Explanation of variability and removal of confounding factors from data through optimal transport

Esteban G. Tabak * Giulio Trigila †

December 13, 2016

Abstract

A methodology based on the theory of optimal transport is developed to attribute variability in data sets to known and unknown factors and to remove such attributable components of the variability from the data. Denoting by x the quantities of interest and by z the explanatory factors, the procedure transforms x into filtered variables y through a z-dependent map, so that the conditional probability distributions $\rho(x|z)$ are pushed forward into a target distribution $\mu(y)$, independent of z. Among all maps and target distributions that achieve this goal, the procedure selects the one that minimally distorts the original data: the barycenter of the $\rho(x|z)$. Connections are found to unsupervised learning and to fundamental problems in statistics such as conditional density estimation and sampling. Particularly simple instances of the methodology are shown to be equivalent to k-means and principal component analysis. An application is shown to a time-series of ground temperature hourly data across the United States.

Keywords: optimal transport, barycenter, confounding factors, batch effect, conditional density estimation, principal component analysis, ENSO.

1 Introduction

Real world observations are often highly individualized. Medical data aggregates samples of patients having each a unique combination of age, sex,

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

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

diet, prior conditions, prescribed drugs, and theses samples are often collected and analyzed at facilities with different equipment and personnel. Meteorological observations are performed at different locations, heights, times and seasons, with a variety of apparatus and evolving technologies. Financial data aggregates assets of many categories, companies of various sizes, stock prices at different markets and days of the week. Economic data covers regions of diverse history, development, connectivity, geographic location, demographics. Virtually every data-rich field has similar explanatory variables, which enrich and confound the variability in the quantities of interest.

The individualized nature of data can be both a blessing and a curse. On the one hand, it opens the way to personalized medicine, to economic policies tailored to local conditions, to accurate micro-local weather forecasts, to reduced risk and increased predictability. On the other, the existence of so many *confounding factors* poses severe challenges to statistical analysis: how can one determine or even define the effect of a treatment, when the outcome may be influenced and confounded by a multitude of factors such as age and prior conditions?

Two complementary tasks arise in connection to these individual factors: determining how much of the variability in the quantities of interest can be attributed to them, and filtering this attributable variability from the data. The need for filtering is clearest for confounding factors with no domain-related relevance, such as the lab where an analysis was performed. However, even for relevant factors, such as age as an explanatory factor for blood pressure, filtering the attributable component of the variability facilitates seeking further, possible unknown variability sources. Which brings in a third, related task: to find previously unknown factors that explain variability in the data. Classical examples include clustering, which explains part of the variability in a dataset by the class assigned to each observation, and principal component analysis, which explains variability in high-dimensional data through variables in a smaller-dimensional manifold.

This article presents a new methodology for the explanation of variability, based on the mathematical theory of optimal transport. The connection to optimal transport presents itself naturally in the context of removing from observations $\{x_i\}$ the variability attributable to a factor z. The existence of such attributable variability means that the conditional distribution $\rho(x|z)$ depends on z. Removing the attributable variability is therefore tantamount to estimating a set of maps $x \to y = Y(x; z)$ such that the resulting distribution $\mu(y)$ is independent of z, so that none of the variability remaining in y can be attributed to it. In addition, one wants these maps to dis-

tort the data minimally, so that the remaining variability in x, unrelated to z, is not affected by the transformation from x to y. In the language of optimal transport, one seeks the barycenter μ of the set of distributions $\rho(:|z)$ under a cost associated with our measure of data distortion (since the barycenter is defined precisely as the distribution that minimizes the total cost of transporting all $\rho(:|z)$ to μ.)

For clarity of exposition, rather that presenting a general theory from the beginning of the article, we build it gradually. Thus we start with the removal of discrete confounding factors, such as batch effects, where the need for filtering and its connection to optimal transport are clearest. We then consider the amalgamation of datasets and the removal of other discrete factors whose explanatory nature may be of interest in itself, unlike batch effects, which represent just a nuisance for the analysis. It is at this point that we make our first technical stop, discussing how the problem can be posed and solved in terms of the observations available, as opposed to the unknown underlying distributions $\rho(x|z)$. This is most naturally done at the level of the dual of a Kantorovich–like formulation of the filtering problem.

Next we consider the extension to continuous and vectorial (multiple) factors. At this point, we introduce a pair of "poor-man-solutions", which restrict the maps over which transport is optimized to a minimal set of affine maps or even rigid translations. Although far more restrictive than the general setting, these reductions make the arguments more concrete and the computational algorithms required much simpler. Finding the Wasserstein barycenter of a set of distributions $\rho_k(x)$ is a challenging numerical problem; even though there are numerical algorithms available for this task (see for instance [10]), the formulation of algorithms specifically designed for the statistical analysis of confounding factors is the subject of current research.

We then consider factor discovery: the explanation of variability in terms of unknown factors. We show that the methodology extends naturally to this setting, and that it includes as nearly trivial examples such powerful tools as principal component analysis and clustering through k-means, which allows one to generalize these classical procedures in a number of ways. We also discuss broad areas of applicability of the methodology, such as conditional density estimation and sampling.

We illustrate the article's main ideas through synthetic examples and through an application to real data: the explanation of the variability in temperature across the United States. Using hourly temperature observations gathered from 48 stations over more than a decade, we first filter from these the effects of the time of the day and seasonality, a station at a time, then the effects of latitude and elevation across stations, and finally perform

a smooth principal component analysis that uncovers the effects of El Niño Southern Oscillation.

The problem of attributing variability in a quantity of interest to covariates –also called confounding and explanatory variables– has a long-standing and fruitful history in statistics and in specific data-rich fields. Classical methodologies include ANOVA, ANCOVA, MANCOVA ([20]), the many variations of factor analysis ([18]) (all of these involve linear regression combined with statistical models), stratification ([8]), de-trending of time-series, various methodologies for the removal of batch effects ([7, 16]), and a large number of specialized techniques developed within individual fields –notoriously bio-statistics, psychology, sociology, econometrics and the environmental sciences. An alternative methodology involves the statistical elimination of confounding effects by the design of randomized experiments ([17]), as opposed to the observational studies where covariates are unavoidable. This article is not the right place to summarize all these widely used procedures; we refer the reader to the references cited and the vast literature on the field. Our methodology differs from those mentioned in some fundamental ways. It provides a general conceptual framework to discuss the explanation of variability, illustrated in this article through some simple instances (maps limited to affine and even to rigid translations) but not limited to these. By contrast, all the procedures above rely on strong hypotheses. For instance stratification, where the data set is divided into groups on which the confounding factor is roughly constant, requires the sample size for each group to be proportional to its variance (for this reason, stratification is well know to work best with a limited number of confounders.) ANCOVA relies on the linearity of the relation between the independent and dependent variables and homogeneity of this dependence among groups ("the assumption of parallel lines".) Most methodologies make assumptions on the errors (making them i.i.d Gaussian variables, etc.) By contrast, our methodology does not conceptualize deviations from the conditional mean as "errors" but just as samples from the conditional probability distribution of the data, which can be very general (we just require that the solution to the Wasserstein barycenter problem exists and is unique, entailing essentially finite second moments for each $\rho(x|z)$). Even though our poor-man solutions involve linear models, they do so differently than in the classical methodologies, in that the linearity is in the maps, not in the models for the parameters of the conditional distributions. This yields different results; for instance, in one-dimension the standard deviation behaves as an additive variable –not the variance as is typical in statistics. These poor-man examples serve only as a proof of concept; they are not intended to show that

the methodology outperforms traditional methods when used in its simplest setting.

The central idea of this article has a number of potential ramifications. Rather than pursuing them all, we limit ourselves here to a description of the driving ideas and a few illustrative examples. Avenues of further research include alternative formulations in terms of the samples available, of spaces richer that the poor-man solutions used for the examples, of cost functions other the squared-distance, and many powerful extensions of principal component analysis, just to mention a few. Field-specific applications, for instance to medicine, climate science and economics, can be better described in individual, focussed articles (The explanation of temperatures across the US is only included to exemplify the use of real data; its proper discussion in a meteorological context requires a separate article informed by the discipline.)

2 Discrete factors

We begin our discussion with the explanation of variability through factors that admit only a discrete set of values. A prominent example is the removal of batch effects, which permeates high-throughput biostatistics. We first address this particular case and then extend the results to the amalgamation of datasets and to the general explanation of variability by discrete variables. After introducing a conceptual framework for the removal of variability attributable to discrete factors, we discuss its implementation in terms of the data points available.

2.1 Removal of batch effects

Due to equipment limitations, high-throughput biological experiments are often performed in batches, each containing a limited number of samples. It has been observed repeatedly that, when the results from the various batches are brought together into a single dataset, much of the variability among samples can be attributed to the batch where each originates. This is the *batch effect*, a persistent confounding factor in biostatistics. A number of techniques have been developed for removing batch effects; see for instance the two reviews in [7, 16]. Here we propose a novel approach based on the theory of optimal transport.

The fact that one can infer from the data at least partial information about the batch to which each sample belongs implies that the distributions $\rho_k(x)$ underlying the data $\{x_i\}, i \in \{1, \dots, m\}$ in the various batches $k \in$

$\{1, \ldots, K\}$ are different. Filtering from the data x all information related to the batch k is transforming the data

$$x \to y, \quad y_i = Y_{z_i}(x_i)$$

so that one cannot infer from y_i the batch $k = z_i$ from which x_i was drawn: the distribution underlying y must be independent of k.

This problem can be phrased naturally in the language of optimal transport [19, 14]. We seek a set of maps $Y_k(x)$ that push forward the distributions $\rho_k(x)$ into a common target distribution $\mu(y)$ (see Figure 1), namely

$$\int_{Y_k^{-1}(A)} \rho_k(x)\, dx = \int_A \mu(y)\, dy \tag{1}$$

for all sets A. Moreover, the maps and target are to be chosen so that the data is minimally distorted, since we want to preserve as much as possible the variability in the data not attributable to the batch. Introducing a cost function $c(x, y)$ that measures the level of distortion from mapping point x into point y, we seek to minimize the "total distortion" :

$$\min_{\mu, Y_k} D = \sum_{k=1}^{K} P_k \int c\left(x, Y_k(x)\right) \rho_k(x)\, dx, \tag{2}$$

where P_k is the proportion of samples in batch k. The target distribution $\mu(y)$ defined through this minimization process is the P-weighted *c-barycenter* of the $\rho_k(x)$ [1][1].

Many applications of optimal transport have a specific cost function $c(x, y)$ appropriate for the problem under consideration. This is not the case here, since the notion of *distortion* admits more than one quantification. Throughout this article, we will use the standard squared-distance cost

$$c(x, y) = \frac{1}{2}\|y - x\|^2, \tag{3}$$

with a rationale similar to the one underlying the use of least-squares for regression: it is a convenient, sensible choice that leads naturally to the use of standard tools in linear algebra. As in regression, there are scenarios where other cost functions would be more appropriate; most of the results

[1]Here our focus is on the removal of variability explainable by the batch of origin. Other relevant tasks in data analysis, such as cleaning the data of outliers, which might be addressed through a differently chosen cost function, fall outside the scope of this article.

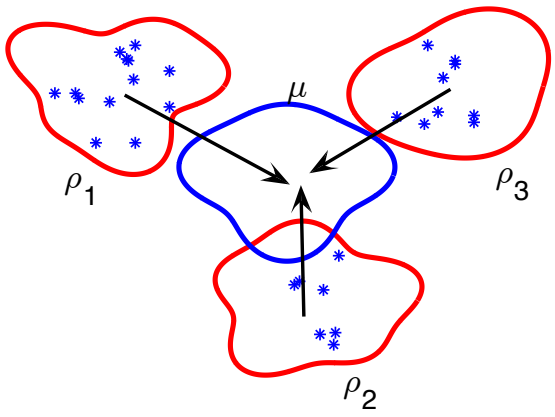

Figure 1: Removal of batch effects. In this example, x is a two dimensional dataset drawn from three batches. The blue stars represent individual samples and the red contours represent the three distributions $\rho_k(x)$ underlying these samples. Three transformations $Y_k(x)$ represented by arrows map these three distributions into a single target $\mu(y)$. Once the maps have been performed, one cannot tell from y_i from which batch the sample x_i has been drawn. The choice of μ that minimizes the total distortion of the data is the barycenter of the ρ_k under the Wasserstein metric associated with the cost $c(x, y)$.

in this paper extend to such more general costs. An additional, *a posteriori* rationale is that the squared-distance cost function leads to curl-free maps $Y_k(x) = \nabla\phi_k(x)$, a quite natural characterization of maps with minimal distortion [4, 2] (One can build examples where this cost would fail to characterize the true maps, for instance when there is significant, batch dependent rotation of the data. To include such scenarios, different cost functions must be chosen.)

