Conditional density estimation and simulation through optimal transport

Esteban G. Tabak¹ · Giulio Trigila² · Wenjun Zhao¹

Received: 11 March 2019 / Revised: 24 October 2019 / Accepted: 11 December 2019 /

Published online: 13 January 2020

© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2020

Abstract

A methodology to estimate from samples the probability density of a random variable x conditional to the values of a set of covariates $\{z_I\}$ is proposed. The methodology relies on a data-driven formulation of the Wasserstein barycenter, posed as a minimax problem in terms of the conditional map carrying each sample point to the barycenter and a potential characterizing the inverse of this map. This minimax problem is solved through the alternation of a flow developing the map in time and the maximization of the potential through an alternate projection procedure. The dependence on the covariates $\{z_I\}$ is formulated in terms of convex combinations, so that it can be applied to variables of nearly any type, including real, categorical and distributional. The methodology is illustrated through numerical examples on synthetic and real data. The real-world example chosen is meteorological, forecasting the temperature distribution at a given location as a function of time, and estimating the joint distribution at a location of the highest and lowest daily temperatures as a function of the date.

Keywords Conditional density estimation · Optimal transport · Wasserstein barycenter · Explanation of variability · Confounding factors · Sampling · Uncertainty quantification

1 Introduction

A very general question in data analysis is to determine how the values of a set of variables x depend on others z , from a set of available observations (x^i, z^i) . Since typically the factors z considered do not fully determine x , the best answer one can hope for adopts the form of

Editor: Pradeep Ravikumar.

✉ Giulio Trigila
giulio.trigila@baruch.cuny.edu

Esteban G. Tabak
tabak@cims.nyu.edu

Wenjun Zhao
wenjun@cims.nyu.edu

¹ Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, USA

² Baruch College, CUNY, 55 Lexington Avenue, New York, NY 10035, USA

a conditional probability distribution, which we shall write in terms of a probability density $\rho(x|z)$. Examples include the effect of a medical treatment, where x comprises measurements of the health of a patient after a treatment (concentration of glucose in the bloodstream, blood pressure, heart rate) and z covariates such as the treatment (type, dosage), the patient (age, weight, habits), lab test results, and others (location, season, social environment). We will illustrate the procedure below with a meteorological example, forecasting the temperature in one site in terms of covariates such as time of day, season and current conditions elsewhere, and estimating the date-dependent joint distribution of highest and lowest daily temperatures. Examples abound in any data-rich field, such as economics and public health.

Among the main challenges that one encounters in conditional density estimation are the following:

1. The problem is highly constrained, as $\rho(x|z)$ needs to be non-negative and integrate to one for all values of z . Addressing this through a parametric approach where the ρ have a specific form with parameters that depend on z (for instance Gaussians with z -dependent mean and covariance) severely restricts the scope of the estimation.
2. The data is scarce, as for each value of z there is typically either a single observation x^i or none. In order to estimate $\rho(x|z)$ for each value of z separately by standard methods, such as Kernel density estimation, one would require a sizable collection of samples for each value of z .
3. The function sought is complex, as the probabilities are typically non Gaussian and their dependence on the covariates is nonlinear. This again excludes most parametric approaches. Moreover, the covariates z can be many and of multiple types (real, vectorial, categorical, distributions, pictures.) Thus one needs to represent multivariable functions in a treatable form, and do it in a way general enough that can handle nearly any data type.

This article proposes a methodology to estimate conditional probabilities based on optimal transport or, more specifically, on a data-based version of the continuous extension of the Wasserstein barycenter problem. Throughout this article, all distributions/measures will always be assumed to be absolutely continuous with respect to the Lebesgue measure, and they will be represented by their density. The difficulties above are addressed as follows:

1. The conditional distribution $\rho(x|z)$ is estimated by mapping it to another distribution $\mu(y)$ (the barycenter of the $\rho(x|z)$) through a z -dependent transformation $y = T(x; z)$, hence all the infinitely many constraints are satisfied automatically if this transformation is one-to-one for all values of z . We will in fact compute both the map and its inverse $x = T^{-1}(y; z)$, given by the gradient (in x) of a convex z -dependent potential $\psi(x; z)$.
2. Making ψ depend smoothly on z effectively links nearby values of z together. Thus the estimation of $\rho(x|z^*)$ is informed by observations with z^i close to z^* . In fact, as we will see below, this closeness needs not be defined by a single distance in z -space, but can be decomposed into distances for each factor z_l . Then the estimation of the dependence of x on a particular factor z_l is informed by all observations z^i with nearby values of z_l , even if the other factors are not close at all. This effectively mitigates the curse of dimensionality in z -space.
3. We use a low-rank tensor factorization, variable separation procedure developed in Tabak and Trigila (2018a) to reduce multivariate functions to sums of products of functions of a single variable. These in turn are approximated as convex combinations of their values on prototypes (Chenyue and Tabak 2018). Since prototypical analysis applies to any space provided with an inner product, the procedure is nearly blind to the type of the various factors z_l .

Conditional probability estimation underlines any data problem where the dependence of some variables on others is sought. Least-square regression can be thought of as a particular instance, where one seeks only the conditional expected value of the distribution $\rho(x|z)$. This article extends the attributable component methodology (Tabak and Trigila 2018a), which is a form of nonlinear regression, to full conditional density estimation. This approach differs considerably from existing methodologies for conditional density estimation, most of which are based on kernel estimators, starting with the work in Rosenblatt (1969). This line of work was further developed in Fan et al. (1996), Hyndman et al. (1996) and Escobar and West (1995), then Bashtannyk and Hyndman (2001) and Fan and Yim (2004) addressed the issue of finding an efficient data-driven bandwidth selection procedure, and De Gooijer and Zerom (2003) enforced the positivity constraint of the estimated conditional density by means of a slight modification of the Nadaraya–Watson (Nadaraya 1964; Watson 1964) smoother. The use of a dual-tree-based approximation (Gray and Moore 2001) of the likelihood function allowed for bandwidth selection in higher dimensional spaces (Holmes et al. 2012). The work in Dutordoir et al. (2018) uses Gaussian processes in a Bayesian approach in which the model's input is extended with latent variables with posterior distribution trained through a neural network.

By contrast, the methodology of this article estimates conditional densities via conditional maps. A map-based density estimation was previously developed in Tabak and Vanden-Eijnden (2010) and Tabak and Turner (2013), with the map computed through a flow in phase-space that ascended the likelihood of the data. A different fluid-like flow formulation based on optimal transport was proposed in Trigila and Tabak (2016). Both flow formulations were developed in the context of single density estimation, while the work in Agnelli et al. (2010) performed clustering and classification by extending the flow methodology in Tabak and Vanden-Eijnden (2010) to a finite number of marginals, which can be thought of as a probability estimation conditioned to a categorical factor. This article considers instead the general conditional probability problem, with factors that can be multiple and continuous, making use of a data-based formulation of the optimal transport barycenter problem.

