\right)\right). \quad (10)$$

Then problem (6) adopts the data-driven form

$$\min_{\{y_i\}} L = \frac{1}{N} \sum_i c(x_i, y_i) + \sum_k \lambda_k \frac{1}{N_k} \sum_{i \in I_k} R(y_i, z_i^k). \quad (11)$$

Notice that we do not need to keep the third term in $R(y_i, z_i^k)$, since it depends only on z and we are minimizing over y. Without the third term, we are minimizing the log-likelihood of $\rho(z|y)$ over the variable y on which we are conditioning. In other words, we are looking for the y such that the observed z, given y, is least likely. It follows from the argument that this y must be independent of z.

In order to guarantee independence between Y and Z, the choice of the covariate subsets $\{Z_k\}$ must satisfy the requirement that, for all subsets of the observations $\{x_i\}$, their maximal common subset of covariates must be one of the $\{Z_k\}$. This is just a minimal requirement though: one can add additional subsets –all the way to those consisting of individual z^l's– in order to enforce a more detailed notion of independence. On the other hand, even the minimal requirement can sometimes be computationally unfeasible. For instance, when covariate values are missing at random, we may be forced to consider all subsets of the $\{Z^l\}$, which grows exponentially with the number of covariates, and which may include subsets with corresponding sample sets I_k with only a handful of available samples. In this case, a relaxation of the problem is appropriate, such as including only those subsets $\{Z_k\}$ with cardinality smaller than a prescribed small number L_{max} and where the number of samples $|I_k|$ is larger than a minimum value N_{min}. The first condition disregards complex dependence between Y and Z involving the non-additive interaction of more than L_{max} factors. Setting $L_{max} = 1$, for instance, would relax the notion of independence between Y and Z to that of independence between Y and each individual Z^l. The second condition addresses possible over-fitting of small subsets of the data.

4.2 Minimization through regularized gradient descent

We minimize the objective function L in (11) through gradient descent, accelerated by a preconditioning procedure that can be conceptualized as a simplified version of implicit gradient descent:

$$y = y - \eta \left(I + \eta H_d\right)^{-1} G, \tag{12}$$

where G with $G_i = \frac{\partial L}{\partial y_i}$ is the gradient of L and H_d is a diagonal matrix containing only the diagonal elements $H_i^i = \frac{\partial^2 L}{\partial y_i^2}$ of the Hessian matrix H. Equation (12) with the full matrix H instead of the diagonal H_d is the building block of implicit gradient descent (see [20] for a similar procedure for minimax problems and [21, 22] for the closely related Levenberg-Marquardt regularization of Newton's method). Keeping just the diagonal elements of H eliminates the need to invert an $n \times n$ matrix, while preserving to a large degree the regularizing effect of the implicit procedure.

The learning rate η in (12) is chosen adaptively according to the strategy described in [20]. Far from the local extreme of the loss function, η is small and (12) reduces to regular gradient descent. Near the minimum, η increases, converging to a (quasi) Newton method with faster convergence rate.

The use of (12) for our problem is greatly facilitated by the fact that the gradient of L in (11) can be robustly approximated by

$$\frac{\partial L}{\partial y_i} = \frac{1}{N}\frac{\partial}{\partial y_i}c(x_i, y_i) + \sum_{k/i\in I_k}\frac{\lambda_k}{N_k}\frac{\partial}{\partial y}\left[\log\left(\frac{\sum_{j\in I_k}K^y\left(y, y_j\right)K^z\left(z_i^k, z_j^k\right)}{\sum_{j\in I_k}K^y\left(y, y_j\right)}\right)\right]_{y=y_i}. \tag{13}$$

The simple form of the gradient of L is due to the fact that the derivative of the mutual information with respect to the second argument of the kernels, namely with respect to the position of the centers, is a random variable with zero mean and vanishing variance as the number of samples grows. We provide here a general argument of why this is the case; a more detailed derivation can be found the appendix.

We focus on the second term in (10), since the result for the first term follows from exactly the same logic. Suppose that the first and second argument of K^y are not necessarily computed at the same set of points $\{y_i\}$. This amount to substituting the estimation of $\mu(y)$, relative to the random variable Y, with kernels that are centered around a set of points $\{y'_j\}$, whose distribution $\xi(y)$ is different from $\mu(y)$. The second term in (10) can then be interpreted as the relative entropy between μ and ξ:

$$\frac{1}{N_k}\sum_{i\in I_k}\log\left(\frac{1}{N'}\sum_{j=1}^{N'}K^y(y_i, y_j)\right) \approx \int \mu(y)\log(\xi(y))dy. \tag{14}$$

Taking the derivative with respect to the second argument of K^y, the position of the centers, can therefore be thought, in the limit of $N' \to \infty$, as computing the variational derivative with respect to ξ of the RHS of (14). Since the relative entropy is maximized when $\xi = \mu$ almost everywhere, then we have that

$$\frac{\delta}{\delta\xi}\int \mu(y)\log(\xi(y))dy\Big|_{\xi=\mu} = 0.$$

(The true relative entropy has an additional term involving the entropy of μ, but this is immaterial for differentiation with respect to ξ.)