2.2 Other confounding discrete factors and supervised unlearning

The batch effect of biostatistics is just one instance of a more general problem of wide applicability: to compile an integrated dataset from data collected under different circumstances. These can be data from different labs or hospitals, different censuses, different experiments. Taken separately, these datasets may lack enough samples to yield statistically significant results. Taken together, a large fraction of their variability may be attributable to their study of origin. Hence the need to filter from the data during amalgamation those characteristics that are idiosyncratic to each study.

Such amalgamation of datasets can be viewed more generally as *explanation of variability*: the component of the variability in x attributable to the batch of origin has been explained away, leaving behind in the transformed variables y the remaining, unexplained variability. This is the language used for instance in principal component analysis, where the variability is measured by the variance and the amount explained by each component is given by the square of the corresponding singular value. When viewed in these terms, the procedure extends far beyond the compilation of disjoint datasets. It can be used, for instance, to explain away the patient's gender in microarray data, the state of a switch in a controlled system, the day of the week in trading data.

The procedure remains completely unchanged: given a set of samples x_i and corresponding discrete labels $z_i \in \{1, \ldots, K\}$, one seeks the set of maps $Y_k(x)$ pushing forward the distributions $\rho_k(x)$ underlying the data into a common target distribution $\mu(y)$ while minimizing the total distortion D. Yet the new perspective suggests a new language: in supervised learning, given features x_i and corresponding labels z_i, one seeks an assignment function $z(x)$. In unsupervised learning, no label z_i is provided for observation x_i, and one seeks structure in the features x alone, for instance clustering the samples into classes, thus assigning them previously unknown labels. In the problem treated in this section, the labels z_i are given, but our goal is to explain them away, effectively removing all signal of the labels from the data. Hence one might denote the procedure "supervised unlearning". If all effects from the labels are removed from the data, there is no way in which an unsupervised procedure applied to the filtered data y_i can recover the original labels z_i, or a supervised procedure can determine the correct assignment function $z(x)$.

A further generalization is brought about by *factor discovery*, where the labels to remove from the data are unknown. This can be thought of as

an instance of clustering, whereby the samples are divided into classes, followed by filtering, whereby any information related to the newly-discovered classes is removed from the original data. Factor discovery, both discrete and continuous, is discussed in a unified setting in section 5.

3 Posing the problem in terms of samples

We have formulated the problem of discrete factor removal in terms of the probability densities $\rho_k(x)$ underlying the data for each value k of the factor. Yet these underlying densities are not known: we only have a set of sample points x_i drawn from them (Even with known densities, solving the optimal transport problem is not trivial; see for instance [6, 3, 11, 22, 12, 9] for some of the numerical methods developed in recent years.) In addition, our problem involves not only optimal transport between each ρ_k and μ, but also the search for the optimal μ, the barycenter of the ρ_k.

We are provided with data consisting of the features $x_i \in R^n$ and the labels $z_i \in \{1, \ldots, K\}$, with $i \in \{1, \ldots, m\}$. The goal is to formulate and solve the problem of transforming each x_i into $y_i = Y_{z_i}(x_i)$ so as to remove the variability in x explained by z. We will first sketch a general framework to address this problem, and then describe two "poor man" solutions that restrict consideration to affine maps and to rigid translations. These and generalizations to continuous, multiple and unknown factors are the procedures that we will use to illustrate this paper's proposal with synthetic and real examples.

3.1 Kantorovich formulation and its dual

In order to pose the problem in (1,2) in terms of sample points rather than distributions, we start with Kantorovich relaxation from maps $Y_k(x)$ pushing forward ρ_k to μ to joint distributions $\pi_k(x, y)$ coupling ρ_k and μ, namely:

$$\min_{\mu, \pi_k} \sum_{k=1}^{K} P_k \int c\,(x, y)\, \pi_k(x, y)\; dxdy, \tag{4}$$

subject to

$$\int \pi_k(x, y)\; dy = \rho_k(x), \quad \int \pi_k(x, y)\; dx = \mu(y). \tag{5}$$

For the squared Euclidean distance cost function (3), the optimal solutions to the original (1,2) and relaxed (4,5) problems agree ([25, 5], see also (8)

below):

$$\pi_k(x,y) = \rho_k(x)\delta\left(y - Y_k(x)\right).$$

The relaxed formulation in (4,5) is an infinite dimensional linear programming problem, with dual

$$\max_{\phi_k,\psi_k} \sum_{k=1}^{K} \int \phi_k(x)\rho_k(x)dx, \tag{6}$$

$$\forall x, y \ \phi_k(x) + \psi_k(y) \leq P_k \ c(x,y), \quad \forall y \ \sum_{k=1}^{K} \psi_k(y) \geq 0. \tag{7}$$

For the squared-distance cost (3), the Kantorovich's dual and Monge's primal solutions are linked through

$$Y_k(X) = x - \frac{1}{P_k}\nabla\phi_k(x) \tag{8}$$

with the implication that the coupling $\pi_k(x,y)$ is supported on a curl-free map ([25]).

One can simplify the dual problem further, noticing that its constraints are equivalent to[2]

$$\forall x = \begin{pmatrix} x^1 \\ \dots \\ x^K \end{pmatrix}, \ \sum_{k=1}^{K} \phi_k(x^k) \leq \min_y \sum_{k=1}^{K} P_k \ c(x^k, y). \tag{9}$$

Thus, introducing the y-independent cost function

$$C(x^1, \dots, x^K) = \min_y \sum_{k=1}^{K} P_k \ c(x^k, y), \tag{10}$$

the dual problem adopts the form

$$\max_{\phi_k} \sum_{k=1}^{K} \int \phi_k(x)\rho_k(x)dx, \quad \forall x = \begin{pmatrix} x^1 \\ \dots \\ x^K \end{pmatrix}, \sum_{k=1}^{K} \phi_k(x^k) \leq C(x^1, \dots, x^K). \tag{11}$$

[2]The implication (7) $\Rightarrow$ (9) follows by summing over k and taking the minimum over y. Since ψ does not appear in the objective function (6), the reverse implication follows from verifying that, given (9), the function $\psi_k(y) = \min_x(P_k c(x,y) - \phi(x))$ satisfies (7).

Notice that the y minimizing (10) is the weighted barycenter of the x^k under the square-distance $c(x, y)$. In fact, for our specific cost function, the cost C can be written without reference to y and hence without an inner optimization problem:

$$C(x^1, \ldots, x^K) = \frac{1}{2} \sum_{k,l} P_k P_l \|x^k - x^l\|^2.$$

3.2 Data-driven formulation

The dual formulation in (11) involves the distributions $\rho_k(x)$ only through integrals representing the expected values of the $\phi_k(x)$ under ρ_k (a fact exploited in [24] to implement regular data-driven optimal transport.) Thus, if only samples from ρ_k are available, it is natural to replace these expected values by empirical means:

$$\max_{\phi_k} \sum_{k=1}^{K} \frac{1}{m_k} \sum_{i \in S_k} \phi_k(x_i)$$

$$\forall x = \begin{pmatrix} x^1 \\ \ldots \\ x^K \end{pmatrix}, \; \sum_{k=1}^{K} \phi_k(x^k) \leq C(x^1, \ldots, x^K), \; \phi_k \in F, \tag{12}$$

where S_k denotes the set of samples $\{i\}$ with $z_i = k$. Replacing the averages with empirical means is the standard Monte Carlo approximation, with error that goes to zero as $\mathrm{var}(\phi_k)/N$, where N is the number of sample points. In (12), m_k is the number of samples with label k and F is the space of functions over which the ϕ_k are restricted in order not to overfit the samples, for instance by placing a delta function around each.

It follows from the dual problem (6,7) that the ψ_k are the Legendre transforms of the ϕ_k, so restricting the latter to a set F of convex functions is equivalent to restricting the ψ_k to the dual set F^* of their Legendre transforms. A particularly simple example is the space F of convex quadratic functions, which agrees with its dual F^*. In this case, recomputing the primal problem from the dual with ϕ_k and ψ_k restricted to F and the expected values of the ϕ_k replaced by their empirical means yields the original objective function (4) but with the constraints in (5) replaced by their weak

counterpart

$$\forall \phi, \psi \in F, \int \left[\frac{1}{m_k} \sum_{i \in S_k} \delta\left(x - x_i\right) - \int \pi_k(x, y)\, dy \right] \phi(x)\, dx = \quad 0 \quad (13)$$

$$\int \left[\mu(y) - \int \pi_k(x, y)\, dx \right] \psi(y)\, dy = \quad 0, \quad (14)$$

meaning that the x-marginal of the π_k and the empirical distribution replacing $\rho_k(x)$ are not required to agree pointwise, but only to yield identical expected values for all test functions $\phi(x)$ in F, and similarly the y-marginals of the various π_k are only required to yield the same expected values for all test functions $\psi(y)$ in F. Since F is the space of convex quadratic functions, only the mean and covariance matrix of the x-marginal of π_k need to agree with their corresponding empirical values under $\rho_k(x)$, and only the mean and covariance of the target μ need to be the same for all K maps.

3.3 Poor man solutions

The data-based barycenter problem can be solved with a level of accuracy that adjusts to the number of samples available by defining the space F of allowed test functions adaptably: in areas where there are more sample points, the functions in F should have finer bandwidths. For instance, one could extend the flow based procedure developed for the data-driven optimal transport problem [24]. For what remains of this article, we will instead adopt two much simpler, non-adaptive "poor man" procedures. The point of these simple-minded approaches is that they permit exploring at ease a number of interesting scenarios (such as continuous, multiple or unknown factors) that would become more complex and computationally expensive in a more accurate, adaptive setting. In addition, we shall see that even the simplest of these settings includes and greatly extends well-established procedures such as principal component analysis and autoregressive models. Hence we postpone the exploration of adaptive procedures to further work.

The richer of the two poor man solutions that we propose is based on the example at the end of the prior subsection: optimizing the dual, data-based problem over test functions ϕ_k from the space F of quadratic polynomials. As reasoned above, this choice only captures the empirical mean and covariance matrices of the distributions ρ_k. Moreover, from the relation in (8) between Kantorovich's dual and Monge's formulations, this choice yields affine maps $Y_k(x) = \alpha_k x + \beta_k$.

To fix ideas, consider first the case where x is one-dimensional. Let $(\bar{x}_k,\ \sigma_k)$ be the empirical mean and standard deviations of the $\rho_k(x)$, and $(\bar{y},\ \sigma_y)$ be the unknown mean and standard deviation of the target $\mu(y)$. Substituting $\phi(x) = \left(x, x^2\right)$ and $\psi(y) = \left(y, y^2\right)$ in the marginal constraints (13), (14) yields

$$\alpha_k = \frac{\sigma_y}{\sigma_k}, \quad \beta_k = \bar{y} - \alpha_k \bar{x}_k. \tag{15}$$

Therefore, after imposing these constraints, the empirical version of the objective function D in (2) depends only on $\bar{y}$ and σ_y:

$$D = \sum_{k=1}^{K} P_k \frac{1}{m_k} \sum_{i \in S_k} \left[\left(\frac{\sigma_y}{\sigma_k} - 1 \right) x_i + \frac{\sigma_k \bar{y} - \sigma_y \bar{x}_k}{\sigma_k} \right]^2, \tag{16}$$

which is minimized by

$$\bar{y} = \sum_k P_k \bar{x}_k, \quad \sigma_y = \sum_k P_k \sigma_k, \quad D_{min} = \sum_k P_k \left[(\bar{x}_k - \bar{y})^2 + (\sigma_k - \sigma_y)^2 \right].$$

Then α_k and β_k are determined from (15), and the filtered signal is given by

$$y_i = \alpha_{z_i} x_i + \beta_{z_i}.$$

The multidimensional case follows a similar pattern, though with less straightforward algebra and no explicit expression for the parameters of the barycenter μ. The input data are the empirical vectorial means $\bar{x}_k$ and covariance matrices Σ_k of the ρ_k, and the unknown parameters of the barycenter are $\bar{y}$ and Σ_y. In terms of these, the map $Y_k(x) = \alpha_k x + \beta_k$ (here α_k is a matrix and β_k a vector) has parameters [15]

$$\alpha_k = {\Sigma_k}^{-\frac{1}{2}} \left({\Sigma_k}^{\frac{1}{2}} \Sigma_y {\Sigma_k}^{\frac{1}{2}} \right)^{\frac{1}{2}} {\Sigma_k}^{-\frac{1}{2}} \quad \beta_k = \bar{y} - \alpha_k \bar{x}_k. \tag{17}$$

Minimizing D yields

$$\bar{y} = \sum_k P_k \bar{x}_k = \bar{x}$$

for $\bar{y}$ as before, and the implicit condition for Σ_y

$$\Sigma_y = \sum_k P_k \left({\Sigma_y}^{\frac{1}{2}} \Sigma_k {\Sigma_y}^{\frac{1}{2}} \right)^{\frac{1}{2}}, \tag{18}$$

which can be solved effectively using the following iterative scheme [15]:

- Propose an initial guess for Σ_y, such as

$$\Sigma^0 = \left[\sum_{k=1}^{K} \left(P_k \Sigma_k{}^{\frac{1}{2}}\right)\right]^2 \quad \text{or} \quad \Sigma^0 = \sum_{k=1}^{K} P_k \Sigma_k, \quad \text{and}$$

- given Σ^n, write

$$\Sigma^{n+1} = \sum_{k=1}^{K} P_k \left((\Sigma^n)^{\frac{1}{2}} \, \Sigma_k \, (\Sigma^n)^{\frac{1}{2}}\right)^{\frac{1}{2}}.$$

  This sequence converges to Σ_y.