This article is structured as follows. After this introduction, Sect. 2 describes conditional density estimation as a Wasserstein barycenter problem, and develops a sequence of formulations of the latter leading to a sample-based minimax formulation suitable to the form of the available data. Section 3 relates this formulation to the attributable component estimation of conditional expectation, showing how the latter arises from the former when the maps are restricted to rigid translations. Section 4 then extends the attributable methodology so that it can be applied to estimate and simulate full conditional densities. Section 5 exemplifies the procedure through its application to synthetic and meteorological data. Finally, Sect. 6 summarizes the work and discusses possible extensions.

2 Problem setting

Given samples $\{x^i, z_1^i, \dots, z_L^i\}$ of a variable of interest x and covariates z_l , we seek to estimate or simulate the conditional probability distribution

$$\rho(x|z_1, \dots, z_L).$$

Here $x \in \mathbb{R}^d$, and each of the factors z_l can be of nearly arbitrary type, including real scalars or vectors, categorical variables, probability densities and pictures.

Fig. 1 Conditional density simulation as a Wasserstein barycenter problem. For easy visualization, this example has $x \in \mathbb{R}^2$ and a single categorical covariate z with 3 possible values. On the left, the samples x^i of the conditional probabilities $\rho_i(x) = \rho(x|z = i)$ are mapped through a z -dependent map $y = T(x^i; z^i)$ to samples y^i of the z -independent barycenter $\mu(y)$. In order to produce additional samples $\tilde{x}^i$ of $\rho(x|z^*)$ for any specific value z^* of z , one maps back the y^i under the inverse of $T(x; z^*)$

We pose this conditional density estimation as a Wasserstein barycenter problem (Agueh and Carlier 2011), whose solution pushes the probability densities $\rho(x|z_1, \dots, z_L)$ to their barycenter $\mu(y)$ through a z -dependent map $y = T(x; z_1, \dots, z_L)$ with inverse $x = T^{-1}(y; z_1, \dots, z_L)$. Then an estimation of μ provides the desired estimation of the $\rho(x|z)$ via the change of coordinates formula. More directly, the simulation of μ using all the $y^i = T\{x^i; z_1^i, \dots, z_L^i\}$ followed by the map $x = T^{-1}(y^i; z^*)$ allows us to immediately simulate $\rho(x|z^*)$ under any choice z^* for the factors z . This formulation is illustrated in Fig. 1.

As a simple conceptual illustration, consider estimating the dependence $\rho(x|z)$ of the blood pressure x on the age z from a set of samples (x^i, z^i) . After finding the conditional map $T(x; z)$, one obtains samples $y^i = T(x^i; z^i)$ of the barycenter $\mu(y)$ of the $\rho(x|z)$. In order to simulate the distribution $\rho(x|z^*)$ of blood pressure for a particular age z^* , one produces samples thereof $\tilde{x}^i = T^{-1}(y^i; z^*)$. Notice that this produces n samples of a distribution $\rho(x|z^*)$ for which we may not have had any observation to start with!

2.1 The Wasserstein barycenter problem: a sequence of formulations

The original optimal transport problem is posed in terms of densities, a property inherited by the barycenter problem (Agueh and Carlier 2011). Yet we do not know the conditional densities $\rho(x|z)$, but only a set of samples $\{x^i, z^i\}$ thereof. This subsection highlights a sequence of formulations of the optimal transport barycenter problem, to obtain one that seeks the family of maps $T(x; z)$ directly from the set of data pairs $\{x^i, z^i\}$ and a given cost function $c(x, y)$.

1. Monge formulation

The following formulation of the barycenter problem follows the original optimal transport problem due to Monge (1781), extended to situations with possibly infinitely many marginals (Pass 2013). Given a family of probability densities $\rho(x|z)$, an extra probability density $\nu(z)$ underlying the factors z and a transportation cost function $c(x, y)$, find the distribution $\mu(y)$ and the corresponding family of maps $y = T(x; z)$ pushing forward $\rho(x|z)$ to $\mu(y)$ so that the total transportation cost

$$C(T, \mu) = \int \left[\int c(x, T(x; z)) \rho(x|z) dx \right] v(z) dz$$

is minimized:

$$[T, \mu] = \arg \min C(T, \mu), \text{ s.t. } \forall z: X \sim \rho(\cdot|z) \Rightarrow Y = T(x; z) \sim \mu. \quad (1)$$

The assumption that $v(z)$ and $\rho(x|z)$ derive from probability densities was made just to give a concrete form to $C(T, \mu)$. Nothing changes here or in what follows if, for more general probability distributions, we define

$$C(T, \mu) = E [c(X, T(X; Z))],$$

since v and ρ appear only in the calculation of the expected value of functions.

2. Kantorovich formulation

For our data problem, we do seek a family of maps $T(x; z)$ as above. As noted in Trigila and Tabak (2016), relaxing these to conditional couplings $\pi(x, y|z)$ according to what described in Kantorovich (1942, 1948) leads to a dual formulation, which will allow us to replace the conditional $\rho(x|z)$ and $v(z)$ by samples thereof. In terms of these conditional couplings, the cost C to minimize adopts the form

$$C(\pi, \mu) = \int \left[\int c(x, y) \pi(x, y|z) dx dy \right] v(z) dz,$$

and the problem becomes

$$[\pi, \mu] = \arg \min C(\pi, \mu), \quad \text{such that } \pi, \mu \geq 0, \quad \text{and} \\ \forall z: \int \pi(x, y|z) dy = \rho(x|z), \quad \int \pi(x, y|z) dx = \mu(y). \quad (2)$$

From the very definition of the barycenter, we should expect the random variables Y and Z to be independent. The map $y = T(x; z)$ is designed precisely to remove the variability in X due to the covariates in z ; if there were any dependence left between Y and Z , such removal would not have been fully achieved. We can verify independence directly from the second constraint in (2). If $\Phi(y, z)$ is the joint distribution of Y and Z , and $P(y|z)$ is the conditional distribution of Y given z , we have that

$$\Phi(y, z) = P(y|z) v(z) = \left[\int \pi(x, y|z) dx \right] v(z) = \mu(y) v(z),$$

confirming that Y and Z are indeed independent.

3. Dual Kantorovich formulation

The problem in (2) is an infinitely dimensional linear programming problem. Introducing Lagrange multipliers $\phi(x, z)$ and $\psi(y, z)$ for the first and second integral constraints respectively, and the Lagrangian

$$L(\pi, \mu, \phi, \psi) = C(\pi, \mu) \\ - \int \left[\int \pi(x, y|z) dy - \rho(x|z) \right] \phi(x, z) dx v(z) dz \\ - \int \left[\int \pi(x, y|z) dx - \mu(y) \right] \psi(y, z) dy v(z) dz,$$

yields the alternative formulation

$$\min_{\pi, \mu \geq 0} \max_{\phi, \psi} L(\pi, \mu, \phi, \psi).$$

Performing the minimization first yields the dual problem

$$\begin{aligned} \max_{\phi, \psi} \int \left[\int \phi(x, z) \rho(x|z) dx \right] v(z) dz, \quad \text{such that} \\ \phi(x, z) + \psi(y, z) \leq c(x, y), \quad \forall y: \int \psi(y, z) v(z) dz \geq 0. \end{aligned} \quad (3)$$

