CONDITIONAL DENSITY ESTIMATION AND SIMULATION THROUGH OPTIMAL TRANSPORT*

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

Abstract. A methodology to estimate from samples the probability density of a random variable x conditional to the values of a set of covariates $\{z_l\}$ 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 $\{z_l\}$ is formulated in terms of convex combinations, so that it can be applied to variables of nearly any type, including real, categorial 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.

Key words. Conditional density estimation, optimal transport, Wasserstein barycenter, explanation of variability, confounding factors, sampling, uncertainty quantification.

AMS subject classifications. 68Q25, 68R10, 68U05

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 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

*Submitted to the editors DATE.

Funding: This work was funded by NSF.

†Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, USA, tabak@cims.nyu.edu

‡Baruch College, CUNY, 55 Lexington avenue, New York, NY 10035, USA, giulio.trigila@baruch.cuny.edu

§Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, USA, wenjun@cims.nyu.edu

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. The difficulties above as 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 $y = Y(x; z)$, 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 $x = X(y; z)$, 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 [15] 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 ([4]). Since prototypal 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 one seeks only the conditional expected value of the distribution $\rho(x|z)$. This article extends the attributable component methodology [15], 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 [14]. This line of work was further developed in [8], [10] and [6], then [3] and [9] addressed the issue of finding an efficient data-driven bandwidth selection procedure, and [5] enforced the positivity constraint of the estimated conditional density by means of a slight modification of the Nadaraya-Watson smoother.

By contrast, the methodology of this article estimates conditional distributions via conditional maps. A map-based density estimation was previously developed in [18] [17], 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 [19]. Both flow formulations were developed in the context of single density estimation, while the work in [1] performed clustering and classifica-

tion by extending the flow methodology in [18] to a finite number of distributions, 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, section 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 distributions. Section 5 exemplifies the procedure through its application to synthetic and meteorological data. Finally, section 6 summarizes the work and discusses possible extensions.

2. Problem setting. Given samples $\left\{x^i, z_1^i, \ldots, z_L^i\right\}$ of a variable of interest x and covariates z_l, we seek to estimate or simulate the conditional probability distribution

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

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

We pose this conditional density estimation as a Wasserstein barycenter problem [2], whose solution pushes the distributions $\rho(x|z_1, \ldots, z_L)$ to their barycenter $\mu(y)$ through a z-dependent map $Y(x|z_1, \ldots, z_L)$ with inverse $X(y|z_1, \ldots, z_L)$. 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 $y^i = Y\left\{x^i|z_1^i, \ldots, z_L^i\right\}$ followed by the map $X(y^i|z^*)$ allows us to immediately simulate $\rho(x|z^*)$ under any choice z^* for the factors z. This formulation is illustrated in figure 1.

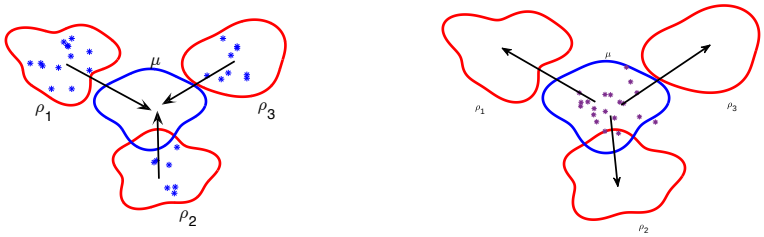

FIG. 1. *Conditional density simulation as a Wasserstein barycenter problem. For easy visualization, this example has $x \in R^2$ and a single categorical covariate z with 3 possible values. On the left, the samples x^i of the conditional probabilities $\rho(x|z)$ are mapped through a z-dependent map $Y(x^i; z^i)$ 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 $Y(x; z^*)$.*

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 n samples (x^i, z^i). After finding the conditional map $Y(x;z)$, one obtains samples $y^i = Y(x^i; z^i)$ 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 $\tilde{x}^i = X(y^i; z^*)$, where $X = Y^{-1}$. 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 is posed in terms of distributions, a property inherited by the barycenter problem [2]. Yet we do not know the conditional distributions $\rho(x|z)$, but only a set of samples $\left\{x^i, z^i\right\}$ thereof. This subsection develops a sequence of formulations of the optimal transport barycenter problem, to obtain one that seeks the family of maps $Y(x;z)$ directly from the set of data pairs $\left\{x^i, z^i\right\}$ 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 [12], extended to situations with possibly infinitely many marginals [13]. Given a family of distributions $\rho(x|z)$, an extra distribution $\nu(z)$ underlying the factors z and a transportation cost function $c(x,y)$, find the distribution $\mu(y)$ and the corresponding family of maps $y = Y(x;z)$ pushing forward $\rho(x|z)$ to $\mu(y)$ so that the total transportation cost
   $$C(Y,\mu) = \int \left[\int c(x, Y(x;z)) \rho(x|z) dx \right] \nu(z) dz$$
   is minimized:
   $$[Y,\mu] = \arg\min C(Y,\mu), \text{ s.t. } \forall z: \; x \sim \rho(:|z) \Rightarrow y = Y(x;z) \sim \mu. \tag{2.1}$$
   The assumption that the distributions $\nu(z)$ and $\rho(x|z)$ derive from probability densities was made just to give a concrete form to $C(Y,\mu)$. Nothing changes here or in what follows if, for more general distributions, we define
   $$C(Y,\mu) = E_{x,z}\left[c\left(x, Y(x;z)\right)\right],$$
   since ν 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 $Y(x;z)$ as above. However, as noted in [19], relaxing these to conditional couplings $\pi(x,y|z)$, in an extension of Kantorovich formulation [11] of the optimal transport problem, leads to a dual formulation, which will allow us to replace the conditional distributions $\rho(x|z)$ and $\nu(z)$ by samples thereof. In terms of these conditional couplings, the cost C to minimize adopts the form
   $$C(\pi,\mu) = \int \left[\int c(x,y) \pi(x,y|z) \; dx \; dy \right] \nu(z) \; dz,$$
   and the problem becomes
   $$[\pi,\mu] = \arg\min C(\pi,\mu), \quad \text{such that} \quad \pi,\mu \geq 0, \quad \text{and}$$
   $$\forall z: \; \int \pi(x,y|z) \; dy = \rho(x|z), \; \int \pi(x,y|z) \; dx = \mu(y). \tag{2.2}$$

