Noname manuscript No.
(will be inserted by the editor)

An adaptive linear programming methodology for data driven optimal transport

Weikun Chen · Esteban G.Tabak

Received: date / Accepted: date

Abstract An adaptive methodology is proposed to solve the data–driven optimal transport problem: to find the coupling between two distributions that minimizes a transportation cost, when the distributions are only known through samples. The methodology bypasses density estimation, using adaptive kernel functions to reduce the problem to a sequence of finite linear programming problems.

Keywords optimal transport · linear programming · data analysis

1 Introduction

The optimal transport problem has received a considerable interests in recent years, due in part to its wide scope of applicability in fields that include econometrics, fluid dynamics, automatic control, transportation, statistical physics, shape optimization, expert systems and meteorology [12,2].

In a modern formulation of Monge's original statement of the optimal transport problem, one has two probability density functions $\rho(x), x \in \mathbb{R}^d$ and $\mu(y), y \in \mathbb{R}^d$. A map M from $\mathbb{R}^d$ to $\mathbb{R}^d$ is said to preserve mass (or to push forward ρ into μ) if, for all bounded subset $S \subset \mathbb{R}^d$,

$$\int_{x\in S} \rho(x)dx = \int_{y\in M(S)} \mu(y)dy \ . \tag{1}$$

For smooth one-to-one maps, this yields the differential condition

$$|\det(\nabla M(x))|\, \mu(M(x)) = \rho(x) \ . \tag{2}$$

Weikun Chen
251 Mercer St, New York, NY 10012
E-mail: wc906@nyu.edu

Esteban G.Tabak
251 Mercer St, New York, NY 10012

Among all mass preserving maps, one seeks the *optimal map* that minimizes a cost such as the Wasserstein distance

$$d_p(\rho(x), \mu(y)) = \left(\inf_M \int \|M(x) - x\|^p \rho(x) dx \right)^{1/p}, \tag{3}$$

where $p \geq 1$ is fixed.

Monge's formulation, which seeks a map $M(x)$ connecting $\rho(x)$ and $\mu(y)$, is not easy to work with, as it leads to problems that are nonlinear and not always have solutions. In addition, many applications do not require the transport between $\rho(x)$ and $\mu(y)$ to be one-to-one. Kantorovich [9] proposed a relaxation where the map $y = M(x)$ is replaced by a more general *coupling* $\pi(x, y)$:

$$\begin{aligned} \min_{\pi(x,y)} \quad & \int c(x,y)\pi(x,y)dxdy \\ s.t \quad & \int \pi(x,y)dx = \mu(y) \\ & \int \pi(x,y)dy = \rho(x) \\ & \pi(x,y) \geq 0. \end{aligned} \tag{4}$$

Here $c(x, y)$ is a pointwise cost function such as $\|y - x\|^p$; the unknown is the joint distribution $\pi(x, y)$ with marginals $\rho(x)$ and $\mu(y)$. It has been shown that, under suitable conditions on the cost function and the marginals, the solution to Kantorovich's problem also solves Monge's, i.e. the coupling $\pi(x, y)$ is concentrated on a line $y = M(x)$ [5].

Among the methods that have been proposed to solve the optimal transport problem numerically, Benamou and Brenier [7] proposed a fluid mechanics framework, constructing an optimal path from $\rho(x)$ to $\mu(y)$ by solving an optimization problem with a partial differential equation as a constraint. Haber, Rehman and Tannenbaum [6] proposed a modification of the objective function, discretized both the objective function and the constraints and solved the problem using sequential quadratic programming. Other numerical procedures can be found in[1,4,10,11].

All these procedures assume that the marginal distributions $\rho(x)$ and $\mu(y)$ are known explicitly. Yet this is not the case in the great majority of applications, where these distributions are only known through a finite set of samples $\{x_i\}$ and $\{y_j\}$ from $\rho(x)$ and $\mu(y)$ respectively. A formulation of the optimal transportation problem in terms of samples was proposed in [3], which also developed a methodology for its numerical solution following a gradient flow in feature-space that pushes $\rho(x)$ to the target distribution $\mu(y)$ through a time-dependent map $z(x; t)$, with $z(x; 0) = x$ and $z(x; \infty) = y(x)$.

In this article, we propose an alternative methodology for solving the sample-based optimal transportation problem, using Kernel functions in both primal and dual space to discretize the problem into a finite linear programming one. The Kernel functions are chosen adaptively in a sequence of such problems, where the active constraints in the solution to the current dual problem provide the basis to build the Kernel functions that define the next problem in the sequence, so as to improve the accuracy of the numerical solution without overly increasing the problem's size.

The outline of the paper is as follows. In section 2, we formulate the problem using data points directly and use kernel functions to relax the problem to a finite linear programming problem. In section 3, we propose an algorithm to adaptably update the kernel functions. Section 4 contains some numerical results, and a summary is provided in section 5.

2 Problem formulation

We start with Kantorovich's formulation of the optimal transport problem:

$$\begin{aligned} \min \quad & \int c(x,y)\pi(x,y)dxdy \\ s.t \quad & \int \pi(x,y)dx = \mu(y) \\ & \int \pi(x,y)dy = \rho(x) \\ & \pi(x,y) \geq 0 \end{aligned} \tag{5}$$

Introducing Lagrange multipliers $\varphi(x)$ and $\psi(y)$, one can write the problem in an equivalent unconstrained formulation:

$$\begin{aligned} \min_{\pi(x,y)\geq 0} \max_{\varphi(x),\psi(y)} \int c(x,y)\pi(x,y)dxdy - \int \varphi(x)\left[\int \pi(x,y)dy - \rho(x)\right]dx \\ - \int \psi(y)\left[\int \pi(x,y)dx - \mu(y)\right]dy, \end{aligned} \tag{6}$$

from which the dual problem can be found by exchanging the order of the minimization and maximization:

$$\begin{aligned} \max_{\varphi(x),\psi(y)} \min_{\pi(x,y)\geq 0} \int (c(x,y) - \varphi(x) - \psi(y))\pi(x,y)dxdy \\ + \int \varphi(x)\rho(x)dx + \int \psi(y)\mu(y)dy \end{aligned} \tag{7}$$

yielding

$$\begin{aligned} \max \quad & \int \varphi(x)\rho(x)dx + \int \psi(y)\mu(y)dy \\ s.t. \quad & \varphi(x) + \psi(y) \leq c(x,y). \end{aligned} \tag{8}$$

In most applications, the distributions $\rho(x)$ and $\mu(y)$ are only known through samples, hence the need to develop a formulation that uses these samples directly. If one has two sets of samples: $\{x_i\}_{i=1}^{n_x}$ from $\rho(x)$ and $\{y_j\}_{j=1}^{n_y}$ from

$\mu(y)$, it is natural to modify the dual problem 8, replacing the expected values of $\varphi(x)$ and $\psi(y)$ by their empirical means:

$$\begin{aligned} \max \quad & \frac{1}{n_x}\sum_i \varphi(x_i) + \frac{1}{n_y}\sum_j \psi(y_j) \\ s.t. \quad & \varphi(x) + \psi(y) \leq c(x,y) \end{aligned} \tag{9}$$

Introducing Lagrange multiplier again, one could find the primal problem for which 9 is the dual:

$$\begin{aligned} \min \quad & \int c(x,y)\pi(x,y)dxdy \\ s.t \quad & \int \pi(x,y)dx = \frac{1}{n_y}\sum_j \delta(y-y_j) \\ & \int \pi(x,y)dy = \frac{1}{n_x}\sum_i \delta(x-x_i) \\ & \pi(x,y) \geq 0. \end{aligned} \tag{10}$$

But this is again a Kantorovich optimal transport problem, where the original distributions $\rho(x)$ and $\mu(y)$ have been replaced by discrete distributions with uniform probability among the sample points available. Clearly this is not the problem we would like to pose: the points $\{x_i\}$ and $\{y_j\}$ are random samples from presumably smooth distributions $\rho(x)$ and $\mu(y)$, not their discrete support.