This justifies the approximation in (13),

$$\begin{aligned}\frac{\partial L}{\partial y_i} &= \frac{1}{N}\frac{\partial}{\partial y_i}c(x_i, y_i) \\ &+ \sum_{k/i\in I_k}\frac{\lambda_k}{N_k}\left[\frac{1}{\sum_{j\in I_k}K^y(y, y_j)K^z(z_i^k, z_j^k)}\frac{\partial}{\partial y}\left(K^y\left(y, y_j\right)K^z\left(z_i^k, z_j^k\right)\right)\right]_{y=y_i} \\ &- \frac{\lambda_k}{N_k}\left[\frac{1}{\sum_{j\in I_k}K^y(y, y_j)}\frac{\partial}{\partial y}K^y\left(y, y_j\right)\right]_{y=y_i}.\end{aligned}$$

which neglects to consider the derivative of the objective function with respect to the Kernel's second argument.

Similarly, the diagonal second order derivatives in H_d^i are well-approximated by

$$\begin{aligned}\frac{\partial^2 L}{\partial y_i^2} &= \frac{1}{N}\frac{\partial^2}{\partial y_i^2}c(x_i,y_i) \\ &+ \sum_{k/i\in I_k}\frac{\lambda_k}{N_k}\left[\frac{\partial}{\partial y}\left(\frac{1}{\sum_{j\in I_k}K^y(y,y_j)K^z(z_i^k,z_j^k)}\frac{\partial}{\partial y}\left(K^y\left(y,y_j\right)K^z\left(z_i^k,z_j^k\right)\right)\right)\right]_{y=y_i} \\ &- \sum_{k/i\in I_k}\frac{\lambda_k}{N_k}\left[\frac{\partial}{\partial y}\left(\frac{1}{\sum_{j\in I_k}K^y(y,y_j)}\frac{\partial}{\partial y}K^y\left(y,y_j\right)\right)\right]_{y=y_i}.\end{aligned}$$

4.3 Prediction through map inversion

The solution of the barycenter problem consists of two elements: the barycenter $\mu(y)$ itself and the map $T(x;z)$ pushing forward the conditional distribution $\rho(x|z)$ to μ. In our numerical solution, neither μ nor T are given in closed form. Instead, they are both represented by the set $\{y_i\}$, where each y_i is both an independent sample from the barycenter $\mu(y)$ and the value that the map T adopts when applied to the sample pair (x_i,z_i).

The barycenter has much value, as elaborated through an example in Section 5.4. Yet our original goal was not to explain variability away from x, but to simulate $\rho(x|z^*)$ for a given target value z^*. For this, we need to compute the inverse map $T^{-1}(\cdot,z^*)$ and use it to push back all the points $\{y_i\}$ in the barycenter to obtain $N=\sum_p N_p$ samples $x_i(z^*)$ from $\rho(x|z^*)$. It would appear at first that inverting a map $y=T(x,z)$ known only through samples (x_i,z_i,y_i) should require numerical interpolation, e.g. using near-neighbors, kernel regression or neural networks. Yet it turns out that the structure of the minimization problem giving rise to T provides a simple closed-form solution for $x=T^{-1}(y,z)$, obtained from the first order condition $\nabla_Y L=0$ of the objective function in (6).

The condition that $\frac{\partial L}{\partial y_i}=0$ in (13) adopts the form

$$\frac{\partial}{\partial y}c(x_i,y)\bigg|_{y=y_i} = -\frac{\partial}{\partial y}\sum_{k/i\in I_k}\frac{\lambda_k N}{N_k}\log\left(\frac{\sum_{j\in I_k}K^y\left(y,y_j\right)K^z\left(z^k,z_j^k\right)}{\sum_{j\in I_k}K^y\left(y,y_j\right)}\right)\bigg|_{\substack{y=y_i\\z=z_i}}.$$

Notice that only the left-had side of this identity depends on x_i. It follows that, if $\frac{\partial c(x,y)}{\partial y}=a$ is invertible as $x=f(y,a)$, then it provides a closed expression for x in terms of y and z, strictly valid for every sample pair (y_i,z_i). Since this expression is smooth as a function of y and z, it provides a natural, closed form expression for $x=T^{-1}(y,z)$ for pairs (y,z) not in the sample set. In particular, under the canonical cost $c(x,y)=\|x-y\|^2/2$, we obtain

$$x=T^{-1}(y,z)=y+\frac{\partial}{\partial y}\sum_{k/i\in I_k}\frac{\lambda_k N}{N_k}\log\left(\frac{\sum_{j\in I_k}K^y\left(y,y_j\right)K^z\left(z^k,z_j^k\right)}{\sum_{j\in I_k}K^y\left(y,y_j\right)}\right). \tag{15}$$

4.4 Hyper-parameter tuning

One needs to choose a set of hyper-parameters for the data-driven formulation: bandwidths of kernels in z^k and y spaces and penalty coefficients $\{\lambda_k\}$. The following two subsections discuss their choice in the current framework.