From the very definition of the barycenter, we should expect the random variables y and z to be independent. The map $y = Y(x; z)$ is designed precisely to remove the variability in x due to the covariates in z; if there was 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 (2.2). 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

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

confirming that y and z are indeed independent.

3. **Dual Kantorovich formulation**

The problem in (2.2) 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

$$\begin{aligned} L(\pi, \mu, \phi, \psi) &= C(\pi, \mu) \\ &- \int \left[\int \pi(x, y|z)dy - \rho(x|z)\right] \phi(x, z)dx\ \nu(z)dz \\ &- \int \left[\int \pi(x, y|z)\ dx - \mu(y)\right] \psi(y, z)dy\ \nu(z)dz, \end{aligned}$$

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

$$\max_{\phi,\psi} \int \left[\int \phi(x, z)\ \rho(x|z)\ dx\right] \nu(z)dz, \quad \text{such that}$$

$$\phi(x, z) + \psi(y, z) \leq c(x, y), \quad \forall y : \int \psi(y, z)\ \nu(z)dz \geq 0. \tag{2.3}$$

4. **Conversion to a minimax problem through conjugate duality**

In problem (2.3), if ψ is given, it follows that

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

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

$$\max_{\psi} \int \left[\int \min_y \left[c(x, y) - \psi(y, z)\right] \rho(x|z)\ dx\right] \nu(z)\ dz,$$

$$\text{where} \quad \forall y : \int \psi(y, z)\ \nu(z)\ dz = 0,$$

or

$$\max_{\psi}\ \min_{Y(x;z)} \int \left[c(x, Y) - \psi(Y, z)\right] \gamma(x, z)\ dx\ dz,$$

$$\forall y : \int \psi(y, z)\ \nu(z)\ dz = 0, \tag{2.4}$$

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

$$\max_{\psi} \min_{Y(x,z)} E_\gamma \left[c(x,Y) - \psi(Y,z) \right], \quad \forall y : \; E_\nu \left[\psi(y,z) \right] = 0. \tag{2.5}$$

Notice that, in the solution to this dual problem, the random variables $y = Y(x;z)$ and z are still independent. Otherwise, the dual problem would be unbounded, as we could find a function $\psi(y,z)$ such that $\forall y : \; E_\nu \left[\psi(y,z) \right] = 0$, but $E_\gamma \left[\psi(Y(x;z),z) \right] \neq 0$. 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 $y = Y(x;z)$ such that y and z are independent equals $\min E_\gamma \left[c(x,Y) \right]$, which agrees with the solution to the primal problem.

5. **Sample based formulation**

The fact that the distributions γ and ν appear in problem (2.5) 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

$$\max_{\psi} \min_{\{y^i\}} \sum_i \left[c(x^i,y^i) - \psi(y^i,z^i) \right], \quad \forall y : \sum_i \psi(y,z^i) = 0, \tag{2.6}$$

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

Cost: for concreteness, we will adopt the canonical quadratic cost

$$c(x,y) = \frac{1}{2}\|x-y\|^2, \tag{2.7}$$

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 [15] fits into the framework developed here. This will allow us 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 (2.6) yields

$$x^i = y^i - \nabla_y \psi(y^i, z^i). \tag{3.1}$$

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

$$\psi(y;z) = -y \cdot Z(z), \tag{3.2}$$

we have

$$y^i = x^i - Z(z^i), \tag{3.3}$$

a z-dependent rigid translation.

Replacing (2.7), (3.2) and (3.3) into (2.6), we obtain the following variational problem for $Z(z)$:

$$\max_Z \sum_i \left[\frac{1}{2}\|Z(z^i)\|^2 + (x_i - Z(z_i)) \cdot Z(z^i)\right], \quad \sum_i Z(z^i) = 0,$$

or

$$\min_Z \sum_i \frac{1}{2}\|Z(z^i) - x_i\|^2, \quad \sum_i Z(z^i) = 0. \tag{3.4}$$

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

$$Z(z) = \bar{x}(z) - \bar{x}, \quad \bar{x} = \frac{1}{m}\sum_i \bar{x}(z^i). \tag{3.5}$$

For convenience, we can remove the empirical mean of x from the observations ab initio, in which case $Z(z) = \bar{x}(z)$, and we do not need to take into account the constraint in (3.4), as it is satisfied automatically (if allowed by the family of functions $Z(z)$ considered.)