From a data-analysis viewpoint, the difficulty above can be attributed to over-fitting: with only a finite number of sample points, one cannot optimize over an unrestricted, infinite-dimensional space of functions π, φ and ψ, else one over-fits the samples, placing a delta function around each. Instead, one should restrict the space of functions available for optimization. A natural way to do this that preserves the linearity of the problem is to introduce a family of kernel functions and restrict the space of solutions to those of the form

$$\varphi(x) = \sum_{i=1}^{N_x} \beta_i^x G(c_i^x, \alpha_i^x, x)$$

$$\psi(y) = \sum_{j=1}^{N_y} \beta_j^y G(c_j^y, \alpha_j^y, y)$$

and

$$\pi(x,y) = \sum_{k=1}^{N_c} \lambda_k \tilde{G}(\tilde{C}_k, \Sigma_k, x, y),$$

where for concreteness we have adopted as kernel functions $G(c,\alpha,x)$ and $\tilde{G}(\tilde{C},\Sigma,x,y)$, Gaussian distributions with means c and $\tilde{C}$ and covariance matrices α and Σ respectively.

With the solutions restricted in this way and the constraints relaxed accordingly, the modified dual problem (9) becomes

$$\begin{aligned}
\max_{\beta^x,\beta^y} \quad & \sum_{i=1}^{N_x} \beta_i^x \frac{1}{n_x} \sum_{l=1}^{n_x} G(c_i^x, \alpha_i^x, x_l) + \sum_{j=1}^{N_y} \beta_j^y \frac{1}{n_y} \sum_{l=1}^{n_y} G(c_j^y, \alpha_j^y, y_l) \\
s.t \quad & \sum_{i=1}^{N_x} \beta_i^x \iint G(c_i^x, \alpha_i^x, x) \tilde{G}_k(\tilde{C}_k, \Sigma_k, x, y)\, dxdy \\
& + \sum_{j=1}^{N_y} \beta_j^y \iint G(c_j^y, \alpha_j^y, y) \tilde{G}_k(\tilde{C}_k, \Sigma_k, x, y)\, dxdy \\
& \leq \iint c(x,y) \tilde{G}_k(\tilde{C}_k, \Sigma_k, x, y)\, dxdy \qquad \forall k,
\end{aligned} \tag{11}$$

with corresponding primal problem

$$\begin{aligned}
\min_{\lambda_k > 0} \quad & \sum_{k}^{N_c} \lambda_k \iint c(x,y) \tilde{G}(\tilde{C}_k, \Sigma_k, x, y)\, dxdy \\
s.t \quad & \sum_{k}^{N_c} \lambda_k \iint \tilde{G}(\tilde{C}_k, \Sigma_k, x, y) G(c_j^y, \alpha_j^y, y)\, dxdy = \frac{1}{n_y} \sum_{l}^{n_y} G(c_j^y, \alpha_j^y, y_l) \qquad \forall j \\
& \sum_{k}^{N_c} \lambda_k \iint \tilde{G}(\tilde{C}_k, \Sigma_k, x, y) G(c_i^x, \alpha_i^x, x)\, dxdy = \frac{1}{n_x} \sum_{l}^{n_x} G(c_i^x, \alpha_i^x, x_l) \qquad \forall i.
\end{aligned} \tag{12}$$

An advantage of choosing Gaussian kernel functions is that, when the cost function $c(x, y)$ is the squared distance between x and y –the one most frequently used– all the integrals in problem (11) can be found analytically and written in closed form. Then (11) becomes a conventional linear programming problem that can be solved through standard procedures, such as simplex or interior point. The question left therefore is how to choose the kernel functions: restricting them to Gaussians still leave open their number, means and covariance matrices (centers and generalized bandwidths in the language of machine learning.) This choice must be necessarily adaptive, as it depends on the number of sample points available and their distribution.

Since under suitable conditions the solution to Kantorovich's problem (5) is concentrated on a map $y = y(x)$, the choice of the kernel functions $\tilde{G}(\tilde{C}, \Sigma, x, y)$ is particularly challenging, as the distribution they need to resolve has singular support, which is not known a priori but rather emerges as part of the solution. Thus the covariance matrices Σ adopted should be highly anisotropic, but with principal directions that one initially does not know. A natural way to address this is through an iterative procedure that solves a sequence of linear programming problems of the form (11), each with kernel functions determined using the solution to the previous problem in the sequence. This procedure is discussed in detail in the following section.

3 Adaptive algorithm

We solve a sequence of linear programming problem of the form 11. The active constraints in the solution to each of these problems are indicative of the support of $\pi(x,y)$, a fact that we use to update the family of kernels $\tilde{G}(\tilde{C}_k, \Sigma_k, x, y)$, making it progressively more attuned to the underlying solution.

3.1 Initialization

At the onset of the algorithm, the only information at our disposal are the data points, which characterize the distributions $\rho(x)$ and $\mu(y)$ implicitly, while leaving quite open the unknown joint-distribution $\pi(x,y)$. Thus the choice of Kernel functions $G(c_i^x, \alpha_i^x, x)$ and $G(c_j^y, \alpha_j^y, y)$ for the marginal distributions can be based on solid information, while that of the initial Kernels $\tilde{G}(\tilde{C}_k, \Sigma_k, x, y)$ can barely go beyond a rough counting of unknowns and active constraints (hence the need for the iterative procedure.)

1. Initialization of c_i^x and c_j^y.
   In order to find centers that capture the marginal distributions underlying the sample points, we divide the latter into several clusters –numbering N_x and N_y for the $\{x_i\}$ and $\{y_j\}$ respectively– through K-means [8] and adopt the mean of each cluster as a center. This way, each "class" of sample points, presumably captured by an individual cluster, is represented among the Kernel functions. Since the first step is exploratory and cannot provide a very accurate solution to the original problem, we use a small number of clusters, adopting $N_x = N_y = 5$ in the numerical examples below.
2. Initialization of α_i^x and α_j^y
   The α's play the dual role of covariance matrices and bandwidths, and must therefore be estimated as local covariance matrices from the data, but with a degree of locality that allows the various Kernel functions to overlap smoothly. Thus one cannot simply calculate covariance matrices using the data in each cluster separately, since then contiguous Kernel functions would not overlap. Instead, we use all sample points to calculate α_i^x –an identical procedure applies to α_j^y– but weighting the points according to their proximity to x_i :

   $$\alpha_i^x = \sum_k \tilde{w}_{ik}(x_k - c_i^x)(x_k - c_i^x)^T.$$

   With i fixed, rather than computing a weight $\tilde{w}_{ik}$ for each sample point x_k, we save computational cost by computing a weight per cluster K:

   $$\tilde{w}_{ik} = \hat{w}_{iK(k)} = \frac{w_{iK(k)}}{\sum_K w_{iK} \times N_K^x}, \quad w_{iK} = e^{-(\frac{d_{iK}}{d_i})^2}.$$

   Here $K(k)$ denotes the cluster to which the kth sample belongs, N_K^x is the number of points in cluster K, d_{iK} is the distance between c_i^x and c_K^x, and d_i is the smallest distance d_{iK} among all clusters $K \neq i$.

