DISCUSSION PAPER SERIES

SISTA: Learning Optimal Transport Costs under Sparsity Constraints

Guillaume Carlier
Arnaud Dupuy
Alfred Galichon
Yifei Sun

DISCUSSION PAPER SERIES

SISTA: Learning Optimal Transport Costs under Sparsity Constraints

Guillaume Carlier
CEREMADE, Université Paris Dauphine and PSL

Arnaud Dupuy
University of Luxembourg and IZA

Alfred Galichon
New York University, Sciences Po and IZA

Yifei Sun
New York University

ABSTRACT

SISTA: Learning Optimal Transport Costs under Sparsity Constraints

In this paper, we describe a novel iterative procedure called SISTA to learn the underlying cost in optimal transport problems. SISTA is a hybrid between two classical methods, coordinate descent ("S"-inkhorn) and proximal gradient descent ("ISTA"). It alternates between a phase of exact minimization over the transport potentials and a phase of proximal gradient descent over the parameters of the transport cost. We prove that this method converges linearly, and we illustrate on simulated examples that it is significantly faster than both coordinate descent and ISTA. We apply it to estimating a model of migration, which predicts the flow of migrants using country-specific characteristics and pairwise measures of dissimilarity between countries. This application demonstrates the effectiveness of machine learning in quantitative social sciences.

Corresponding author:
Arnaud Dupuy
University of Luxembourg
6, rue Richard Coudenhove-Calergi
L-1359 Luxembourg

E-mail: arnaud.dupuy@uni.lu

1 Introduction

Optimal transport has received a great deal of attention recently, see [22], [23], [21], and [11]. The discrete version of the problem in a nutshell is as follows. Consider a $N \times N$ matrix c_{ij} called a *transport cost*, which is the cost of pairing $i \in \{1, ..., N\}$ with $j \in \{1, ..., N\}$, and consider two marginal probability distributions p and q with support $\{1, ..., N\}$. The optimal transport problem is the problem of finding the optimal *transport plan*, the joint probability distribution π_{ij} over $\{1, ..., N\}^2$ which is optimal in the sense that $\sum_{1 \leq i,j \leq N} \pi_{ij} c_{ij}$ is minimal over all the probability π with marginal distributions p and q. A variant of this problem where one seeks to minimize the cost with a regularizing entropy term, namely $\sum_{1 \leq i,j \leq N} \pi_{ij} c_{ij} + T \pi_{ij} \ln \pi_{ij}$, with $T > 0$ being a temperature parameter, has been proposed by various authors and has been found to enjoy attractive computational and statistical properties, see e.g. [16] and references therein.

Most of the literature on optimal transport takes the transport cost matrix c as given and seeks to compute the optimal transport plan π. However, in some situations, one *observes* π and would like to *learn* the cost c. This problem, which belongs to the general class of "inverse problems," see [15], arises in particular in economics, see [12], where one observes optimally matched agents and one would like to infer why they have chosen to match. The problem of learning the transport cost has received relatively little attention in the machine learning literature. Some exceptions are [8], who use a Difference of Convex (DC) functions approach, [14], who use a strategy based on Neighborhood Components Analysis (NCA), and [10], who compute a rank-regularized moment matching estimator of the ground distance using proximal gradient descent.

In this paper, we assume that the transportation cost c_{ij} takes a parametric form c_{ij}^{β}, where the parameter vector β is assumed to be sparse. We introduce an iterative procedure to estimate β by minimizing a convex loss function under an l_1 penalization based on the dual formulation of the optimal transport problem. Our procedure is a hybrid between coordinate descent and proximal gradient descent: indeed, a phase of exact minimization with respect to the transport potentials alternates with a phase of proximal gradient descent (ISTA, see [1]) with respect to the parameters of the transport cost. This procedure is thus a natural extension of the celebrated Sinkhorn algorithm (see an account in [19] and a historical overview in [16]), a.k.a. the Iterative Proportional Fitting Procedure (IPFP), which is a coordinate descent algorithm with respect to the transport potentials, for a fixed value of the parameter vector. One important advantage of the Sinkhorn algorithm compared with alternative methods is that it is fast, parameteter-free and can be naturally parallelized, as documented for example in [7] and [6].