4. Conversion to a minimax problem through conjugate duality

In problem (3), if ψ is given, it follows that

$$\phi(x, z) = \min_y [c(x, y) - \psi(y, z)],$$

so the problem can be cast in terms of ψ alone:

$$\begin{aligned} \max_{\psi} \int \left[\int \min_y [c(x, y) - \psi(y, z)] \rho(x|z) dx \right] v(z) dz, \\ \text{where } \forall y: \int \psi(y, z) v(z) dz = 0, \end{aligned}$$

or

$$\begin{aligned} \max_{\psi} \min_T \int [c(x, T(x; z)) - \psi(T(x; z), z)] \gamma(x, z) dx dz, \\ \forall y: \int \psi(y, z) v(z) dz = 0, \end{aligned} \quad (4)$$

where $\gamma(x, z) = \rho(x|z)v(z)$ is the joint distribution of X and Z . Again, for distributions that cannot be described in terms of densities, we have

$$\begin{aligned} \max_{\psi} \min_T E_{\gamma} [c(X, T(X; Z)) - \psi(T(X; Z), Z)], \\ \forall y: E_{\nu} [\psi(y, Z)] = 0. \end{aligned} \quad (5)$$

Notice that, in the solution to this dual problem, the random variables $Y = T(X; Z)$ and Z are still independent. Otherwise, the dual problem would be unbounded, as we could find a function $\psi(y, z)$ such that $\forall y: E_{\nu} [\psi(y, Z)] = 0$, but $E_{\gamma} [\psi(T(X; Z), Z)] \neq 0$. Multiplying this function by an arbitrary constant we could make the objective function arbitrarily large. But the dual problem can only be unbounded if the primal is unfeasible, which is not the case for the optimal transport barycenter problem.

It follows from this independence that there is no duality-gap, as the optimal objective function over those functions $Y = T(X; Z)$ such that Y and Z are independent equals $\min E_{\gamma} [c(X, T(X; Z))]$, which agrees with the solution to the primal problem.

5. Sample based formulation

The fact that the distributions γ and ν appear in problem (5) only in the calculation of the expected value of functions, allows us to switch to a sample-based formulation, where these expected values are replaced by the corresponding empirical means over the samples provided. In terms of these samples (x^i, z^i) , the problem becomes

$$\max_{\psi} \min_{\{y^i\}} \sum_i [c(x^i, y^i) - \psi(y^i, z^i)], \quad \forall y: \sum_i \psi(y, z^i) = 0, \quad (6)$$

where we have written y^i for $T(x^i; z^i)$.

The problem in (6) is formulated exclusively in terms of available data, the samples (x^i, z^i) . To pose the problem completely requires specifying over which families of functions are the maps $T(x; z)$ and the potentials $\psi(y, z)$ optimized. The choice of families that we make below is based on the fact that, rather than studying the empirical distribution determined by the samples, we think of these as drawn from an underlying conditional distribution with smooth density $\rho(x|z)$. This smoothness is indirectly characterized by the choice of the families of functions from which we choose the maps $T(x; z)$. See for instance in Caffarelli (2003) how the map's smoothness relates to a finite second moment and compact support of the source and target distributions. Enforcing the map's smoothness requires methods for the solution of the barycenter problem different from those used in the discrete setting where the $\rho(x|z)$ are assumed to be convex linear combinations of Dirac deltas (Solomon et al. 2015).

Cost For concreteness, we will adopt the canonical quadratic cost

$$c(x, y) = \frac{1}{2} \|x - y\|^2, \quad (7)$$

though much of what follows can be extended to more general cost functions.

3 Conditional expectations

In this section, we solve a scaled-down problem: instead of the conditional probability $\rho(x|z)$, we seek its conditional expectation $\bar{x}(z) = E_{\rho(x|z)}[X]$. We do this in order to show how the *attributable component* methodology (Tabak and Trigila 2018a) fits into the framework developed here. This will allow us in Sect. 4 to extend the low-rank factorizations used in attributable components to capture the full conditional dependence of X on Z .

The minimization over y^i in (6) yields

$$x^i = y^i - \nabla_y \psi(y^i, z^i). \quad (8)$$

In particular, if we restrict consideration to functions ψ that are linear in y ,

$$\psi(y; z) = -y \cdot \zeta(z), \quad (9)$$

we have

$$y^i = x^i - \zeta(z^i), \quad (10)$$

a z -dependent rigid translation.

Substituting (7), (9) and (10) into (6), yields the following variational problem for $\zeta(z)$:

$$\max_{\zeta} \sum_i \left[\frac{1}{2} \|\zeta(z^i)\|^2 + (x^i - \zeta(z^i)) \cdot \zeta(z^i) \right], \quad \sum_i \zeta(z^i) = 0,$$

or

$$\min_{\zeta} \sum_i \frac{1}{2} \|\zeta(z^i) - x^i\|^2, \quad \sum_i \zeta(z^i) = 0. \quad (11)$$

Hence $\zeta(z)$ is the empirical conditional expectation of $X|z$, displaced so that its expected value over Z vanishes:

$$\zeta(z) = \bar{x}(z) - \bar{x}, \quad \bar{x} = \frac{1}{m} \sum_i \bar{x}(z^i). \quad (12)$$

For convenience, we can remove the empirical mean of x from the observations ab initio, in which case $\zeta(z) = \bar{x}(z)$, and we do not need to take into account the constraint in (11), as it is satisfied automatically (if allowed by the family of functions $\zeta(z)$ considered.)

3.1 Attributable components

If one leaves the function $\zeta(z)$ completely unrestricted, the solution to (11) is given by the trivial $\zeta(z^i) = x^i$ when all z^i 's are different, and by $\zeta(z) = \text{mean}(x^i)$ over the x^i such that $z^i = z$, when some z^i are repeated. This solution is fine when the factors z are categorical and the number of their combinations is small compared to the number of observations, but otherwise it may severely overfit the data and it is not informative on the value of $\zeta(z)$ for values of z not in the dataset.

One could propose a parametric ansatz, such as

$$\zeta(z) = \sum_k \beta_k \zeta_k(z),$$

with $\{\zeta_k(z)\}$ a given set of functions (the “features”), and optimize over the parameters β . Instead, Tabak and Trigila (2018a) proposed the low-rank tensor factorization (or separated variable approximation, depending on whether one approaches it through linear algebra or multivariable calculus)

$$\zeta(z) = \sum_{k=1}^r \prod_{l=1}^L \zeta_l^k(z_l). \quad (13)$$

This decomposes the multivariable function $\zeta(z)$ into r components, each a product of single-variable functions $\zeta_l^k(z_l)$. Here by “single-variable” we mean “single z_l ”, as each variable z_l can be of virtually any type, including vectorial.