4.4.1 Bandwidths in z^k-space

These hyper-parameters must be chosen first, as their values remain constants throughout the procedure, since the z_i are kept fixed during the iteration pushing forward x to $T(x, z)$.

We chose the bandwidths h_z in z space based on Silverman's rule-of-thumb [23, 24] and the cardinality of the subset of the the data for which each specific component of z is known. Then the bandwidth for $K^z(z_i^k, z_j^k)$ with $i, j \in I_k$, is given by a diagonal covariance matrix with elements d_{kk} given by

$$\sqrt{d_{kk}} = \sigma_{z^k} \left(\frac{4}{d+2} \right) |I_k|^{\frac{-1}{d+4}}, \tag{16}$$

where σ_{z^k} is the estimated standard deviation over the $\{z_i^k : i \in I_k\}$, $|I_k|$ is the cardinality of I_k and d is the dimension of z.

4.4.2 Penalty coefficients $\{\lambda_k\}$ and bandwidths in y space

These two sets of hyper parameters are related, since both control the extent to which we resolve the independence of y and z: the magnitude of $\lambda_k > 0$ quantifies the weight assigned to the vanishing of the mutual information $MI(Y, z^k)$, and the bandwidths set the accuracy of the estimation of $MI(Y, z^k)$ via the kernels K^y in (9).

We tune λ_k and h_y via cross validation. To this end, we set aside from the data a validation set $\{x_{i,val}, z_{i,val}\}$, and use as objective function for the cross validation the log-likelihood of this set:

$$\overline{\log \mathcal{L}} = \frac{1}{N_{val}} \sum_{i=1}^{N_{val}} \log \hat{\rho}(x_{i,val}|z_{i,val}), \tag{17}$$

where for each choice of the hyper-parameters λ and h_y, $\hat{\rho}(\cdot|z_{i,val})$ is estimated in following way: we solve the barycenter problem using the training data, obtain samples $x_{i,val}^j \sim \rho(\cdot|z_{i,val})$ through the inversion of T (described in the following subsection) on each y_j in the barycenter and each $z_{i,val}$, and then estimate $\rho(\cdot|z_{i,val})$ through kernel density estimation on those $x_{i,val}^j$.

We adopt the following strategy in order to facilitate the exploration of a hyper parameter space that would otherwise be too high dimensional when k takes more than 2 different values. All λ_k are determined by a single tunable parameter λ, through

$$\lambda_k = \lambda r_i \quad \text{where} \quad r_i = |I_k| MI^{est}(X, Z^k). \tag{18}$$

The rationale for this choice is that the penalty coefficient should increase with the degree of dependence between X and Z^k, quantified by their mutual information, and to the significance of the corresponding subset I_k, which its naturally quantified by its cardinality $|I_k|$.

Additional information regarding hyperparameter tuning will be provided when discussing concrete examples in Section 5.

5 Numerical Examples

This section illustrates the use of the HB on two common classes of problems in the machine learning literature: prediction on data sets with missing values and on data sets with structured cofactors. In both cases, the output of our analysis are samples $\{x_i^*\}$ drawn from the conditional density $\rho(x|z^*)$, obtained through the procedure described in Section 4.

5.1 Missing Data - Synthetic Example

Our first synthetic example uses a synthetic dataset $\{(x_i, z_i)\}_{i=1,\dots,N}$, where $x_i \in \mathbb{R}$ is sampled from a normal distribution $\mathcal{N}(f(z_i), g(z_i)^2)$ with mean $f(z)$ and standard deviation $g(z)$, and $z_i = (z_i^1, z_i^2) \in \mathbb{R}^2$ is drawn from a bivariate uniform distribution $U_{[0,1]^2}$ with independent components. The dataset is then divided into three subgroups:

- I_1, with x_i and only z_i^1 observed,
- I_2, with x_i and only z_i^2 observed, and
- I_3, with x_i and both z_i^1 and z_i^2 observed.

The subsets I_1 and I_2 have 80 points each, while I_3 contains only 20 points. We also keep an additional set of 20 points for cross validation, as described in section 4.4, with the values of all variables known.

In this case, the objective function (6) has the form

$$\min_{y_i} \sum_{i=1}^{N} c(x_i, y_i) + \lambda_1 MI^{est}(y, z^1) + \lambda_2 MI^{est}(y, z^2) + \lambda_3 MI^{est}(y, (z^1, z^2)) + \lambda_0 MI^{est}(y, w), \tag{19}$$

where w_i is a categorical variable indicating whether data point x_i belongs to either I_1, I_2 or I_3. As described in Section 2.2, including w precludes the resulting barycenter from becoming a mixture of three distinct distributions, since the variability due to w must be explained away.