The remainder of the paper is organized as follows. Section 2 presents the optimal transport problem, recalls duality results, and introduces the inverse problem of learning the parameter vector of the transport cost under an l_1 penalization. Section 3 describes the SISTA algorithm, states results on its linear convergence, and

benchmarks its speed of convergence via numerical experiments. Section 4 applies this procedure to a model of migration based on country-specific characteristics and pairwise measures of dissimilarity between countries. Section 5 concludes the paper.

2 Optimal transport with entropic regularization

2.1 Problem formulation

Consider two discrete probability distributions p and q supported by N points[1]. The optimal transport problem between p and q, see [22], is classically defined as

$$(2.1) \qquad \min_{\pi \in \Pi} \sum_{1 \le i,j \le N} \pi_{ij} c_{ij}$$

where c_{ij} is the transport cost from i to j, and Π is the set of probability mass vectors π with margins p and q, that is

$$(2.2) \qquad \Pi = \left\{ \pi_{ij} \ge 0 : \sum_{j=1}^{N} \pi_{ij} = p_i \text{ and } \sum_{i=1}^{N} \pi_{ij} = q_j \right\}.$$

Recently, the literature (see for instance [12], where a microfoundation using a random utility model is given, and [7]) has considered the entropic regularization of (2.1), that is

$$(2.3) \qquad \min_{\pi \in \Pi} \sum_{1 \le i,j \le N} \pi_{ij} c_{ij} + T \pi_{ij} \ln \pi_{ij}$$

where $T > 0$ is a temperature parameter, so that $T \to 0$ recovers the previous object. The dual version of (2.3) is

$$(2.4) \qquad \max_{u,v} \sum_{i=1}^{N} p_i u_i + \sum_{j=1}^{N} q_j v_j - T \sum_{1 \le i,j \le N} \exp\left(\frac{u_i + v_j - c_{ij}}{T} \right).$$

Note that by homogeneity, the solution of problem (2.3) is left invariant by dividing c and T by the same constant. Taking this constant to be T, we can without loss of generality set

$$(2.5) \qquad T = 1,$$

which we will do in the sequel. After taking the negative, we obtain the following dual optimization problem

$$(2.6) \qquad \min_{u,v} \sum_{1 \le i,j \le N} \exp\left(u_i + v_j - c_{ij} \right) - \sum_{i=1}^{N} p_i u_i - \sum_{j=1}^{N} q_j v_j.$$

---

[1] The analysis extends without modification to the case when the support of p and q don't have the same number of points.

By first order conditions, the optimal π in (2.3) and the optimal (u,v) in (2.6) are such that

$$\text{(2.7)} \qquad \pi_{ij} = \exp\left(u_i + v_j - c_{ij}\right),$$

where u and v are adjusted so that the primal constraints are satisfied, i.e. $\pi \in \Pi$ as in (2.2).

The dual problem (2.6) is a convex minimization problem, which can be solved by gradient descent. However, it is well-known that blockwise coordinate descent over u and v iteratively, a procedure called Sinkhorn's algorithm, see [19] and references therein, is a preferable alternative. Given (u^t, v^t), the minimization of problem (2.6) with respect to u_i yields, by first order conditions,

$$\text{(2.8)} \qquad \exp\left(u_i^{t+1}\right) = \frac{p_i}{\sum_{j=1}^{N} \exp\left(v_j^t - c_{ij}\right)},$$

while the minimization of (2.6) with respect to v_j yields

$$\text{(2.9)} \qquad \exp\left(v_j^{t+1}\right) = \frac{q_j}{\sum_{i=1}^{N} \exp\left(u_i^{t+1} - c_{ij}\right)}.$$

Alternating the minimization steps with respect to u_i and v_j therefore yields to alternating between the closed-form updating formulas (2.8) and (2.9).

