Information and Inference: A Journal of the IMA (2017) Page 1 of 27
doi:10.1093/imaiai/drn000

Conditional expectation estimation through attributable components

ESTEBAN G. TABAK*,
Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, USA
*tabak@cims.nyu.edu
GIULIO TRIGILA
Baruch College - CUNY, One Bernard Baruch Way, New York, NY 10010, USA
giulio.trigila@baruch.cuny.edu

[Received on 8 December 2017]

A general methodology is proposed for the explanation of variability in a quantity of interest x in terms of covariates $z = (z_1, \ldots, z_L)$. It provides the conditional mean $\bar{x}(z)$ as a sum of components, where each component is represented as a product of non-parametric one-dimensional functions of each covariate z_l that are computed through an alternating projection procedure. Both x and the z_l can be real or categorical variables; in addition, some or all values of each z_l can be unknown, providing a general framework for multi-clustering, classification and covariate imputation in the presence of confounding factors.
The procedure can be considered as a preconditioning step for the more general determination of the full conditional distribution $\rho(x|z)$ through a data-driven optimal-transport barycenter problem. In particular, just iterating the procedure once yields the second order structure (i.e. the covariance) of $\rho(x|z)$.
The methodology is illustrated though examples that include the explanation of variability of ground temperature across the continental United States and the prediction of book preference among potential readers.

Keywords: Principal component analysis, optimal transport, conditional density estimation

2000 Math Subject Classification: 34K30, 35K57, 35Q80, 92D25

1. Introduction

A broad class of problems in data analysis focus on explaining variability in a quantity of interest x in terms of the values of a set of covariates z. For instance, one can attribute part of the variability in blood pressure to factors such as age and exercise. One can go one step further and seek hidden factors: new covariates z with values to be determined as part of the problem so as to account for a significant share of the variability in x. Examples are clustering, which explains variability in terms of discrete classes, and principal component analysis, which does so in terms of continuous variables z of dimension lower than the original x.

The authors have recently developed Tabak and Trigila [2017] a unified framework for the explanation of variability, based on extensions of the mathematical theory of optimal transport. In this framework, one estimates the conditional probability distributions $\rho(x|z)$ by mapping them to their Wasserstein barycenter Agueh and Carlier [2011]. It was shown in Tabak and Trigila [2017] that principal components emerge naturally from the methodology's simplest setting, with maps restricted to rigid translations, and hence capturing not the full conditional probability distribution $\rho(x|z)$ but only its conditional expectation $\bar{x}(z)$. From here stems one of this article's legs: if one thinks of principal components in terms of explanation of variability, it is natural to consider more general scenarios, where the covariates, though still with values a priori unknown, are associated with particular attributes such

as space, time or activity networks, and hence required to be smooth in the topologies associated with these attributes.

For explaining variability, it is the subspace spanned by the principal components that matters, not the individual components. Seeking such subspace is a low-rank factorization problem, which can be computed with particular ease through alternating projections. This is the article's second leg: we develop far-reaching extensions of principal component analysis that can be computed efficiently through variations of the alternating projection methodology.

A third leg is the occurrence of missing and unstructured data. Frequently some –and sometimes most– of the observations x^j are incomplete. This is for instance the case of the Netflix problem, where each viewer only ranks a small subset of all movies, and each movie is only rated by a small fraction of viewers. In addition, observations may not take place on regular grids. For instance, a person –the "observation" in a census– has individual, values of age, height and weight drawn from a continuum, not a lattice. It turns out that both the extensions of principal components discussed here and their implementation in terms of alternating projections adapt very naturally to such scenarios, so they are developed without assuming that the data is complete or structured.

At the core of the methodology proposed lies a formal analogy between the low-rank factorization of tensors and the expansion of multivariate functions as sums of products of single-variable ones, as in the classical construction of solutions to linear partial differential equations via separation of variables and in multivariate Fourier and Taylor expansions. Thus, when a variable x depends on many covariates z, we write it as a sum of "components", each consisting of the product of functions of the individual covariates. In low-rank factorization, these one-variable functions are vectors (which can be thought of as functions of the row or column of the matrix entry being approximated), and hence completely unrestricted in the values they can adopt. When explaining variability, this is still the case for categorical covariates, such as sex or blood type. By contrast, in the separated variable solution to PDEs, the individual functions are typically given (sines, exponentials, special functions) and it is only the coefficients multiplying them that are fit to the data. Our procedure adopts free vectors, fitted to the data through the alternating projection methodology as in low-rank tensor factorization, but penalizes their non-smoothness as functions of those covariates z for which a distance can be defined (which includes continuous variables such as time or blood pressure, and ordinal ones such as rankings.) Thus the methodology unifies the handling of categorical and non-categorical covariates.

Describing multivariate functions in terms of sums of products of single-variable ones avoids the curse of dimensionality as the number of explanatory variables grows. Describing each of these single-variable functions by the values that it adopts on a grid further makes the computational cost largely independent of the number of observations available, particularly when combined with online learning and stochastic descent. In addition, the attribution of observations to grid points has a probabilistic interpretation that leads to a natural classification and clustering procedure, also applicable to the assignment of missing covariate values. The simultaneous consideration of the other covariates turns this into a general approach to multi clustering and classification in the presence of confounding factors.

Finding the conditional expectation is a regression problem, with a vast literature to which we cannot possibly do justice within the short frame of this introduction. Hence we shall only comment on methodologies that overlap significantly with the attributable components developed here. Least-square error regression Friedman et al. [2001] is the natural starting point, as we also employ its characterization of the conditional mean as the minimizer of the variance (though in our case the computation of the conditional mean is framed within the general setting of optimal transport, where least-squares emerges naturally from the use of the 2-Wasserstein distance.) Least-square regression is typically parametric, with linear models being by far the most commonly used, though these can be extended to handle non-

linearity, either explicitly through the introduction of nonlinear feature functions, or implicitly via the "kernel trick" Mohri et al. [2012]. Smoothing splines Hastie and Tibshirani [1990] are a less parametric alternative with a penalization of non-smoothness similar to ours, though without the penalization weighting system based on the number of observations partially attributed to each node.

Principal component analysis and, more generally, low-rank matrix factorization can be regarded as least-square regressions in terms of two categorical variables, the row and column of a matrix entry. Their calculation through alternate projections Golub and Van Loan [2012] is the computational basis for our extension to higher dimensional and less structured scenarios, while low-rank tensor factorization Grasedyck et al. [2013] underlies our treatment of more than two categorical covariates. Our subsequent extension to real covariates has formal similarities to the multivariable adaptive regression splines (MARS)Friedman [1991], though its elementary components are more general than MARS' hinges and the conceptual and computational framework of the two methodologies are substantially different.

For clarity of exposition, rather than posing the proposed procedure from the start, this article follows its conceptual evolution through a series of simple steps. Thus, after this Introduction, section 2 first briefly summarizes the relation between explanation of variability and optimal transport developed in Tabak and Trigila [2017] (subsection 2.1) and restricts the maps defining the Wasserstein barycenter to rigid translations (subsection 2.2). Section 3 first considers the estimation of conditional means for tensors and arbitrary sets of categorical covariates, yielding a matrix completion problem solved through a low rank factorization procedure. Then we consider covariates with an underlying metric, with components that are required to be smooth functions of these covariates, an extension motivated by the formal analogy between the components of a low-rank factorization and the single-variable factors in multivariate series such as Taylor's and Fourier's. Sub-section 3.1 further parameterizes these single-variable functions in terms on the values they adopt not on the observed samples but on an externally prescribed grid. Among the advantages of this switch is that it allows us to refine our characterization of smoothness to include higher derivatives (section 4). Perhaps more importantly, it permits the inclusion of latent and not fully available covariates, extending to procedure to handle clustering, multi-clustering, classification and covariate inference, all in the presence of confounding factors (section 5).

The description so far addresses the explanation of variability in real variables x. Section 6 demostrates that the procedure remains fundamentally unchanged when the x are categorical instead, as a simple embedding into Euclidean space allows one to define their conditional expected value $\bar{x}(z)$. Section 7 addresses the simple but important issue of post-processing of the results so as to make them interpretable by the user, providing the dependence of x on a selected number of covariates z through marginalization over all the others. Section 8 shows how iterating the procedure once allows one to capture not just the conditional expectation but also the full conditional covariance structure of x.

Section 9 includes a number of examples, some synthetic and some real applications, which display the versatility and power of the attributable component methodology. Finally, some conclusions and further extensions and drawn in section 10.

2. Removal of variability attributable to known factors

2.1 *The optimal transport barycenter problem*

Consider a variable $x \in R^n$ and a set of d covariates z. Removing the variability in x explainable by z is equivalent to transforming x through a z-dependent map $y = Y(x;z)$, so that y is independent of z. In order not to remove additional variability in x not attributable to z, one imposes the additional condition that the maps deform the data minimally. This process finds a natural formulation in the language of

optimal transport Tabak and Trigila [2017]:

$$\min_{Y,\mu} C = \int \left[\int c\,(x, Y(x;z))\, \rho(x|z)\, dx \right] \nu(z)\, dz, \quad y = Y(x;z) \sim \mu. \tag{2.1}$$

Here the cost $c(x,y)$ is a measure of the deformation brought about by the map, for which we will adopt the squared distance

$$c(x,y) = \|x-y\|^2, \tag{2.2}$$

and $\nu(z)$ represents the distribution of the covariates z. The resulting optimal $\mu(y)$, independent of z, is the ν-weighted *c-barycenter* of the conditional distributions $\rho(x|z)$ (Tabak and Trigila [2017]). For this particular cost function, the maps Y are given by the gradient of a potential function: $Y(x;z) = \nabla_x \phi(x;z)$ Brenier [1991].

In applications, one does not know the distributions $\rho(x|z)$ and $\nu(z)$: the data consists solely of a finite set of samples $\{x^i, z^i\}$, $i \in [1 \ldots m]$, in terms of which the objective function becomes

$$C = \frac{1}{m} \sum_{i=1}^{m} c\left(x^i, Y(x^i; z^i)\right).$$

Then one needs to restrict the space of allowable maps $Y(x;z)$ –else one would overfit the data– and weaken the requirement that all $\rho(x|z)$ be pushed into a single $\mu(y)$, since a) this is not achievable within a restricted family of maps and b) the $\rho(x|z)$ are only known through samples and $\mu(y)$ through their corresponding transformed points.