Then each of these functions is approximated by the convex combinations of an array of unknown values V to be optimized:

$$\zeta_l^k(z_l^i) = \sum_j \alpha(l)_i^j V(l)_j^k,$$

where the $\alpha(l)_i^j$ are computed once at the beginning of the procedure and satisfy

$$\alpha(l)_i^j \geq 0, \quad \sum_j \alpha(l)_i^j = 1.$$

For example, if z_l is a single real-variable, given a grid $\{z_{gl}^j\}$ (not necessarily uniform), and interpret the $V(l)_j^k$ as $\zeta_l^k(z_{gl}^j)$, the value of the function on the grid points, and $\alpha(l)_i^j$ as the piecewise linear functions that interpolate z_l^i on the grid. Notice that the $\alpha(l)_i^j$ can be computed straightforwardly for each value of z_l^i once a grid is chosen, and that they satisfy the convexity requirements above. Moreover, in this scenario only two of the $\alpha(l)_i^j$ are non-zero for each l and i . Many criteria can be used for the choice of grid points. In the numerical examples below, we have used an equispaced grid. Another choice is use an adaptive grid with mesh size inversely proportional to the local density of points. The purpose of adopting a grid is to approximate the one dimensional function $\zeta(z)$. The approximation error incurred by regular grids, or by Chebyshev's points to handle boundary effects optimally, is a central topic in function approximation, which is discussed, for instance, in Trefethen (2000).

If the z_l is instead categorical, then the $\{z_{gl}^j\}$ contain all the values that z_l may adopt, and we simply have $\alpha(l)_i^j = 1$ when $z_l^i = z_{gl}^j$, and zero otherwise. More generally, if the value z_l^i of covariate l for observation i is not known, then the corresponding $\alpha(l)_i^j$ represents the

probability that it adopt the value z_{gl}^j . The same framework applies to more general types of covariates (probability distributions, photographs) via prototypal analysis (Chenyue and Tabak 2018): given a set of n samples z_l^i of z_l , we seek $m_l \ll n$ prototypes

$$z_{gl}^j = \sum_{i=1}^n \beta(l)_j^i z_l^i, \quad \beta(l)_j^i \geq 0, \quad \sum_{i=1}^n \beta(l)_j^i = 1$$

such that the z_l^i can be well approximated by convex combinations

$$z_l^i \approx \sum_{j=1}^{m_l} \alpha(l)_i^j z_{gl}^j$$

of a subset of the $\{z_{gl}^j\}$ that are local, i.e. not far from z_l^i . The procedure to find the α can be formulated exclusively in terms of inner products in z -space, so it applies to any space where inner products are defined.

Finally, we add to (11) a penalty term to enforce the smoothness or control the variability of the functions $\zeta_l^k(z_l)$:

$$\begin{aligned} \min_V \left\{ \sum_i \frac{1}{2} \left\| x^i - \sum_k \prod_{l=1}^L \sum_j \alpha(l)_i^j V(l)_j^k \right\|^2 \right. \\ \left. + \sum_{l=1}^L \lambda_l \sum_k \left(\prod_{b \in [1:L], b \neq l} \|V(b)^k\|^2 \right) V(l)^{k^t} C^l V(l)^k \right\}. \end{aligned} \quad (14)$$

Typically, the quadratic form $V(l)^{k^t} C^l V(l)^k$ is chosen, for real z_l , as the square norm of a finite difference approximation to the first or second derivatives of $\zeta_l^k(z_l)$, and for categorical z_l , as the variance of $\zeta_l^k(z_l)$. The prefactor $\prod_{b \in [1:L], b \neq l} \|V(b)^k\|^2$ is included to balance the two terms in the objective function. Otherwise, the smoothness requirement on one $\zeta_l^k(z_l)$ could be bypassed by making that ζ_l^k smaller by a constant factor while keeping $\zeta(z)$ constant by enlarging other ζ_b^k less constrained. The objective function in (14) is quadratic in each matrix $V(l)$, so it can be optimized through an alternate-direction procedure, in which one minimizes over one $V(l)$ at the time through the solution of a linear system.

In order to estimate the conditional expectation not of X but of some function $F(X)$, it suffices to replace x^i by the corresponding $F(x^i)$. In particular, calculating first $\bar{x}(z)$, subtracting it from the observations and taking products among the resulting zero-mean quantities, one captures the conditional second order structure of the data or covariance, and taking the square of their Fourier coefficients and adding the mode as an explanatory factor, the conditional energy spectrum.

4 The full barycenter problem

This section addresses the numerical solution to the full minimax problem in (6). In order to move from conditional expectation to the full conditional density estimation, one must allow a nonlinear dependence of ψ on y . Then the expression in (8) determines y^i only implicitly, so one cannot replace it straightforwardly in (6) as with (10).

We solve this problem as in Tabak and Trigila (2018c), through an alternate iterative procedure that updates the values of y for fixed $\zeta(z)$ and vice versa, linearizing each time

the y dependence of ψ at the current values of y^i . This can be thought of as a primal–dual approach, which updates in one step the dual variable ϕ and in the other the primal map $T(x; z)$. In order to move from the rigid translations of attributable components to more general maps, we expand the factorization in (13) from only the z -dependence to both z and y :

$$\psi(y, z) = - \sum_{k=1}^r \zeta^k(z) \eta_k(y),$$

leaving temporarily aside how each of the $\zeta^k(z)$ and $\eta_k(y)$ is defined. Again, this can be thought of as a truncated separable variable approximation to a multivariate function, or as an extension to continuous indices of the low-rank factorization of matrices

$$A_i^j \approx \sum_k u_i^k v_k^j.$$

Since the equation in (8) determines y only implicitly, we replace it by the local approximation

$$y^i = x^i + \nabla_y \psi(y, z^i) \Big|_{y=y_n^i} = x^i - \sum_k \zeta^k(z^i) J_k^i, \quad (15)$$

where y_n^i represents the state of y^i at step n —as opposed to the step $n + 1$ at which y^i is being presently computed—and $J_k^i = \nabla_y \eta_k(y) \Big|_{y=y_n^i}$. Consistently, we approximate

$$\psi(y^i, z^i) \approx - \sum_k \left(\eta_k(y_n^i) + J_k^{i^t} \left(x^i - \sum_c \zeta^c(z^i) J_c^i - y_n^i \right) \right) \zeta^k(z^i). \quad (16)$$

Replacing into (6) yields

$$\begin{aligned} \max_{\zeta} \sum_i \left[\frac{1}{2} \left\| \sum_k J_k^i \zeta^k(z^i) \right\|^2 \right. \\ \left. + \sum_k \left(\eta_k(y_n^i) + J_k^{i^t} \left(x^i - \sum_c \zeta^c(z^i) J_c^i - y_n^i \right) \right) \zeta^k(z^i) \right], \end{aligned}$$

or

$$\min_{\zeta} \sum_i \left[\frac{1}{2} \left\| \left(x^i - y_n^i \right) - \sum_k J_k^i \zeta^k(z^i) \right\|^2 - \sum_k \eta_k(y_n^i) \zeta^k(z^i) \right]. \quad (17)$$

subject to the conditions

$$\forall y \sum_k \left(\eta_k(y) \sum_i \zeta^k(z^i) \right) = 0. \quad (18)$$

For functions $\eta_k(y)$ that are independent, the conditions in (18) are equivalent to

$$\forall k \sum_i \zeta^k(z^i) = 0. \quad (19)$$

We impose this stronger requirement, easier to implement, even when the independence of the η_k does not hold, without loss of generality, since the non-independence of the η_k makes the choice of $\zeta(k)$ non-unique, with degrees of freedom that exactly balance the extra requirements in (19).