2.2 Learning the transport cost

We now turn to the inverse problem of recovering the transport cost c_{ij}^{β} based on the observed transport plan $\hat{\pi} \in \Pi$. To achieve this, we assume that the cost function can be represented by a linear combination of basis functions d^k, $k \in \{1, ..., K\}$ as

$$\text{(2.10)} \qquad c_{ij}^{\beta} = \sum_{k=1}^{K} \beta_k d_{ij}^k,$$

where each d_{ij}^k is some measure of dissimilarity between i and j, and $\beta = (\beta_1, \cdots, \beta_K)^\top$ is a parameter vector to be learned.

We present a few choices of dissimilarity measures d_{ij}^k.

(a). If x^i and y^j are vectors of characteristics for i and j, one may set $d_{ij}^k = \left(x_k^i - y_k^j\right)^2$, where the subscript k denotes the k-th component of a vector.

(b). Similarly, if $k = (r,s)$ stands for a pair of indices, one may set $d_{ij}^{rs} = \left(x_r^i - y_s^j\right)^2$ as is done in [9], in order to capture off-diagonal interactions between characteristics. The matrix parameter β_{rs} to be learned measures the interaction between two distinct characteristics.

(c). If, like in our application, i and j are countries, then d_{ij}^k may be the share of inhabitants of country i who do not speak the language of country j. Note that this measure has no reason to be symmetric since $d_{ij}^k \neq d_{ji}^k$ in general.

Our learning procedure is based on looking for β such that $\pi_{ij}^{\beta} = \exp\left(u_i + v_j - c_{ij}^{\beta}\right)$ solves the margin constraints $\pi^{\beta} \in \Pi$, and matches the moments of d^k for $k = 1,...,K$, that is

$$(2.11) \qquad \sum_{i=1}^{N} \pi_{ij}^{\beta} = q_j, \sum_{j=1}^{N} \pi_{ij}^{\beta} = p_i, \sum_{1 \le i,j \le N} \pi_{ij}^{\beta} d_{ij}^{k} = \sum_{1 \le i,j \le N} \hat{\pi}_{ij} d_{ij}^{k}.$$

Following the argument in [12], equations (2.11) can be interpreted as the first order conditions of a maximum likelihood estimation procedure, which can be recast as the following minimization problem

$$(2.12) \qquad \min_{u,v,\beta} F(u,v,\beta),$$

where

$$(2.13) \qquad F(u,v,\beta) := \sum_{1 \le i,j \le N} \exp\left(u_i + v_j - c_{ij}^{\beta}\right) + \sum_{1 \le i,j \le N} \hat{\pi}_{ij}\left(c_{ij}^{\beta} - u_i - v_j\right)$$

is a convex function.

Now if agents match only on a small number of characteristics, because a large number of them are irrelevant in the matching process, then β, the parameter vector, will be sparse. One way to handle this problem is to add the l_1 penalty term in (2.12), namely considering the following problem:

$$(2.14) \qquad \min_{u,v,\beta} \Phi(u,v,\beta) := F(u,v,\beta) + \gamma|\beta|_1,$$

where γ is a regularization parameter. This regularization is similar in spirit to [10], who have introduced a rank regularization technique using the nuclear norm in a problem where β is a matrix. In the next section, we provide the SISTA algorithm for the computation of (2.14).

3 SISTA

3.1 Algorithm

The SISTA algorithm consists of the minimization of $\Phi(u,v,\beta)$ in (2.14) by iterating three alternating phases:

- an exact minimization of with respect to u, holding β and v constant;
- an exact minimization of with respect to v, holding β and u constant;
- a proximal gradient descent step with respect to β, holding u and v constant.