Our "poorest man" solution has the maps restricted even further to rigid translations of the form

$$Y_k(x) = x + \beta_k.$$

From the means $\bar{x}_k$, one obtains $\bar{y} = \sum_k P_k \bar{x}_k = \bar{x}$ as before, and hence

$$\beta_k = \bar{x} - \bar{x}_k,$$

which moves the means $\bar{x}_k$ of each class to the global mean $\bar{x}$. The reason to even mention this seemingly vastly under-resolving procedure is that, as we shall see, when extended to unknown factors, it is rich enough to include and generalize widely used tools such as principal components and K-means.

4 Continuous and multiple factors

Factors that explain variability are not necessarily discrete as in the examples considered so far. In medical studies, the patient's age could be a confounding factor; in climate studies, the level of solar radiation or of atmospheric CO_2; in financial studies, indicators such as liquidity or company size. A large proportion of the variability factors found in applications are continuous.

Extending the procedure above to continuous factors naively has the problem of excessive granularity: since each sample x_i has its own unique factor's value z_i, one would need to estimate values for the mean and covariance from just one sample, a hopelessly over-resolving task. Instead, we replace the K maps $Y_k(x)$ by the continuous family of maps

$$y = Y(x; z),$$

with the dependence on z constrained so as to avoid over-resolution: making this dependence smooth links together samples with nearby values of z_i. In particular, for the affine maps of our first poor man solution, one has

$$y = \alpha(z)x + \beta(z).$$

One can either model the map parameters $\alpha(z), \beta(z)$ directly –a procedure that will be developed elsewhere– or in terms of a model for the z-dependence for the mean and covariance of the data:

$$\alpha(z) = \Sigma(z)^{-\frac{1}{2}} \left(\Sigma(z)^{\frac{1}{2}} \Sigma_y \Sigma(z)^{\frac{1}{2}}\right)^{\frac{1}{2}} \Sigma(z)^{-\frac{1}{2}}, \;\; \beta(z) = \bar{y} - \alpha(z)\bar{x}(z), \quad (19)$$

$$\bar{x} = \bar{x}(z;\gamma), \quad \Sigma = \Sigma(z;\phi),$$

where γ and ϕ are sets of parameters fitted to the data (x_i, z_i). A simple way to frame this fitting procedure is through maximal likelihood, assuming $\rho(x|z)$ to be a Gaussian with mean $\bar{x}(z)$ and covariance $\Sigma(z)$, which yields the loss function (minus the log-likelihood)

$$L = \sum_i \left[\frac{1}{2}\log|\Sigma(z_i)| + \frac{1}{2}\left[x_i - \bar{x}(z_i)\right]' \Sigma^{-1}(z_i)\left[x_i - \bar{x}(z_i)\right]\right] \quad (20)$$

to minimize over (γ, ϕ). This is not meant to imply that restricting the maps $x \to y$ to affine corresponds to a Gaussian-mixture model for the data: the Gaussian-based maximal likelihood is just one convenient framework for parameter fitting, and the minimizer of L in (20) is a regular estimator for the mean and covariance independently of the distribution underlying the samples.

One simple model for scalar x is

$$\bar{x}(z) = az + b, \quad \sigma(z) = e^{cz+d},$$

depending on the four scalar parameters a, b, c, d (Here the exponential function is introduced to guarantee the positivity of $\sigma(z)$.) An analogous proposal for vectorial x has the standard deviation σ replaced by the covariance matrix Σ, and the scalar parameters $\gamma = a, b$ and $\phi = c, d$ by vectors and symmetric matrices respectively. This model can be generalized to

$$\bar{x}(z) = \sum_j a_j f_j(z) + b, \quad \Sigma(z) = e^{\sum_j c_j g_j(z) + d}, \quad (21)$$

where the f_j, g_j are arbitrary feature functions, designed to capture non-linear dependence of $\rho(x|z)$ on z. The Cholesky decomposition provides an

alternative parameterization of positive $\Sigma(z)$ different from the exponential in (21) [21].

Real data typically depend on more than one external factor (for instance on a patient's age, sex and ethnicity in medical studies, and on time of the day, time of the year, latitude and elevation in whether prediction.) The procedure above can be generalized to such multi-factor cases, replacing the scalar factor z by a vectorial one. In fact, this is already a particular instance of (21), if one introduces as feature functions $(f_j(z), g_j(z))$, the components z_j of z.

We consider now three examples that illustrate the versatility of the methodology just described: evaluation of whether a variable differs across groups in the presence of covariates, which we compare with ANCOVA, time series analysis, and the explanation of ground temperature variation across the United States.

4.1 Comparison with ANCOVA

To illustrate how the methodology developed in this article compares with those used in statistical practice for the removal of confounding factors, we chose the analysis of covariance (ANCOVA). This is used to evaluate whether the mean of a continuous variable depends on a categorical one, while controlling for a possible affine dependence on covariates. For the comparison, we create synthetic data that fits the formal frameworks of both ANCOVA and our poor-man solution filtering, through the following procedure:

1. Draw N i.i.d. samples $z_i \sim \mathcal{N}(2, 1)$ of a confounding factor z.

2. Divide these N points into 3 groups identified by the categorical variable $s \in \{1, 2, 3\}$, with a random assignment that depends on z, defined through

$$s_i = \arg\min_s \left(q_i(s)\right), \quad \text{where} \quad q_i(s) = |\xi_i - s|, \quad \xi_i \sim \mathcal{N}(z_i, 0.25).$$

3. Sample the z_i-dependent variable $x_i \sim \mathcal{N}\left(Az_i + B, \exp(Cz_i + D)\right)$, with $A = 0.742$, $B = -1.35$, $C = 0.4924$, $D = -0.542$.

The procedure above generates samples x_{is} (z_{is} respectively) where x_{is} is the i-th observation from group s. The goal is to compare x across the three groups while controlling for the confounding factor z. For our example, a successful test would reveal that the difference in the distribution of x among

the 3 groups is due solely to the effect of the confounding factor z. In other words, by removing the effect of the confounding factor we should see no difference among the three groups.

We will denote by ρ_s the distribution of the data points x_{is} in the group s, and by μ_s the filtered distribution obtained with our poor man solution.

We first use ANCOVA with the x_{is} and z_{is}. Figure 2 displays 1000 sample points from $\rho_{s=1,2,3}$ with their respective regression lines (top-left panel). With an F statistics associated with the interaction term $x * z$ of 0.9 and a corresponding p-value of 0.4, ANCOVA finds that the slopes of the regression lines associated with each group are not significantly different. Hence enforcing the same slope on the three regression lines, ANCOVA computes the three intercepts, and concludes –correctly– that, after excluding the effect of the covariate z, the mean values of x in the three groups are not significantly different, with p-values of 0.75, 0.36 and 0.78.

Using our methodology, we compare instead the filtered probability densities $\mu_{s=1,2,3}$ showed in Figure 2 (bottom-left panel). For this, we perform an ANOVA test, with results displayed through a notched box plot in the bottom-right panel. In addition, we perform a two-sample F-test for equal variances with the built-in `vartest2` MATLAB function. With p =0.5 for the three means being equal and p =0.99, 0.4802 and 0.4830 for the pairwise difference between the three variances, the tests finds the differences among $\mu_{s=1,2,3}$ not to be significant.

Even though the two procedures are successful in this example, the procedure of this paper gives richer results, as it verifies not only that the mean of x does not vary among the three populations, but also that the variance is the same (allowing maps richer than our poor man solution's would provide further information on the dependence of the distributions $\rho(x|z)$ on s, and would not require the dependence of x on z to be affine.)

In our example, not only the mean of x but also its standard deviation depends on z, with largest spreads for $s = 3$, since this population has typically larger values of z and hence of σ. This makes ANCOVA particularly sensitive to the resulting outliers as the number of sample points decreases. Figure 3 shows an experiment with 20 sample points. In this case, ANCOVA finds the hypothesis of parallel lines not to hold (the p value of the interaction term $x * z$ is $p = 0.0012$. This makes it conclude –wrongly– that the relation between z and x depends on s. By contrast, the ANOVA analysis and the two-sample F-test for equal variances reveal that the means and the variances of the filtered $\mu_{s=1,2,3}$ are not significantly different from each other: the p values of the two-sample F-test are $p = 0.29$ (comparing μ_1 and μ_2), p =0.83 (comparing μ_1 and μ_3), p =0.42 (comparing μ_2 and μ_3).

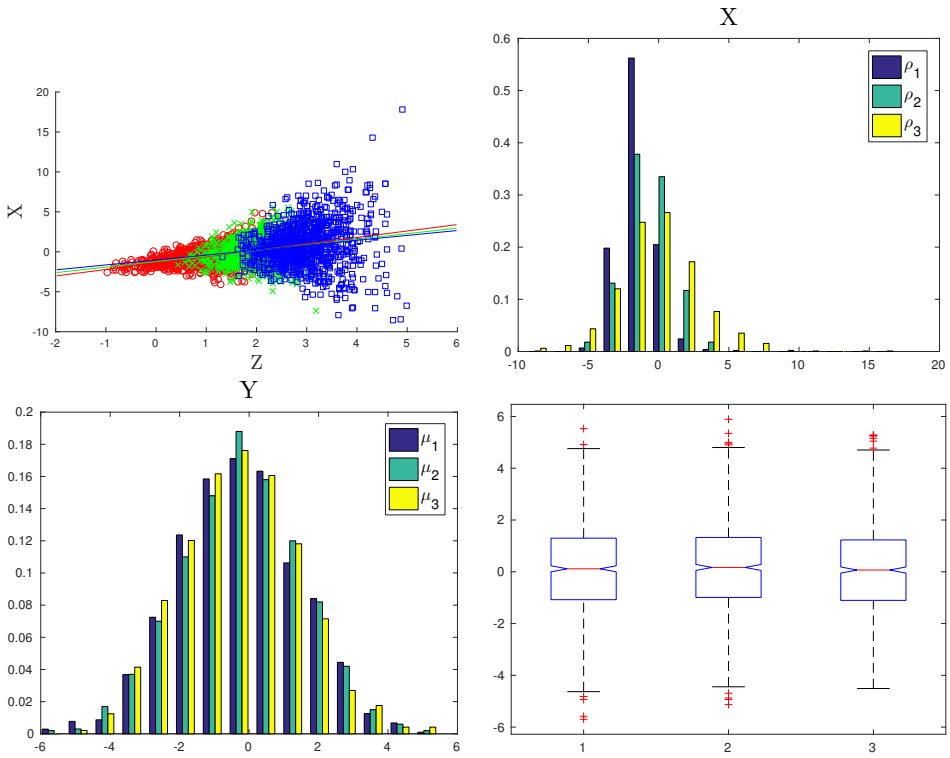

Figure 2: 1000 points data set. Top-left: Scatter plot of the x_{is} as functions of the z_{is}, with points colored according to the value of s. Top-right: histogram of $\rho_{s=1,2,3}$. Bottom-left: histogram of $\mu_{s=1,2,3}$ obtained by filtering z from x. Bottom-right: Notched box plot of the sample points from $\mu_{s=1,2,3}$ produced by the `anova1` MATLAB function.