3. Initialization of $\tilde{C}_k$ and Σ_k.
   Since initially one lacks any information on the joint distribution $\pi(x, y)$, one needs to base the choice of initial parameters for the corresponding Kernel functions on the available samples from the marginal distributions $\rho(x)$ and $\mu(y)$. The simplest choice for the covariance matrix Σ is a uniform one, given by
$$\Sigma = \frac{1}{N_x} \begin{pmatrix} \Sigma_x & 0 \\ 0 & \Sigma_y \end{pmatrix}. \tag{13}$$
   Here Σ_x and Σ_y are the empirical covariance matrices of the marginal distributions:
$$\Sigma_x = \frac{1}{n_x} \sum_k (x_k - \bar{x})(x_k - \bar{x})^T, \quad \Sigma_y = \frac{1}{n_y} \sum_k (y_k - \bar{y})(y_k - \bar{y})^T,$$
   where $\bar{x}$ and $\bar{y}$ denote the means of the $\{x_i\}$ and $\{y_j\}$. The choice of the prefactor $\frac{1}{N_x}$ in (13) stems from the fact that $O(N_x)$ constraints will be active in the solution, and one would like the corresponding kernels to have bandwidths that would make them overlap with each other.
   For the kernel centers, a natural choice would be the Cartesian product c_{ij} of the c_i^x and c_j^y. However, this does not provide enough constraints on the problem 11, whose solution is therefore typically unbounded. The reason is that one needs the $N_x + N_y$ constraints that will be typically active to lie in a neighborhood of the unknown support $y(x)$ of the solution $\pi(x, y)$. Yet from a rectangular $N_x \times N_y$ grid of centers, only $O(N_x)$ will lie near $y(x)$. To enlarge the number of constraints, we sample n points from each Gaussian distribution centered at c_{ij} with covariance matrix Σ and use these $N_c = n \times N_x \times N_y$ samples as centers for the Kernel functions. Rather than drawing these samples randomly, we found that it yields better results to use more regularly distributed samples. A simple, semi-random procedure that works well is the following: in order to obtain N^2 samples from a standard multivariate Gaussian distribution, we sample N points from each of N spherical surfaces, where the radii of the spheres are obtained from N equispaced points on a segment via the Box-Muller transformation. The N points from each spherical surface are sampled randomly. Next we transform these samples from a standard Gaussian into samples from the desired distribution with center C and covariance matrix Σ though the linear map
$$x \to Ax + C, \quad \text{where} \quad A = \Sigma^{\frac{1}{2}}.$$

3.2 Update of the Kernel parameters

After the first step, the information available on the marginal distributions does not change –it is always provided by the samples that define the problem– but much can be learned about the joint distribution $\pi(x, y)$: the constraints that are active in the solution to the previous problem in the sequence point to the support of $\pi(x, y)$. This permits a progressive refinement of the choice of Kernel parameters to more accurately represent the solution $\pi(x, y)$.

1. Update of the marginal Kernel parameters c_i^x, c_j^y, α_i^x and α_j^y
The procedure for updating the centers c_i^x and c_j^y and covariances α_i^x and α_j^y is the same as for their initialization, except that, as a more accurate solution is sought, the numbers N_x and N_y of clusters need to be progressively increased. In the numerical examples below, we increase both by three in each step, unless this leads to unbounded solutions, in which case the increase in cluster numbers is made smaller.
2. Update of the centers $\tilde{C}_k$ for the joint distribution
We consider the active constraints from the previous step, i.e. those with nonzero Lagrange multipliers λ_k, and obtain from each n new centers $\tilde{C}_l$ by drawing n samples from the corresponding Kernel $\tilde{G}(\tilde{C}_k, \Sigma_k, x, y)$, using the same semi-random sampling procedure as in the initialization. The choice of n is based on the total number of centers one would like to have, which together with the number of Kernel functions discretizing the marginal distributions accounts for the complexity of the resulting linear programming problem. Our choice in the experiments below has been to leave this complexity nearly constant among iterations, increasing only gradually as one seeks a higher resolution. In this way, a more accurate solution is obtained not because the number of Kernel functions grows along the procedure, but because these become progressively better distributed along the support of $\pi(x, y)$.
3. Update of the Kernel covariance matrices Σ_k for the joint distribution
For the covariance matrix Σ_k, we use a procedure similar to the one used to update the marginal covariance matrices α_i^x. We first group the centers $\tilde{C}_k$ into N_x clusters using K-means. For support center $\tilde{C}_k$, we define

$$\begin{aligned} d_k = \max(\, & \max(d_{ki},\ \text{for } \tilde{C}_i \text{ in the same cluster as } \tilde{C}_k), \\ & \min(d_{kj},\ \text{for } \tilde{C}_j \text{ in clusters different from } \tilde{C}_k)) \end{aligned}$$

where d_{ki} is the Euclidean distance between $\tilde{C}_k$ and $\tilde{C}_i$. Here d_k quantifies the distance between cluster k and its closest neighboring cluster.
Then weights for all centers sampled from $\tilde{C}_l$ are given by:

$$w_{kl} = e^{-\beta\left(\frac{d_{kl}}{d_k}\right)^2}.$$

Then these weights are normalized and the covariance matrix sought is given by the empirical weighted covariance matrix as for the kernels for the marginal distributions. β is an adjustable parameter that depends on the dimensionality of the problem in hand.

4 Numerical experiments

We illustrate this article's procedure through numerical experiments with marginal distributions in spaces of dimensions 1 and 2. Since the procedure provides an estimate of the solution $\pi(x, y)$ to Kolmogorov's formulation of the problem, we need a way to estimate from this coupling the optimal map $y(x)$ of Monge's formulation, for comparison with a known solution in the examples here, and more generally because this map is the answer sought in a number of applications.

A simple estimate for $y(x)$ is the expected value of y under $\pi(x, y)$, conditional on x:

$$y(x) = \int y \frac{\pi(x, y)}{\int \pi(x, y) dy} dy. \tag{14}$$

Notice that, for the Gaussian kernels adopted, all these integrals can be computed in closed form, so the estimated $y(x)$ is an explicit function of the parameters λ_k in the optimal solution.

4.1 1D Gaussian distribution to 1D Gaussian distribution

For a first numerical example, we chose the simplest setting of two one-dimensional Gaussian distributions with means μ_x, μ_y and standard deviations σ_x, σ_y, for which the exact solution to Monge's problem is the linear map

$$y(x) = \mu_y + \frac{\sigma_y}{\sigma_x} (x - \mu_x) .$$

In the example shown, $\mu_x = 0$,$\mu_y = 100$, $\sigma_x = 3$ and $\sigma_y = 2$. The data consists of 10000 samples randomly drawn from each distribution.

Figure 1 displays the evolution of the estimated joint distribution $\pi(x, y)$ through ten steps of the algorithm. We can see how this estimate, which is quite bumpy and coarse in the first step, rapidly becomes smooth and more slender, approximating very accurately the [singular] exact solution.