Each of these three phases is straightforward. The exact minimization with respect to u is done in closed-form by the updating formula (2.8) and for v done by (2.9). The proximal gradient descent step with respect to β yields

$$\beta^{t+1} = \text{prox}_{\rho\gamma|\cdot|_1}\left(\beta^t - \rho\nabla_{\beta}F\left(u^{t+1}, v^{t+1}, \beta^t\right)\right),$$

which is explicit and whose components are given by the well-known soft-thresholding formula:

$$\text{prox}_{\rho\gamma|\cdot|}(z) = \begin{cases} z - \rho\gamma & \text{if } z > \rho\gamma \\ 0 & \text{if } |z| \leq \rho\gamma \\ z + \rho\gamma & \text{if } z < -\rho\gamma \end{cases}.$$

Combining these three steps yields SISTA, which we describe in algorithm 1.

Algorithm 1 The SISTA algorithm

Input: Initial guess of parameter vector β^0, of potentials u^0 and v^0, step size ρ, and dissimilarity measures d_{ij}^k

while not converged **do**

(Sinkhorn step). Set $c_{ij}^{\beta^t} := \sum_{k=1}^K \beta_k^t d_{ij}^k$ and update:

$$\begin{cases} \exp\left(u_i^{t+1}\right) = \frac{p_i}{\sum_{j=1}^N \exp\left(v_j^t - c_{ij}^{\beta^t}\right)} \\ \exp\left(v_j^{t+1}\right) = \frac{q_j}{\sum_{i=1}^N \exp\left(u_i^{t+1} - c_{ij}^{\beta^t}\right)} \end{cases}$$

(ISTA step). Let $\pi_{ij}^{\beta^t} := \exp\left(u_i^{t+1} + v_j^{t+1} - c_{ij}^{\beta^t}\right)$. For $k = 1, \ldots, K$,

$$\beta_k^{t+1} = \text{prox}_{\rho\gamma|\cdot|}\left(\beta_k^t - \rho \sum_{1 \leq i,j \leq N}\left(\hat{\pi}_{ij} - \pi_{ij}^{\beta^t}\right)d_{ij}^k\right).$$

end while
Return: β

3.2 Convergence

We introduce two assumptions on the dissimilarity measures d_{ij}^k:

$$(3.1) \qquad \forall k = 1, \ldots, K, \ \sum_{i=1}^N d_{ij}^k = 0, \ \forall j, \ \sum_{j=1}^N d_{ij}^k = 0, \forall i,$$

and

(3.2)   the matrices $\{d^1, \ldots, d^K\}$ are linearly independent.

Note that (3.1) is without loss of generality, see section 3.3. In addition, we assume

$$(3.3) \qquad \hat{\pi}_{ij} > 0, \ \forall (i,j) \in \{1, \ldots, N\}^2.$$

We are now ready to state our main theorem of the paper:

Theorem 3.1. *Assume (3.1)-(3.2)-(3.3). The sequence $x^t := (u^t, v^t, \beta^t)$ generated by the SISTA scheme in algorithm 1 converges to the solution $x^* = (u^*, v^*, \beta^*)$ of*

(2.14) as $t \to +\infty$ provided the step $\rho > 0$ is chosen small enough. Moreover, in this case, there exists $\delta > 0$ such that

$$(3.4) \qquad \Phi(x^t) - \Phi(x^*) \leq \frac{\Phi(x^0) - \Phi(x^*)}{(1+\delta)^t}, \ \forall t \in \mathbb{N}.$$

3.3 Proof of convergence

We present a full proof of Theorem 3.1 in this section.

Preliminaries

Our aim is to solve the convex problem

$$(3.5) \qquad \inf_{(u,v,\beta) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^K} \Phi(u,v,\beta) := G \circ \Lambda(u,v,\beta) + \gamma |\beta|_1,$$

where Λ is the linear map $\mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^K \to \mathscr{M}_N$ defined entrywise by

$$(3.6) \qquad (\Lambda(u,v,\beta))_{ij} := u_i + v_j - c_{ij}^\beta, \ \forall (i,j),$$

and G is the (smooth, strictly convex and separable) function defined by:

$$(3.7) \qquad G(\lambda) := \sum_{1 \leq i,j \leq N} (\exp(\lambda_{ij}) - \hat{\pi}_{ij}\lambda_{ij}), \ \forall \lambda \in \mathscr{M}_N.$$