3.1. Attributable components. If we leave the function $Z(z)$ completely unrestricted, the solution to (3.4) is given by the trivial $Z(z^i) = x^i$ when all z^i's are different, and by $Z(z) = \text{mean}(x^i)$ 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 $Z(z)$ for values of z not in the dataset.

One could propose instead a parametric ansatz, such as

$$Z(z) = \sum_k \beta_k Z_k(z),$$

with $\{Z_k(z)\}$ a given set of functions (the "features"), and optimize over the parameters β, but this suffers from the pitfalls of all parameterizations, particularly when the number L of factors z_l is large.

Instead, we proposed in [15] the low-rank tensor factorization (or separated variable approximation, depending on whether one approaches it through linear algebra or multivariable calculus)

$$Z(z) = \sum_{k=1}^{r}\prod_{l=1}^{L} Z_l^k(z_l). \tag{3.6}$$

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

Then we modeled each of these functions as convex combinations of an array of unknown values V:

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

where the $\alpha(l)_i^j$ are given, 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, we can adopt a grid $\left\{ {z_{gl}}^j \right\}$ (not necessarily uniform), and interpret the $V(l)_j^k$ as $Z_l^k \left({z_{gl}}^j \right)$: 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. If the z_l is instead categorical, then the $\left\{ {z_{gl}}^j \right\}$ are 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 ${z_{gl}}^j$. More general types of covariates (probability distributions, photographies) can be made to fit into the same framework via prototypal analysis ([4]): given a set of n samples z^i of z_l, we seek m prototypes

$$y_j = \sum_{i=1}^{n} \beta_j^i x^i, \quad \beta_j^i \geq 0, \quad \sum_i \beta_j^i = 1$$

such that the objective function

$$L = \sum_i \left\| x^i - \sum_j \alpha_j^i y^j \right\|^2 + P, \quad \alpha_j^i \geq 0, \quad \sum_j \alpha_j^i = 1$$

is minimized. Keeping only the first sum in L corresponds to archetypal analysis (XXX): one seeks a set of archetypes $\{y_j\}$, convex combinations of the $\left\{x^i\right\}$, such that approximating the x by convex combinations of the y produce the smallest L^2 error. The added penalty term yields prototypes instead, where the y used via convex combination to approximate each x are should be close to x. This is what is required to approximate functions of x via local convex combination of their values on the y.

Because the objective function L is written in terms of squared norms, the procedure to find the α can be formulated exclusively in terms of inner products, so that it applies to any space where inner products are defined. For probability distributions, for instance, one can use the inner product corresponding to the Energy norm (XXX).

Finally, we add to (3.4) a penalty term to enforce the smoothness or control the variability of the functions $Z_l^k(z_l)$:

$$\min_V \left\{ \sum_i \frac{1}{2} \left\| x^i - \sum_k \prod_{l=1}^{L} \sum_j \alpha(l)_i^j V(l)_j^k \right\|^2 + \right. \tag{3.7}$$

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

For instance, when z_l is a real variable, the quadratic form ${V(l)^k}^t C^l V(l)^k$ may be chosen to represent the square norm of a finite difference approximation to the first or

second derivatives of $Z_l^k(z_l)$, and when z_l is categorical, it may be chosen to represent the variance of $Z_l^k(z_l)$. The prefactor $\prod_{b\in L, b\neq l} \|V(b)^k\|^2$ is included to balance the two terms in the objective function. Otherwise, the smoothness requirement on one $Z_l^k(z_l)$ could be bypassed by making that Z_l^k smaller by a constant factor while keeping $Z(z)$ constant by enlarging other Z_b^k less constrained. The objective function in (3.7) 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 to 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. 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 (3.1) determines y^i only implicitly, so we cannot replace it straightforwardly in (2.6) as with (3.3).

We solve this problem as in [16], through an alternate iterative procedure where we update the values of y for fixed $Z(z)$ and vice versa, linearizing each time the y dependence of ψ at the current values of y^i. Notice that this can be though of as a primal-dual approach, where we update in one step the dual variable ϕ and in the other the primal map $Y(x|z)$. In order to perform the linearization, we expand the factorization in (3.6) from only the z-dependence to all of ψ:

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

leaving temporarily aside how each of the $Y_k(y)$ and $Z^k(z)$ is defined. Then we replace (3.1) by the local approximation

$$y^i = x^i + \nabla_y \psi(y, z^i)\Big|_{y=y_n^i} = x^i - \sum_k Z^k(z^i)\ J_k^i, \tag{4.1}$$

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 $J_k^i = \nabla_y Y_k(y)\Big|_{y=y_n^i}$. Consistently, we approximate

$$\psi(y^i, z^i) \approx -\sum_k \left(Y_k^i + {J_k^i}^t \left(x^i - \sum_c Z^c(z^i)\ J_c^i - y_n^i \right) \right) Z^k(z^i), \tag{4.2}$$

with $Y_k^i = Y_k(y_n^i)$. Replacing into (2.6) yields

$$\max_Z \sum_i \left[\frac{1}{2} \left\| \sum_k J_k^i\ Z^k(z^i) \right\|^2 \right.$$

$$\left. + \sum_k \left(Y_k^i + {J_k^i}^t \left(x^i - \sum_c Z^c(z^i)\ J_c^i - y_n^i \right) \right) Z^k(z^i) \right],$$

or

$$\min_Z \sum_i \left[\frac{1}{2} \left\| (x^i - y_n^i) - \sum_k J_k^i \; Z^k(z^i) \right\|^2 - \sum_k Y_k^i Z^k(z^i) \right]. \tag{4.3}$$

subject to the conditions

$$\forall y \; \sum_k \left(Y_k(y) \sum_i Z^k(z^i) \right) = 0. \tag{4.4}$$

If the $Y_k(y)$ are independent functions, (4.4) is equivalent to

$$\forall k \; \sum_i Z^k(z^i) = 0. \tag{4.5}$$

We will impose this stronger requirement, easier to implement, even when the independence of the Y_k does not hold. There is no loss of generality in this, since the non-independence of the Y_k makes the choice of $Z(k)$ non-unique, with degrees of freedom that exactly balance the extra requirements in (4.5).