We run experiments with three different choices for the functions $f(z)$ and $g(z)$:

- Test 1 - Additive mean: $f(z) = 4z^1(1 - z^1) + 0.5z^2, \quad g(z) = 0.2.$
- Test 2 - Non-additive mean: $f(z) = 4z^1(1 - z^1) + 0.5z^1z^2, \quad g(z) = 0.2.$
- Test 3 - Heterogeneous standard deviation: $f(z) = 4z^1(1 - z^1) + 0.5z^2, \quad g(z) = 0.25(\sqrt{z^1} + \sqrt{z^2})$.

The idea behind these tests is to quantify the procedure's accuracy under increasing complex levels of dependence between the random variables X and Z.

We compare our methodology with three alternative procedures for conditional density estimation, computing in all cases the Kullback-Leibler (KL) divergence ([25]) between the estimated and the exact density. Since the densities under consideration are Gaussian, we can use the following closed form for the KL divergence for a given z (see for instance [26]):

$$KL(\hat{\rho}(\cdot|z)||\rho(\cdot|z)) = \log\left(\frac{\sigma_\rho}{\sigma_{\hat{\rho}}}\right) + \frac{\sigma_{\hat{\rho}}^2 + (\mu_\rho - \mu_{\hat{\rho}})^2}{2\sigma_\rho^2} - 1/2,$$

where σ_ρ and μ_ρ are the standard deviation and the expected values of ρ respectively (similarly for $\hat{\rho}$). We then average $KL(\hat{\rho}(\cdot|z)||\rho(\cdot|z))$ over different values of z chosen on a uniform grid covering the support of the density of underlying z. The procedures under comparison are:

1. Hierarchical Barycenter (HB), the procedure described in Section 4.
2. Benchmark 1 (B1), a regular barycenter problem which uses only the points in I_3 for which the values of both covariates (z^1, z^2) are known.
3. Benchmark 2 (B2), which first imputes the missing values for the covariates z's (based on the nearest neighbor in x space) and then solves a regular barycenter problem with no missing values (see subsection 2.1)
4. Benchmark 3 (B3), which solves the classical barycenter problem without hiding any of the values of either z^1 or z^2. This is the best possible scenario, since everything is known and the error in the estimate of the conditional density is exclusively due to the Monte Carlo approximations used to compute the Mutual Information that enforces the independence between y and $z = (z^1, z^2)$.

The results are summarized in the following table:

Procedures	HB	B1	B2	B3
Test 1	**0.1997**	0.7366	0.2189	0.1605
Test 2	**0.1890**	0.7014	0.2335	0.1597
Test 3	**0.2746**	0.5963	0.2817	0.1998

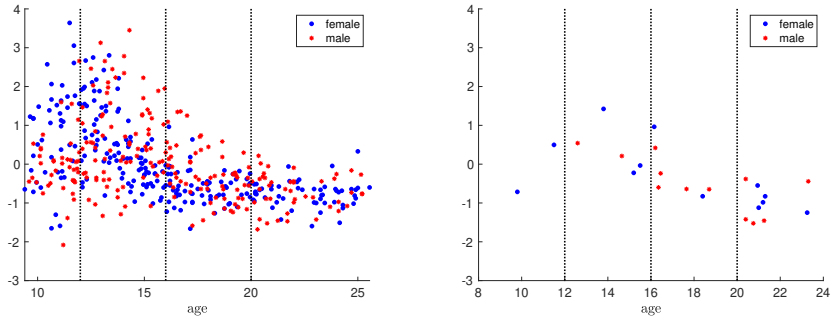
Table 1: KL divergence for the three synthetic tests, which estimate the conditional density using the hierarchical barycenter (HB) and three different benchmarks (B1-B3) described in subsection 5.1.

Table 1 shows that the lowest KL values between the exact and the estimated densities are obtained for the hierarchical barycenter, achieving values close to those obtained using B3, where no values of the cofactors z were hidden. The estimate using B1 has the largest error, since it does not use the information contained in I_2 and I_3

5.2 Missing Data - Bone Mineral Density

This section analyzes a data set with spinal bone mineral density measurements on 485 North American adolescents [6].

In order to adapt this data set to our purpose, we divide the data set into 3 subgroups and hide for each group the values of different cofactors, recreating the same setting of the synthetic example in subsection 5.1. In this context, z^1 and z^2 represent the gender and the age of each individual and x represents the bone mineral density. The three subgroups $I_{1,2,3}$ have the same meaning as in subsection 5.1. While $I_{1,2}$ both contain 218 points, the subgroup I_3, with no hidden values for z^1 or z^2, contains 24 points. In addition, we have a validation set of 25 points. Figure 1 depicts the data set, highlighting the much smaller subset where all information is available.