2.2 *Map restriction to rigid translations*

Arguably the simplest setting restricts the maps to rigid translations depending on z:

$$Y(x;z) = x - \beta(z). \tag{2.3}$$

For rigid translations, the only constraint that one can impose on the distribution of the range of Y is that its expected value $\bar{y}$ should be independent of z, with the implication that

$$\beta(z) = \bar{x}(z) - \bar{y}, \quad \text{with } \bar{y} = \int \bar{x}(z)\ \nu(z)\, dz. \tag{2.4}$$

In addition to its explanatory value, removing the expectation $\bar{x}(z)$ from the conditional distribution $\rho(x|z)$ is an effective pre-conditioner for the full optimal transport barycenter problem, as the following theorem shows:

Theorem: Consider the solution $Y_{full}(x;z)$ to the optimal transport barycenter problem (2.1) with the quadratic cost function (2.2), and the family of rigid translations $Y_{mean}(x;z)$ given by (2.3, 2.4). The latter pushes forward the conditional distributions $\rho(x|z)$ into new conditional distributions $\rho_*(y|z)$, with common mean $\bar{y}$ (by contrast, $Y_{full}(x;z)$ pushes forward the $\rho(x|z)$ into the z-independent barycenter μ.) Solve now again the full problem (2.1), this time using the $\rho_*(x|z)$ as input distributions, and denote the corresponding optimal map $Y_{nl}(x;z)$. Then

$$Y_{full}(y;z) = Y_{nl}(Y_{mean}(y;z);z). \tag{2.5}$$

In other words, the barycenter $\mu(y)$ of the $\rho_*(x|z)$ agrees with that of the original distributions $\rho(x|z)$, and the optimal maps pushing forward the $\rho(x|z)$ onto $\mu(y)$ agree with the composition of the rigid

translations and the optimal maps from the $\rho_*(x|z)$ to $\mu(y)$. Finding Y_{nl} is typically much easier computationally than finding Y_{full}, as the $\rho_*(x|z)$ are closer to each other than the $\rho(x|z)$, since they share the same expected value for all values of z.

Proof: The statement above follows from two basic ingredients: a characterization of the barycenter distribution $\mu(y)$ and the inverse $X(y;z)$ of $Y(x;z)$ as the unique satisfiers of the two properties Álvarez-Esteban et al. [2016], Kuang and Tabak [2017]:

1. Each point y is the barycenter of its images under X:

$$y = \int X(y;z)\ \nu(z)\ dz$$

2. For each value of z, $X(y;z)$ maps optimally $\mu(y)$ onto $\rho(x|z)$

and a characterization of optimal maps under the square-distance cost as gradients of a convex function (Caffarelli [2003]):

$$X(y;z) = \nabla_y \psi(y;z), \quad \psi(\cdot;z) \text{ convex.}$$

The composition of the two inverse maps is

$$X_{full}(y;z) = X_{mean}(X_{nl}(y;z);z) = X_{nl}(y;z) + \beta(z) = \nabla_y \psi_{nl}(y;z) + \beta(z),$$

with $\psi_{nl}(y;z)$ convex in y, and therefore

$$X_{full}(y;z) = \nabla_y \psi_{full}(y;z), \quad \text{with } \psi_{full}(y;z) = \psi_{nl}(y;z) + \beta(z)y \text{ also convex,}$$

so it is optimal, as required by property 2. Also

$$\begin{aligned} \int X_{full}(y;z)\ \nu(z)\ dz &= \int [X_{nl}(y;z) + \beta(z)]\ \nu(z)\ dz \\ &= y + \int [\bar{x}(z) - \bar{y}]\ \nu(z)\ dz = y, \end{aligned}$$

thus satisfying property 1 and concluding the proof.

3. A low rank tensor factorization, separated variable procedure

Given a set of m observations x^i and L corresponding covariates z_l^i, the task of removing the variability associated with z from x through the z-dependent rigid translation

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

reduces to regression, i.e. to estimating the conditional mean $\bar{x}(z)$. This can be characterized as the minimizer of the variance:

$$\bar{x}(z) = \arg\min_f \sum_i \left\| x^i - f(z^i) \right\|^2 \tag{3.1}$$

over a proposed family of functions $f(z)$.

Any multivariable function $f(z)$ can be approximated to arbitrary accuracy by the superposition of products of functions f_l of the individual components z_l of z:

$$f(z) \approx \sum_k \prod_l f_l^k(z_l).$$

Classical examples are the power series when $z \in \mathbb{R}^L$,

$$f(z) \approx \sum_k a_k \prod_{l=1}^{L} z_l{}^{s_l^k}, \quad s^k \in \mathbb{N}^L,$$

the Fourier series when z is in the 2π-periodic L-dimensional torus,

$$f(z) \approx \sum_k a_k e^{i\sum_l \xi_l^k z_l}, \quad \xi^k \in \mathbb{Z}^L$$

and the singular value decomposition when $z = (i, j) \in \mathbb{N}^2$,

$$x(i, j) \approx \sum_k \sigma_k \, u^k(i) v^k(j).$$

This suggests proposing the approximation

$$\bar{x}(z) \approx \sum_{k=1}^{d} \prod_{l=1}^{L} V(l)^k(z_l). \tag{3.2}$$

If x is vectorial, we can make it scalar by including the index as one more factor z_l. Then (3.1) reduces to minimizing

$$L = \sum_i \left(x^i - \sum_{k=1}^{d} \prod_{l=1}^{L} V(l)^k(z_l^i) \right)^2 \tag{3.3}$$

over the degrees of freedom available in the specification of the functions $V(l)^k(z)$.

Consider first the particular case where all z_l are categorical variables –such as the vector index above– with a finite number of possible values. Then, for each l, $V(l)$ is a matrix with components $V(l)^k_{z_l}$, and the minimization of L becomes the low-rank factorization of the tensor $x(z_1, \ldots, z_L)$ from the available –possibly repeated– entries $\{x^i\}$. This problem can be solved through an alternating direction methodology, minimizing L alternatively over each $V(l)$. This minimization yields, for each value j of z_l,

$$V(l)_j = \left(\sum_{i \in I_j} x^i \prod_{b \neq l} V(b)_{z_b^i} \right) \left[\sum_{i \in I_j} \left(\prod_{b \neq l} V(b)_{z_b^i} \right)^T \left(\prod_{b \neq l} V(b)_{z_b^i} \right) \right]^{-1}, \tag{3.4}$$

where

$$I_j = \left\{ i : z_l^i = j \right\}.$$

Such tensor-completion through low-rank factorization has a broad range of applications. Consider the following two typical examples:

1. A blood test, measuring various quantities such as cholesterol level and white cell counts. Here the variable x –a scalar– is the value of a measurement. Some of the categorical covariates z_l that one would use to account for variability in x are:

(a) The quantity being measured (the row in regular low-rank matrix factorization),

(b) The facility where the test was performed,

(c) Individual characteristics of the patient (sex, ethnicity, treatment, etc.)

Excluded so far are covariates with continuous values, such as age, weight, time of the day and results of prior tests; these will be discussed below. Further down we will also include latent covariates, such as distinct groups into which the patients can be divided in light of their test results, a case of clustering in the presence of confounding factors.

2. The "Netflix problem": x is the rating given by viewers to movies, z_1 the viewer and z_2 the movie. Further extensions below allow for the inclusion of more granular characteristics, such as movie type and year of release, and viewer's age and location.

In tensor-completion, there is only a finite set of possible arguments $j = z_l^i$ for $V(l)(z_l)$, and the corresponding values $V(l)_j^k$ are free for the minimization to determine. By contrast, when using separation of variables for solving PDE's, the form of the $V(l)^k(z)$ are known ab-initio (sines, exponentials, special functions), with only their amplitudes available to fit initial or boundary data. In a data-analysis context, one would call such model *parametric*. Without a specific principle to guide the selection of the functions $V(l)(z)$, this approach suffers from the arbitrariness of any such choice, which could lead to poor, overfitted or uninterpretable results.

Instead, we can preserve the freedom in the assignment of values to each $V(l)^k_{z_l^i}$ while enforcing a smooth dependence of $V(l)^k$ on z_l by adding to the objective function a penalization for non-smoothness, proportional to

$$\sum_{i,j>i} C^l_{i,j} \left(V(l)^k_{z_l^i} - V(l)^k_{z_l^j} \right)^2 ,$$

with $C^l_{i,j}$ inversely proportional to a measure of the distance between z_l^i and z_l^j.

The logic behind this choice is the following:

- When z_l^i and z_l^j are close to each other, the corresponding $V(l)^k$ need also be close, as otherwise they would incur a penalty. This is how one enforces smoothness of $V(l)^k(z)$.
- As the penalty term is quadratic in each $V(l)$, solving for each $V(l)$ with all other $V(b)$ fixed amounts to inverting a linear system, as in conventional alternating projections.

The new objective function has the form

$$\begin{aligned} L \;=\; & \sum_i \left(x^i - \sum_k \prod_l V(l)^k_{z_l^i} \right)^2 + \\ & \sum_{l\in M} \lambda_l \sum_{k=1}^{d} \left(\prod_{b\in L, b\neq l} \|V(b)^k\|^2 \right) \sum_{i,j>i} C^l_{i,j} \left(V(l)^k_{z_l^i} - V(l)^k_{z_l^j} \right)^2 , \end{aligned} \tag{3.5}$$

where M is the set of factors with an underlying metric, for which the smoothness of functions is required. The pre-factors to the penalty terms, products of squares of the norms of the $V(b)^k$, are included so that the objective function is invariant under re-scalings of the $V(l)^k$ that preserve their

product. Otherwise, the penalty terms could be made arbitrarily small not by enforcing smoothness, but by multiplying each $V(l)$ by a small factor and absorbing this factor into other V's less constrained by the smoothness requirement. The constants λ_l quantify the amount of penalization associated to smoothness for each factor.