As before, we propose for ζ^k the factorization

$$\zeta^k(z) = \prod_{l=1}^L \zeta_l^k(z_l), \quad \zeta_l^k(z_l^i) = \sum_j \alpha(l)_i^j V(l)_j^k,$$

and add to (17) a penalty term of the form

$$\frac{1}{2} \sum_{l=1}^L \lambda_l \sum_k \left(\prod_{b \in L, b \neq l} \|V(b)^k\|^2 \right) V(l)^{k^t} C^l V(l)^k. \quad (20)$$

Yet there is one more consideration to make: for the approximations (15) and (16) to be valid, we need y^i and y_n^i to be close to each other, i.e. to make the optimization steps small. To this end, we add a second penalty of the form

$$\frac{1}{2} v_z \sum_{l=1}^L \|V(l) - V(l)_n\|^2, \quad (21)$$

where $V(l)_n$ stands for the current value of $V(l)$.

The procedure above describes how the $\zeta^k(z)$ are updated. Regarding the $\eta_k(y)$, there are two possibilities: they can be given externally, with form and number depending on the complexity of the maps sought, or updated as well through the maximization in (6), proposing for them either a parametric representation or a factorization similar to the one for ζ^k :

$$\eta_k(y) = \prod_{j=1}^n \eta_j^k(y_j).$$

For the numerical examples of this article, we adopted a parametric proposal of the form

$$\eta_k(y) = \sum_s \beta_k^s G_s(y), \quad (22)$$

with given functions $G_s(y)$, thus extending the attributable component procedure, which had only the function $G(y) = y$. Then we add to the objective function the penalty term

$$\frac{1}{2} v_y \|\beta - \beta_n\|^2, \quad (23)$$

and denote by O the objective function resulting from the sum of (17), (20), (21) and (23).

$$\begin{aligned} O(V, \beta) = & \sum_i \left[\frac{1}{2} \left\| \left(x^i - y_n^i \right) - \sum_k J_k^i \zeta^k(z^i) \right\|^2 - \sum_k G_k^i \zeta^k(z^i) \right] \\ & + \frac{1}{2} \sum_{l=1}^L \lambda_l \sum_k \left(\prod_{b \in L, b \neq l} \|V(b)^k\|^2 \right) V(l)^{k^t} C^l V(l)^k \\ & + \frac{1}{2} v_z \sum_{l=1}^L \|V(l) - V(l)_n\|^2 + \frac{1}{2} v_y \|\beta - \beta_n\|^2, \end{aligned} \quad (24)$$

where

$$\zeta^k(z^i) = \prod_{l=1}^L \sum_j \alpha(l)_i^j V(l)_j^k, \quad (25)$$

$$G_k^i = \sum_k \beta_k^s G_s(y_n^i), \quad (26)$$

and

$$J_k^i = \sum_k \beta_k^s \nabla_y G_s(y_n^i). \quad (27)$$

The procedure goes as follows: given the samples $\{x^i, z_1^i, \dots, z_L^i\}$, the grids $\{z_l^g\}$ with corresponding interpolating parameters $\alpha(l)_i^j$ and penalty matrix C^l , the number r of components sought, the proposed set of functions $G_s(y)$, and the penalty coefficients λ, ν ,

1. Initialize $y_0^i = x^i$, $\beta_k^s = 0$, $V(l)_j^k$ arbitrarily.
2. Iterate to convergence the following procedure:
 - (a) For each l , minimize O over $V(l)$ subject to (19) and update the $\{y^i\}$ via (15).
 - (b) Minimize O over the $\{\beta_k^s\}$ and update the $\{y^i\}$ via (15).

The minimization over each of the $V(l)$ has the general form of a quadratic optimization with linear constraints:

$$\min_x \frac{1}{2} x^t A x + B x \quad \text{subject to} \quad C x = 0.$$

Introducing a vector of Lagrange multipliers λ , this constrained optimization reduces to solving the linear system

$$\begin{pmatrix} A^t & C^t \\ C & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} -B^t \\ 0 \end{pmatrix}.$$

Regarding the choice of the hyperparameters λ , ν_z and ν_y , numerical experiments, not included here for brevity, display close to zero sensitivity to their choice. This lack of sensitivity can be explained as follows: the parameter λ enforces smoothness of the z -dependence of the map. Yet this smoothness is already present in the data, so even without penalization it will hold, unless the grids are so fine as to over-resolve the noise in the observations. The parameters ν_z and ν_y , on the other hand, are intended to limit the step size, so as not to violate the value in the approximation of functions by the first two terms in their Taylor expansions. Yet, for the quadratic cost adopted, the problem is locally convex, so the parameters will not move very far even without penalization parameters. We include those parameters, rather than setting them to zero, to address the infrequent scenarios where the combination of few sample points and significant outliers may disrupt the local convexity of the objective function.

5 Examples

In order to illustrate the methodology proposed, we use one simple synthetic example and a more complex, data-based meteorological one. In addition, in the synthetic example, we compare our approach with conditional density estimator (C-KDE).

Fig. 2 Original data x versus filtered data y as a function of the covariate z on the left panel, and their PDFs (marginalized over z) on the right. One sees how the z -dependence of the distribution of x is gone in y , and how this results in a reduced total variability

5.1 Synthetic example, comparison with KDE

For visual clarity, we choose a synthetic example with a one-dimensional variable x depending on a single, one dimensional real variable z . However, we make both the conditional probability densities $\rho(x|z)$ and their dependence on z highly nonlinear.

To generate the data, we choose a distribution $\nu(z)$ uniform in the interval $[0, 1]$, and draw 4000 random samples $\{z^i\}$ from it. For $\rho(x|z)$, we choose the third power of a Gaussian:

$$\tilde{x}(z) \sim \mathcal{N}(\sin(2\pi(z - 0.5)), 0.02), \quad x(z) = \tilde{x}(z)^3.$$

This distribution has the advantage of being both highly nonlinear and easily sampleable, as for each z^i we can draw one $\tilde{x}(z)$ from the corresponding Gaussian distribution and then cube it to produce x^i .

The parameters that we have used for the algorithm are the following: for the features $\eta_k(y)$, monomials up to 5th order in the variable y , each repeated twice, giving a total of $r = 12$ components. The z -dependence of each component is determined through a piecewise linear function over a uniform 30 point grid. Rather than tuning the penalization coefficient λ by cross-validation, we picked an arbitrary value $\lambda = 3$, as experiments showed little sensitivity of the results to values of λ within a range spanning two orders of magnitude.

Figure 2 shows the x^i displayed in terms of the z^i , and the corresponding filtered y^i from the barycenter. We can see the high z -variability of $\rho(x|z)$, in mean, variance and skewness, which is absent in the barycenter $\mu(y)$. The pdfs of the marginal $\int \rho(x|z)\nu(z)dz$ and of $\mu(y)$ show the decrease in variability of the latter, as all variability due to z has been filtered out by the procedure.