(a) Complete data set (b) Data contained in I_3

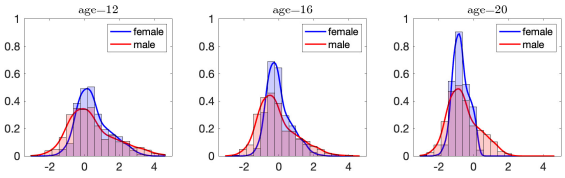
Fig. 1: Bone mineral density (y-axis) as a function of age (x-axis) and gender (color). The left panel displays the full data set –including those covariate values that will not be made available to the analyst– while the right panel displays only the data in subset I_3, where the values of both z^1 and z^2 are known.

Since there is no ground truth to compare the results with, we assess the accuracy of the estimated conditional densities through the corresponding likelihood of the test set. In order to have more robust statistics, we repeat the experiment 30 times, each time hiding the values of the cofactors for a different subset of points. The results reported in the Table 2 contain the average likelihood over those different realizations of $I_{1,2,3}$. Again, other than B3, which uses complete information supposedly not available, the hierarchical barycenter yields the smallest error.

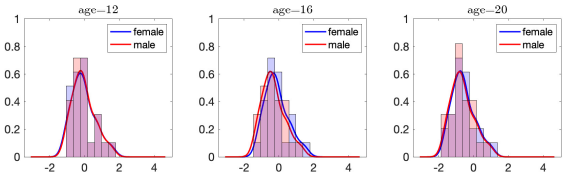
Approaches	HB	B1	B2	B3
	-1.0790	-1.3671	-1.2573	-0.8596

Table 2: Average likelihood of the validation set for the bone mineral dataset (The higher the better.)

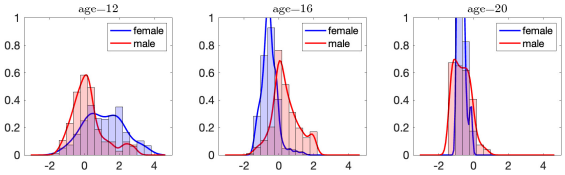
Figure 2 visualizes some of the conditional densities obtained for three different ages, using the hierarchical barycenter and Benchmarks 1 and 3. Even though it is



(a) Results from HB



(b) Results from B1



(c) Results from B3

Fig. 2: Conditional densities $\rho(x|z^*)$ estimated for three different ages and both genders of bone mineral density dataset. The conditional densities are visualized through histograms and a kernel density estimator using the samples from $\rho(x|z^*)$ obtained with the hierarchical barycenter and the benchmark procedures 1 and 3 respectively.

hard to asses visually from this particular instance whether HB is performing better than B1 when B3 is used as surrogate for the ground truth, one feature that HB seems to reproduce better than B1 is the heteroscedasticity of the conditional distribution for females, namely the decrease of its standard deviation as age increases.

5.3 Structured cofactors - synthetic examples

5.3.1 Different cofactors

This section tests the hierarchical barycenter on datasets where different subsets have different cofactors. In particular, we consider a dataset divided into two subgroups:

- I_1, where x_i is drawn from $\mathcal{N}(f_1(z), 0.2^2)$, with $z \in \mathbb{R}^2$ and $f_1(z) = 4z^1(1-z^1) + \alpha(z^2 - 1/2)$
- I_2, where x_i is drawn from $\mathcal{N}(f_2(z^1), 0.2^2)$, with $z^1 \in \mathbb{R}$ and $f_2(z^1) = 4z^1(1-z^1)$

The variables z^1 and z^2 are drawn independently from $U_{[0,1]}$. The value of $\alpha \geq 0$ controls how much information is shared between the two subgroups. The goal is to estimate the conditional density of points in I_1 with aid from the information available in I_2. For large values of α, one would expect little gains from using the information contained in I_2 in order to estimate the conditional density of points in I_1, an effect balanced by the fact that the set I_1 contains only 40 sample points, while I_2 contains 100, yielding potential value to its use. Hyper-parameter tuning is performed via cross validation over the points in I_1. Figure 3 displays the KL divergence between the true value of $\rho(x|z)$ and its estimation via the hierarchical barycenter using the points in both I_1 and I_2, and the classical barycenter using only the points in I_1. In order to mitigate the effect of specific realizations of the training set, we average the value of the KL over 30 realizations of the noise and over all the values of z in I_1. As expected, the advantage of using the information contained in I_2 decreases as α increases.

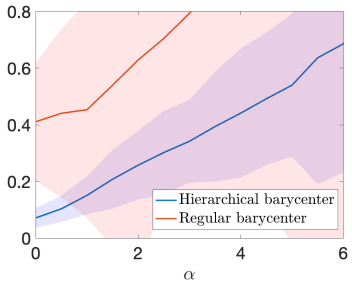


Fig. 3: The solid lines, relative to hierarchical barycenter and the regular barycenter respectively, represent the average KL value over different realizations of the noise. The shaded area corresponds to mean plus minus one standard deviation over 30 different realizations of the noise.