As before, we propose for Z^k the factorization

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

and add to (4.3) a penalty term of the form

$$\frac{1}{2} \sum_{l=1}^{L} \lambda_l \sum_k \left(\prod_{b \in L, b \neq l} \|V(b)^k\|^2 \right) V(l)^{k\,t} C^l V(l)^k. \tag{4.6}$$

Yet there is one more consideration to make: for the approximations (4.1) and (4.2) 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 can add a second penalty of the form

$$\frac{1}{2} \nu_z \sum_{l=1}^{L} \|V(l) - V(l)_n\|^2, \tag{4.7}$$

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

The procedure above describes how the $Z^k(z)$ are updated. Regarding the $Y_k(y)$, 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 (2.6), proposing for them either a parametric representation or a factorization similar to the one for Z^k:

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

A sensible parametric proposal adopts the form

$$Y_k(y) = \sum_s \beta_k^s \tilde{Y}_s(y), \tag{4.8}$$

with given functions $\tilde{Y}_s(y)$, thus extending the attributable component procedure, which had only the function $\tilde{Y}(y) = y$. Then we add to the objective function the penalty term

$$\frac{1}{2}\nu_y\|\beta - \beta_n\|^2, \tag{4.9}$$

and denote by O the objective function resulting from the sum of (4.3), (4.6), (4.7) and (4.9).

$$\begin{aligned} O(V,\beta) = &\sum_i \left[\frac{1}{2} \left\| \left(x^i - y_n^i\right) - \sum_k J_k^i \; Z^k(z^i) \right\|^2 - \sum_k Y_k^i Z^k(z^i) \right] \\ &+ \frac{1}{2}\sum_{l=1}^{L} \lambda_l \sum_k \left(\prod_{b\in L, b\neq l} \|V(b)^k\|^2 \right) V(l)^{k^t} C^l V(l)^k \\ &+ \frac{1}{2}\nu_z \sum_{l=1}^{L} \|V(l) - V(l)_n\|^2 + \frac{1}{2}\nu_y\|\beta - \beta_n\|^2, \end{aligned} \tag{4.10}$$

where

$$Z^k(z^i) = \prod_{l=1}^{L} \sum_j \alpha(l)_i^j V(l)_j^k, \tag{4.11}$$

$$Y_k^i = \sum_k \beta_k^s \tilde{Y}_s(y_n^i), \tag{4.12}$$

and

$$J_k^i = \sum_k \beta_k^s \nabla_y \tilde{Y}_s(y_n^i). \tag{4.13}$$

The procedure goes as follows: given the samples $\left\{x^i, z_1^i, \ldots, z_L^i\right\}$, 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 $\tilde{Y}_s(y)$, 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 (4.5) and update the $\{y^i\}$ via (4.1).
   (b) Minimize O over the $\{\beta_k^s\}$ and update the $\{y^i\}$ via (4.1).

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

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

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

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

5. Examples. In order to illustrate the methodology proposed, we use one simple synthetic example and a more complex, data-based meteorological one.

5.1. Synthetic example. 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 $Y_k(y)$, monomials up to 5th order y^n, $n = 0, 1, \ldots, 5$, 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 $X(y^i; z)$, and compare the results with the true $\rho(x|z)$ underlying the data. The left panel of figure 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 the recovered histograms with 4000 points fit the corresponding conditional distributions very well. The empirical moments where 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.

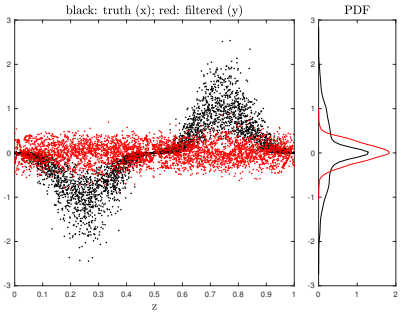

FIG. 2. *Original data x vs. filtered data y as a function of the covariate z on the left panel, and their PDFs (marginalized over z) on the right. One sees how the z-dependence of the distribution of x in gone in y, and how this results in a reduced total variability.*

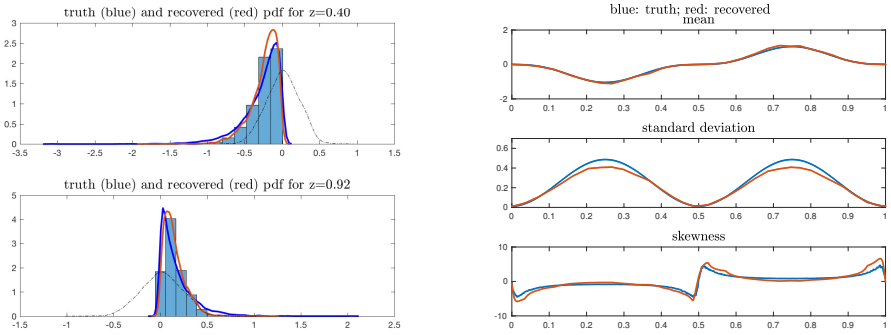

FIG. 3. *Left: true distribution (blue line) vs. histogram of recovered samples and their fitted pdf (red line) for $z = 0.40$ and $z = 0.92$. The dotted line displays the barycenter $\mu(y)$. Right: True and recovered mean, standard and skewness as functions of z.*

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 where we have data available since at least since 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 Celcius. 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.