Since

$$(3.8) \qquad \nabla G(\lambda) = \exp(\lambda) - \hat{\pi}, \ D^2 G(\lambda) = \mathrm{diag}(\exp(\lambda))$$

(where $\exp(\lambda)$ denotes the matrix with entries $\exp(\lambda_{ij})$, and $\mathrm{diag}(\exp(\lambda))$ is the $N^2 \times N^2$ diagonal matrix having entries $\exp(\lambda_{ij})$ on its diagonal), we also have (setting $x = (u,v,\beta)$ to shorten notations):

$$(3.9) \qquad \nabla F(x) = \Lambda^T(\exp \Lambda(x) - \hat{\pi}), \ D^2 F(x) = \Lambda^T \mathrm{diag}(\exp(\Lambda(x)))\Lambda.$$

Properties of F of course strongly rely on properties of Λ. Recall the two assumptions (3.1) and (3.2):

$$\forall k = 1, \ldots, K, \ \sum_{i=1}^N d_{ij}^k = 0, \ \forall j, \ \sum_{j=1}^N d_{ij}^k = 0, \forall i,$$

and

$$\text{the matrices } \{d^1, \ldots, d^K\} \text{ are linearly independent.}$$

Note that (3.1) is not really a restriction. Indeed, if we introduce the new matrices $\widetilde{d}^k$

$$\widetilde{d}_{ij}^k := d_{ij}^k - a_i^k - b_j^k$$

where

$$a_i^k := \frac{1}{N} \sum_{j=1}^N d_{ij}^k, \ b_j^k := \frac{1}{N} \sum_{i=1}^N d_{ij}^k - \frac{1}{N^2} \sum_{1 \leq p,q \leq N} d_{p,q}^k$$

then, obviously $\{\widetilde{d}^1, \dots, \widetilde{d}^K\}$ satisfies the row and columns zero sum conditions from (3.1). Defining

$$\widetilde{c}^\beta := \sum_{k=1}^K \beta_k \widetilde{d}^k, \ \forall \beta \in \mathbb{R}^K$$

as well as the linear map, $\widetilde{\Lambda}$:

$$\widetilde{\Lambda}(\widetilde{u}, \widetilde{v}, \beta)_{ij} := \widetilde{u}_i + \widetilde{v}_j - \widetilde{c}^\beta_{ij},$$

we immediately see that $\Lambda(u, v, \beta) = \widetilde{\Lambda}(u - \sum_{k=1}^K \beta_k a^k, v - \sum_{k=1}^K \beta_k b^k, \beta)$. In other words, there is no loss of generality in replacing the matrices d^k by the matrices $\widetilde{d}^k$ in our minimization problem. Observing finally that for every constant vector m, $\Lambda(u + m, v - m, \beta) = \Lambda(u, v, \beta)$, we can remove this invariance by imposing a normalization conditon on u or v, we will impose the simplest one, namely $u_1 = 0$. Let us then set

(3.10) $E := \{(u, v, \beta) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^K \ : \ u_1 = 0\} \simeq \mathbb{R}^{2N-1+K}.$

Lemma 3.2. Assume (3.1)-(3.2), then Λ is an injective map from E to $\mathscr{M}_N$. In particular Φ is strictly convex on E.

Proof. Let $(u, v, \beta) \in E$ be in the null space of Λ i.e.

(3.11) $u_i + v_j = \sum_{k=1}^K \beta_k d_{ij}^k, \forall i, j.$

Summing over j and using (3.1) give $Nu_i + \sum_{j=1}^N v_j = 0$ and then taking $i = 1$ gives $\sum_{j=1}^N v_j = 0$ hence $u = 0$ and $v = 0$ as well. Therefore (3.11) becomes $\sum_{k=1}^K \beta_k d^k = 0$ so that $\beta = 0$ thanks to (3.2). $\square$