Figure 2 shows, for the same ten steps, the centers that the algorithm adopts for the kernels for the joint distribution, and which of these become active constraints in the numerical solution. The initial nearly uniform distribution of centers, a reflection of our initial ignorance of the support of $\pi(x, y)$, is rapidly replaced by a set of kernels that adaptably capture the support of the actual solution.

Figure 3 displays the evolution of the estimation (14) of the optimal map $y(x)$, which in this linear case becomes quite indistinguishable from the exact solution far before the corresponding joint distribution develops a truly slender support.

Finally, figure 4 displays the x-marginal of the estimated $\pi(x, y)$ and compares it to the actual Gaussian density $\rho(x)$ underlying the data.

4.2 1D Gaussian distribution to 1D uniform distribution

As a second one-dimensional example, we chose for $\rho(x)$ a Gaussian distribution with mean 0 and standard deviation 3 and for $\mu(y)$ the uniform distribution on the segment $[-10, 10]$, drawing from each again 10000 samples. Even

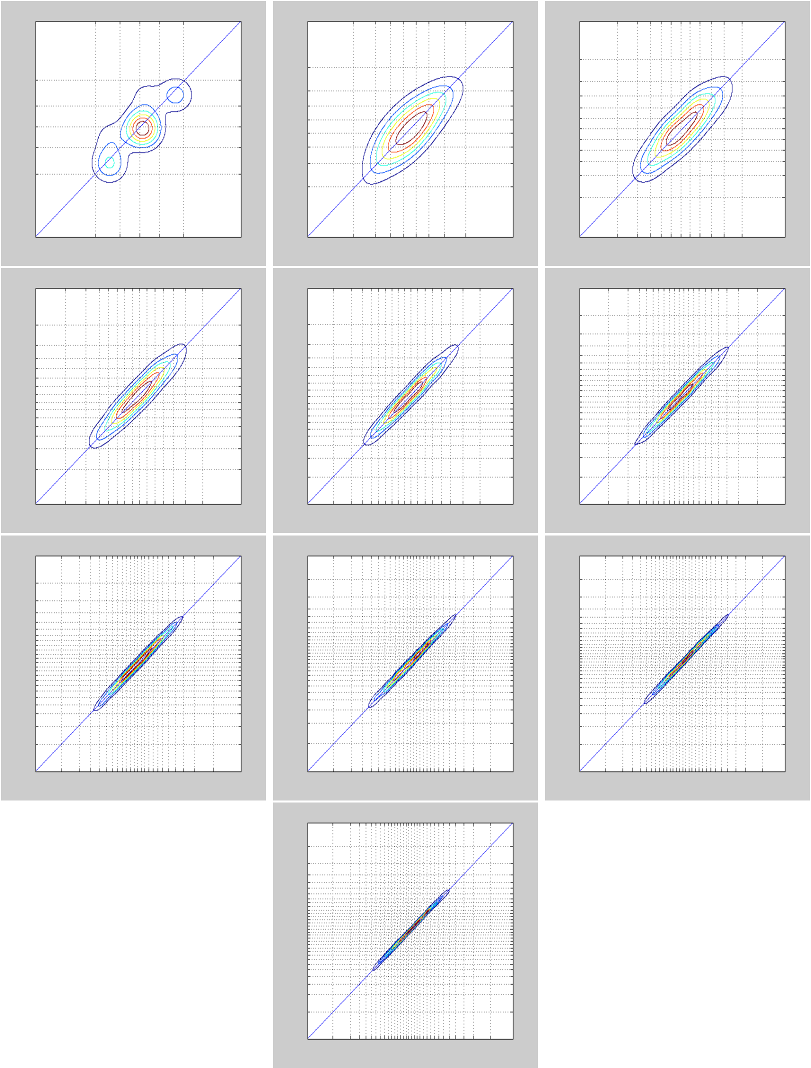

Fig. 1 Evolution of joint distribution $\pi(x, y)$ through 10 steps of the algorithm. The blue line represents the exact solution, and the dotted grid lines correspond to the kernel centers (c_i^x and c_j^y) for the marginal distributions.

though the exact solution this time is nonlinear, the results are equally good, as displayed in figures 5, 6, 7, 8 and 9. It is interesting to observe in figure 9 a Gibbs-like phenomenon associated with the discontinuities of $\mu(y)$ at the two ends of its supporting segment.

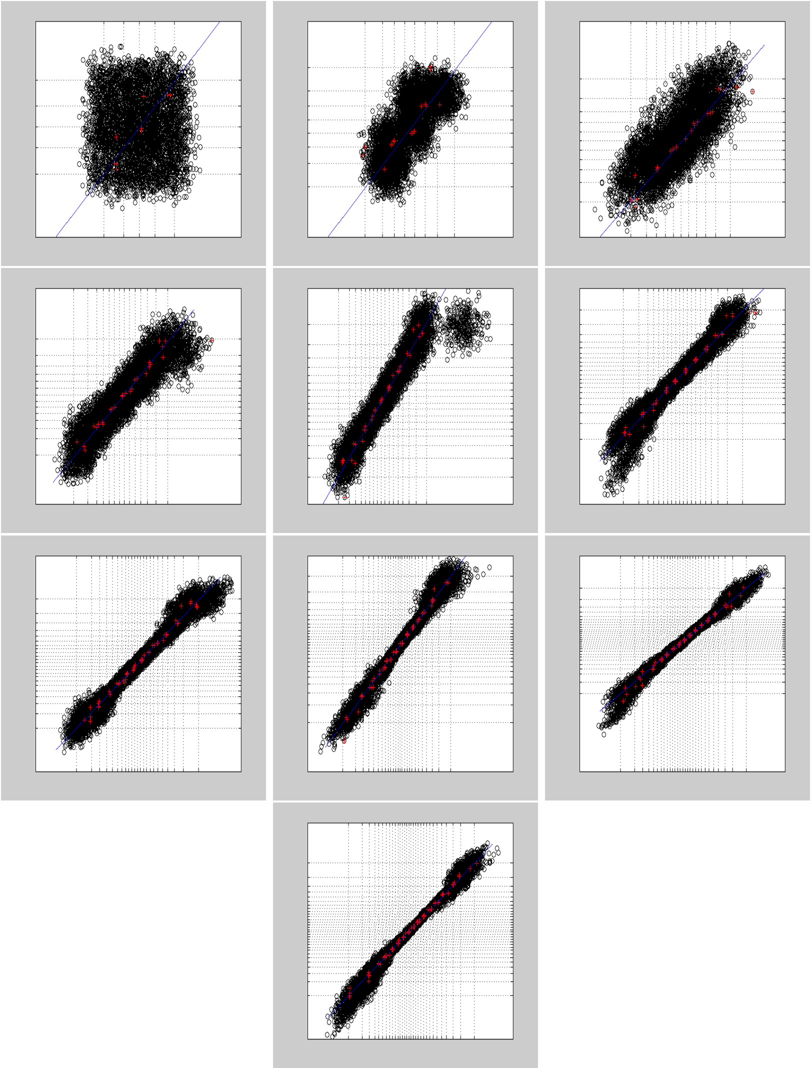

Fig. 2 Evolution of the centers $\tilde{C}_k$ corresponding to the primal solution –or equivalently constraints on the dual problem– through 10 steps of the algorithm. Represented in black circles are the centers for the current step, and in red crosses the ones that become active constraints, revealing the support of the current estimate for $\pi(x, y)$.