4.2 Time series analysis

When analyzing a time series x^n, one may propose a Markov model

$$x^{n+1} = F\left(x^n, z^{n+1}_{known}, w^{n+1}, t^{n+1}\right), \tag{22}$$

where t is the time, z_{known} represents known factors (external to the model) that influence x, and w represents unknown sources of variability, typically including unknown external factors and random effects. Hence the available data consists of the time series x^n, the times t^n and the z^n_{known}; both the w^n and the function F are unknown.

The form of (22) suggests explaining away through optimal transport the

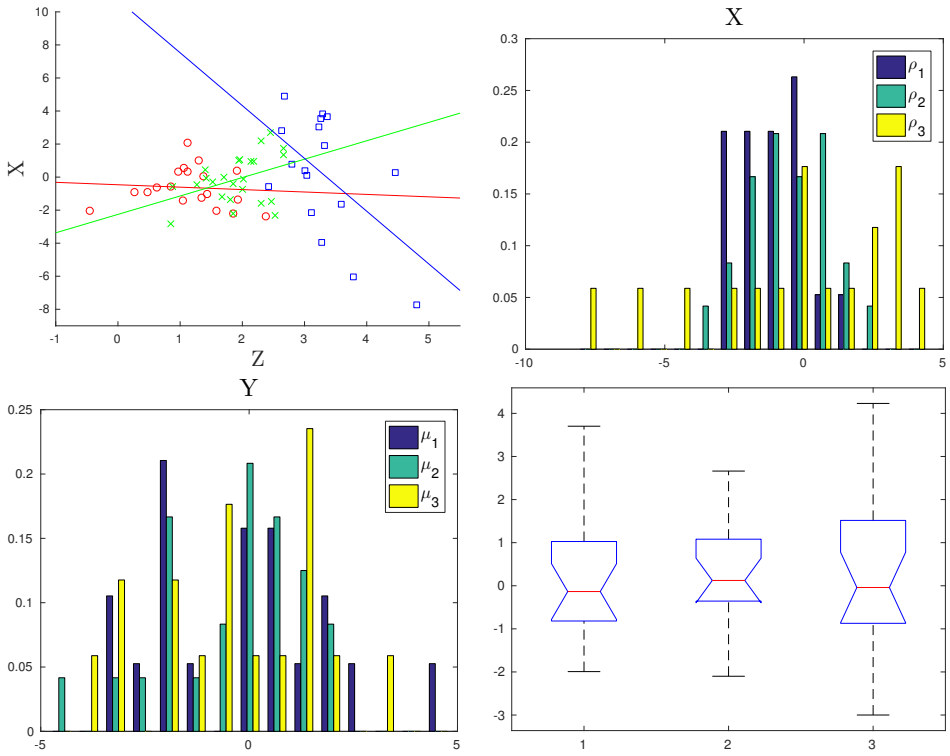

Figure 3: 20 points data set. Top-left: Scatter plot of the x_{is} as functions of the z_{is}, with points colored according to the value of s. Top-right: histogram of $\rho_{s=1,2,3}$. Bottom-left: histogram of $\mu_{s=1,2,3}$ obtained by filtering z from x. Bottom-right: Notched box plot of the sampled points from $\mu_{s=1,2,3}$.

variability in x^{n+1} attributable to $z^{n+1} = [x^n, t^{n+1}, z^{n+1}_{known}]$, with emphasis placed on the x^n: the prior element in the time series acts as an explanatory factor for x^{n+1}. After the filtering map

$$y^n = Y\left(x^n; z^n\right),$$

one has $y^n = G\left(w^n\right)$, meaning that we will have uncovered the hidden source of variability w.

As a simple example, we analyze synthetic data generated from the following two-dimensional Markov process:

$$x^n = \left(Ax^{n-1} + b\right) + \Sigma(x^{n-1})w^n, \qquad n \in [1 \ldots m] \tag{23}$$

where $A, \Sigma \in \mathbb{R}^{2\times 2}$ and $b, x, w \in \mathbb{R}^{2\times 1}$. The source of variability w has the form (unknown to the analyzer)

$$w^n = \begin{pmatrix} \sin(2\pi n/m) w_1^n \\ \cos(4\pi n/m) w_2^n \end{pmatrix}, \tag{24}$$

where w_1^n and w_2^n are i.i.d. normally distributed scalars. The inverse of the matrix Σ has the form

$$\Sigma^{-1} = \begin{pmatrix} a_1^2 & \varepsilon_{12} a_2 a_2 \\ \varepsilon_{12} a_2 a_2 & a_2^2 \end{pmatrix} \tag{25}$$

with $a_i = \exp(\alpha_i^T x + \beta_i)$, $\alpha_i \in \mathbb{R}^2$, $\beta_i \in \mathbb{R}$ and $\varepsilon_{12} \in [0, 1]$. Overall the model depends on 13 parameters that, together with the initial value x^0 and the realization of white noise in the w_i^n, were chosen randomly to generate the data represented in Figure 4.

The input to the algorithm is the time series $\{x^n\}$, which automatically generates the explanatory $\{z^n\} = \{x^{n-1}\}$. The parametric form used to model the $\Sigma(z)$ is the same of the one used to generate the data. The parameters are estimated via log-likelihood maximization and are used to compute the affine map described by eq. (19).

The goal is to compare the filtered signal y generated by the algorithm with the unknown source of variability w used in the model and displayed on the left panel of Figure 4. Figure 5 compares the filtered signal y (upper

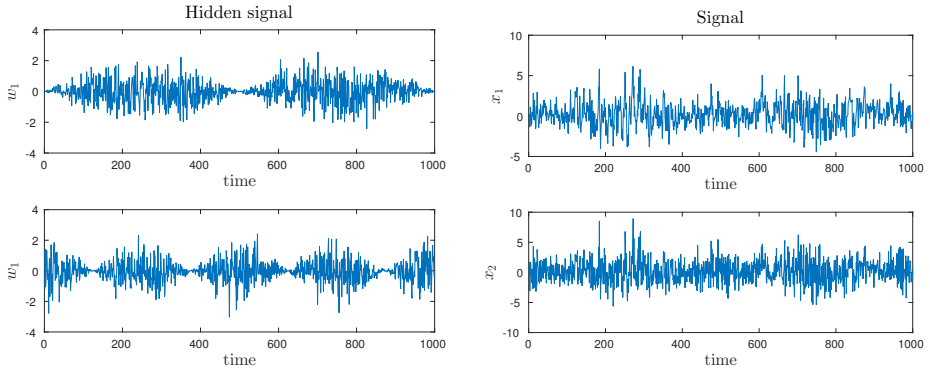

Figure 4: The left panel displays the hidden signal used to generate the time series displayed on the right panel according to the model in (23)

right panel) and the original signal x (left panel) together with the unknown

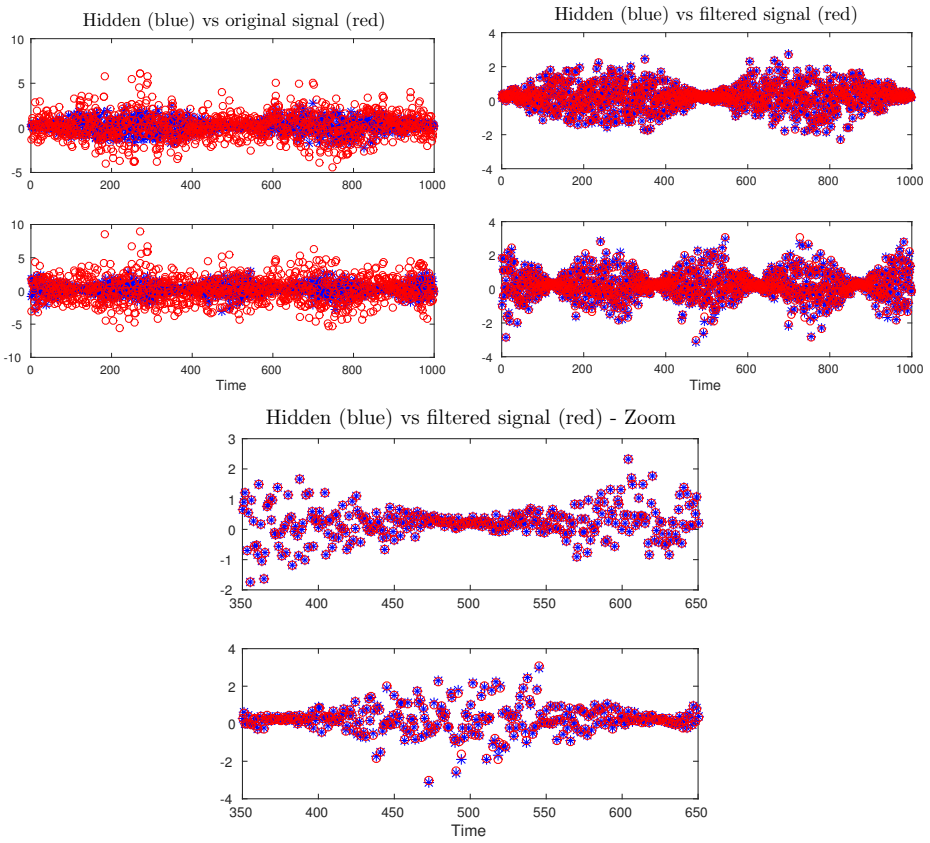

Figure 5: Upper left panel: the two components (at the top and bottom) of the hidden ($w_{1,2}$) and original ($x_{1,2}$) signals. Upper right panel: hidden (w) and filtered (y) signals. After explaining away the variability in x^{n+1} attributable to x^n, one should obtain $y^n = G(w^n)$. Because the poor-man solution uses only affine maps, G is also affine. We performed a linear least square fit between w and x (respectively y) to determine G. As one can see, while the best linear fit between w and x does not reveal any particular relation between the two variables, a linear fit between w and y, brings y to overlap with w, confirming that $y^n = G(w^n)$. Bottom panel: zoom of the upper right panel.

source of variability w. Because the model used to generate the data is the same as the one implemented in the filtering algorithm, the filtered signal recovers almost exactly the unknown source of variability as expected.

By contrast, Figure 6 shows an example in which the covariance matrix $\Sigma(z)$ used by the filtering algorithm and the one used to generate the time series in (23) have two different parametric forms. The covariance matrix used to generate the data is

$$\Sigma = \begin{pmatrix} \cos(a_1^2)^2 & d \\ d & a_2^{1/4} \end{pmatrix} \tag{26}$$

with $a_1 = \exp(\alpha^T(\cos(z)^2 + \beta z) + \gamma)$, $a_2 = \delta^T \sin(z)^2$ where $\alpha, \gamma \in \mathbb{R}^2$, $\beta, \gamma \in \mathbb{R}$, and the functions $\sin(z)^2, \cos(z)^2$ are defined entry-wise for $z \in \mathbb{R}^2$. The functional form for the hidden variable w is still given by (24). The model for the covariance matrix used by the filtering algorithm is (25) as before. Even though in this case we cannot expect to recover exactly the hidden source of variability w by looking at the filtered signal y, we should still be able to see that these two variables are linearly related to a good approximation. This is shown in Figure (6), which displays moving averages of y and w over small time intervals.

4.3 A real world example: explanation of temperature variability across the United States

As an example of the use of our methodology on real data, we develop an application to meteorology: the explanation of ground temperature variability in the continental United States. We use data publicly available in `http://www1.ncdc.noaa.gov/pub/data/uscrn/products/hourly02`, consisting of hourly-measured temperatures at various stations across the United States and one station in Ontario, Canada. We use only those 48 stations (see map in figure 7) for which data is available since at least 2005, thus covering over a decade of hourly data[3].

4.3.1 Filtering time of the day, season and global trends from each station

Figure 8 displays the temperature series available for Boulder, Colorado, at four levels of resolution: over 12 years, one year, one month and one week. The first two show a clear seasonal effect, with warm summers and cold winters, and the fourth a diurnal cycle of alternating warm days and cold

[3]Some of these stations provide data as early as 2003; we use this additional data when filtering factors individually from each station but not for global explanations of variability.