3.1 *Semi-parameterization*

Now we extend the procedure above to simultaneously address various issues:

1. As the number of observations grows, so does the number of unknowns $V(l)^k_{z^i_l}$, turning the solution to the linear system from each step into a computational bottleneck. Yet the underlying functions $V(l)^k(z)$ do not change in any fundamental way; we are just seeking their values at a larger set of points $\{z^i_l\}$. Ultimately, we want to resolve these functions to a level of detail determined by our needs, not by the number of observations available.

   Also, the observed values z^i_l may not provide an adequate grid for resolving the $V(l)^k(z)$, as they may cluster in some locations and under-resolve others. As a consequence, the cost matrix C^l_{ij} may be highly unbalanced.

2. Frequently, values of a covariate z_l are repeated systematically across the dataset. Two typical scenarios are the following:

   - The variable z_l can only adopt a finite number of values (as in rankings), which are then necessarily repeated many times.
   - The dataset consists of observations drawn from various individuals, each observed many times. The characteristics z of each individual then appear repeatedly, once for each such observation. Examples: patient's age in medical studies where each patient is followed over time, a meteorological station's latitude and elevation in datasets comprising observations from more than one station.

   In such situations, the most sensible choice is to adopt just one value of $V(z)$ for each value of z, rather than penalizing their variation.

3. When some values z^i_l are missing from the dataset –even though x^i and possibly other covariates z^i_b have been recorded–, applying the procedure requires a way to assign values to the corresponding $V(l)^k_{z^i_l}$. Sometimes the z^i_l are unknown ab-initio, as they are *latent* variables to be assigned by the algorithm to further explain variability in an unsupervised fashion (We will address latent variables in section 5 using the tools developed here.)

4. The procedure is non-parametric, with the functions $V_l(z)$ defined only by their values on the sample points z^i_l and with no specified functional form, just constrained by the penalization on non-smoothness. Yet there are reasons why one may want to have a more explicit description of these functions, both within and after the procedure:

   (a) **Prediction and imputation:** One typical goal of data analysis is to predict the value of x or its probability distribution for a new set of values of the covariates z. This requires evaluating the $V(l)^k(z_l)$ at the new values provided.

(b) **On-line learning:** When new data keeps flowing in, as in weather forecasting, one should not start the procedure from scratch every time. Avoiding this requires a current estimation of the $V(l)^k(z_l)$ for the newly arrived values of z_l.

5. Even though the procedure as described so far combines naturally categorical and non-categorical variables, the latter are penalized for non-smoothness, while the former are not. Hence the algorithm will be biased to explain as much variability as possible in terms of the categorical variables, a possibly undesired feature. In fact, the exact opposite is often required, as some of these categorical variables represent "idiosyncratic" variability (as do the "viewer" variable in the Netflix problem, the "patient" variable in a medical study and the "stock" variable in finance), which one would like to use to explain only the variability left once more descriptive factors (age, weight, industry) have had their say.

In order to address the issues above, we introduce for each l a grid (not necessarily regular) of values $z_g(l)^j$, and the linear interpolation scheme

$$z_l^i = \sum_j \alpha(l)_i^j z_g(l)^j, \quad \text{so that we can posit} \quad \tilde{V}(l)^k_{z_l^i} = \sum_j \alpha(l)_i^j V(l)_j^k. \tag{3.6}$$

Here the $\tilde{V}_{z_l^i}$ are the values associated with observation i, while the V_j are the smaller set of unknowns associated with the grid $z_g(l)^j$. Then the problem becomes

$$\min_V \sum_i \left(x^i - \sum_k \prod_{l\in L} \sum_j \alpha(l)_i^j V(l)_j^k \right)^2 + \sum_{l=1}^{L} \lambda_l \sum_k \left(\prod_{b\in L, b\neq l} \|V(b)^k\|^2 \right) \sum_{i,j>i} C_{i,j}^l \left(V(l)_i^k - V(l)_j^k \right)^2, \tag{3.7}$$

which includes (3.5) as a particular case with the $\alpha \in \{0,1\}$. Let us see how this modification addresses each of the issues raised before:

1. The number of unknowns $V(l)_j^k$ is controlled by the externally supplied grid, not the number of observations available. Moreover, this grid can be designed so as to be well-balanced, with neither clusters nor holes, and with as fine a mesh as desired in regions where higher resolution is sought.

2. A repeated value of z_l^i is assigned through the same coefficients α to the same grid points $z_g(l)^j$. When the number of values that z_l can adopt is finite, those are by default the values given as grid points, with corresponding $\alpha \in \{0,1\}$.

3. When the value of z_l for observation i is missing, one should marginalize over it, adopting for $V(l)^k_{z_l^i}$ the expected value of $V(l)^k(z)$. This corresponds to assigning as $\alpha(l)_i^j$ the mean value $\alpha(l)_i^j = \frac{1}{n_t}\sum_{t=1}^{n_t} \alpha(l)_t^j$ of α over the observations z_l^t for which z_l is known. If there is a surrogate for z_l that allows us to infer that the missing z_l^i is closer to some of the observed z_l^j than others –often the time t adopts this role–, a correspondingly weighted mean of α should be adopted.

4. In order to evaluate $\tilde{V}(l)^k(z)$ for a new value of z, one just needs to find the values of α that interpolate z in the grid and apply (3.6).

5. One can extend the penalization to categorical variables by adding a new, dummy value of z to the grid, with corresponding unknown $V(l)^k_o$, and making each $C^l_{oj} = 1$ and all other $C^l_{ij} = 0$. This penalizes variability in the $V(l)^k$ without reference to any underlying metric. The final value of $V(l)^k_o$ will be the empirical mean of $V(l)^k(z)$, and the quantity penalized, its variance.

4. Higher orders of smoothness

Our characterization of smoothness has been restricted so far to continuity, enforced through the penalization of large differences in $V(l)$ when the corresponding z_l are close. With the procedure extended to incorporate well-balanced grids for each z_l, we can refine this characterization, penalizing higher derivatives of $V(l)$. In addition to providing a smoother fit to the data, this refinement is also useful for extrapolation. Consider a situation where one would like to estimate $V(z)$ for a value of z outside the range of the data, say $z > z_{\max}$. The choice most consistent with the current algorithm is to set $V(z) = V(z_{\max})$, as this is the most "continuous" choice. If instead we were minimizing $\|V''\|^2$, then we would extrapolate linearly, following the slope at $z_{\max}$. An intermediate choice follows from minimizing $a^2\|V'\|^2 + |V''\|^2$, as now beyond $z_{\max}$, not being required to explain any data, one would we solving the ODE

$$\frac{\delta}{\delta V'} \int \left(a^2\|V'\|^2 + |V''\|^2\right) dz = 0 \to \frac{d^2V'}{dz^2} - a^2V' = 0,$$

with solution

$$V'(z) = V'(z_{\max})e^{-a(z-z_{\max})},$$

where the slope V' agrees with its value at $z_{\max}$ but then decays to zero.

One simple recipe for higher-order smoothness is to penalize the squared norm of a finite difference approximation to a derivative, as in the early versions of smoothing splines (Hastie and Tibshirani [1990]). In addition, it is convenient to weight this norm, so that each value $V(l)^k_j$ is rewarded for its explanatory value and penalized for non-smoothness in a balanced way. In particular, if values of z_g are added beyond the range of the data, these should not affect the fit within the range. For concreteness, we develop here the procedure to penalize the first and second derivatives of $V(l)$; extending this to higher derivatives is straightforward.

Given a sorted grid $\{z^j\}$, the three-point finite-difference approximation to the first and second derivatives of V at point z_j is given by

$$V'_j \approx aV_{j-1} + bV_j + cV_{j+1}, \quad V''_j \approx AV_{j-1} + BV_j + CV_{j+1},$$

with

$$a = -\frac{{\Delta_+}^2}{D}, \quad b = -\frac{{\Delta_+}^2 - {\Delta_-}^2}{D}, \quad c = \frac{{\Delta_-}^2}{D},$$

$$A = \frac{\Delta_+}{D}, \quad B = -\frac{\Delta_+ - \Delta_-}{D}, \quad C = -\frac{\Delta_-}{D},$$

$$\Delta_+ = z^{j+1} - z^j, \quad \Delta_- = z^{j-1} - z^j, \quad D = \frac{1}{2}\Delta_+\Delta_-\left(\Delta_- - \Delta_+\right).$$

Defining the weight w_j of grid point z^j_g by the sum of its contributions to the data points $\{z^i\}$:

$$w_j = \varepsilon + \sum_{i=1}^{m} \alpha^j_i,$$

where $\varepsilon \ll 1$ is the weight assigned to grid values away from the data, one can propose a penalty for non-smoothness of the form

$$\lambda \sum_j w_j \left(\delta \|V_j'\|^2 + (1-\delta)\|V_j''\|^2\right) = \lambda V'CV.$$

(The C so defined is a non-negative-definite tridiagonal matrix.) Here $\delta \in [0,1]$ measures the relative weight given to the first and second derivatives of V. For notational clarity, we are absorbing into the constant λ the product of the square norms of the $V(b)_k$ with $b \neq l$ in (3.7). When written in full, the problem becomes

$$\min_V \sum_i \left(x^i - \sum_k \prod_{l \in L} \sum_j \alpha(l)_i^j V(l)_j^k \right)^2 + \sum_{l=1}^{L} \lambda_l \sum_k \left(\prod_{b \in L, b \neq l} \|V(b)^k\|^2 \right) V(l)^{k\prime} C^l V(l)^k. \tag{4.1}$$

5. Clustering, classification and posterior assignment of missing covariate values

The previous section discussed how to handle situations in which some, but not all values of a given covariate z_l where missing. When no value is known, z_l is a latent variable, whose values should be assigned by the algorithm to further explain the variability in the data. This section extends the procedure to handle the discovery of latent categorical variables. Assigning a value z_l^i to data point x^i is a clustering problem, which the presence of other covariates z_b, $b \neq l$ turns it into clustering in the presence of confounding variables.

In our setting, this gives rise to a mechanism analogous to k-means and Expectation Maximization (Friedman et al. [2001]). In order to assign z_l^i, one first sets a number n_{cl} of clusters, and defines the corresponding "grid" $z_g(l)^j$, with $j = 0 \ldots n_{cl}$ (with 0 standing for the dummy element to which no point is assigned for the penalization of functions of categorical variables).