5.3.2 Extrapolation

We modify the experiment in sub-section 5.3.1, setting $\alpha = 1$ and changing the distribution of the z as follows: for the points in I_1, $z_1 \sim U_{[0,0.5]}$ and $z_2 \sim U_{[0,1]}$ while for the points in I_2 we have $z_1 \sim U_{[0,1]}$. The goal is to estimate $\rho(x|z_*)$ underlying the points in I_1 when $z_*^1 > 0.5$ and $z_*^2 \in [0, 1]$. The samples in I_1 are missing information for such estimate, since I_1 does not contain any points with $z_1 \in [0.5, 1]$. Since such information is instead contained in I_2, one can hope to use the points in I_2 to extrapolate $\rho(x|z_*)$. Figure 4 displays the data set (blue points) together with the two points used for the interpolation and extrapolation. Figure 5 compares the estimate of $\rho(x|z_*)$ underlying the model used to generate I_1 for $z_a = (0.85, 0.25)$ and $z_b = (0.25, 0.25)$. As expected, the estimate of $\rho(x|z_b)$ is close to the truth when using only the data in I_1 as opposed to the estimate of $\rho(x|z_b)$ (see second column of the Figure 5). When we use the solution of the regular barycenter problem to estimate $\rho(x|z_a)$ computed only on the data relative I_1 the estimate is far from the truth (lower left panel). We

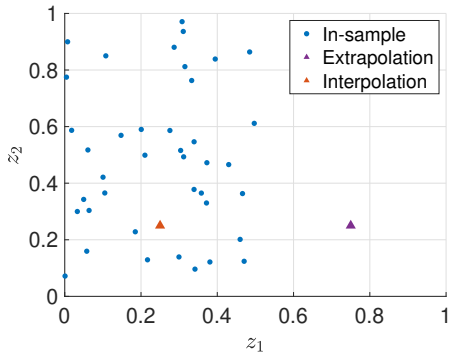
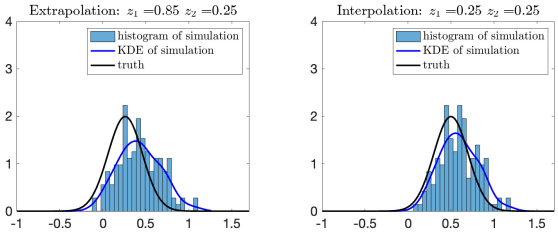
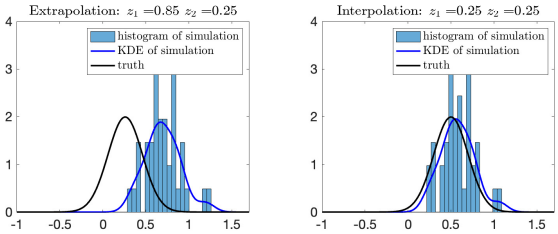


Fig. 4: Scatter plot of in-sample observations relative to (z_1, z_2) in I_1. The triangular markers indicate the target values z^* for which the estimation of $\rho(x|z^*)$ is sought.

need to extend the barycenter to the hierarchical barycenter to also use the points in I_2 (using the procedure described in Section 4) in order to improve substantially the estimate of $\rho(x|z_a)$ (upper left panel). As in the previous section, the hyper-parameter tuning is done through cross-validation on I_1.



(a) Hierarchical barycenter using points in $I_1 \cup I_2$



(b) Regular barycenter using only points in I_1

Fig. 5: The histograms are relative to simulated points through the barycenter (a) and the hierarchical barycenter (b) respectively (see description of the procedure in Section 2). The blue curves are obtained performing kernel density estimation on the simulated points. The ground truth, known in closed form in this case, is represented by the black solid curve.

5.4 Hierarchical barycenter and hidden factors

A significant byproduct of the hierarchical barycenter procedure is the barycenter itself. This is a distribution containing only the variability in x that cannot be explained by the known cofactors z ([18]). Consider the following modification of the example described in Section 5.3.1:

- In I_1, $x = 4z^1(1-z^1) + (z^2 - 1/2) + 0.2\epsilon$,
- In I_2, $x = 4z^1(1-z^1) + 0.2\epsilon$,

with both random variables z^1 and z^2 uniform in $[0,1]$, and

$$\epsilon \sim N(z_{hidden}, 0.25^2), \quad z_{hidden} \sim \frac{1}{3}\delta_{-1} + \frac{2}{3}\delta_1.$$

This example is very similar to the previous ones. The main difference is that the noise involved is no longer Gaussian but it has a bimodal distribution with two modes centered at $+1$ and -1 due to the z_{hidden}, a *latent* random variable, i.e. one whose values are not measured.