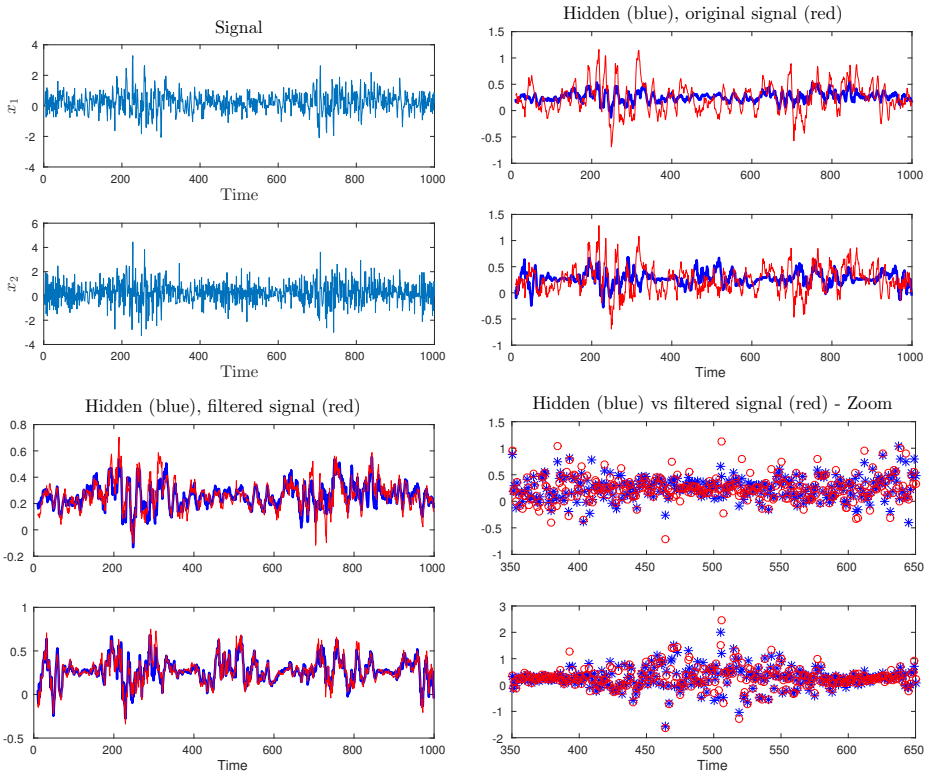

Figure 6: Top left panel: Signal x generated using the model in (23) with w and Σ given by (24) and (26) respectively. Top right panel: best linear fit between the moving averages of $x(t)$ and $w(t)$, with averaging time window set to 10 time-steps. Bottom left panels: moving averages of the best linear fit between $w(t)$ and the filtered signal $y(t)$ (with correlation values of 0.91 and 0.92 for the two components.) Bottom right panel: zoom of the best fit between w^n and y^n.

nights. The third has a less regular level of organization, yet one sees the typical weather systems running through the mid-latitudes with time-scales of around 5 days.

In order to filter the seasonal and diurnal variability, we introduce variability factors z that are periodic in time, with periods of one year and one day respectively, through Fourier series truncated after the fourth harmonic. All possible products of these two series are included too, as the parameters for daily variation may depend on the time of the year. In addition, we

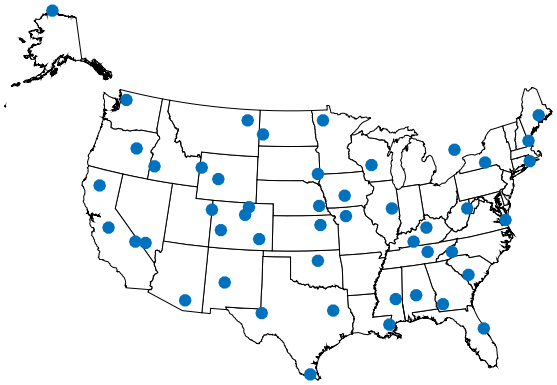

Figure 7: Location of the 48 stations used for the explanation of ground-temperature variability

include a trend linear in time, to account for possible long term effects such as global warming. Thus the matrix z adopts the form

$$z = \begin{pmatrix} t_y \\ \cos(t_y) \\ \sin(t_y) \\ \dots \\ \cos(4t_y) \\ \sin(4t_y) \\ \cos(t_d) \\ \sin(t_d) \\ \dots \\ \cos(4t_d) \\ \sin(4t_d) \\ \cos(t_y)\cos(t_d) \\ \dots \\ \cos(2t_y)\sin(3t_d) \\ \dots \\ \sin(4t_y)\sin(4t_d) \end{pmatrix},$$

where t represents the time in days, $t_d = 2\pi t$ and $t_y = \frac{2\pi t}{365.25}$.

The resulting time-series y of filtered temperature values are displayed

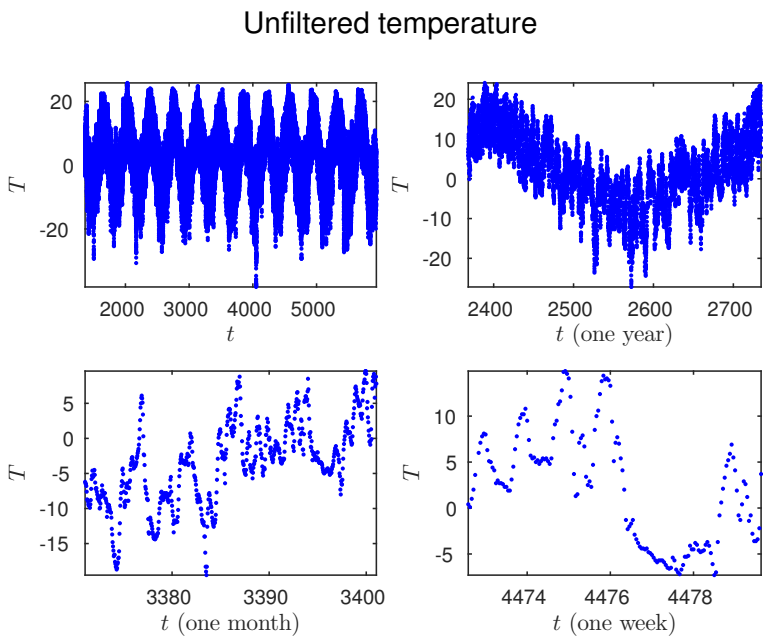

Figure 8: Hourly ground temperature measurements at Boulder (CO) at various degrees of resolution: full time series, one year, one month and one week. One can see seasonal effects in the first two panels, diurnal effects in the last panel, and less regular weather systems lasting a few days in the third. The time t is measured in days elapsed since January 1st 2000.

in figure 9 at the same levels of resolution of figure 8. As one can see, the seasonal and diurnal effects have been explained away, not so the aperiodic weather systems in the one-month plot. Figure 10 contains histograms of x (temperature) and y showing the reduction in variability that filtering the factors brings. Figure 11 displays, over an interval of 10 days, the original temperature $x(t)$, the filtered $y(t)$, and the estimated mean $\mu(t)$ plus/minus one standard deviation $\sigma(t)$. Here one sees the shape of the mean daily profile for this particular time of the year (late winter), and the effect of filtering this from the raw data. The patterns are quite different in figure 12 for Barrow, Alaska, where the daily signals are very small and the strong seasonal effects dominate.

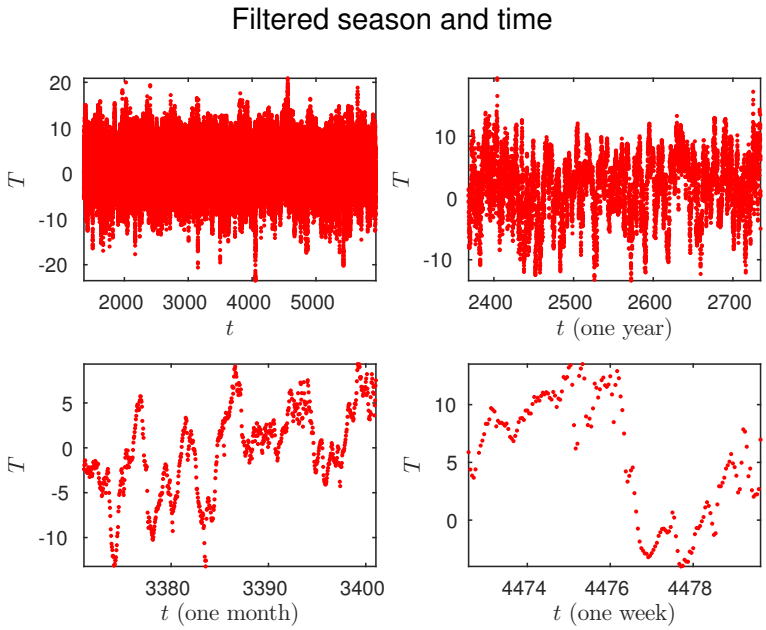

Figure 9: Filtered hourly ground temperature measurements at Boulder (CO) at the same degrees of resolution as in figure 8. The diurnal and seasonal effects have been eliminated, while the weather systems persist. The time t is measured in days elapsed since January 1st 2000.

4.3.2 Further filtering of latitude and elevation: repeated factor values

We have at this point a set of filtered time-series $\{y^i(t_j)\}$, where i stands for the station where the measurements have been performed and from which we have removed the variability associated with the diurnal cycle, the seasons and linear global trends. We switch focus now to the variability among stations, attempting to remove the fraction explainable by two external factors z: latitude and elevation.

The procedure is considerably simplified by the fact that many observations share the same factor values: all entries in the time series $\{y^i(t_j)\}$ share the same factors z_i. This simplification results from the partition of

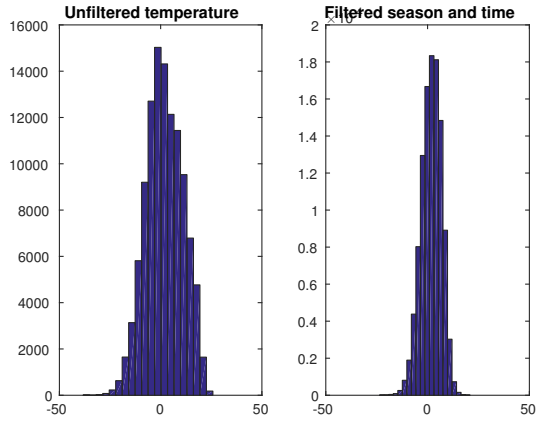

Figure 10: Histograms of hourly temperature data for Boulder (CO), with the original signal on the left, and the signal with seasonal and diurnal effects filtered on the right.

the variability into within and between stations in the loss function L:

$$\begin{aligned} L &= \sum_i \sum_j \frac{1}{2} \left[\frac{y^i(t_j) - \bar{y}(z_i)}{\sigma(z_i)} \right]^2 + \log(\sigma(z_i)) \\ &= \sum_i \sum_j \frac{1}{2} \left[\frac{\left(y^i(t_j) - \bar{y^i}\right) + \left(\bar{y^i} - \bar{y}(z_i)\right)}{\sigma(z_i)} \right]^2 + \log(\sigma(z_i)) \\ &= m \sum_i \frac{1}{2} \left[\frac{\sigma^{i2} + \left(\bar{y^i} - \bar{y}(z_i)\right)^2}{\sigma(z_i)^2} \right] + \log(\sigma(z_i)), \end{aligned} \tag{27}$$

where m is the number of times t_j. Here we have made a distinction between the mean and standard deviation of each time-series $(\bar{y^i}, \sigma^i)$ and their modeled values as functions of z, $(\bar{y}(z_i), \sigma(z_i))$. If the functional dependence of the latter on z were left completely free, minimizing L would make these two sets of values agree. However, for parametric or otherwise constrained models, this is generally not the case: two stations with nearby values of latitude and altitude will have nearby modeled parameters, while the parameters of their two time-series may be completely different due to other variability factors not included in the model, such as their proximity to the sea. The minimization of L from (27) requires only the mean and standard deviation for each station as opposed to the full time series, yielding a huge reduction in computational expense.

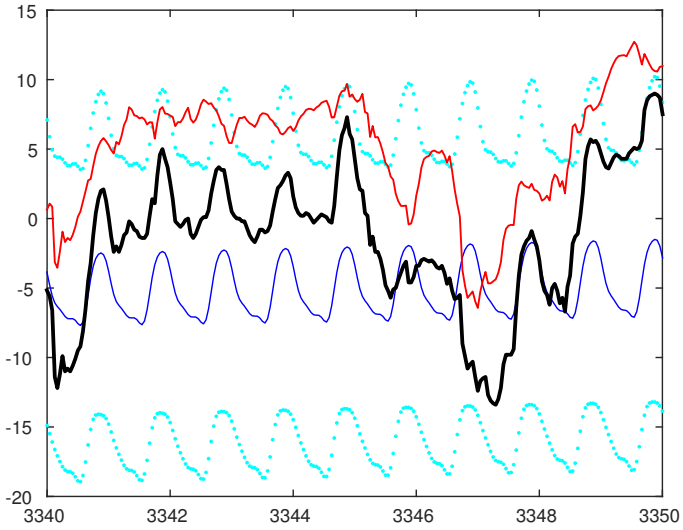

Figure 11: Temperature evolution over ten days in Boulder (CO), with the original signal in black, the modeled time-dependent mean in blue, the modeled standard deviation in light blue (added and subtracted to the mean), and the filtered signal in red. The diurnal signal is absent in the filtered signal, which lies above the original data because the ten days displayed are in winter time.