From eq. (3.6), the z_l^i sought follows from the values of the $\alpha(l)_i^j$ used to interpolate on the grid $z_g(l)$. These determine automatically to which l-variable cluster the observation x^i belongs. When clustering through hard assignments, all entries $\alpha(l)_i^j$ but one equal zero, except the one specifying to which class j the data point x^i is assigned. The values of $\alpha(l)_i^j$ are linked to the corresponding values of $V(l)_j^k$ in a two step procedure similar to Lloyd's algorithm for k-means:

- Given the values of $V(l)_j^k$, we assign z_l by choosing the j for which $\alpha(l)_i^j = 1$ so that the resulting predicted mean $\bar{x}(z^i)$ is closest to the observation x^i (in other words, to decrease as much as possible the first term in 4.1). This corresponds to the reassignment step based on proximity within a given centroid in Lloyd's algorithm.
- Once the alpha are known the values of $V(l)_j$ are updated as before by minimizing L (eq. 4.1). This step finds the mean of x within each class defined by the values of z_l and corresponds to the update of the centroid in the Lloyd's algorithm.

A version with soft assignments can be developed along similar lines, with each observation x^i having a probability α_i^j of belonging to class j. Then the $V(l)$ are updated as before, and the α_i^j can be assigned through the following Bayesian procedure:

1. For each cluster j, compute the square distances d_i^j between the cluster's mean $\bar{x}(z_1^i,\dots,z_l^j,\dots z_L^i)$ and each observation x_i:

$$d_i^j = \left(x^i - \sum_k \left(\prod_{b\neq l} \sum_j \alpha(b)_i^j V(b)_j^k \right) V(l)_j^k \right)^2$$

and the standard deviation

$$\sigma_j = \left(\frac{\sum_i \alpha(l)_i^j d_i^j}{\sum_i \alpha(l)_i^j} \right)^{\frac{1}{2}}.$$

2. Denoting $Q(l)_s$ the s'th set of n_s observations $\{i\}$ constrained to share the same assignments $\alpha(l)_i$, compute the average square distance to each cluster's mean:

$$D_s^j = \frac{1}{n_s} \sum_{i\in Q(l)_s} d_i^j.$$

3. Update $\alpha_i^j = P_s^j$ $(i \in Q(l)_s)$ through Bayes formula, modeling the clusters as one-dimensional Gaussians and adopting P_s^j itself as prior:

$$\rho^j \propto \frac{1}{\sigma_j} \exp\left(-\frac{D_s^j}{2{\sigma_j}^2} \right),$$

$$P_s^j = \frac{P_s^j \rho^j}{\sum_h P_s^h \rho^h},$$

a choice with positive feedback that leads to convergence to rigid assignments $P_s^j \in \{0,1\}$. If soft assignments are desired throughout, one should use in Bayes formula the non-informative prior $P_{prior}^j = \frac{1}{n_{cl}}$ instead of P_s^j.

In classification problems, the α_i are known (and fixed) for the training set, while they evolve as above for the testing set. To be consistent with the Baysian approach, one should in this case use either the fixed prior

$$P_{prior}^j = \frac{1}{m_k} \sum_{i=1}^{m_k} \alpha_i^j,$$

where the sum is taken over the m_k observations for which z^i is known, or, in the spirit of EM, the evolving prior

$$P_{prior}^j = \frac{1}{m} \sum_{i=1}^{m} \alpha_i^j$$

computed over all observations, including the ones for with α is being updated.

Notice that the procedure just described applies without changes when the latent variable z_l is non-categorical. Here we would not call the procedure "classification", as this refers to categorical classes; a more appropriate name would be "posterior assignment of unknown covariate values". It applies, for instance, to softly assign an age to individuals for which it has not been recorded.

5.1 *Use for complexity reduction*

In large datasets, an idiosyncratic variable z_l (the viewer, the movie, the patient, the station, the stock) can adopt a very large number of values. Since idiosyncratic variables are categorical, we cannot reduce the number of unknowns $V(l)^k$ simply by interpolating on a grid $z_g(l)$ via (3.6), since one would not know, for instance, how to interpolate one patient among others when the more informative variables –age, weight, etc.– are already considered separately. However, one can still use (3.6) to the same effect, but with the values of α unknown beforehand, i.e. performing clustering. In the example just mentioned, we would replace the idiosyncratic variable "patient" by "patient type", with the number of types decided based on the level of resolution sought.

Below, we will show an application of this clustering procedure to a matrix completion problem for a book recommendation system, where the idiosyncratic variables "reader" and "book" are clustered into "reader type" and "book type" (different from the book genre or any other covariates taken into account explicitly.)

6. Categorical dependent variables

The methodology allows for covariates z of any type: categorical, ordinal, real, periodic. Does a comparable level of generality extend to the variable x whose unexplained variability one attempts to minimize? In particular, there are many scenarios of interest where x is categorical: has person z_1 read book z_2, will patient z_1 develop illness z_2, which party does person z_1 vote for, etc.

The object that the procedure seeks, the conditional expectation $\bar{x}(z)$, can still be assigned a meaning when x is categorical. For instance, for a binary x, one can assign the values $\{0,1\}$ to the two possible outcomes, and define

$$\bar{x}(z) = 0 \cdot P(0|z) + 1 \cdot P(1|z) = P(1|z),$$

where the $P(x|z)$ are the z-conditional probabilities for the outcome x. In doing so, we have embedded the space of outcomes x into a metric space, i.e. the real line. The resulting $\bar{x}$ has a clear meaning in terms of the underlying probability distribution P. Similarly, if x has three possible outcomes, we can embed those into the two-dimensional plane, as vertices of an equilateral triangle. The important notion here is that all the points so assigned be equidistant, not to affect the categorical nature of the variable through a metric-induced preferential ordering. Then, following the attributed component methodology, the two dimensions of x in this example are captured through a covariate z with values in $\{1,2\}$. More generally, a number n of outcomes is embedded into an $n-1$-dimensional space.

A simpler alternative would embed a variable x with n outcomes into an n-dimensional space, as vertices of the corresponding simplex, i.e. the 1 of K encoding. This has more straightforward interpretability, as each component of $\bar{x}$ represents the probability of the corresponding outcome. On the other hand, it has one more dimension to consider and it requires imposing the constraint that $\sum_i \bar{x}_i(z) = 1$, which increases –albeit slightly– the computational complexity of the algorithm.

7. Interpretability through marginalization

The procedure provides an estimation of the conditional mean of the form

$$\bar{x}(z_1,\ldots,z_L) = \sum_{k=1}^{d} \prod_{l\in L} \sum_{j} \alpha(l)^j(z_l) V(l)^k_j,$$

where the $V(l)^k_j$ are known, and $\alpha(l)^j(z_l)$ are the coefficients that interpolate z_l in the given grid $z_g(l)^j$. It is straightforward to use this expression to find the estimate for $\bar{x}$ for a new value of z. On the other hand, determining how x depends on each individual z_l under average conditions for the other covariates z_b is a question of marginalization: find

$$\begin{aligned}\bar{x}(z_l) &= \int x\,\rho(x|z)\,dz_1\ldots dz_{l-1}dz_{l+1}\ldots dz_L \\ &= \int \bar{x}(z)\,dz_1\ldots dz_{l-1}dz_{l+1}\ldots dz_L. \end{aligned} \tag{7.1}$$

Evaluating at $z_l = z_l^*$ and replacing expectations by the corresponding empirical means, one has

$$\begin{aligned}\bar{x}(z_l^*) &= \frac{1}{m}\sum_{i=1}^{m}\bar{x}(z_1^i,\ldots,z_l^*,\ldots,z_L^i) \\ &= \sum_{k=1}^{d}\frac{1}{m}\sum_{i=1}^{m}\left[\prod_{b\neq l}\sum_j \alpha(b)_i^j V(b)_j^k\right]\sum_j \alpha(l)^j(z_l^*)V(l)_j^k.\end{aligned}$$

Similarly, one can compute the marginal of $\bar{x}$ over s variables $\{z_{l_h}\}$, through

$$\bar{x}(z_{l_1}^*,\ldots,z_{l_s}^*) =$$
$$\sum_{k=1}^{d}\frac{1}{m}\sum_{i=1}^{m}\left[\prod_{b\notin\{l_h\}}\sum_j \alpha(b)_i^j V(b)_j^k\right]\prod_{h=1}^{s}\sum_j \alpha(l_h)^j(z_{l_h}^*)V(l_h)_j^k. \tag{7.2}$$

8. Further explanation of variability: the variance

The estimation of the conditional mean $\bar{x}(z)$ extends effortlessly to second order statistics. Introducing the filtered variable y through

$$y = x - \bar{x}(z), \tag{8.1}$$

one can for instance estimate the z-dependent variance $\sigma^2(z)$ by applying the procedure not to x but to $w = y^2$, since

$$\sigma^2(z) = \bar{w}(z).$$

Often it is a covariance matrix we are after, with one of the z_l representing the "entry" or "variable" in x. If two entries $z_l = i$ and $z_l = j$ are always observed simultaneously –as are, for instance, two quantities in a blood test–, then one writes $w = y(z_l = i, z_b)y(z_l = j, z_b)$ with z_b comprising all covariates except z_l, in order to estimate $\Sigma_i^j = \bar{w}$ as a function of z_b. Rather than estimating each entry (i, j) at a time though, it is better to estimate the whole covariance matrix $\Sigma(i, j, z_b)$ at once, as this will use all the information available and link the various entries of Σ, so that in particular for each value of z_b, $\Sigma(:,:,z_b)$ is a non-negative definite matrix.

Yet this procedure extends to much more general situations. One might seek, for instance, the correlation between x at two locations/times z_l as a function of the distance/interval Δz_l between them. For this, we introduce all products $w = x(z_l^i)x(z_l^j)$ and a new covariate $\Delta z = z_l^i - z_l^j$. For the remaining covariates z_b, one may take the average between their values at the two observations. Alternatively, one may restrict the products considered to those where the two values of z_b are not too far from each other, or include the Δz_b as extra covariates. Similar examples abound: correlation between height or weight at different ages, between precipitation at different locations as a function of time-lag, etc.