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 $m = 87600$. The functions $\tilde{Y}_s(y)$ adopted are monomials up to the 4th degree, each repeated 6 times, yielding a total of $r = 30$ components.

The upper-left panel of figure 5 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 $x = X(y; z)$, 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 hours 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 figure 5. 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-

[1] https://www1.ncdc.noaa.gov/pub/data/uscrn/products/hourly02/

informed meteorologically however, as the weather over the US continent does not stay put in one location but travels instead from from west to east following the thermal wind. For instance, the left panel in figure 4 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 hours, and it beats significantly the correlation of Ithaca with itself for lapses larger than a day. Hence we shall use for extra covariates not the 1-day old record in Ithaca, but the normalized temperatures 36 hours before in Des Moines and two other stations (Stillwater, OK and Goodridge, MN) displayed on the map on the right of figure 4.

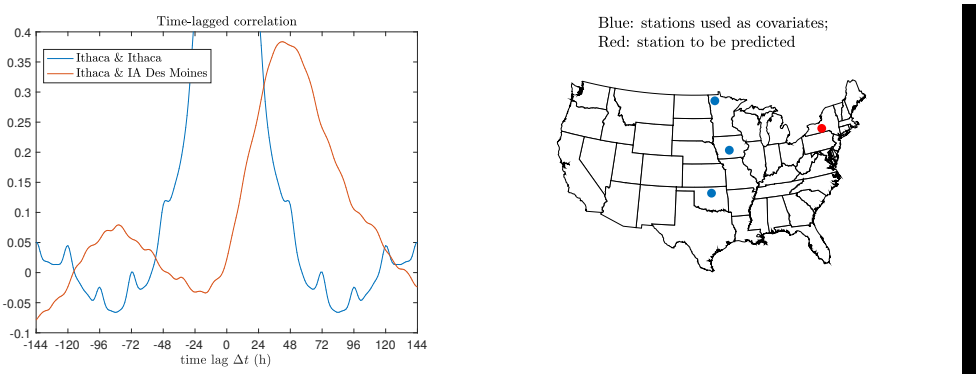

FIG. 4. *Left: time-lagged correlation between Ithaca and Des Moines (red) and autocorrelation of Ithaca itself (blue). Right: choice of stations (blue) with strong time-lagged correlation with Ithaca (red).*

We use as before $r = 30$ components with monomials up to the 4th degree for the $\tilde{Y}_s(y)$. 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 figure 5. They are far more sharply adjusted to the observations that 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 conventional mean squared deviation, given by

$$SMD^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_i)^2,$$

where μ_i is the predicted mean, 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 figure 5) using the three sets of covariates are shown in table 1. 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

410 covariates (XXX). As expected, the third set of covariates gives the smallest error by
411 both measurements and the smallest unexplained variability.

	set 1	set 2	set 3
SMD	4.8607	4.4606	3.9205
log likelihood	2.9418	2.8567	2.7575
Var(y)	22.2498	18.7307	14.8699

TABLE 1
Error measurements with three sets of covariates.

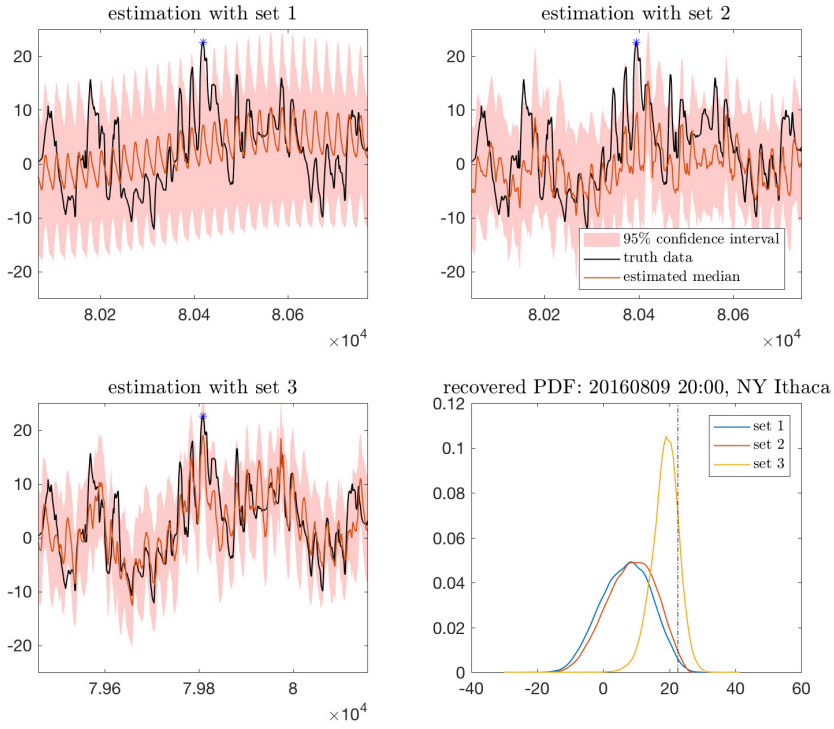

FIG. 5. *Estimated median (red), truth data (black) and* 95% *confidence interval (shaded with pink) in one month. Upper left: prediction with set* 1. *Upper right: prediction with set* 2. *Lower left: predicted with set* 3. *Lower right: recovered probability distribution function (marked with blue 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.*