From now on, we will assume that (3.1)-(3.2) hold. Then, thanks to Lemma 3.2 and formulas (3.8)-(3.9), we see that for every $M > 0$ there exist $v = v(M) > 0$ and $\alpha = \alpha(M)$ such that whenever $x := (u, v, \beta)$ and $y := (u', v', \beta')$ are in E and $\max(\|\Lambda(u, v, \beta)\|_\infty, \|\Lambda(u', v', \beta')\|_\infty) \leq M$, one has the ellipticity condition

(3.12) $F(x) \geq F(y) + \nabla F(y)(x - y) + \dfrac{v}{2} \|x - y\|_2^2$

(one can just take $v(M) := e^{-M} \sigma_{\min}$ where $\sigma_{\min} > 0$ is the smallest eigenvalue of $\Lambda^T \Lambda$) as well the Lipschitz estimate

(3.13) $\|\nabla F(x) - \nabla F(y)\|_2 \leq \alpha \|x - y\|_2, \text{ with } \alpha = e^M \|\Lambda\|_2 \|\Lambda^T\|_2,$

where $\|A\|_2$ denotes the 2-operator norm of a matrix A.

Our last assumption (3.3) is

$$\hat{\pi}_{ij} > 0, \ \forall (i, j) \in \{1, \dots, N\}^2.$$

Rewriting our initial optimization problem with the normalization constraint $u_1 = 0$ as

$$(3.14) \qquad \inf_{(u,v,\beta)\in\mathbb{R}^{2N-1+K}} \Phi(u,v,\beta) = G(\Lambda(u,v,\beta)) + \gamma|\beta|_1$$

we then have:

Proposition 3.3. Under assumptions (3.1)-(3.2)-(3.3), we have:

- Φ is coercive on $\mathbb{R}^{2N-1+K}$, i.e. its sublevels are bounded,
- problem (3.14) admits a unique solution $x^* := (u^*, v^*, \beta^*)$
- $x^* := (u^*, v^*, \beta^*)$ is characterized by the optimality conditions

$$(3.15) \qquad \nabla_u F(x^*) = 0, \ \nabla_v F(x^*) = 0, \ -\nabla_\beta F(x^*) \in \gamma\partial|.|_1(\beta^*).$$

Proof. Observe that whenever $p > 0$ one has $e^t - pt \geq p - p\log(p)$ for every $t \in \mathbb{R}$ (Young's inequality). Obviously, if $\lambda < 0$ we have $e^\lambda - \hat{\pi}_{ij}\lambda \geq \hat{\pi}_{ij}|\lambda|$. If $\lambda \geq 0$, $e^\lambda - \hat{\pi}_{ij}\lambda = \hat{\pi}_{ij}|\lambda| + e^\lambda - 2\hat{\pi}_{ij}\lambda$, using Young's inequality and the fact that $\hat{\pi}_{ij} \in [0,1]$, we get

$$e^\lambda - \hat{\pi}_{ij}\lambda \geq \hat{\pi}_{ij}|\lambda| - 2\log(2).$$

Hence

$$(3.16) \qquad \Phi(u,v,\beta) \geq \sum_{1\leq i,j\leq N} \hat{\pi}_{ij}|\Lambda(u,v,\beta)_{ij}| - 2N^2\log(2) + \gamma|\beta|_1.$$

Thanks to (3.3) and the injectivity of Λ, the coercivity of Φ follows. Since Φ is continuous, there therefore exists a solution for (3.14) which is unique by strict convexity. Since Φ is the sum of the smooth term F and the l_1 norm of the last component, it is easy to see that the optimality condition $0 \in \partial\Phi(u^*, v^*, \beta^*)$ is exactly the system (3.15). $\qquad\square$

Proof of convergence

Starting from $(u^0, v^0, \beta^0) \in \mathbb{R}^{2N-1+K}$, inductively define the sequence (u^t, v^t, β^t) via:

$$(3.17) \qquad u^{t+1} = \operatorname{argmin} F(.,v^t,\beta^t), \ v^{t+1} = \operatorname{argmin} F(u^{t+1},.,\beta^t).$$