9. Examples

This section includes examples of application of the attributable component methodology. Though one of the examples uses real data –hourly ground temperatures across the continental US–, the purpose here is not to highlight any particular application, but to illustrate various aspects of the workings of the procedure. Thus we start with a couple of simple synthetic examples that directly compare the components discovered by the algorithm with the actual variable-separated solution underlying the data. Then we consider the real example of ground temperatures, which combines different kinds of covariates: categorical, real and periodic, is big enough to require the use of stochastic descent, and also illustrates different uses of the iteration of the procedure that captures second order structure underlying the data. Finally, we consider a synthetic Netflix-like problem, posed in terms of book ratings by readers, that illustrates other capabilities of the methodology, including bi-clustering for complexity reduction in the presence of confounding factors and the use of idiosyncratic variables (reader, book) in combination with more descriptive –but less specific– covariates such as age, location and book type.

9.1 *Numerical separation of variables*

Our first two examples analyze data generated using synthetic models where the distribution underlying the samples takes the variable-separated form that our algorithm adopts as ansatz. The first of these models has three covariates z_l, two periodic and one real, and takes the form

$$x^i = \bar{x}(z^i) + w^i, \quad \bar{x}(z) = \frac{1}{5}\cos(z_1)\sin(z_2)\,z_3, \quad w \sim \mathcal{N}(0,1),$$

where the $m = 1000$ samples from z_1, z_2 and z_3 are drawn uniformly in the interval $[\pi,\pi]$. Figure 1 displays the results of applying the procedure with $d = 1$ component and $\lambda_l = 0.01,\ l = \{1,2,3\}$. The initial values of $V(l)^k_j$ (here and in all examples in this paper) are taken as small, symmetry-breaking perturbations of the uniform $V(l)^k_j = 1$.

The second model, with one less covariate and one more component, is

$$x^i = \bar{x}(z^i) + w^i, \quad \bar{x}(z) = 4\cos(z_1)\sin(z_2) + 0.3\ z_1 {z_2}^2, \quad w \sim \mathcal{N}(0,1). \tag{9.1}$$

Here we also draw $m = 1000$ samples of z_1 and z_2 uniformly in the interval $[\pi,\pi]$ and adopt $\lambda_l = 0.01,\ l = \{1,2\}$, setting this time the number of components to $d = 2$. The purpose of this example is to display the non-uniqueness of the separation into components: as shown in Figure 2, the two components recovered by the algorithm, though fitting $\bar{x}(z)$ to perfection, do not agree with those used to generate the data in (9.1). This non-uniqueness of the decomposition into components is general whenever $L \geqslant 2$. For instance, for any f_1, f_2, g_1, g_2, we can write

$$f_1(x)g_1(y) + f_2(x)g_2(y) = \tilde{f}_1(x)\tilde{g}_1(y) + \tilde{f}_2(x)\tilde{g}_2(y),$$

with

$$\tilde{f}_1(x) = f_1(x), \quad \tilde{g}_1(y) = g_1(y) + g_2(y), \quad \tilde{f}_2(x) = f_2(x) - f_1(x), \quad \tilde{g}_2(y) = g_2(y).$$

The algorithm picks, among all equivalent decompositions, the "smoothest" one, in the sense of minimizing the second term in (4.1) penalizing non-smoothness. Notice though that $\bar{x}(z)$ itself and all marginals are still uniquely determined, so the non-uniqueness of the decomposition affects neither the numerical predictions nor the interpretability of the results.

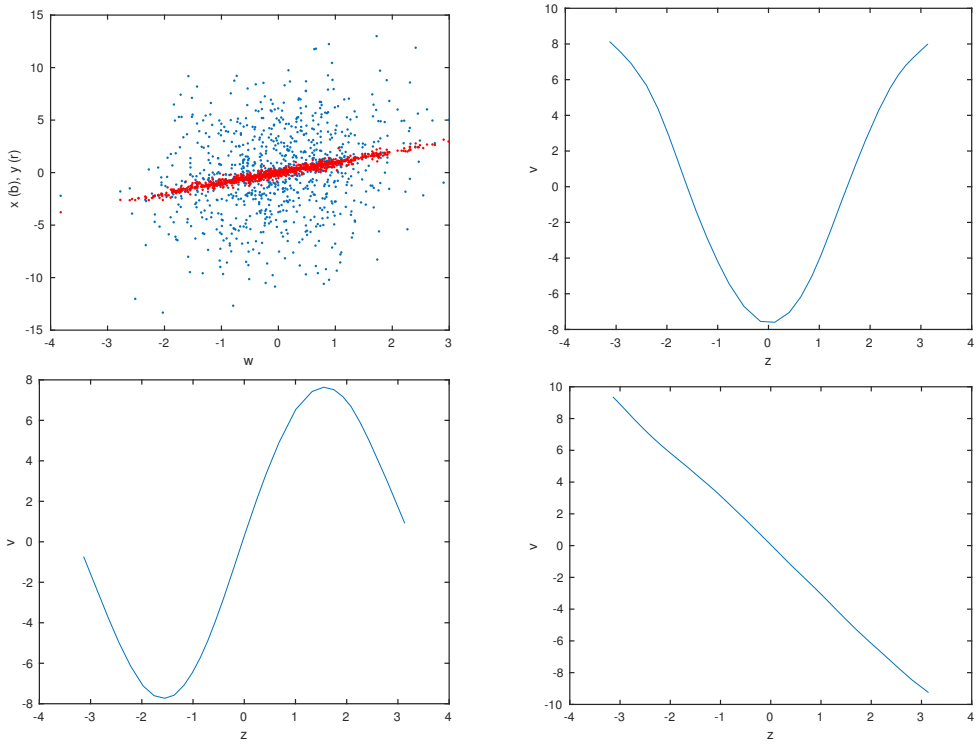

FIG. 1. Upper left: original data x (in blue) and the filtered signal y (in red) as a function of the noise w. As one can see, y is a function of the noise, i.e. of the unexplained variability, as all explainable variability has been filtered away. Upper right: first (and only) component as a function of z_1. Lower left: same component as a function of z_2. Lower right: same component as a function of z_3. As one can see, the true separated solution underlying the data has been fully recovered.

9.2 *Ground temperature variability in the continental US*

For a second example, we use hourly measurements of the ground-level temperature at 47 stations located across the continental United States and one in Canada, publicly available from NOAA [1]. We picked an array of stations that covers the US roughly uniformly and have data available since at least the year 2005 to the present. In this example, the variable x to explain is the temperature itself, measured in degrees Celcius, and we have chosen as covariates z_l the following four quantities:

1. The station itself, a categorical variable with values $z_1 \in [1,2,\ldots,48]$.
2. The local time of the day $z_2 \in [0,24]$, periodic. We adopted a grid with 24 elements, with one grid point per hour.
3. The season, described as day of the year, $z_3 \in [0,365.25]$, periodic, also with a grid of 24 points.
4. Time in years $z_4 \in [2005,2017.3]$, real, included to account for climate variability in a multi-year scale, with a grid of 75 points.

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

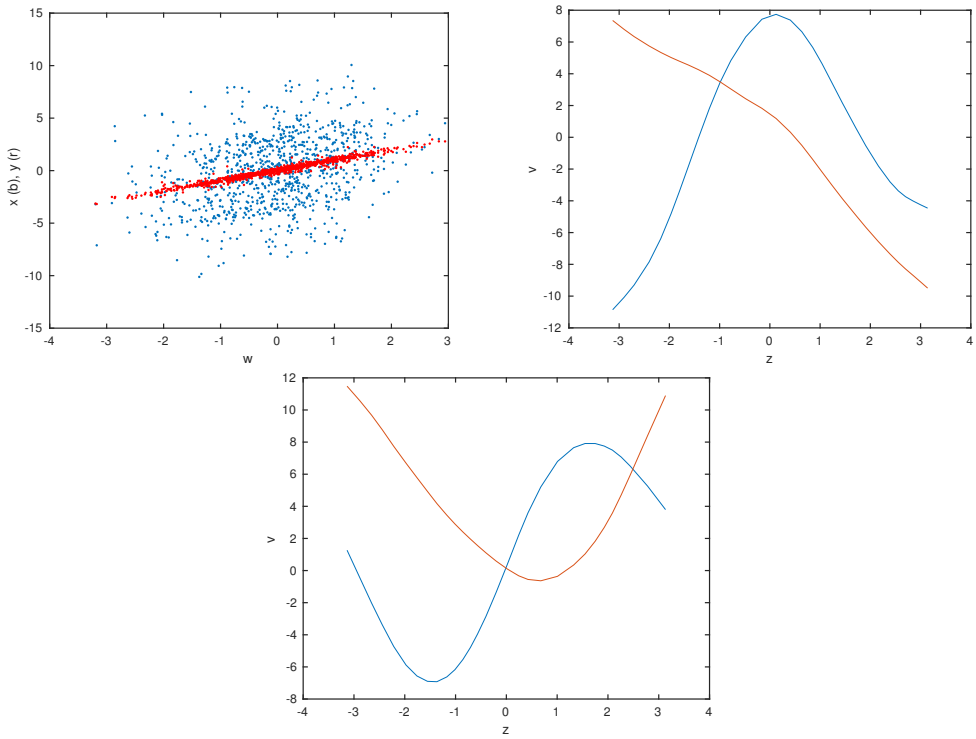

FIG. 2. Upper left panel: original data x (in blue) and the filtered signal y (in red); the latter is a function of the unexplainable variability, i.e. the noise w. Upper right panel: the two components as functions of z_1. Lower left: same as functions of z_2. This separated solution does not agree with the one proposed to generate the data, yet it is completely equivalent, though smoother from the perspective of the objective function.