4.3 2D Gaussian distribution to 2D Gaussian distribution

For our last example, we chose the optimal coupling between two two-dimensional Gaussian distributions, both with mean $(0, 0)$ and with covariance matrices $\begin{pmatrix} 4 & 1 \\ 1 & 16 \end{pmatrix}$ and $\begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix}$ respectively.

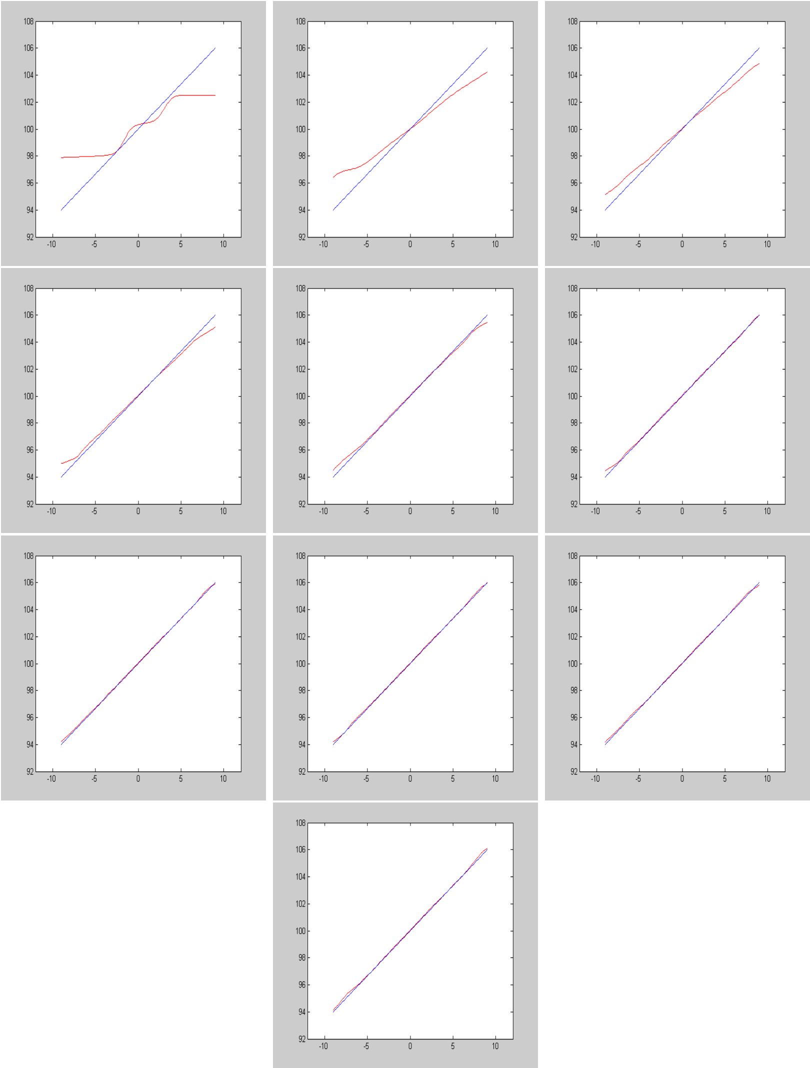

Fig. 3 Evolution of the estimated optimal map (in red) displayed over the underlying exact solution (in blue).

In order to display the four dimensional estimated solution $\pi(x_1, x_2, y_1, y_2)$, we selected six points (x_1, x_2) and displayed in figure 10 the corresponding π's as functions of (y_1, y_2). The exact solution in this case are six delta functions, with locations shown as red dots. As the steps progress, the estimated solution concentrates increasingly sharply around these centers.

To complement this figure, figure 11 displays the evolution of the two two-dimensional marginal distributions of the estimated $\pi(x_1, x_2, y_1, y_2)$ through 12 steps of the algorithm.

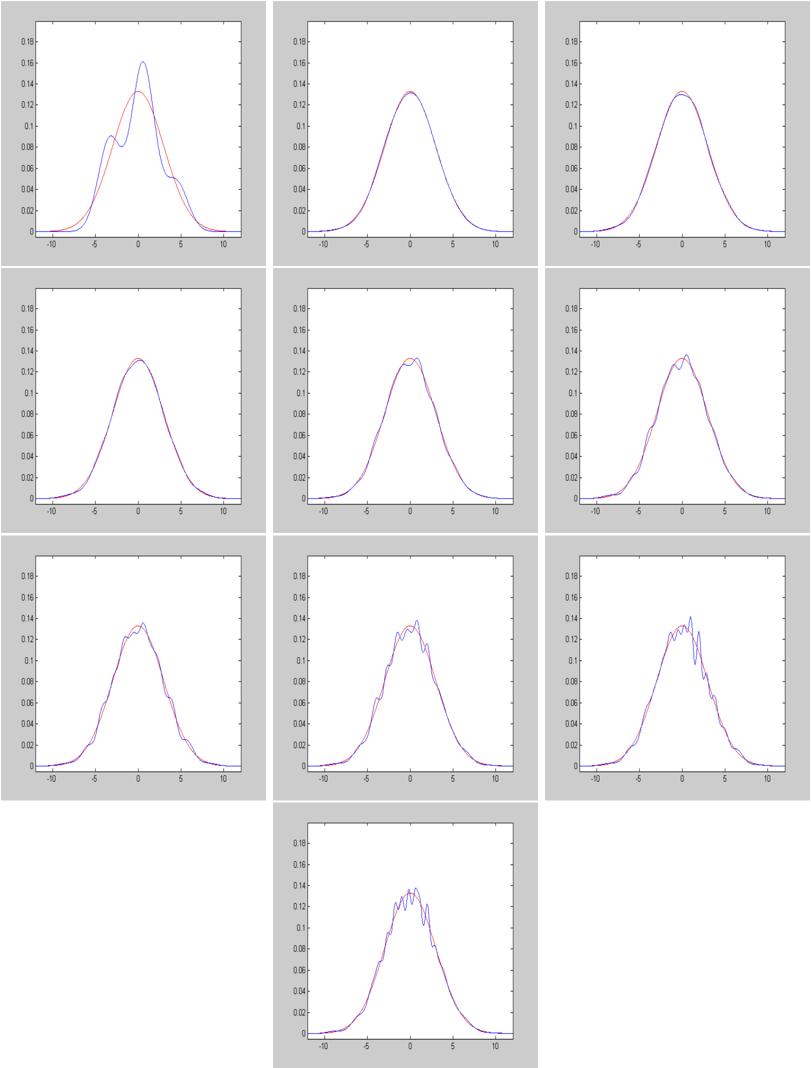

Fig. 4 Evolution of the x-marginal distribution of the estimated $\pi(x, y)$ (in blue) compared with the actual $\rho(x)$ underlying data (in red). The good agreement depends on two factors: having centers and variances for the Kernel functions associated with the x-marginal distribution that sufficiently capture the marginal constraints, and parameters for the Kernel functions associated with the joint distribution $\pi(x, y)$ that do not overfit these constraints.