Having illustrated how the procedure explain 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

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

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

$$\psi_l(y; z_l) = \int \psi(y; z) N_l\left(z_1, \ldots, z_{l-1}, z_{l+1}, \ldots, z_L\right) dz_1, \ldots, dz_{l-1}, dz_{l+1}, \ldots, dz_L$$

$$\approx -\sum_{k=1}^{r} \left(\frac{1}{m} \sum_i \prod_{b \neq l} Z_b^k(z_b^i) \right) Y^k(y) Z_l^k(z_l) \tag{5.1}$$

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

Figure 6 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 7 shows the marginalized median and 95% confidence interval over the filtered temperature 36 hours before at the 3 other stations.

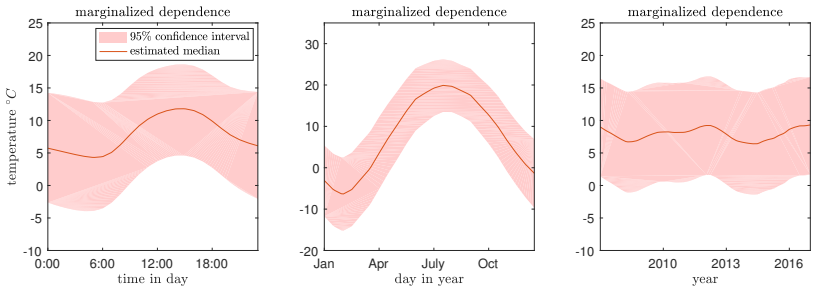

FIG. 6. *Marginalized dependence (median and confidence interval estimation) over static factors.*

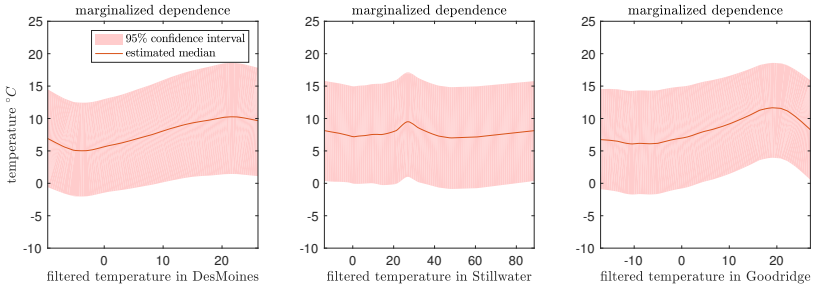

FIG. 7. *Marginalized dependence (median and interval estimation) over filtered temperature at Des Moines, Stillwater and Goodridge* 36 *hours before.*

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 one year before. Since we observed a nearly 4-year cycle in the third covariate, we 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 figure 8, where we can see that they adjust quite accurately to the true observations.

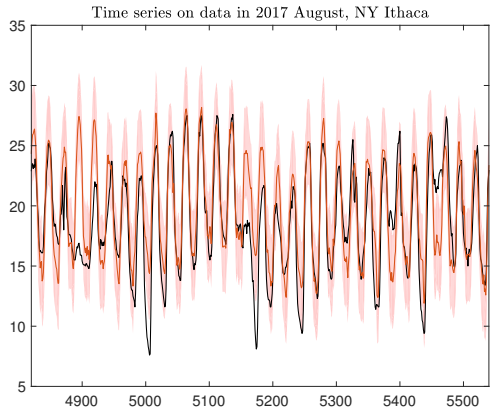

FIG. 8. *Time series on test set in* 2017 *August, NY Ithaca based on the result for the past* 10 *years, with the mapping of prediction generated from covariate set* 3.

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 $Y_s(y)$ 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 to 15.2777 (highest temperature). The time series of observed data and predicted mean are shown in figure 9. 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.

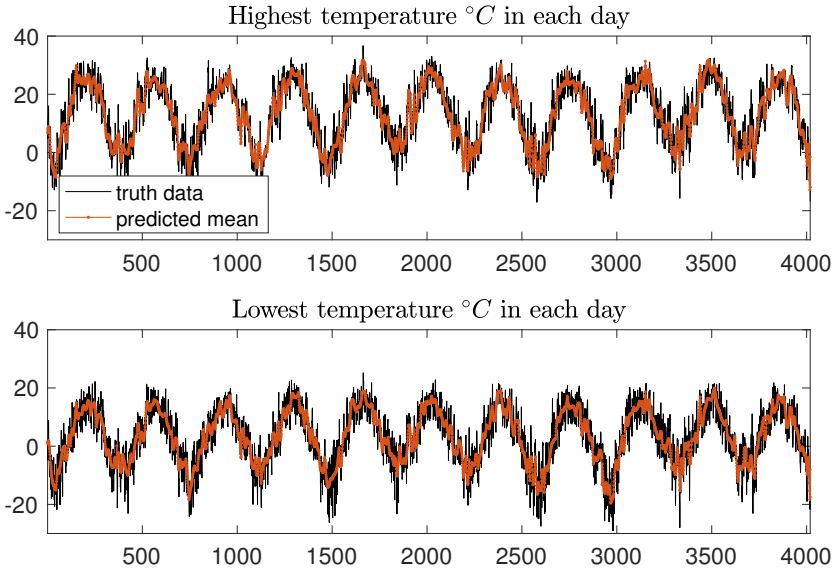

FIG. 9. *Truth data (black) and predicted mean (red) for daily highest and lowest temperature.*

The overall distribution of highest/lowest temperature for winter and summer have very different regimes (see figure 10). 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 any choice of the covariates z. We plot four such snapshots

of the pdf in figure 11. 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.