The total number of observations $m \approx 5.7 \cdot 10^6$ was large enough to justify the use of stochastic gradient descent, which we performed by choosing at each step of the algorithm a random subset of 10^5 entries of x. We adopted $d = 8$ components and penalization parameters $\lambda = 0.0001$ for the station and $\lambda = 0.005$ for all other covariates. The reduction in variability can be quantified by the variance, which moved from 151.7 for x to 23.8 for the transformed y, corresponding to a reduction in standard deviation from 12.3 to 4.9 degrees Celcius. Figure 3 (upper left panel) shows the 8 periodic components associated with z_2, the time of the day. Similarly, there are 8 components, not displayed, associated with each of the other 3 covariates. The set of all these components constitute the output of the algorithm, from which all predictions are made. Figure 3 (upper right panel) displays the easier to interpret marginalized prediction, where the dependence of $\bar{x}$ on the time of the day is marginalized over time of year, station and multi-year variability. Similarly, the lower left panel of the same figure displays the marginalized dependence of $\bar{x}$ on the time of the year, with cold winters and hot summers, and the lower right panel displays through color the marginalized dependence of $\bar{x}$ on the station (this dependence is marked mostly by latitude, but also by other station characteristics such as elevation and distance from the sea.)

Figure 4 (left panel) shows the data (in blue) and predicted mean (in red) individualized for a particular station: Manhattan, Kansas, at four levels of detail: global, yearly, monthly and weekly. We see

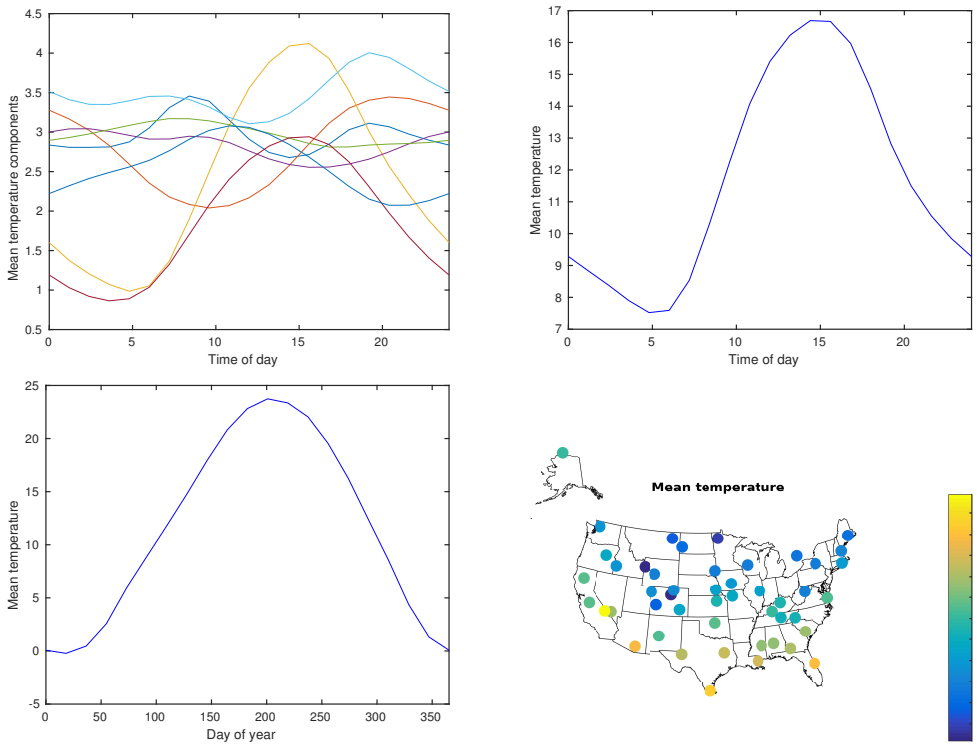

FIG. 3. Upper left panel: mean temperature components as functions of the time of the day (see eq. (3.6)). Upper right panel: mean temperature as function of the time of the day, marginalized over time of the year, climate variability and station. Lower left panel: mean temperature as a function of day of the year, marginalized over station, climate variability and time of the day . Lower right panel: mean temperature as a function of the station marginalized over time of day, day of year and climate variability.

how the prediction "explains" the cycle of seasons and hours of the day, but not the less regular weather systems, of typically around five days, that flow through the region. The right panel displays similar results for Barrow, at the northern tip of Alaska. An interesting observation here is that the estimated mean daily cycle captures two peaks. It is winter in Alaska, so the sun does not rise at all. What we are observing is a manifestation of the thermal tides Chapman and Lindzen [1970], with their two characteristic peaks, most often observed in the pressure but manifested here through their temperature signal.

Finally, figure 5 displays in blue the climate variability captured through the marginalized dependence of $\bar{x}$ on time in a multi year scale. To illuminate its possible meaning, we have superimposed in red the El Niño Index, which measures the moving three-month average temperature of the Indian Ocean. Noticeably, the two signals have a very similar amplitude of two to three degrees and time-scale of roughly three-four years. Clearly, this factor has captured a global phenomenon in its manifestation on the north American continent.

As described in section 8, we can iterate the procedure once to capture the variance as function of the same factors. Figure 6 displays the marginalized dependence of the standard deviation σ on the time of the year, with maximal variability in the winter and smallest in the late summer–early fall.

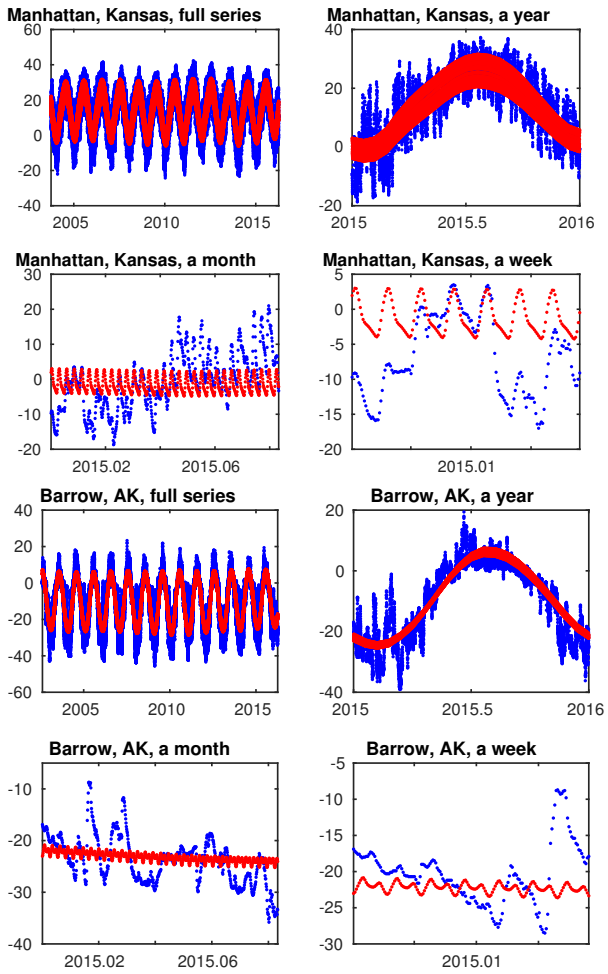

FIG. 4. Upper 4 panels: data (in blue) and predicted mean (red) for Manhattan, Kansas, displayed over the length of one year, one month and a week. Unlike the diurnal and yearly cycle, the weather systems (most noticeable on the second row, left column panel) are not "explained" by the procedure, as no covariate z representing them has been included. Lower 4 panels: data (in blue) and predicted mean (red) for Barrow, Alaska. Notice the two-peaked thermal tide in the predicted diurnal signal (4th row, right column panel).

After normalizing the y through the transformation $y^i \to \frac{y^i}{\sigma(z^i)}$, using their standard deviation $\sigma(z)$ so estimated, we can also estimate the time-lagged correlation between two stations in the following way: we input as a new variable x the product $x = y_1\, y_2$, where $y_{1,2}$ are the normalized observations at stations 1 and 2 respectively. As explanatory factors, we take the time of day and of year at station 1, and the time lag Δt between the two observations. Figure 7 shows the results of applying this procedure to Ithaca, NY and Des Moines, IA, through the dependence of the correlation on the time-lag, marginalized over time of day, but drawn specifically for one summer and one winter day. In both cases, we see the correlation peaking at a lag of two days, consistently with the time it takes for an average weather system to travel

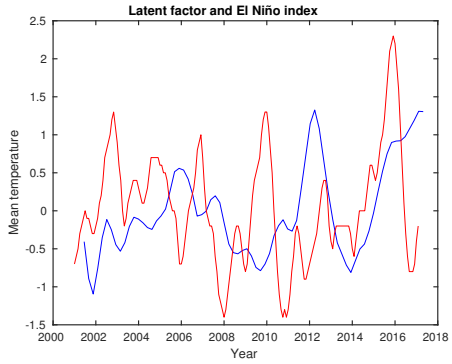

FIG. 5. El Niño index (in red) and $\bar{x}$ (in blue), marginalized over time of day, day of year and station.

from the midwest to the eastern US. This suggests improving the prediction at Ithaca by explaining its

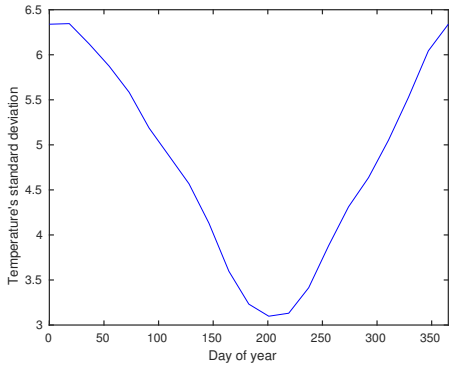

FIG. 6. Standard deviation as a function of the time of the year, marginalized over station, time of day and climate variability. Early winters have about twice the variability of the late summer–early falls.

already normalized signal y using as an additional explanatory factor the normalized y at Des Moines two days before, and possibly also the normalized temperatures at other stations in the past. Figure 8 shows the result of explaining the temperature in Ithaca using those in Des Moines (IA), Crossville (TN), Egbert (ON), Monahans (TX), Selma (AL) and Ithaca itself two days before, and that in Boulder (CO) four days before. We can see that the weather systems, not captured at all through the previously considered cofactors, are now accounted for to a significant degree.

Clearly, many more things can be done with this data along similar lines; our purpose here though is not to perform a systematic data-based study of the weather and climate or to develop a complete methodology for time-series analysis, but just to illustrate the workings of the attributable component methodology in a realistic setting.