5 Summary

An adaptive methodology was proposed and developed to solve the data-driven optimal transport problem: to find the coupling $\pi(x, y)$ between two distributions $\rho(x)$, $\mu(y)$ that minimizes the expected value of a pairwise transportation

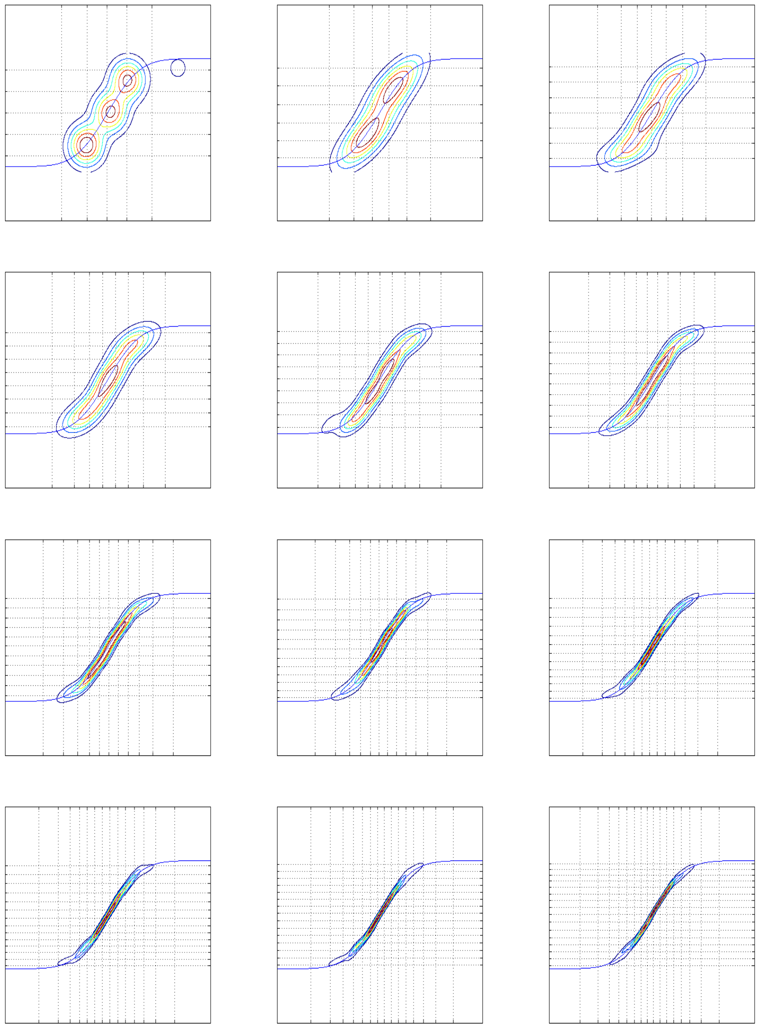

Fig. 5 Contours of the evolution of the estimate for the joint distribution $\pi(x, y)$ through 12 steps of the algorithm, with the exact solution to the underlying problem displayed as a blue line.

cost $c(x, y)$, where the marginal distributions are only known through a finite collection of samples.

The methodology replaces the marginal distributions by their samples through the dual of Kantorovich's formulation of the problem, where empirical means over the samples replace the expected values of the corresponding dual variables. Then a collection of Kernel functions is introduced to turn this problem into a finite linear programming one, and these kernels are evolved through an adaptive algorithm that updates their number and parameters

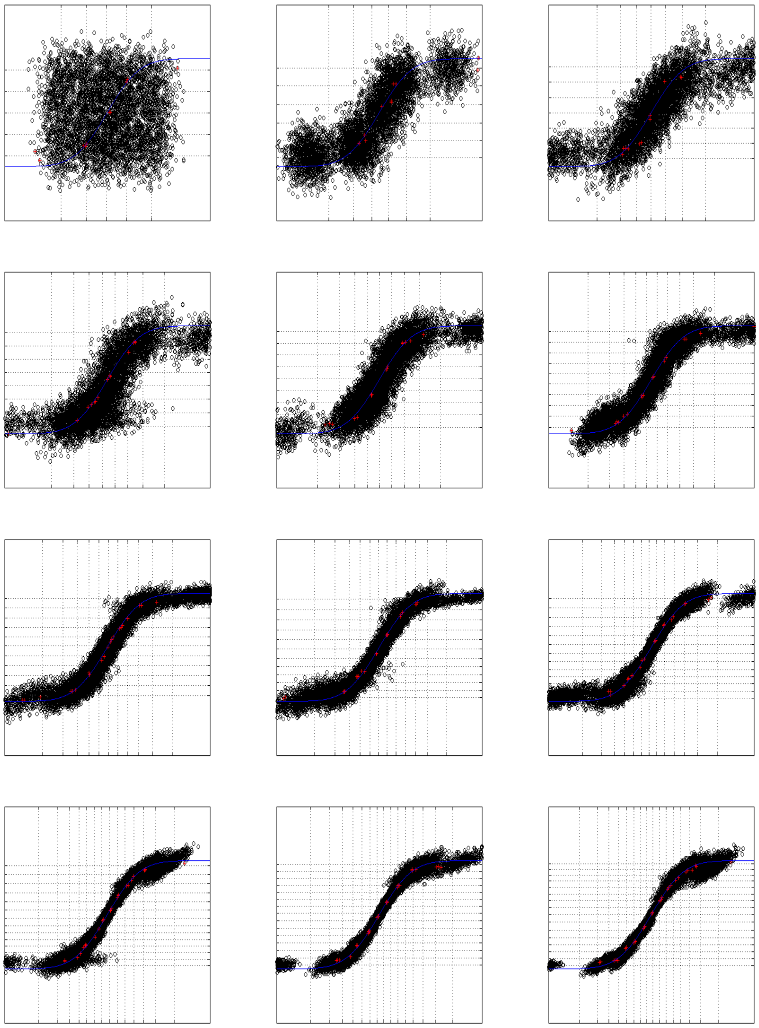

Fig. 6 Evolution of the centers $\tilde{C}_k$ through 12 steps of the algorithm. Represented in black circles are the centers for the current step, and in red crosses the ones that become active constraints, revealing the support of the current estimate for $\pi(x, y)$.

based on the solution to the previous step in a sequence of increasingly refined problems.

For the standard square-distance cost, adopting Gaussian Kernel functions is particularly convenient, as all the integrals defining the sequence of problems and the estimation of the optimal map $y(x)$ from the coupling $\pi(x, y)$ can be computed in closed form.