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 $y = Y(x; z)$ 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: $Y(x; z)$ via its values $y^j = Y(x^j; z^j)$ on the available observations, and $X(y; z)$ through a potential function $\psi(y; z)$ such that $x = \nabla_y [c(x, y) - \psi(y; z)]$ (This implicit characterization of the inverse map $X(y; z)$ has an explicit solution for the standard squared-distance cost.)

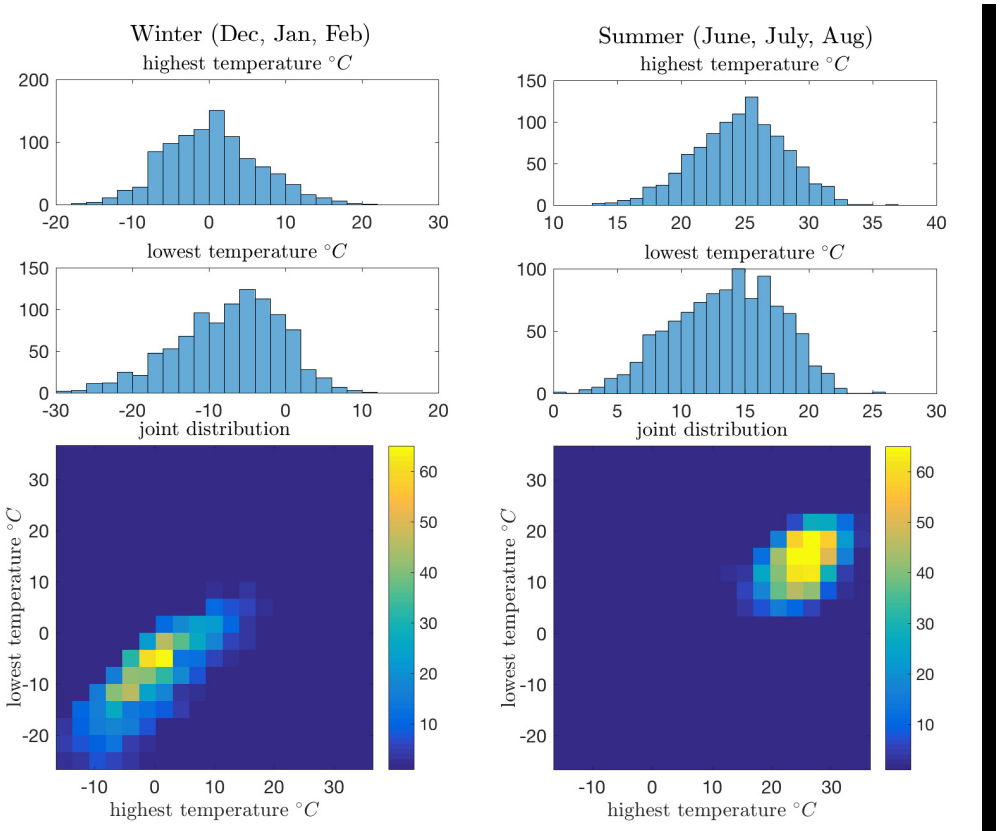

FIG. 10. *Histograms for highest, lowest and joint temperatures during winter and summer. The 2D joint distribution can not have finer grids, as there are only around* 1000 *data for each season.*

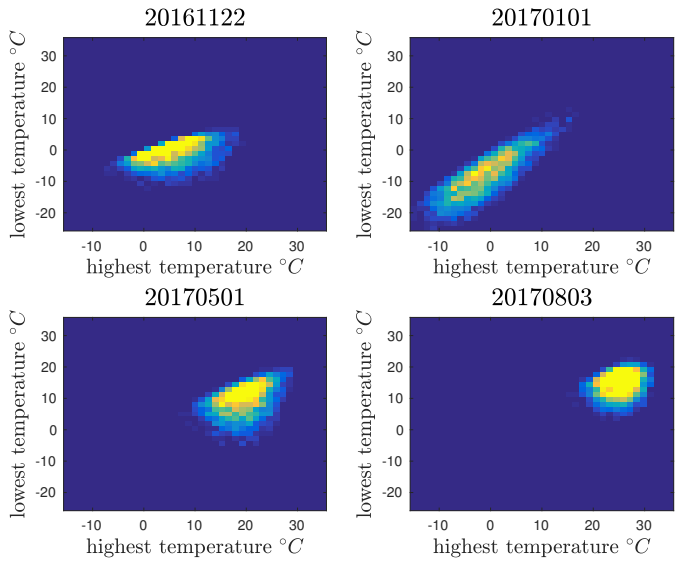

FIG. 11. 4 *regimes of full distribution in 2D space of highest/lowest temperature.*

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 section 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 $\tilde{Y}_s(y)$, which in the examples of section 5 was restricted to a set of monomials up to the fourth degree. This extends the attributable component methodology [15] quite significantly, as the latter uses only $Y_s = y$ 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 [7]. Another is to replace the features $\tilde{Y}_s(y)$ 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 (2.6) evolve as the y^i flow from x^i to their final converged values.