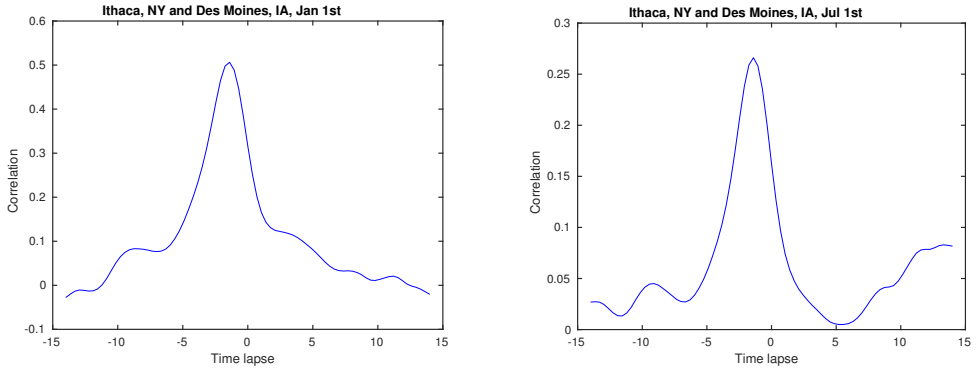

FIG. 7. Time-lagged correlation between Ithaca, NY and Des Moines, IA as a function of the time-lag, marginalized over time of day and climate variability, but drawn specifically for one summer and one winter day. Both peak at a lag of two days, but the correlation is significantly bigger in the winter.

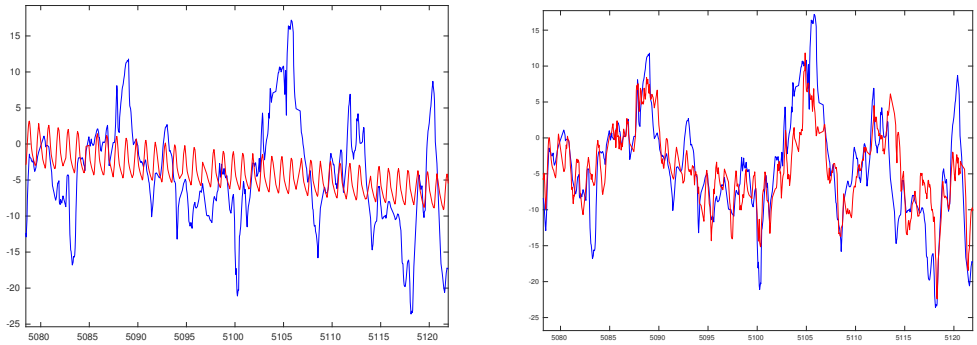

FIG. 8. Observed (in blue) and predicted (in red) temperature at Ithaca, NY over 45 days. Left panel: prediction using time of day, day of year and climate variability. Right panel: prediction using as additional cofactors the temperature in various stations 2 and 4 days before, thus capturing much of the weather systems that dominate the weather at this particular location and time.

9.3 *Book preference prediction*

In this section we test the algorithm on a preference learning problem. We create a synthetic data set representing book ratings by readers, where each book is assigned an integer score between 1 and 5 by each person that reads it. Each reader and book are characterized by categorical and non-categorical cofactors: we know each reader's age (a real cofactor) and location through its geographical longitude (a periodic cofactor), and each book's type (a categorical cofactor with ten possible values) and number of pages (a real cofactor). In addition, each reader and book have idiosyncratic traits, i.e. individual characteristics not captured by the descriptive cofactors just mentioned.

The goal is to learn each reader's book preferences based on previously rated books, so as to be able to predict the rating of a book that has not been read yet by a particular reader. This problem belongs to the category of matrix completion problems (Candes and Recht [2012]): the matrix with entries R_{ij} containing the rating that reader i assigns to book j is only partially filled, since not every reader has

read every book in the data set. The task is to predict the expected value of the missing entries.

To build the dataset, we first input the number of readers and books n_{users} and n_{items}. For each reader, the cofactors are defined as follows:

z_1: the reader, $z_1 \in \{1,\ldots,n_{users}\}$.

z_2: the age, drawn uniformly in the interval $[10,70]$.

z_3: the longitude, drawn from a binary Gaussian mixture with centers located on the American and Euro-Asian continents.

z_4: the book, $z_4 \in \{1,\ldots,n_{items}\}$.

z_5: the book type, an integer drawn uniformly from $[1,\ldots,10]$.

z_6: the number of pages, drawn from a Gaussian distribution centered around 70 and truncated below at $z_6 = 10$.

In order to create data that do not have the same form as the components sought by the algorithm, we first decompose z_3 into two: $c_3 = \cos(z_3)$ and $s_3 = \sin(z_3)$ and re-center and normalize all z_i, c_3 and s_3 so that they lie in the interval $[1/2,3/2]$. Then we define a set of functions g_j of the z_i as follows:

- $g_1 = \sin(\pi * z_1 * z_2 * c_3)$
- $g_2 = \cos(\pi/(s_3 * z_4 * z_5 * z_6))$
- $g_3 = 1/(z_1 + z_2 + c_3 + s_3 + z_4 + z_5 + z_6)$
- $g_4 = z_1/z_2 + z_4/c_3 + z_5/z_6$

These functions g_i are used to create a preliminary version of the data set through $x = \sum_{i=1}^{4} a_i \frac{g_i - \mu_i}{\sigma_i}$, where μ_i, σ_i are the mean and standard deviation of g_i, and the coefficients a_i are drawn from the normal distribution. Then we map x onto a uniform distribution in $[1,5]$ through sorting, and add to it normally distributed noise so as to displace in average each value of x by one unit. Finally, we truncate each value of x to obtained integer numbers from 1 to 5 in equal proportions. The result is a matrix $R_{z_1=i}^{z_4=j}$ with values between 1 to 5, depending on z_i, $i = 1,...,6$ in a coupled and nonlinear way.

Filling the entries of R for which the pair (z_1, z_4) is absent from the data set is most challenging when R is very sparse, a setting where the methodology of this article is particularly useful. On the one hand, the availability of qualifying cofactors, such as a reader's age or a book's type, groups entries together, thus partially overcoming the sparsity of entries per reader and per book. On the other, even in the absence of such additional cofactors, one can still group books and readers through bi-clustering, again overcoming to a certain degree the sparsity of data for individual readers or books. In order to test this, we have created data sets with both small and large numbers of readers and books with an average of about two book ratings per reader. To measure the quality of the results, we divided the dataset into two parts, one used for training the model and one for testing it. We then computed the root mean squared deviation (RMSD) on both the training set (in-sample error) and in the testing set (out of sample error). The sample used for testing consists of 100 entries randomly selected from the dataset, after making sure that no rows or columns were left empty.

Figure 10 shows predictions for a dataset of 500 readers and 100 books. The rating matrix R, with the sparsity pattern displayed in Figure 9, contains 955 non-zero entries, or roughly 2 books per reader.

To assess the effect of clustering, we run the algorithm by using as predictors only the row (z_1) and column (z_4) indices with and without clustering. The upper panels of Figure 10 shows in-sample and out of sample error without clustering, while to obtain the two lower panels of the figure the algorithm partitioned both readers and books into 20 possible classes. Notice that both the in-sample and out of sample RMSD are smaller in the run where clustering was used.

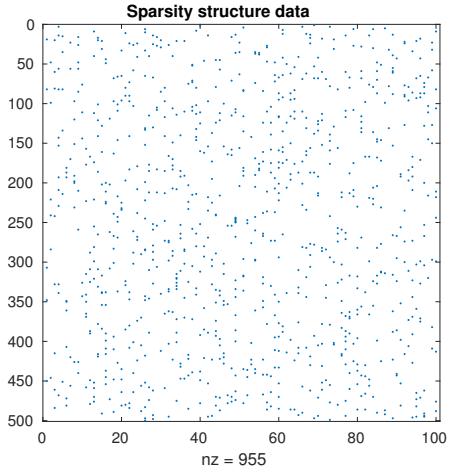

FIG. 9. Sparsity structure of R

Figure 11 shows results for a larger dataset with 50000 users, 10000 books and about 106000 preferences. The purpose of this experiment is to show how information beyond the idiosyncratic row and column index can improve the prediction, substantially reducing the out of sample error. This is a case in which clustering the users and books (into 20 classes each as before), besides improving the predictions, is also useful from a computational viewpoint, as the number of unknowns associated for instance to the readers decreases by a factor of 250, thus challenging the linear solvers to a much smaller degree. Stochastic gradient descent is not strictly required for the roughly $m = 10^5$ entries available, but we have applied it nonetheless, with $m/5$ random entries used per time-step, since for even larger datasets this would be the only practical option available.

As one can see in the results, using as predictors only reader and book (via clustering) or only their qualifying cofactors (age, location, type, length), yields poorer results than combining the two into a maximally explanatory set. In order to assess the degree of accuracy of the predictions obtained including reader, book and all their qualifying cofactors, the RMSD obtained in this case (0.79) should be compared with the RMSD obtained in two extreme cases:

1. All the entries to predict are assigned a value of 3, right in the middle of the rating scale. In this case, one can easily see that RMSD $= \sqrt{2} = 1.4142\ldots$
2. All the entries to predict are assigned the value obtained from the model underlying the data, but without adding the normally distributed noise described in the last step of the construction of our dataset, which cannot possibly be predicted. In this case, we obtain RMSD $= 0.4$.

Thus the root mean square error of the predictions, RMSD ≈ 0.8, is only twice the amount directly attributable to the noise.