We instead are given samples (x_i, z_i^1, z_i^2) and the goal is to see whether the bimodal pattern can be detected by looking at the barycenter or, in other terms, if we can characterize the variability of x that is not due to the known z^1 and z^2. This goal can be achieved via the numerical procedure developed in [18, 19] computing the -classical- barycenter of distributions $\rho(x|z^1, z^2)$ from samples contained in I_1. The question here is to see if we can integrate the information contained in I_2 via the hierachical barycenter. In the following, the numerical experiments are performed with 40 points in I_1 and 50 points in I_2.

Figure 6 shows the histogram relative to the data set $I_1 \cup I_2$. As expected the histogram does not look bimodal since the variability in x derived from z^1 and z^2 hides the one due to z_{hidden}.

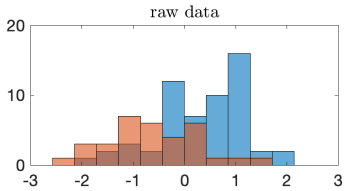


Fig. 6: Histogram relative to $I_1 \cup I_2$, colored according to the hidden binary value.

Figure 7 shows the barycenter obtained with the procedure described in Section 4.

We close this section with a numerical experiment comparing the ordinary barycenter computed without using the samples in I_2 and the hierachical barycenter that instead uses $I_1 \cup I_2$. Figure 8 displays the classical barycenter and, as it can be noticed, it is harder in this case to detect the bimodality of this distribution. This is due to the small size of I_1, showing that the information contained in I_2 in this case improved the detection of z_{hidden}.

6 Conclusions

The problem of inferring from data how a set of variables of interest x depends on covariates z, is frequently formulated under the assumption that the observations consist of a set of identically distributed data pairs $\{z_i, x_i\}$. Yet the population of samples for real data is often strongly heterogeneous; in particular, the kind and number of the covariates $\{z_i\}$ may depend on the observation i. This arises for instance when data sets from various sources are aggregated, each with its own set of observed covariates, when some covariate values have not always been observed or recorded, and when the covariates have a hierarchical structure, so that only a subset of them is defined for each observation. This article proposes and develops a methodology to address data analysis under such scenarios, simulating the conditional distribution $\rho(x|z)$ through an extension of the optimal transport barycenter problem to heterogeneous and not fully observed covariates z.

Applying this methodology to a data set produces, in addition to a simulation of $\rho(x|z_*)$ for any target value z_* –which may itself be incompletely observed– samples $\{y_i\}$ from the barycenter μ of $\rho(:|z)$. The barycenter has additional applications, such as facilitating the detection and identification of hidden covariates. The corresponding variable Y is defined as the one with minimal transportation cost from X among all random variables independent of the covariates Z, where the latter includes additional markers of missing data. The numerical procedure developed in this article uses as a measure of independence the mutual information between Y and Z. This is not the

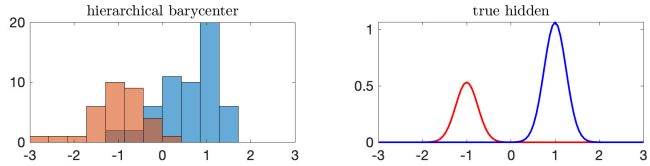


Fig. 7: Left: histogram of hierarchical barycenter computed using both I_1 and I_2, after re-scaling the data. Right: true distribution of ϵ. The two modes with $z_{hidden} = \pm 1$ are indicated via colors. One can see how the hierarchical barycenter makes more evident the bimodality of the true distribution, which is hidden in the original data by the known covariates.

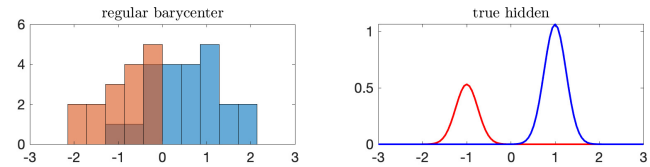


Fig. 8: Left: histogram of barycenter barycenter computed using only the samples in I_1, after taking z-score (just to normalize the scale). Right: true distribution of ϵ. The two modes with $z_{hidden} = \pm 1$ are indicated via colors. The regular barycenter does not show two modes very clearly because of the small sample size

only possible choice: other quantifiers of independence were briefly described in section 2, and still others are under development. The article used the numerical procedure developed to illustrate the broad applicability of the hierarchical barycenter concept through both real and synthetic examples.

Declarations

Funding

The work of E. G. Tabak is supported in part by the Office of Naval Research, through grants # N00014-19-1-2407 and # N00014-22-1-2192.

Author contribution

E. G. Tabak, G. Trigila and W. Zhao, designed the research, analysed the data and wrote the manuscript. W. Zhao performed the simulations.