The results of this further filtering process are displayed in Figure 13, showing the filtered and furthered-filtered signals from all 48 stations over a period of 40 days. While the individually filtered signals show significant stratification and high variability among stations, the jointly-filtered ones have both highly reduced. Concatenating the data from all stations together, the total variance decreases from 63.8 to 30.2.

The natural next step would be to explain the joint variability across stations through global climate and weather patterns. This involves factor discovery, which we study in the following section.

5 Factor discovery

We have concentrated so far on the explanatory power of known factors. Yet one often seeks hidden factors to explain variability in data, with clustering and principal component analysis as typical examples. There are various ways to think of factor discovery. One is dimensionality reduction, where

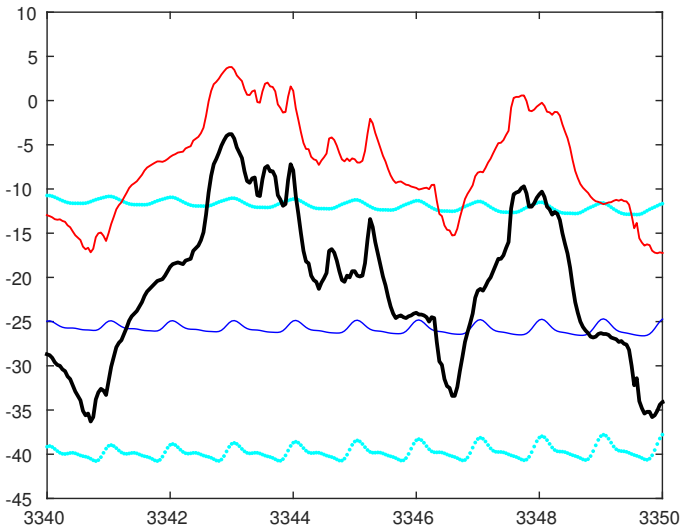

Figure 12: Temperature evolution over ten days in Barrow (AK), with the original signal in black, the modeled time-dependent mean in blue, the standard deviation in light blue (added and subtracted to the mean), and the filtered signal in red. The diurnal signal is much smaller at this extreme latitude, where seasonal effects dominate. Because these are ten winter days, the red signal, with seasonal effects filtered, lies well above the unfiltered black one.

one replaces the original data x by a smaller-dimensional, possibly discrete set of variables z that captures much of the original variability. Another is true explanation: the discovery of hidden underlying causes for the observed variability, such as a biological condition or a set of dynamical modes. Then there are areas intermediate between factor discovery and explanation of variability by known factors, including softly assigned factors and situations where only some values for the factors are known, for instance when certain illnesses are known to underlay a set of symptoms, but other yet to be found biological conditions are conjectured to account for much of the remaining variability.

The methodology in this article extends naturally to factor discovery. One proposes as before filtering maps

$$y_i = Y(x_i, z_i),$$

where the z_i are now extra unknowns, additional to the parameters that

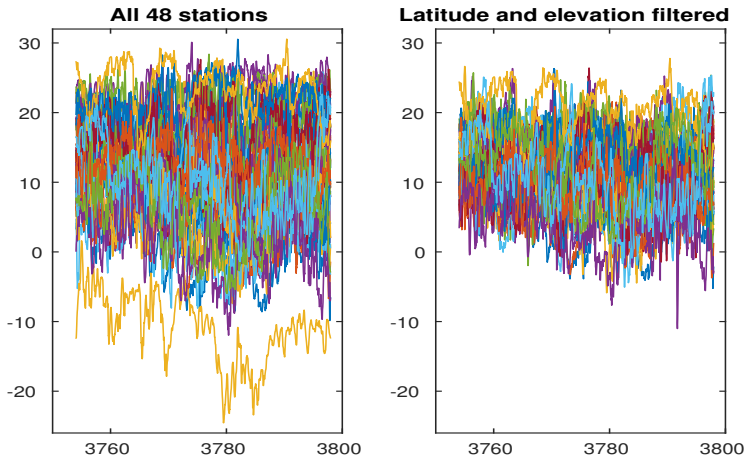

Figure 13: Results of filtering latitude and elevation effects from the already station-wise filtered temperature signals, displayed over 40 days. On the left, the filtered signals $y^i(t)$ from all 48 stations. On the right, the signals further filtered of elevation and latitude, which yields a cross-sectional reduction of variability.

determine the map Y. In order to assign the z_i, one requires that the variability left in the filtered variable y be minimal, thus making the z_i maximally explanatory. A natural quantification of the variability in y is given by its empirical variance,

$$\mathrm{Var}(y) = \frac{1}{m^2} \sum_{i,j} \|y_i - y_j\|^2,$$

so we propose

$$\min_{\{z_i\}} \mathrm{Var}\left(\{y_i = Y(x_i, z_i)\}\right)$$

The next subsections show how this principle yields k-means for z discrete and principal component analysis for z continuous when the maps $Y(x; z)$ are restricted to rigid translations, the poorest man solution of section 3.3. Hence these two procedures can be generalized broadly through the use of more general models.

5.1 Discrete factor discovery and k-means

Discrete factor discovery is tantamount to clustering: assigning a discrete label $z_i \in \{1, 2, \ldots, K\}$ to each observation amounts to dividing the dataset

into K classes. Mapping all these classes into their barycenter further filters from the data the variability associated with the class assignment.

Consider the very specific situation where the original variables x are in R^n and one restricts the maps to rigid translations (see end of section 3.3)

$$y_i = x_i + \beta_{z_i},$$

where

$$\beta_{z_i} = \bar{y} - \bar{x}_{z_i}, \quad \bar{y} = \frac{1}{m}\sum_i \bar{x}_{z_i}.$$

In this setting, the variance of the filtered signal can be rewritten as:

$$\begin{aligned} \mathrm{Var}(y) =& \frac{1}{m^2}\sum_{i,j}\|y_i - y_j\|^2 = \frac{1}{m^2}\sum_{i,j}\|(x_i - \bar{x}_{z_i}) - (x_j - \bar{x}_{z_j})\|^2 = \\ &= \frac{2}{m}\sum_i \|x_i - \bar{x}_{z_i}\|^2 = \frac{2}{m}\sum_k \sum_{i\in S_k}\|x_i - \bar{x}_k\|^2 \end{aligned} \tag{28}$$

where we have used the fact that

$$\frac{1}{m}\sum_i (x_i - \bar{x}_{z_i}) = 0, \tag{29}$$

and have reorganized the result into sums by class, denoting by S_k the set of samples $\{i\}$ with $z_i = k$.

It follows that minimizing $\mathrm{Var}(y)$ over all the possible assignments z_i agrees with k-means, which divides the data into K classes, where the observations in class k are closer to their class mean $\bar{x}_k$ than to the means of all other classes. Hence the procedure replicates k-means, followed by K rigid translations that move all centers $\bar{x}_k$ into their barycenter $\bar{y}$, thus removing the variability associated with the newly-discovered classes.

5.2 Continuous factor discovery and principal component analysis

In the continuous setting, the poorest man solution yields

$$y = x + \beta(z), \tag{30}$$

with

$$\beta(z) = \bar{y} - \bar{x}(z), \quad \bar{y} = \int \bar{x}(z)\,\nu(z)dz, \quad \bar{x}(z) = \int x\rho(x|z)dx,$$

where $\nu(z)$ and $\rho(x|z)$ indicate the distribution of z and the conditional distribution of x given z. As in the previous subsection we have that

$$\int\int \|y_1 - y_2\|^2 \mu(y_1)\mu(y_2)dy_1dy_2 = 2\int\int \|x - \bar{x}(z)\|^2 f(x,z)dxdz \quad (31)$$

where $f(x,z)$ is the joint probability distribution of x and z and we have used the fact that

$$\int (x - \bar{x}(z))f(x,z)dxdz = 0.$$

Hence, in terms of samples, minimizing the variance of y over the assignments z_i corresponds to minimizing

$$\frac{1}{m}\sum_i \|x_i - \bar{x}_{z_i}\|^2.$$

The link with principal component analysis arises if we model $\bar{x}$ as a affine function of z:

$$\bar{x}(z) = Az + b,$$

where $A \in R^{n\times d}$, $b \in R^n$. Then the $z_i \in R^d$ are minimizers of

$$L = \sum_i \|x_i - (Az_i + b)\|^2. \quad (32)$$

Notice though that L is also the loss function that A and b minimize when the z_i are given (it defines the mean $\bar{x}_z$), so factor assignment and filtering jointly satisfy the variational principle

$$\min_{A,b,\{z_i\}} L.$$

To understand the nature of the solution to this minimization problem, assume that A and b have already been found. Then $Az + b$ determines a d-dimensional hyperplane, and minimizing (32) over each z_i finds the point $Az_i + b$ on that hyperplane that is closest to x_i, i.e. the projection of x_i onto the hyperplane. Hence minimizing (32) over A and b corresponds to finding the hyperplane of dimension d that minimizes the sum of the square distances between the x_i and their projections onto that hyperplane. But this is also the definition of the subspace generated by the first d principal components of the matrix $X = [x_1 \dots x_m]$.

Then we have "rediscovered" principal component analysis as a particularly simple instance (affine $\beta(z)$) of our "poorest man solution" of factor

discovery through optimal transport. To understand how principal components are interpreted from this viewpoint, notice that the resulting map $Y(x;z) = x - \bar{x}(z) + \bar{y}$ removes from each x its projection onto the hyperplane $Az+b$, leaving only the unexplained component normal to this plane, and then adds a uniform $\bar{y}$ that agrees with the mean of all x_i, so as to preserve this mean for the y_i and hence minimally distort the data.

Notice that the optimal hyperplane $Az + b$ and the means $\bar{x}(z_i)$ are generally unique, but the matrix A, the vector b and factor values z_i are not: the gauge transformation

$$z \to C(z - z_0), \quad A \to AC^{-1}, \quad b \to b + Az_0$$

preserves optimality for any invertible matrix C and vector z_0. This is a general feature of factor discovery: any invertible function of the factors z can explain the same variability as z. In the present case, where the model for the dependence on z of the map's parameters is linear, the gauge is limited to invertible linear functions of z. This gauge invariance allows us to choose the representation corresponding to the sorted principal components of x: we perform the QR decomposition $A = QR$, define $Z = Rz$, perform the singular value decomposition $Z = \tilde{U}\Sigma V'$, and finally write

$$x - b \approx Az = U\Sigma V', \quad \text{where} \quad U = Q\tilde{U}.$$

Clearly the procedure can be easily generalized, thus providing powerful extensions of principal components analysis. Here we mention three such extensions; many more will be developed in [23]:

1. Using non-linear models for $\bar{x}(z)$, for instance through the introduction of feature functions $f_j(z)$ as in (21), moves us away from linear components into principal surfaces [13].

2. Moving from rigid translations to general affine maps $y = \alpha(z)x + \beta(z)$ –thus modeling both the mean $\bar{x}$ and the covariance Σ as functions of z– greatly increases the amount of explainable variability.

3. **Smooth principal components:** In our example of ground temperature data, one may be interested in capturing modes of the system that evolve smoothly over time. One way to achieve that is to penalize the non-smoothness of the z in (32) through the modified optimization problem [23]

$$L = \sum_i \|x_i - (Az_i + b)\|^2 + \lambda \sum_{k=1}^{d} \|A^k\|^2 \sum_{j=2}^{m} \|z_{j+1}^k - z_j^k\|^2, \qquad (33)$$

where A^k indicates the k-th column of A and the penalization constant λ can be written in the form

$$\lambda = \left(\frac{T}{\Delta t}\right)^2,$$

with Δt the time interval between observations, and T the time scale below which one penalizes $O(1)$ variations. Figures 14 and 15 show the result of applying this procedure to find the first smooth principal component of the filtered temperature data from subsection 4.3.2, with $T = 180$ days. We see on the map that the spatial pattern (the entries of A) consists of a dipole between the West Coast and the rest of the country, with the strongest signal in the midwest. The entries of z, on the other hand, evolve smoothly –as required– over time, with a typical time scale of roughly 4 years. A natural climate variability factor known to evolve over such scales is the El Niño Southern Oscillation (ENSO). We superimpose on the plot of z the evolution of the El Niño Index (ENI), representing the three month moving average of the temperature over the Indian Ocean. As one can see, there is a clear relation between ENI and z: they follow each other, correlating positively until the late 2010, and negatively since. To our knowledge, this is the first time that this signature of ENSO on the ground level temperature across the US has been detected.