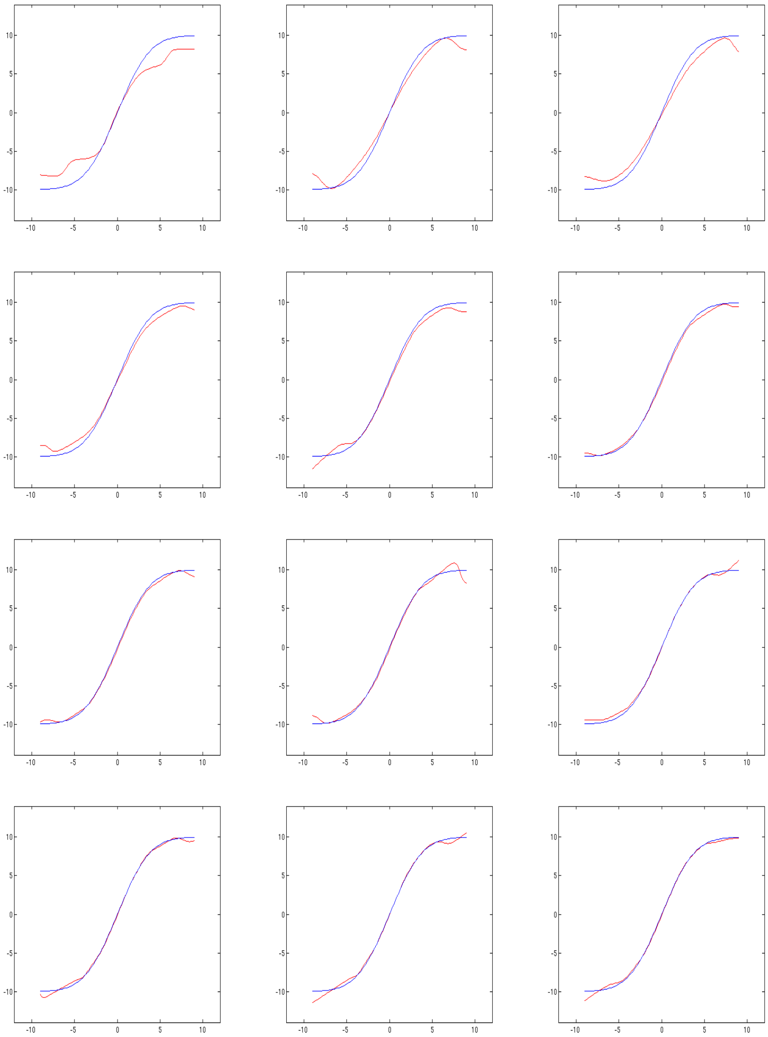

Fig. 7 Evolution of the estimated optimal map $y(x)$ in red, with the exact solution to the underlying problem in blue.

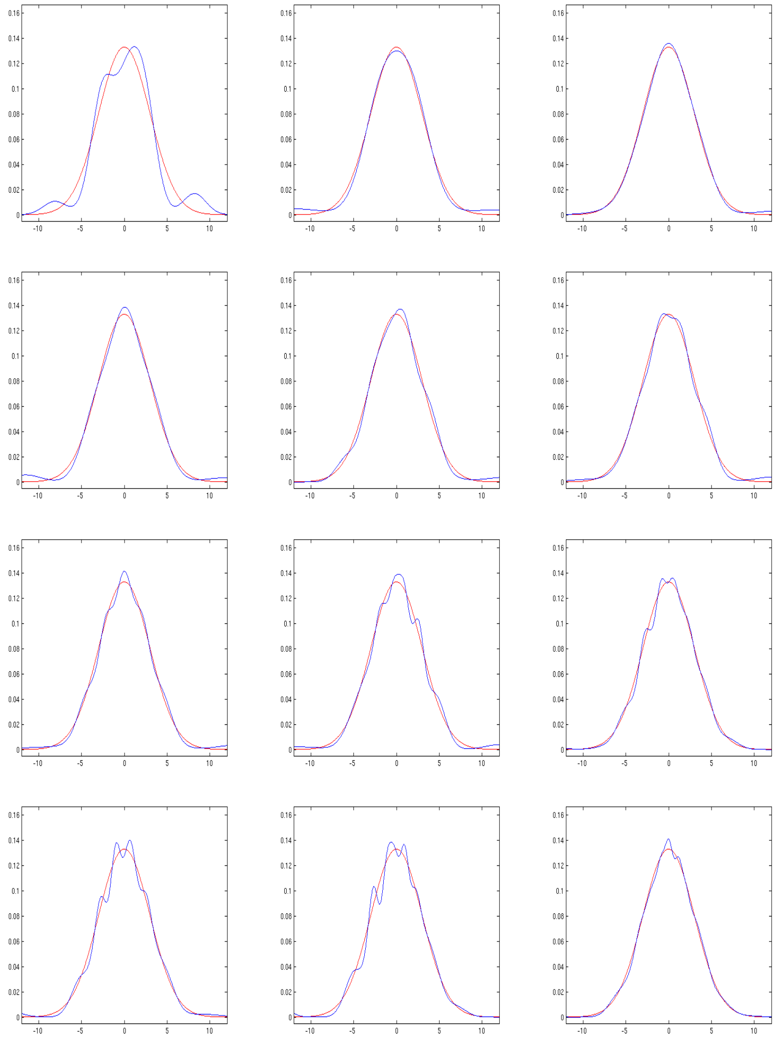

Fig. 8 Evolution of the x-marginal of the estimated distribution $\pi(x, y)$, compared with the Gaussian $\rho(x)$ underlying the data.

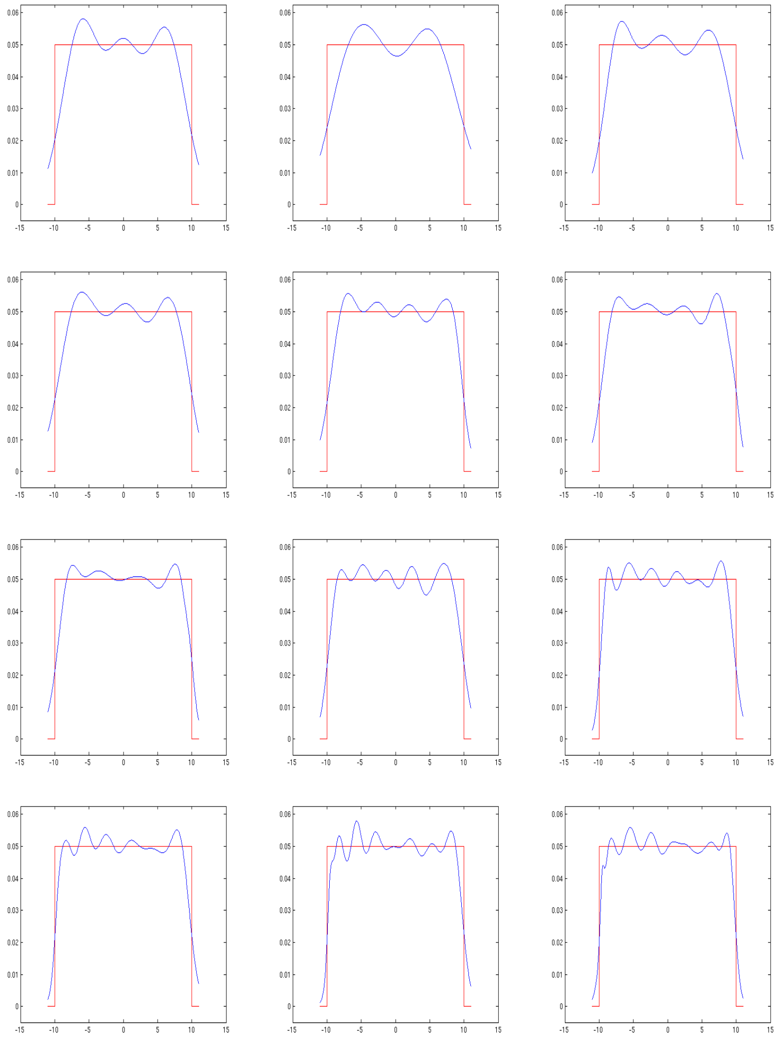

Fig. 9 Evolution of the y-marginal estimated distribution $\pi(x, y)$, compared with the uniform $\mu(y)$ underlying the data.