Next we simulate the $\rho(x|z)$ for various values of z via $T^{-1}(y^i; z)$, and compare the results with the true $\rho(x|z)$ underlying the data. The left panel of Fig. 3 shows this comparison for two values of z , and the right panel the comparison of the empirical mean, standard deviation and skewness of the recovered data with their true values. Notice that there is no sample x in the data corresponding exactly to the two values of z chosen for the left panel, and yet

Fig. 3 Left: true distribution (blue line) versus histogram of recovered samples and their fitted pdf (red line) for $z = 0.40$ and $z = 0.92$. The dotted line displays the barycenter $\mu(y)$. Right: true and recovered mean, standard and skewness as functions of z (Color figure online)

the recovered histograms with 4000 points fit the corresponding conditional distributions very well. The empirical moments were computed on a 10-point grid in z and linearly interpolated in between. One can verify the close agreement throughout, though with an underestimated standard deviation near its maximum values at $z = \frac{1}{4}$ and $\frac{3}{4}$. The reason for this underestimation is that the comparatively larger standard deviation of the corresponding true $\rho(x|z)$ stems from very long tails (we can see a hint of them even at the more moderate values corresponding to the z 's on the left panel), which are severely under-represented in the finite sample of roughly 200 points in the intervals with largest variance.

Next we compare our approach with two different methodologies for the estimation of $\rho(x|z^*)$ for a given a value z^* . The first methodology estimates $\rho(x|z^*)$ through classical kernel density estimation on the set $\{x^i, z^i\}$ of N closest points z^i to z^* . We call this nearest neighborhood conditional kernel density estimation (NN-KDE). The second methodology adopts the Nadaraya–Watson (NW) conditional density estimator (De Gooijer and Zerom 2003):

$$\rho(x|z) = \frac{\sum_i K_{\sigma_1}(\|x - x^i\|) K_{\sigma_2}(\|z - z^i\|)}{\sum_i K_{\sigma_2}(\|z - z^i\|)}$$

where K_{σ_i} represents a kernel, Gaussian in our case, with bandwidth σ_i derived from the rule of thumb in Silverman (1986) for univariate densities and the formula in Bashtannyk and Hyndman (2001) for multidimensional ones.

Figure 4 shows results obtained using Optimal Transport (red line), NN-KDE (yellow line) and C-KDE (purple line). As it is clear from the figure, all three procedures yield good, comparable results. Yet NN-KDE is less accurate in capturing the behavior at the tails of the distribution (left two panels) and C-KDE seems to slightly over-resolve the data. As in Fig. 3, we also plot the mean, standard deviation and median of $\rho(x|z)$ as a function of z , where all three procedures yield sensible results, with the new one being the most accurate and less noisy. The quite noisy results from the NN-KDE are due to the high sensitivity of the standard deviation of $\rho(x|z)$ in terms of the N values of z^i closest to a given value of z . This phenomenon is particularly evident on the peaks of the standard deviation around $z = 0.25$ and $z = 0.75$. Table 1 summarizes the errors incurred by all three methods on each of these statistics via their L^2 and L^∞ norms and the correlation between the estimates and the underlying true answer.

Fig. 4 Left: true distribution (blue line) versus histogram of recovered samples and their fitted pdf (red line) for $z = 0.40$ and $z = 0.92$. The dotted line displays the barycenter $\mu(y)$, while the yellow and the purple line represent the estimates of NN-KDE and C-KDE respectively. Right: True and recovered mean, standard and skewness as functions of z (Color figure online)

Table 1 The L^2 , L^∞ errors and correlation in prediction of mean, standard deviation and skewness with three methods as in Fig. 4

	Mean			SD			Skewness		
	C-KDE	NN-KDE	OT	C-KDE	NN-KDE	OT	C-KDE	NN-KDE	OT
L^2	0.0010	0.0024	0.0015	0.0022	0.0011	0.0017	0.3703	1.0483	0.6332
L^∞	0.0686	0.1860	0.1223	0.0897	0.1345	0.0803	2.7954	4.4073	2.4955
Corr	0.9989	0.9978	0.9991	0.9849	0.9792	0.9978	0.9505	0.8787	0.9637

The best/worst methods are marked with bold/italic respectively

5.2 A meteorological example

Next we consider a meteorological example, using hourly measurements of the ground-level temperature in stations across the continental United States, publicly available from NOAA.¹ We chose stations with data available since at least 2006. We use this data in two ways: to explain and forecast the hourly temperature in one station, and to study the time evolution in one station of the joint probability of the highest and lowest daily temperatures.

5.2.1 A scalar case: hourly temperature forecast

In this example, the variable x to explain is the temperature itself, measured in degrees Celsius. A first natural set of covariates, which we denote “static” and “set 1” are the following:

1. The local time of the day $z_1 \in [0, 24]$, periodic, to capture the diurnal cycle. The corresponding grid is uniform with 24 points, one point per hour.
2. The day in the year to capture the seasonal cycle, $z_2 \in [0, 365.25]$, periodic, also with a uniform grid of 24 points.
3. Time in years, $z_3 \in [2006, 2017]$, real, with a grid of 41 points, 4 points per year. This covariate describes longer term (in our case decadal) temperature variations, such as those caused by El Niño or global warming.

¹ <https://www1.ncdc.noaa.gov/pub/data/uscrrn/products/hourly02/>.

The different time scales of the various static covariates are captured by normalizing them to one over a day, a year and 10 years respectively, while adopting a uniform penalization parameter $\lambda = 0.001$. For each station, the total number of observations is $m = 87,600$. The functions $G_s(y)$ adopted are monomials up to the 4th degree, each repeated 6 times, yielding a total of $r = 30$ components.

The upper-left panel of Fig. 6 displays the results of applying this article's procedure to the hourly temperatures in Ithaca, NY, with results plotted over a month. The line in black shows the actual observed hourly ground temperatures, the line in red the recovered median and the area shaded in pink represents the 95% confidence interval. Since the map between y and x for each value of z is monotonic, the value of x corresponding to any desired percentile can be readily computed from the map $x = T^{-1}(y; z)$, where the y is the value yielding the same percentile for the barycenter (i.e. the value such that the required fraction of the y^i fall below it) and z is the current value of the cofactors (in our case, 3 real numbers, one for each time-scale) for which x is sought. One can observe how the daily and seasonal signals are captured (a month is too short to observe any longer-term trend), while the weather systems, with a typical time-scale of one week, are not, since no covariate z refers to them.

A common-sense attempt to capture weather systems is to include the temperature in Ithaca itself 24 h before as an extra covariate (using this alone corresponds to the simple-minded forecast procedure of repeating the weather observed the day before.) We chose to use as z_4 not x^{i-24} but the corresponding normalized y^{i-24} from a previous run of the algorithm using only the static covariates. The rationale for this is that the covariate should measure deviation from standard conditions the day before, rather than repeat known information about normal conditions for the corresponding time and day. The results from using this second set of covariates are displayed on the upper-right panel of Fig. 6. We can see a pattern that follows the weather systems to some degree, yielding a sharper estimate (a more quantitative comparison will be shown below.)