For comparison, we display in Figures 16 and 17 the results of applying the same procedure with identical parameters to the original, unfiltered data. Here the first smooth principal component captures the seasonal effects, as well as their modulation through the years. On the map, one sees the seasons affecting felt most strongly the areas furthest from the sea. Thus the time-scale of the first smooth principal component of the filtered data agrees with the El Niño Index in time scale (of about 4 years, in contrast to the yearly scale of the unfiltered data); the Pearson correlations of the filtered and unfiltered components with the ENI are 0.26 and -0.065 respectively up to 2010, and -0.18 and -0.07 from 2010 on.

6 Conditional density estimation and sampling

Conditional density estimation is ubiquitous when estimating uncertainty: the conditional distribution $\rho(x|z)$ quantifies the variability in x not attributable to z, as opposed to the unconditioned probability distribution

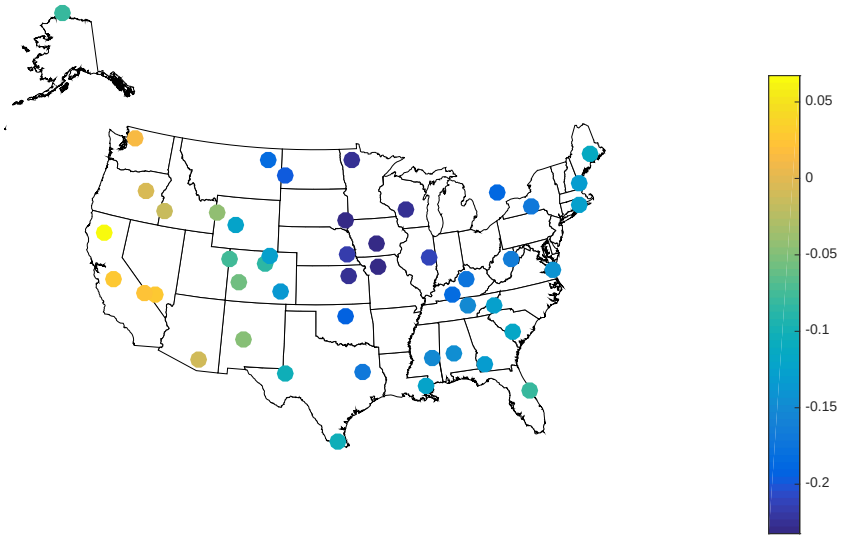

Figure 14: Spatial distribution of the first smooth principal component of the filtered data: a dipole between the west coast and the rest of the country, with signal strongest in the mid-west.

$\rho(x)$ which includes the variability in z. Even conventional tools such as least-square regression can be regarded as estimators of the mean of the variable of interest conditioned on the value of the predictors.

In the framework of this article, conditional density estimation arises as a byproduct of the filtering algorithm. Since the maps $y = Y(x; z)$ push forward the conditional distributions $\rho(x|z)$ into their barycenter $\mu(y)$, it follows that

$$\rho(x|z) = J_z(x)\ \mu\left(Y(x; z)\right),$$

where $J_z(x)$ is the Jacobian determinant of $Y(x; z)$. Thus one only needs to estimate $\mu(y)$, a comparatively much easier task than estimating $\rho(x|z)$ directly, since:

1. There is only one density to estimate, as opposed to one for each value of z.

2. There are as many samples (m in our notation) y_i of μ as original pairs (x_i, z_i), as opposed to few or no samples of $\rho(x|z)$ for individual values of z.

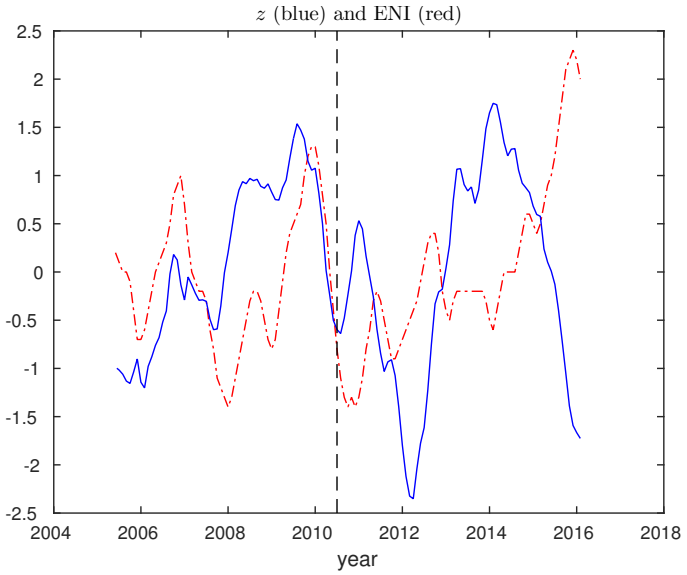

Figure 15: Temporal evolution of the first smooth principal component in blue and El Niño Index (the three-month moving average of the temperature over the Indian Ocean) in red. Until the late 2010, the two signals are positively correlated (implying that the temperature in the mid west correlates negatively with the temperature of the Indian Ocean), while from the late 2010 onwards the two signals are negatively correlated.

3. The distribution $\mu(y)$ has less variability than the unconditioned $\rho(x)$, as the variability attributable to z has been removed.

A task much simpler –and oftentimes more useful– than estimating $\rho(x|z)$ for a given value z^* of z is to sample it. Since the filtering procedure has already produced m samples y_i of μ, one can produce m samples $x_i^{z^*}$ from $\rho(x|z^*)$ by simply inverting the filtering map using that particular value for z:

$$x_i^{z^*} = X(y_i, z^*), \tag{34}$$

where $X(y, z)$ is the inverse map for fixed z of $Y(x; z)$. Thus the procedure transforms m samples x_i from different distributions $\rho(x|z_i)$ into m samples of the single distribution $\rho(x|z^*)$.

For a first example that can be solved in closed form and that adapts naturally to our poor man solution, consider the family of exponential dis-

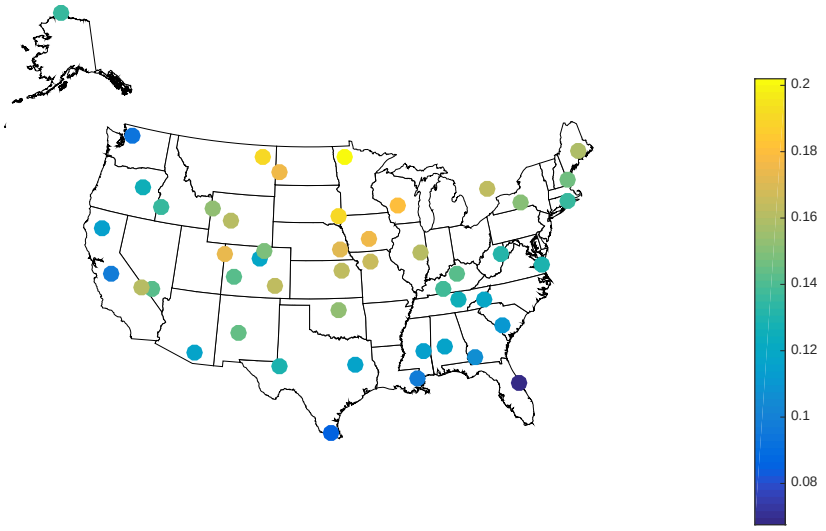

Figure 16: Spatial distribution of the first smooth principal component of the unfiltered data: the seasons are most strongly felt furthest from the sea.

tributions

$$\rho(x|z) = \lambda(z)e^{-\lambda(z)(x-s(z))}, \tag{35}$$

where the rate $\lambda(z) > 0$ and the displacement $s(z)$ are given by smooth functions of z.

The barycenter of this family of distributions $\rho(x|z)$ is again an exponential distribution

$$\mu(y) = \lambda_\mu \; e^{-\lambda_\mu(y-s_\mu)},$$

with rate λ_μ and displacement s_μ given by

$$\frac{1}{\lambda_\mu} = E\left[\frac{1}{\lambda(z)}\right]$$

and

$$s_\mu = E\left[s(z)\right],$$

where the expected value is taken over the distribution $\nu(z)$ underlying z.

To see this, notice that the linear transformations

$$x_z = \frac{\lambda_\mu}{\lambda(z)}\left(y - s_\mu\right) + s(z)$$

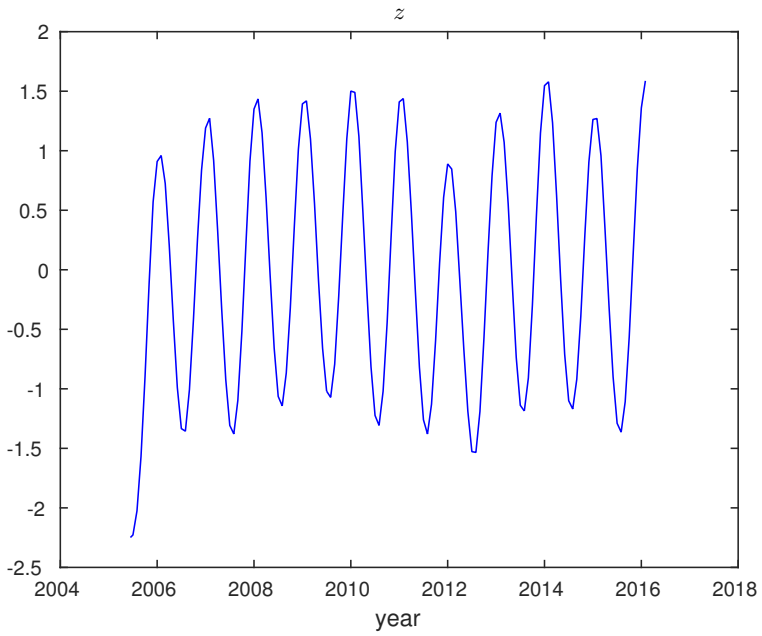

Figure 17: Temporal evolution of the first smooth principal component of the unfiltered data, capturing the seasons as well as their yearly variability.

1. push $\mu(y)$ to $\rho(x|z)$,
2. are optimal, as they are given by the gradient of the convex potential

$$\phi(y) = \frac{1}{2}\frac{\lambda_\mu}{\lambda(z)}\left(y - s_\mu\right)^2 + s(z)y,$$

3. satisfy the property that

$$y = E\left[x_z\right] = \int x_z \nu(z)\, dz,$$

as can be readily verified.

Properties 1, 2 and 3 fully characterize μ as the barycenter of the $\rho(x|z)$ [1].

In this case, restricting the maps to affine, as in our poor man solution, does not introduce any approximation regarding the barycenter μ, so by inverting the map we can recover exactly $\rho(x|z)$ for any value z^* of z. Figure

18 shows a numerical example, where z has a bimodal distribution given by the Gaussian mixture

$$\nu(z) = \frac{1}{3}\left[\mathcal{N}(-0.43, 0.3) + \mathcal{N}(0.8, 0.25) + \mathcal{N}(4, 0.5)\right]$$

where $\mathcal{N}(m, \sigma)$ denotes a Gaussian with mean m and standard deviation σ, and each conditional distribution $\rho(x|z)$ is exponential

$$\rho(x|z) = \begin{cases} \lambda(z)e^{-\lambda(z)(x-s(z))} & \text{if } x > s \\ 0 & \text{otherwise} \end{cases}$$

with $\lambda(z) = \exp(0.91z + 0.76)$ and $s(z) = 0.54z + 0.831$.