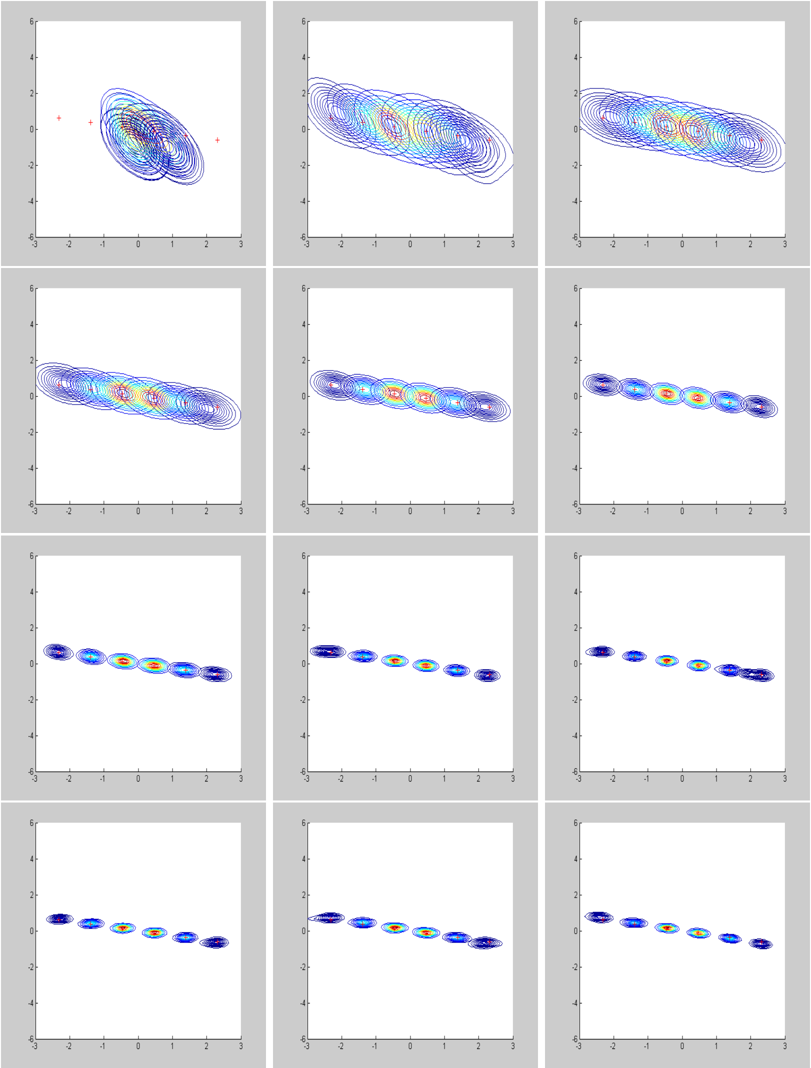

Fig. 10 Evolution of the estimated optimal coupling $\pi(x, y)$ for six specific points y_j. The exact solution is made of six delta functions concentrated on the points displayed in red.

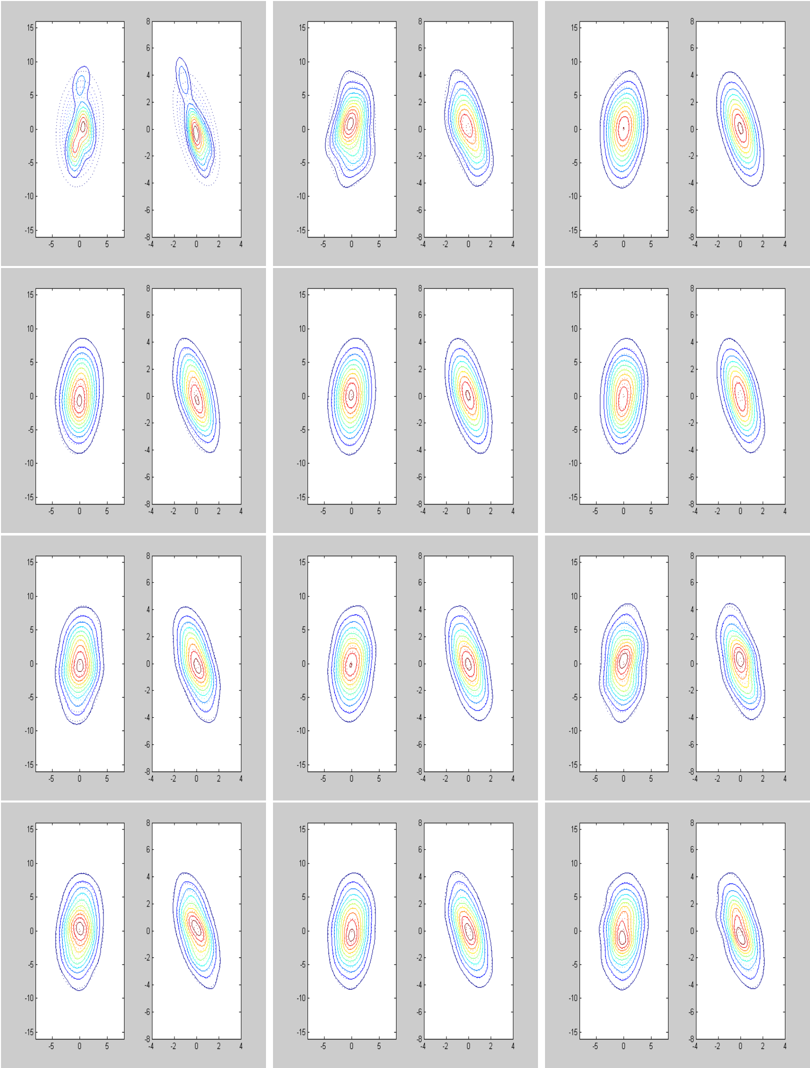

Fig. 11 Evolution through 12 steps of the algorithm of the marginal distributions, left for x and right for y, of the estimated coupling $\pi(x, y)$.

References

1. A.M.Oberman: Wide stencil finite difference schemes for the elliptic Monge-Ampère equation and functions of the eigenvalues of the Hessian. Discrete Contin.Dyn.Syst.Ser.B **10**, 221–238 (2008)
2. C.Villani: Topics in Optimal Transportation. AMS (2003)
3. E.G.Tabak, G.Trigila: Data-Driven Optimal Transport. CPAM (2015)
4. E.J.Dean, R.Glowinski: Numerical methods for fully nonlinear elliptic equations of the Monge-Ampère type. Comput.Methods Appl.Mech.Engrg. **195**, 1344–1386 (2006)
5. Gangbo, W., J.Mccann, R.: The geometry of optimal transportation. Acta Math. **177**, 113–161 (1996)
6. Haber, E., Rehman, T., Tannenbaum, A.: An efficient numerical method for the solution of the L_2 optimal mass transfer problem. Siam J. Sci. Comput. **32**, 197–211 (2010)
7. J.-D.Benamou, Brenier, Y.: A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numer. Math. **84**, 375–393 (2000)
8. J.A.Hartigan, M.A.Wong: A K-Means Clusteing Algorithm. Journal of the Royal Statistical Society, Series C **28**, 100–108 (1979)
9. L.V.Kantorovich: On a problem of Monge. Uspekhi Mat.Nauk **3**, 225–226 (1948)
10. R.Chartrand, K.Vixie, B.Wohlberg, E.Bollt: A gradient descent solution to the Monge-Kantorovich problem. Appl.Math.Sci. **3**, 1071–1080 (2009)
11. S.Angenent, S.Haker, A.Tannenbaum: Minimizing flows for the Monge-Kantorovich problem. SIAM J.Math.Anal. **35**, 61–97 (2003)
12. S.Rachev, L.Rüschendorf: Mass Transportation Problems. Springer-Verlag (1998)