Appendix A Gradient of the objective function

This section describes an alternative argument to the one developed in section 4.2, for why one needs only consider the derivatives with respect to the first argument of the kernel functions. We focus again for brevity on the second term of (10) and consider its derivative with respect to y_l

$$\frac{\partial}{\partial y_l} \sum_{i \in N_k} \log \left(\frac{1}{N_k} \sum_{j \in I_k} K^y(y_i, y_j) \right) = \frac{\partial}{\partial y_l} \left(\log(\tilde{\rho}(y_l)) + \sum_{i \neq l} \log(\tilde{\rho}(y_i)) \right) \tag{A1}$$

where for simplicity we wrote

$$\tilde{\rho}(y_l) = \frac{1}{N_k} \sum_{j \in I_k} K^y(y_l, y_j).$$

Then

$$\frac{\partial}{\partial y_l} \left(\sum_{i \neq l} \log(\tilde{\rho}(y_i)) \right) = \sum_{i \neq l} \frac{1}{\tilde{\rho}(y_i)} \frac{1}{N_k} \frac{\partial}{\partial y_l} \sum_{j \in I_k} K^y(y_i, y_j) \approx$$

$$\approx \frac{\partial}{\partial y_l} \int \frac{1}{\tilde{\rho}(y)} K^y(y, y_l) \tilde{\rho}(y) dy = 0 \tag{A2}$$

where we used the fact that the sum over i approximates the expected value over y with density $\tilde{\rho}$ and where the last equality follows from the fact that the kernel integrates to 1 for every y_l.

References

[1] Fan, J., Yao, Q., Tong, H.: Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. Biometrika **83**(1), 189–206 (1996)

[2] De Gooijer, J.G., Zerom, D.: On conditional density estimation. Statistica Neerlandica **57**(2), 159–176 (2003)

[3] Tabak, E.G., Trigila, G., Zhao, W.: Conditional density estimation and simulation through optimal transport. Machine Learning **109**(4), 665–688 (2020)

[4] Tabak, E.G., Trigila, G., Zhao, W.: Distributional barycenter problem through data-driven flows. Pattern Recognition **130**, 108795 (2022)

[5] Santambrogio, F.: Optimal transport for applied mathematicians. Birkäuser, NY **55**(58-63), 94 (2015)

[6] Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction vol. 2. Springer, ??? (2009)

[7] Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., Hachiya, H., Okanohara, D.: Conditional density estimation via least-squares density ratio estimation. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 781–788 (2010). JMLR Workshop and Conference Proceedings

[8] Agueh, M., Carlier, G.: Barycenter in the Wasserstein space. SIAM J. MATH. ANAL. **43**(2), 094–924 (2011)

[9] Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, ??? (1999)

[10] Zhang, Y., Yang, Q.: A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering **34**(12), 5586–5609 (2021)

[11] Caruana, R.: Multitask learning. Machine learning **28**, 41–75 (1997)

[12] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proceedings of the IEEE **109**(1), 43–76 (2020)

[13] Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big data **3**, 1–40 (2016)

[14] Argyriou, A., Evgeniou, T., Pontil, M.: Multi-task feature learning. Advances in neural information processing systems **19** (2006)

[15] Mejjati, Y.A., Cosker, D., Kim, K.I.: Multi-task learning by maximizing statistical dependence. In: Proceedings of the IEEE Conference on Computer Vision and

Pattern Recognition, pp. 3465–3473 (2018)

[16] Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics **69**(6), 066138 (2004)

[17] Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D.: Mutual information neural estimation. In: International Conference on Machine Learning, pp. 531–540 (2018). PMLR

[18] Tabak, E.G., Trigila, G.: Explanation of variability and removal of confounding factors from data through optimal transport. Communications on Pure and Applied Mathematics **71**(1), 163–199 (2018)

[19] Yang, H., Tabak, E.G.: Conditional density estimation, latent variable discovery, and optimal transport. Communications on Pure and Applied Mathematics **75**(3), 610–663 (2022)

[20] Essid, M., Tabak, E.G., Trigila, G.: An implicit gradient-descent procedure for minimax problems. Mathematical Methods of Operations Research **97**(1), 57–89 (2023)

[21] Hanzely, S., Kamzolov, D., Pasechnyuk, D., Gasnikov, A., Richtarik, P., Takac, M.: A damped newton method achieves global and local quadratic convergence rate. Advances in Neural Information Processing Systems **35**, 25320–25334 (2022)

[22] Moré, J.J.: The levenberg-marquardt algorithm: Implementation and theory. In: Watson, G.A. (ed.) Numerical Analysis, pp. 105–116. Springer, Berlin, Heidelberg (1978)

[23] Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Routledge, ??? (2018)

[24] Duong, T., Hazelton, M.L.: Cross-validation bandwidth matrices for multivariate kernel density estimation. Scandinavian Journal of Statistics **32**(3), 485–506 (2005)

[25] Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22**(1), 79–86 (1951)

[26] Robert, C.P.: Intrinsic losses. Theory and decision **40**, 191–214 (1996)