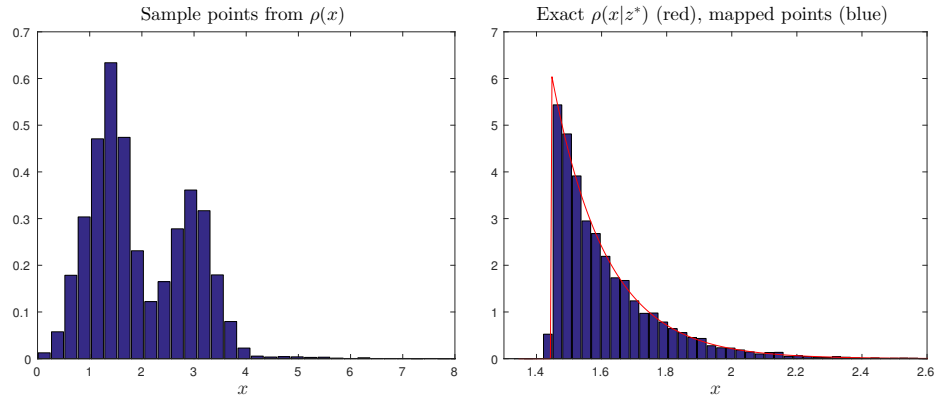

Figure 18: Left panel: after sampling m points z_i from $\nu(z)$, one samples for each z_i a point x_i from $\rho(x|z)$. The left panel shows the distribution $\rho(x) = \int \rho(x|z)\nu(z)dz$, where $\nu(z)$ is a 2 component Gaussian mixture. Right panel: exact distribution $\rho(x|z^*)$ (in red) together with the histogram of the points $x_i^{z^*}$ obtained from mapping each sample y_i from the barycenter μ through $X(y_i, z^*)$.

The perfect results in the example above follow from the fact that the restriction to affine maps in our poor man filtering procedure was in fact not restrictive at all for the exponential family of distributions. An affine map does not change the shape of a distribution, so the fact that we could transform samples from $\rho(x|z_1)$ into samples from $\rho(x|z_2)$ for $z_1 \neq z_2$ was only possible because both distributions had the same shape. This raises the question of how to sample $\rho(x|z)$ when the family of maps allowed is

not big enough to push forward the $\rho(x|z)$ for different values of z into each other.

In order to overcome this limitation, one could only map back those samples that have original values z_i of z not too far from z^*. Better still, one can give each sample $x_i^{z^*}$ a weight $w(z_i)$ that peaks near z^*, such that

$$\frac{\sum_i w_i z_i}{\sum_i w_i} = z^*,$$

thus giving more relevance to those sample points y_i of μ that were mapped from pairs (x_i, z_i) with z_i close to z^*. As a proof of concept, consider the family of distributions

$$\rho(x|z) = \phi(z)\mathcal{N}(2, 0.4) + (1 - \phi(z))\mathcal{N}(5, 0.8), \tag{36}$$

with $\phi(z) = 1 - z$ and $z \in [0, 1]$. The barycenter μ of this family of distribution under affine maps is showed in Figure 19. Even though μ is unimodal, it is still possible to recover the original, bimodal $\rho(x|z^*)$ by mapping the points with the inverse of the affine map $Y(x; z^*)$ and then re-weighting them with $w(z)$. In the numerical example displayed in Figure 19, $w(z)$ was chosen to be a Gaussian with mean z^* and standard deviation $\sigma_w = 10^{-2}$. In general, the width of $w(z)$ is problem specific, depending on the smoothness of $\rho(x|z)$ as a function of z and on how many sample points from $\rho(x|z)$ are available in a neighborhood of z^*.

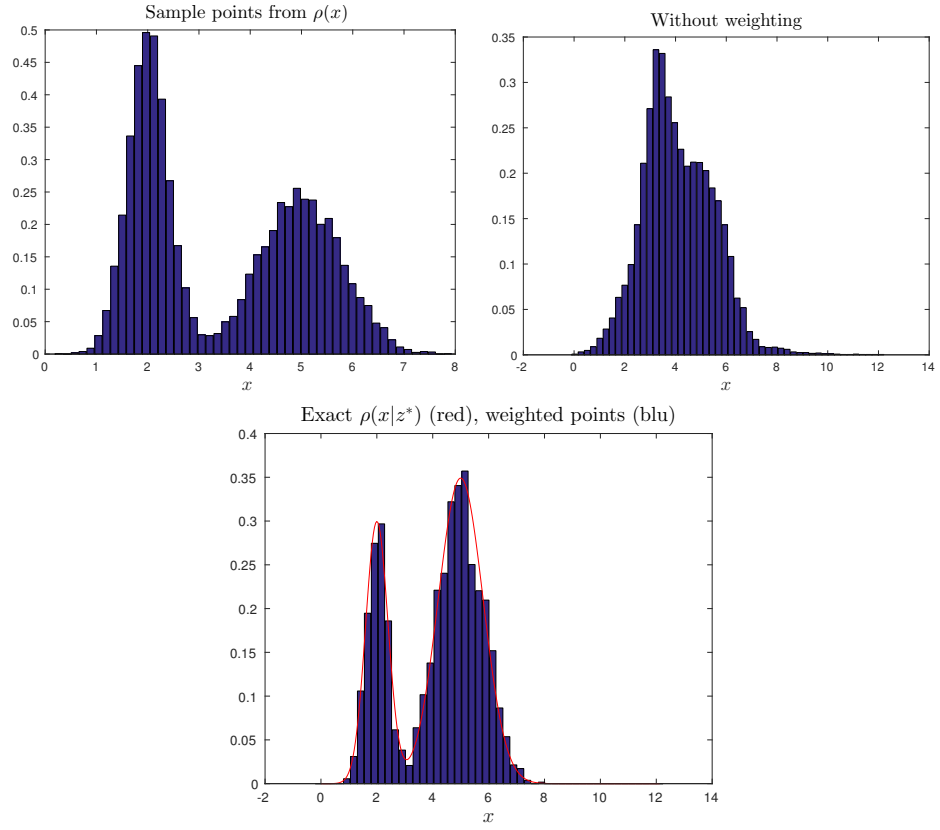

Figure 19: Upper left panel: histogram of points sampled from $\rho(x)$ as in the left panel of Figure 18. Upper right panel: histogram of the points obtained by mapping μ with $y^{-1}(x, z^*)$. Lower panel: histogram of the points obtained by weighting the points mapped from μ as described in the text.

7 Conclusions

This article has introduced and developed a methodology for the attribution of variability in a set of observations $\{x_i\}$ to known and unknown factors $\{z_i\}$, as well as the removal of those attributable components of the variability from the data. In the language of optimal transport, the procedure seeks a family of maps $y = Y(x; z)$ that push forward the conditional probability distributions $\rho(x|z)$ onto their weighted barycenter $\mu(y)$ while minimizing a transportation cost $E[c(x, y)]$ associated with data distortion.

The factors z can be discrete (as in the amalgamation of data sets and the removal from clinical data of the variability associated with a patient's sex) or continuous (as in a patient's age in clinical data and the time of the day in meteorological observations.) Time series analysis can be regarded as a particular instance of the procedure, with factors that include at each time elements from the past.

Uncovering previously unknown discrete factors is tantamount to clustering, followed by the removal of the variability associated with the classes just found. In this discrete case, when the maps $Y(x; z)$ are restricted to rigid, z-dependent translations, the resulting clustering procedure agrees with k-means. When similar rigid translations are made to depend linearly on unknown factors z adopting a continuum of values, one recovers principal component analysis. This immediately suggest a number of generalizations, some of which will be described in [23].

Since the maps $Y(x; z)$ are necessarily invertible, they immediately provide a natural procedure for the estimation and sampling of conditional probability distributions, a task of broad applicability in fields ranging from personalized medicine to financial risk quantification.

This article discussed a general framework and some ramifications of the methodology proposed, but postponed many natural extensions to further work. In particular, the family of maps used for the examples was restricted to at most affine transformations. When extending the procedure to more general maps, a choice has to be made on whether to still model the conditional distributions $\rho(x|z)$ –as when parameterizing $\bar{x}(z)$ and $\Sigma(z)$ in the context of the poor man solution– or to model directly the maps $Y(x; z)$, imposing as a constraint that the resulting $\{y_i\}$ should be independent of the $\{z_i\}$. Work in progress explores various ways of achieving higher generality in the maps, both in parametric and non-parametric ways.

The general methodology developed here has significant applicability in many data-rich fields. To do them justice, these applications must be developed in field-specific work. In particular, this applies to the explana-

tion of ground temperature variability across the United States, which was sketched here to illustrate the use of the methodology on real data. This explanation can be carried much further: for instance, we stopped short of explaining the variability in the data attributable to traveling whether systems through asynchronous principal components, another extension of PCA to be described in [23].

8 Acknowledgments

This work was initially motivated by a collaboration with Dr. Martín Cadeiras and Eleanor Chang from UCLA on the diagnosis of transplanted hearts, which will be reported elsewhere. It also benefitted from discussions with Lyuba Chumakova, Paul A. Milewski, Rubén R. Rosales and Horacio Rotstein.

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

References

[1] Agueh, M.; Carlier, G. Barycenter in the Wasserstein space. *SIAM J. MATH. ANAL.* **43** (2011), no. 2, 094–924.

[2] Angenent, S.; Haker, S.; Tannenbaum, A.; Kikinis, R. On area preserving mappings of minimal distortion. in *System Theory*, pp. 275–286, Springer, 2000.

[3] Benamou, J.-D.; Brenier, Y. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik* **84** (2000), no. 3, 375–393.

[4] Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics* **44** (1991), no. 4, 375–417.

[5] Caffarelli, L. A. The Monge-Ampère equation and optimal transportation, an elementary review. in *Optimal transportation and applications*, pp. 1–10, Springer, 2003.

[6] Chartrand, R.; Vixie, K.; Wohlberg, B.; Bollt, E. A gradient descent solution to the Monge-Kantorovich problem. *Applied Mathematical Sciences* **3** (2009), no. 22, 1071–1080.

[7] Chen, C.; Grennan, K.; Badner, J.; Zhang, D.; E, G.; et al. Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *Plos One* **6** (2011), no. 2, 17 232.

[8] Cochran, W. G. *Sampling techniques*, John Wiley & Sons, 2007.

[9] Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transportation, in *Advances in Neural Information Processing Systems*, 2013 pp. 2292–2300.

[10] Cuturi, M.; Doucet, A. Fast computation of Wasserstein barycenters. *arXiv preprint arXiv:1310.4375* (2013).

[11] Froese, B. D.; Oberman, A. M. Fast finite difference solvers for singular solutions of the elliptic Monge-Ampère equation. *Journal of Computational Physics* **230** (2011), no. 3, 818–834.

[12] Haber, E.; Rehman, T.; Tannenbaum, A. An efficient numerical method for the solution of the L_2 optimal mass transfer problem. *SIAM Journal on Scientific Computing* **32** (2010), no. 1, 197–211.

[13] Hastie, T.; Stuetzle, W. Principal curves. *Journal of the American Statistical Association* **84** (1989), no. 406, 502–516.

[14] Kantorovich, L. V. On the translocation of masses. *Compt. Rend. Akad. Sei* **7** (1942), 199–201.

[15] Knott, M.; Smith, C. On a generalization of cyclic monotonicity and distances among random vectors. *Linear algebra and its applications* **199** (1994), 363–371.

[16] Lu, J.; Schumacher, M.; Scherer, A.; Sanoudou, D.; et al. A comparison of batch effect removal methods for enhancement of prediction performance using maqc-ii microarray gene expression data. *The Pharmacogenomics Journal* **10** (2010), 278–291.

[17] Manly, B. F. *Randomization, bootstrap and Monte Carlo methods in biology*, vol. 70, CRC Press, 2006.

[18] McDonald, R. P. *Factor analysis and related methods*, Psychology Press, 2014.

[19] Monge, G. *Mémoire sur la théorie des déblais et des remblais*, De l'Imprimerie Royale, 1781.

[20] Neter, J.; Kutner, M. H.; Nachtsheim, C. J.; Wasserman, W. *Applied linear statistical models*, vol. 4, Irwin Chicago, 1996.

[21] Pinheiro, J. C.; Bates, D. M. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing* **6** (1996), no. 3, 289–296.

[22] Sulman, M. M.; Williams, J.; Russell, R. D. An efficient approach for the numerical solution of the Monge-Ampère equation. *Applied Numerical Mathematics* **61** (2011), no. 3, 298–307.

[23] Tabak, E. G.; Trigila, G. Around principal components. *In preparation* .

[24] Trigila, G.; Tabak, E. G. Data-driven optimal transport. *Communications on Pure and Applied Mathematics* **66** (2016), 613.

[25] Villani, C. *Topics in optimal transportation*, vol. 58, American Mathematical Soc., 2003.