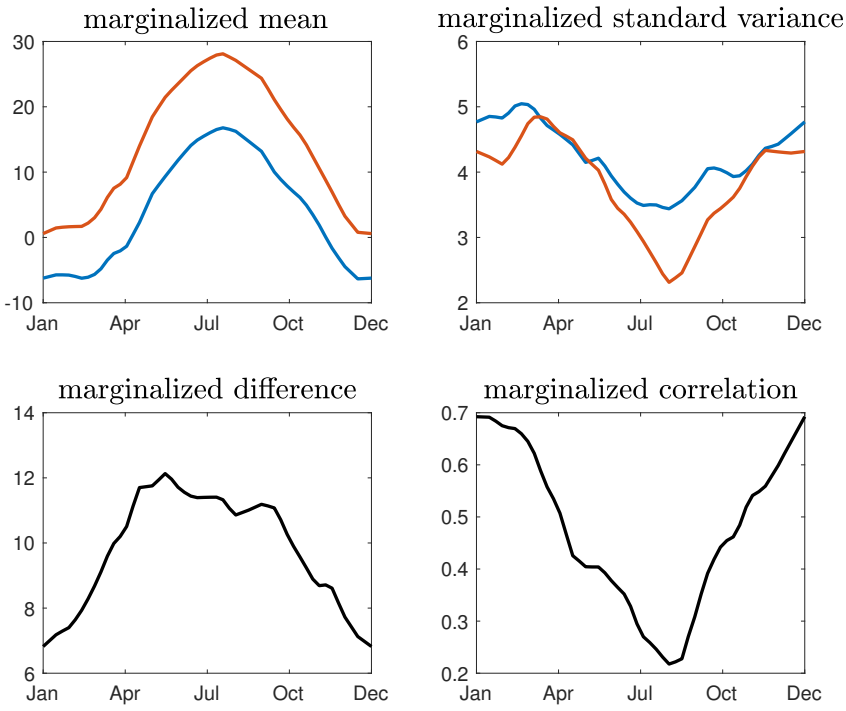

FIG. 12. *Dependence over time of year: marginalized mean and standard deviation of highest/lowest temperature (first row), marginalized difference and correlation between highest/lowest temperature (second row).*

Acknowledgments. 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

[1] J. P. AGNELLI, M. CADEIRAS, E. G. TABAK, C. V. TURNER, AND E. VANDEN-EIJNDEN, *Clustering and classification through normalizing flows in feature space*, SIAM MMS, 8 (2010).
[2] M. AGUEH AND G. CARLIER, *Barycenter in the Wasserstein space*, SIAM J. MATH. ANAL., 43 (2011), pp. 094–924.
[3] D. M. BASHTANNYK AND R. J. HYNDMAN, *Bandwidth selection for kernel conditional density estimation*, Computational Statistics & Data Analysis, 36 (2001), pp. 279–298.
[4] W. CHENYUE AND E. G. TABAK, *Prototypal analysis and prototypal regression*, In preparation, (2018).
[5] J. G. DE GOOIJER AND D. ZEROM, *On conditional density estimation*, Statistica Neerlandica, 57 (2003), pp. 159–176.
[6] M. D. ESCOBAR AND M. WEST, *Bayesian density estimation and inference using mixtures*, Journal of the american statistical association, 90 (1995), pp. 577–588.
[7] M. ESSID, D. LAEFER, AND E. G. TABAK, *Adaptive optimal transport*, Submitted to Information and Inference, (2018).
[8] J. FAN, Q. YAO, AND H. TONG, *Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems*, Biometrika, 83 (1996), pp. 189–206.
[9] J. FAN AND T. H. YIM, *A crossvalidation method for estimating conditional densities*,

Biometrika, 91 (2004), pp. 819–834.
[10] R. J. Hyndman, D. M. Bashtannyk, and G. K. Grunwald, *Estimating and visualizing conditional densities*, Journal of Computational and Graphical Statistics, 5 (1996), pp. 315–336.
[11] L. V. Kantorovich, *On the translocation of masses*, Compt. Rend. Akad. Sei, 7 (1942), pp. 199–201.
[12] G. Monge, *Mémoire sur la théorie des déblais et des remblais*, De l'Imprimerie Royale, 1781.
[13] B. Pass, *On a class of optimal transportation problems with infinitely many marginals*, SIAM Journal on Mathematical Analysis, 45 (2013), pp. 2557–2575.
[14] M. Rosenblatt, *Conditional probability density and regression estimators*, Multivariate analysis II, 25 (1969), p. 31.
[15] E. G. Tabak and G. Trigila, *Conditional expectation estimation through attributable components*, Information and Inference: A Journal of the IMA, 128 (2018).
[16] E. G. Tabak and G. Trigila, *An iterative method for the Wasserstein barycenter problem*, In preparation, (2018).
[17] E. G. Tabak and C. V. Turner, *A family of non-parametric density estimation algorithms*, CPAM, LXVI (2013).
[18] E. G. Tabak and E. Vanden-Eijnden, *Density estimation by dual ascent of the log-likelihood*, Comm. Math. Sci., 8 (2010).
[19] G. Trigila and E. G. Tabak, *Data-driven optimal transport*, Communications on Pure and Applied Mathematics, 69 (2016), pp. 613–648.