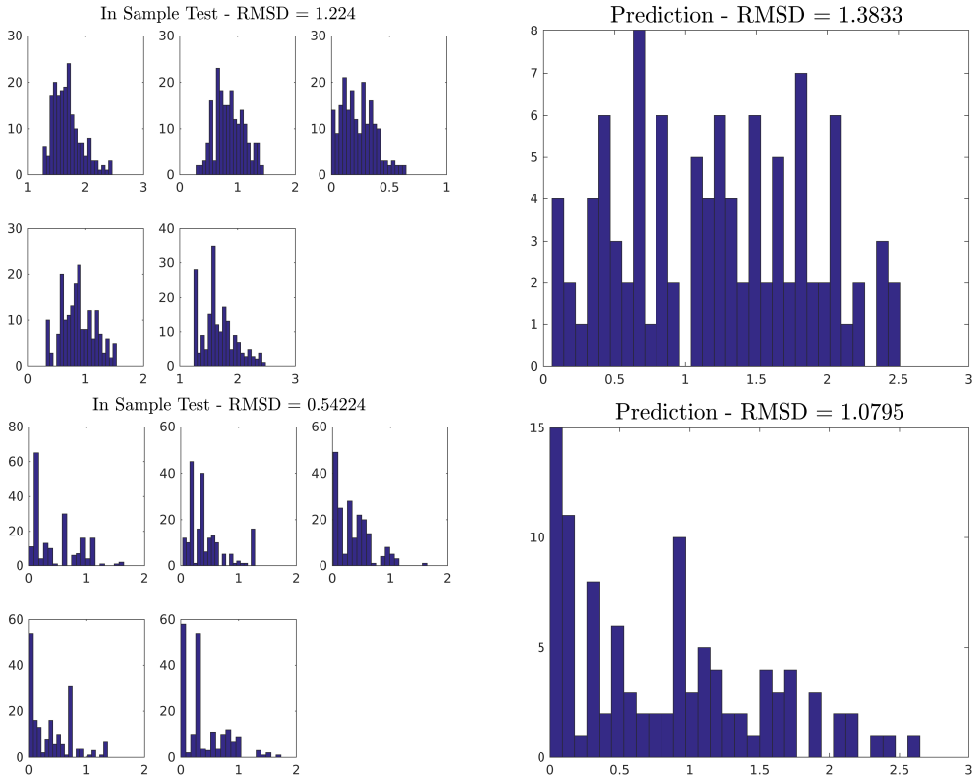

FIG. 10. Synthetic example obtained with 100 books, 500 users and about 955 preferences. The out of sample error is computed on 100 entries randomly selected among the 955. Upper left panels: in-sample error for each value of the score (1 to 5) obtained without clustering and by considering only the idiosyncratic covariates z_1 and z_4; the overall root mean square deviation is reported on top of the figure. Upper right panel: out-of-sample error distribution computed on the 100 entries previously eliminated. Lower left panels: same as upper left but using clustering. Lower right panel: out-of-sample error when clustering rows and columns into 20 classes each.

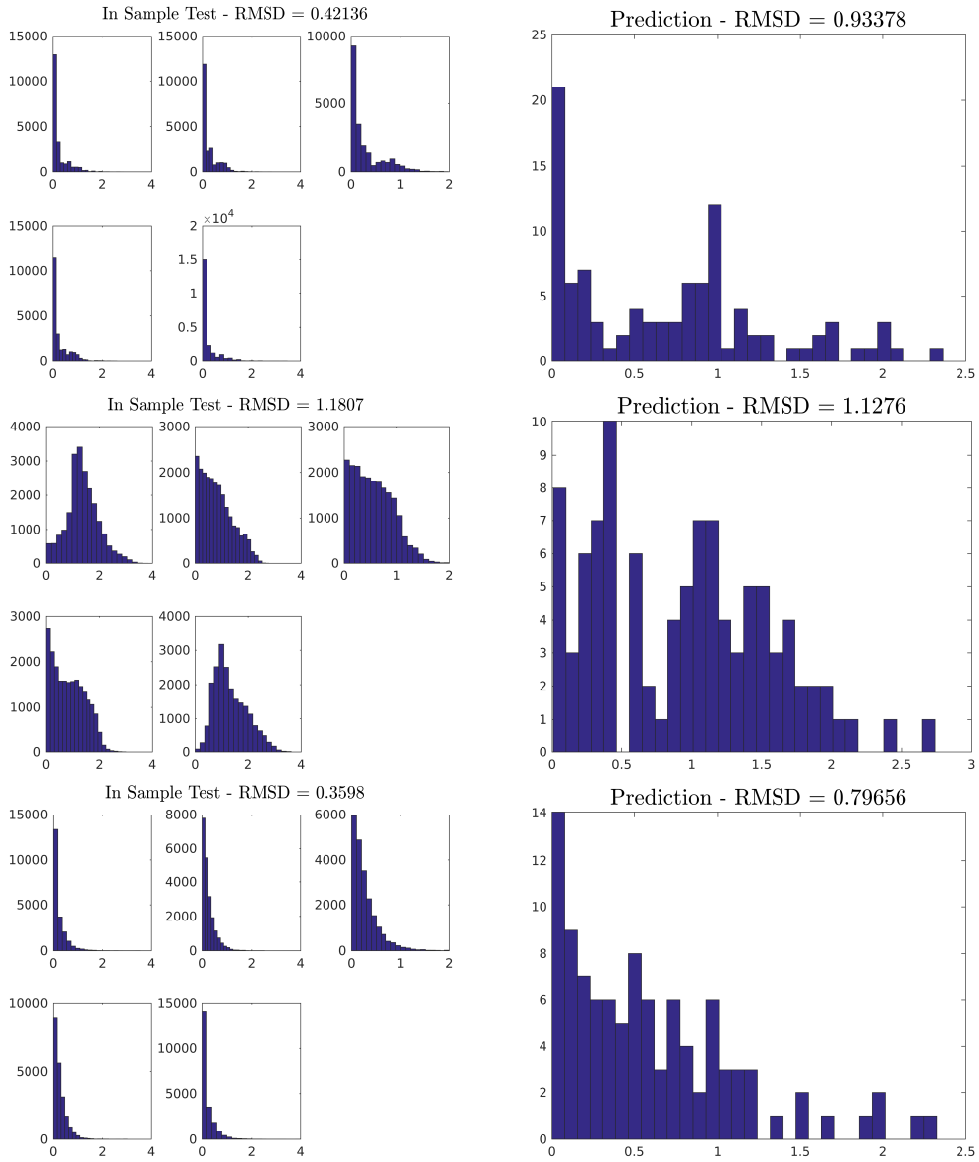

FIG. 11. Synthetic dataset with 10000 books, 50000 users and about 106000 preferences. The out-of-sample error is computed on 100 entries randomly selected among the 106000. First row: in-sample (left) and out-of-sample (right) errors when using only the reader and the book index. Second row: in and out-of-sample errors when using only the age, longitude, book's type and number of pages. Third row: in and out of sample error when using all the available cofactors. In all three scenarios, the readers (z_1) and books (z_4) were clustered into 20 classes each.

10. Conclusions

This article develops "attributable components", a methodology for the non-parametric estimation of the conditional mean of a variable x in terms of a set of covariates $\{z_l\}$. This is framed within a more general setting: the estimation of the full conditional probability $\rho(x|z)$ through a family of maps $y = Y(x;z)$ that push forward the $\rho(x|z)$ onto their Wasserstein barycenter $\mu(y)$. Estimating the conditional mean results from restricting these maps to z-dependent rigid translations $Y = x - \beta(z)$. We prove that these act as pre-conditioners to the full conditional density estimation, in the sense that, if one performs the latter after removing the conditional mean, the composition of the resulting optimal maps and the previous rigid translations yields the solution to the original barycenter problem. Extending the methodology of this article to handle the full conditional density estimation problem is the subject of current work. On the other hand, the conditional second order structure of the data, including the conditional covariance matrix, can be found by a straightforward iteration of the procedure.

The conditional mean $\bar{x}(z)$ is a function of the possibly many co-variates $\{z_l\}$, which may in addition have different types, such as categorical, ordinal and real. The procedure represents this multivariable function as a sum of components, each of which is the product of single-variable functions. Each of these single-variable functions, in turn, is represented by the values that it adopts on a grid, for which smoothness is enforced through penalization. Every observation is assigned a weighted sum of these grid values, through interpolation when the z_l is continuous and through straightforward assignment when it is discrete. When some or all values of z_l are unknown, finding these weights becomes part of the problem, which then produces clustering, classification or continuous covariate assignments. The functions of each covariate z_l are found through alternating descent of the objective function. Since this is quadratic in the functions of each z_l, the descent step consists of the solution to a linear, Sylvester-like system.

The methodology scales well for big datasets, as it adapts straightforwardly to stochastic descent and online learning. Despite its complex, non-parametric form, the predicted $\bar{x}(z)$ is easily rendered interpretable through marginalization over subsets of covariates.

Among the many extensions that one could foresee, those currently being pursued include, in addition to the full conditional probability estimation mentioned before, specific extensions directed at the analysis of time series. As for any methodology for the analysis of data, a wealth of interesting problems and extensions arises from considering particular applications. Not to overstretch the size of this methodological article, we have included here only a handful of representative examples, though we believe that they are diverse enough to convey a feeling for the methodology's versatility and broad applicability.

Acknowledgments

This work was partially supported by grants from the Office of Naval Research and from the NYU-AIG Partnership on Global Resilience.

References

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

Álvarez-Esteban, P. C., del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2016). A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*,

441(2):744–762.

Bartels, R. H. and Stewart, G. W. (1972). A solution of the equation $ax + xb = c$. *Commun. ACM*, 15:820–826.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417.

Caffarelli, L. A. (2003). The Monge-Ampère equation and optimal transportation, an elementary review. In *Optimal transportation and applications*, pages 1–10. Springer.

Candes, E. and Recht, B. (2012). Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119.

Chapman, S. and Lindzen, R. (1970). Atmospheric tides, 200 pp. *D. Reidel, Norwell, Mass*.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

Grasedyck, L., Kressner, D., and Tobler, C. (2013). A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen*, 36(1):53–78.

Hartley, R. and Schaffalitzky, F. (2003). Powerfactorization: 3d reconstruction with missing or uncertain data. In *Australia-Japan advanced workshop on computer vision*, volume 74, pages 76–85.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC press.

Kuang, M. and Tabak, E. G. (2017). Sample-based optimal transport and barycenter problems. *submitted to Communications on Pure and Applied Mathematics*.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of machine learning*. MIT press.

Tabak, E. G. and Trigila, G. (2017). Explanation of variability and removal of confounding factors from data through optimal transport. *Accepted in Communications on Pure and Applied Mathematics*.