Selecting the normalized temperature at Ithaca itself as a covariate is not well-informed meteorologically however, as the weather over the US continent does not stay put in one location but travels instead from west to east following the thermal wind. For instance, the left panel in Fig. 5 shows the time-lagged correlation between the normalized temperatures y^i in Ithaca and Des Moines, Iowa, well to its west. This correlation peaks between 36 and 48 h, and it beats significantly the correlation of Ithaca with itself for lags larger than a day. Hence we shall use for extra covariates not the 1-day old record in Ithaca, but the normalized temperatures 36 h before in Des Moines and two other stations (Stillwater, OK and Goodridge, MN) displayed on the map on the right of Fig. 5.

We use, as before, $r = 30$ components with monomials up to the 4th degree for the $G_s(y)$. For each of the new non-static covariates, we adopt a uniform grid with 30 points. The results from this third set of covariates can be seen on the lower-left panel of Fig. 6. They are far more sharply adjusted to the observations than any of the other two models, even for the outlier temperature plotted in blue. The lower-left panel displays the pdfs fitted to the histograms of $\rho(x|z)$ recovered for the specific value of z corresponding to that extreme observation. We can see that using set 3 allows us to forecast an histogram highly consistent with this unusual observation.

To render this comparison more quantitative, we introduce two measurements of error: the square-root of the normalized mean squared deviation, given by

$$NSMD^2 = \frac{1}{m} \sum_{i=1}^m \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2,$$

Fig. 5 Left: time-lagged correlation between Ithaca and Des Moines (red) and autocorrelation of Ithaca itself (blue). Right: choice of stations (blue) with strong time-lagged correlation with Ithaca (red) (Color figure online)

where μ_i and σ_i are the predicted mean and standard deviation, and (minus) the point-wise empirical log likelihood under a Gaussian assumption

$$-\frac{1}{n} \sum_{i=1}^m \log \rho_i(x_i) = \frac{1}{n} \sum_{i=1}^m \left[\left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 + \log(\sigma_i) \right] + \frac{1}{2} \log(2\pi).$$

These measurements of error (over the full decade of the series, not just the one month plotted in Fig. 6) using the three sets of covariates are shown in Table 2. The table also includes the variance of the barycenter $\mu(y)$ for each set, a measure of the amount of variability left after explaining away the fraction attributable to the covariates (Tabak and Trigila 2018b). As expected, the third set of covariates gives the smallest error by both measurements and the smallest unexplained variability.

Having illustrated how the procedure explains variability attributable to covariates, we switch to the issue of interpretability. One natural question is: can we extract from the results the way in which x depends on each of the six covariates z_l , independently of the others? We address this question through marginalization. If z_l is independent of the other z_b , we can factor the probability density $\nu(z)$ as

$$\nu(z) = \nu_l(z_l) N_l(z_1, \dots, z_{l-1}, z_{l+1}, \dots, z_L),$$

and marginalize the potential $\psi(y; z)$ via

$$\begin{aligned} \psi_l(y; z_l) &= \int \psi(y; z) N_l(z_1, \dots, z_{l-1}, z_{l+1}, \dots, z_L) dz_1, \dots, dz_{l-1}, dz_{l+1}, \dots, dz_L \\ &\approx - \sum_{k=1}^r \left(\frac{1}{m} \sum_i \prod_{b \neq l} \zeta_b^k(z_b^i) \right) \eta^k(y) \zeta_l^k(z_l) \end{aligned} \quad (28)$$

Performing the corresponding z_l -dependent map $T_l(y; z_l) = \nabla_y \psi_l(y; z_l)$ on all the y^i allows us to build the marginalized conditional probability $\rho_l(x|z_l)$.

Figure 7 shows the marginalized median and 95% confidence interval over the static factors. From the marginalized mean over the year, we can see an approximately 4-year cycle with an amplitude of around 2 degrees Celcius consistent with El Niño. Figure 8 shows

Fig. 6 Estimated median (red), truth data (black) and 95% confidence interval (shaded with pink) in one month. Upper left: prediction with set 1. Upper right: prediction with set 2. Lower left: predicted with set 3. Lower right: recovered probability distribution function (marked with blue star in time series), with set 1 (blue), set 2 (red), set 3 (yellow). The black dashed line represents the truth data. As a numerical verification, the probability for the true observations to fall in this empirical 95.0% confidence interval is 94.8%, 95.0% and 94.9% respectively for the three sets of covariates (Color figure online)

Table 2 Error measurements with three sets of covariates

	Set 1	Set 2	Set 3
NSMD	1.0035	1.0008	1.0036
log likelihood	2.9418	2.8567	2.7575
Var(y)	22.2498	18.7307	14.8699

the marginalized median and 95% confidence interval over the filtered temperature 36 h before at the 3 other stations.

So far we have applied our procedure to analysis, not forecast, as all observations were included in the training set. To show that it works nearly equally well in the forecast mode, we now use the components and filtered data y from 2006 to 2016 at NY Ithaca, and run the prediction for the data in 2017, with 8760 data points. We assume that values of all the covariates are known, except for the one corresponding to the year, which cannot be anticipated 1 year before. Since we observed a nearly 4-year cycle in the third covariate, we

Fig. 7 Marginalized dependence (median and confidence interval estimation) over static factors

Fig. 8 Marginalized dependence (median and interval estimation) over filtered temperature at Des Moines, Stillwater and Goodridge 36h before

will use for this factor its average value over the last such cycle available in the training data, 2013–2016. The results of the forecast are displayed for a month in Fig. 9, where we can see that they adjust quite accurately to the true observations.

5.2.2 A vector case: daily observed highest/lowest temperature

Using the same data set as in the prior subsection, the variable x we now analyze is the 2-dimensional vector containing the highest and lowest temperature of each day, i.e. the daily temperature range. The location chosen is again Ithaca, NY, observed from 2006 to 2017, a total of $m = 4019$ days. We adopt 2 static covariates here: the day of the year, $z_1 \in [0, 365.25]$, periodic, with 24 uniformly distributed grid points, and the year, $z_2 \in [2006, 2018]$, real, with a grid of 45 points, 4 points per year. The penalization parameter λ that we use for each covariate is 0.1, and we use the 9 functions $G_s(y)$ given by all non-constant monomials in (y_1, y_2) up to the 3rd order.

After filtering, the individual variances dropped from 97.7251 to 18.5887 (lowest temperature) and 119.5649–15.2777 (highest temperature). The time series of observed data and predicted mean are shown in Fig. 10. We can see that the lowest temperature has many more local extreme values than the highest temperature, which is the reason why its variance decreased less with filtering: it contains more variability that cannot be explained by static factors alone.

Fig. 9 Time series on test set in 2017 August, NY Ithaca based on the result for the past 10 years, with the mapping of prediction generated from covariate set 3

Fig. 10 Truth data (black) and predicted mean (red) for daily highest and lowest temperature (Color figure online)

The overall distribution of highest/lowest temperature for winter and summer have very different regimes (see Fig. 11). In winter, the highest temperature has negative skewness, while the lowest temperature is positively skewed, which indicates the underlying pdf might be non linear and non Gaussian. In summer, the variances are smaller, and the skewness is also weaker. However, as we only have one data point per day, we cannot obtain histograms focused more sharply than on a full season. Even less so for the 2d distribution of highest-lowest temperatures, which displays a clear correlation between both during winter but a much less marked one in the summer.

Instead, our methodology allows us to recover the full PDF for the joint distribution on any specified day, since we have over 4000 filtered data points y^i that can be mapped back to x for

Fig. 11 Histograms for highest, lowest and joint temperatures during winter and summer. The 2D joint distribution can not have finer grids, as there are only around 1000 data for each season

any choice of the covariates z . We plot four such snapshots of the pdf in Fig. 12. We can see that during winter, not only the variance of highest/lowest temperature respectively becomes larger, but also the correlation between them increases—the relation is almost linear! And in the transition between the coldest and hottest seasons in the year, for instance, on 20161202 or 20170401, the histogram is non-Gaussian and highly skewed. Only during summer is the joint distribution close to an isotropic Gaussian, i.e. the two variables become nearly independent with approximately the same variance (Fig. 13).

6 Summary and extensions

This article has developed a conditional density estimation and simulation procedure based on a sample-based formulation of the Wasserstein barycenter problem, extended to a continuum of distributions. This is formulated as a minimax problem where the two competing strategies correspond to the map $y = T(x; z)$ moving point x with covariate value z to the barycenter, and to its inverse $x = X(y; z)$. However, the two maps are represented in very different ways: $T(x; z)$ via its values $y^j = T(x^j; z^j)$ on the available observations, and $T^{-1}(y; z)$ through a potential function $\psi(y; z)$ such that $x = \nabla_y [c(x, y) - \psi(y; z)]$.

Fig. 12 4 regimes of full distribution in 2D space of highest/lowest temperature

Fig. 13 Dependence over time of year: marginalized mean and standard deviation of highest/lowest temperature (first row), marginalized difference and correlation between highest/lowest temperature (second row)

(This implicit characterization of the inverse map $T^{-1}(y; z)$ has an explicit solution for the standard squared-distance cost.)

The Wasserstein barycenter problem provides a natural conceptual framework for conditional probability estimation, and the methodology developed here shows that it leads to practical algorithmic implementations. The factorization of the dependence on cofactors into a sum of products of single-variable functions, plus the characterization of the latter by a finite number of parameters via prototypal analysis, makes the methodology useful even for problems with a large number of potential cofactors of different types. The meteorological examples displayed in Sect. 5 show that the procedure can solve problems seemingly intractable, such as the simulation of the full joint probability distribution of the highest and lowest daily temperatures for a specific day, for which there is at most one sample available in the historical case, and none in forecasting scenarios.

Even though the dependence of the potential ψ on z is made quite general through the use of prototypes, its dependence on y is restricted to the space of functions spanned by the externally provided family $G_s(y)$, which in the examples of Sect. 5 was restricted to a set of monomials up to the fourth degree. This extends the attributable component methodology (Tabak and Trigila 2018a) quite significantly, as the latter uses only $\eta_s = y$ as a feature, and hence can only capture the conditional expectation of $\rho(x|z)$. By contrast, quadratic monomials capture its covariant structure, higher order monomials its kurtosis and higher moments, and additional features can be added to capture other, possibly more localized characteristics. Yet one may wish for a more adaptive approach, that will extract the relevant features from the data without any a priori knowledge of which could be relevant. One possibility is to extend to the barycenter problem the adaptive methodology recently developed for optimal transport in Essid et al. (2018). Another is to replace the features $G_s(y)$ by low-rank factorizations, as is already done for the z -dependence of ψ in the current implementation. Still another possibility is to let the parameterization of ψ in (6) evolve as the y^i flow from x^i to their final converged values.

Acknowledgements The work of E. G. Tabak and W. Zhao was partially supported by NSF Grant DMS-1715753 and ONR Grant N00014-15-1-2355.

References

- Agnelli, J. P., Cadeiras, M., Tabak, E. G., Turner, C. V., & Vanden-Eijnden, E. (2010). Clustering and classification through normalizing flows in feature space. *SIAM Multiscale Modeling & Simulation*, 8, 1784–1802.
- Agueh, M., & Carlier, G. (2011). Barycenter in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2), 094–924.
- Bashtannyk, D. M., & Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36(3), 279–298.
- Caffarelli, L. A. (2003). The Monge–Ampère equation and optimal transportation, an elementary review. In *Optimal transportation and applications*. Lecture Notes in Math (pp. 1–10). Berlin: Springer.
- Chenyue, W., & Tabak, E.G. (2018) Prototypal analysis and prototypal regression. In preparation
- De Gooijer, J. G., & Zerom, D. (2003). On conditional density estimation. *Statistica Neerlandica*, 57(2), 159–176.
- Dutordoir, V., Salimbeni, H., Hensman, J., & Deisenroth, M. (2018). Gaussian process conditional density estimation. In *Advances in neural information processing systems* (pp. 2385–2395).
- Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430), 577–588.
- Essid, M., Laefer, D., & Tabak, E. G. (2018). Adaptive optimal transport. Submitted to Information and Inference.
- Fan, J., Yao, Q., & Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1), 189–206.

- Fan, J., & Yim, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4), 819–834.
- Gray, A. G., & Moore, A. W. (2001). N-body' problems in statistical learning. In *Advances in neural information processing systems* (pp. 521–527).
- Holmes, M. P., Gray, A. G., & Isbell, C. L. (2012). Fast nonparametric conditional density estimation. arXiv preprint [arXiv:1206.5278](https://arxiv.org/abs/1206.5278).
- Hyndman, R. J., Bashtannyk, D. M., & Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5(4), 315–336.
- Kantorovich, L. V. (1942). On the translocation of masses. *Compt. Rend. Akad. Sei*, 7, 199–201.
- Kantorovich, L. V. (1948). On a problem of Monge. *Uspekhi Matematicheskikh Nauk*, 3(2), 225–226.
- Monge, G. (1781). *Mémoire sur la théorie des déblais et des remblais*. Histoire De L'acad mie Royale Des Sciences. Paris.
- Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1), 141–142.
- Pass, B. (2013). On a class of optimal transportation problems with infinitely many marginals. *SIAM Journal on Mathematical Analysis*, 45(4), 2557–2575.
- Rosenblatt, M. (1969). Conditional probability density and regression estimators. *Multivariate Analysis II*, 25, 31.
- Silverman, B. (1986). Density estimation for statistics and data analysis, chapter 2–3
- Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., et al. (2015). Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4), 66.
- Tabak, E. G., & Trigila, G. (2018a). Conditional expectation estimation through attributable components. *Information and Inference: A Journal of the IMA*, 7(4), 727–754.
- Tabak, E. G., & Trigila, G. (2018b). Explanation of variability and removal of confounding factors from data through optimal transport. *Communications on Pure and Applied Mathematics*, 71(1), 163–199.
- Tabak, E. G., & Trigila, G. (2018c). An iterative method for the Wasserstein barycenter problem. In preparation
- Tabak, E. G., & Turner, C. V. (2013). A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2), 145–164.
- Tabak, E. G., & Vanden-Eijnden, E. (2010). Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8, 217–233.
- Trigila, G., & Tabak, E. G. (2016). Data-driven optimal transport. *Communications on Pure and Applied Mathematics*, 69(4), 613–648.
- Trefethen, L. N. (2000). *Spectral methods in MATLAB* (Vol. 10). SIAM.
- Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4), 359–372.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.