Note that these coordinate descent updates are explicit:

$$\exp(u_i^{t+1}) = \frac{p_i}{\sum_{j=1}^N \exp(v_j^t - c_{ij}^{\beta^t})}, \exp(v_j^{t+1}) = \frac{q_j}{\sum_{i=1}^N \exp(u_i^{t+1} - c_{ij}^{\beta^t})}.$$

We now consider the SISTA hybrid method where the updates for β are given by the ISTA algorithm (see [1]) with a constant step $\rho > 0$:

$$(3.18) \qquad \beta^{t+1} = \operatorname{prox}_{\rho\gamma|.|_1}\left(\beta^t - \rho\nabla_\beta F(u^{t+1}, v^{t+1}, \beta^t)\right)$$

i.e. β^{t+1} is obtained by minimizing with respect to $\beta \in \mathbb{R}^K$

$$(3.19) \qquad \beta \mapsto \rho\gamma|\beta|_1 + \frac{1}{2}\|\beta - (\beta^t - \rho\nabla_\beta F(u^{t+1}, v^{t+1}, \beta^t))\|_2^2.$$

Note that the solution of (3.19) is explicit and given by a well-known thresholding formula. So all the steps for algorithm (3.17)-(3.18) are totally explicit. The step size ρ has to be chosen in an appropriate way to ensure convergence. Setting $C = \Phi(u^0, v^0, \beta^0)$, thanks to (3.16) we have

$$\Phi(u, v, \beta) \leq C \Rightarrow |\beta|_1 \leq R := \frac{C + 2N^2 \log(2)}{\gamma}.$$

Defining $A := C + 2N^2 \log(2) + 1$ we thus have

$$(3.20) \qquad\qquad A > \gamma R, \text{ and } F \leq A \text{ whenever } \Phi \leq C.$$

Now let $\theta \in C^\infty(\mathbb{R}, \mathbb{R})$ be a nondecreasing function such that
(3.21)
$$\theta(t) = t \text{ when } t \leq A, \ \theta(t) \geq t \text{ when } t \in [A, 2A] \text{ and } \theta(t) = 2A \text{ when } t \geq 2A.$$

The function $\widetilde{F} := \theta \circ F$ is nonconvex but its gradient is globally Lipschitz. Let $\alpha > 0$ be a Lipschitz constant of $\nabla \widetilde{F}$ (and of ∇F on $F \leq A$), and now take our step ρ so as to satisfy

$$(3.22) \qquad\qquad 0 < \rho \leq \frac{1}{\alpha}.$$

Note that since $\nabla \widetilde{F}$ is α-Lipschitz, we have

$$(3.23) \qquad \widetilde{F}(x) \leq \widetilde{F}(y) + \nabla \widetilde{F}(y) \cdot (x - y) + \frac{\alpha}{2} \|x - y\|_2^2, \ \forall (x, y) \in E^2.$$

Let us now show inductively that for ρ satisfying (3.22), the iterates (3.17)-(3.18) remain in the sublevel set $\Phi \leq C$. Of course, if $\Phi(u^t, v^t, \beta^t) \leq C$ then by (3.17), $C \geq \Phi(u^{t+1}, v^t, \beta^t) \geq \Phi(u^{t+1}, v^{t+1}, \beta^t)$. Proving that $\Phi(u^{t+1}, v^{t+1}, \beta^{t+1}) \leq C$ requires a little more work. First note that

$$F(u^{t+1}, v^{t+1}, \beta^t) = \widetilde{F}(u^{t+1}, v^{t+1}, \beta^t) \text{ and } \nabla F(u^{t+1}, v^{t+1}, \beta^t) = \nabla \widetilde{F}(u^{t+1}, v^{t+1}, \beta^t).$$

Now, it follows from (3.18) that $\beta^t - \beta^{t+1} - \rho \nabla_\beta F(u^{t+1}, v^{t+1}, \beta^t) \in \rho \gamma \partial(|.|_1)(\beta^{t+1})$, hence

$$(3.24) \quad \gamma|\beta^t|_1 \geq \gamma|\beta^{t+1}|_1 + \frac{1}{\rho}\|\beta^{t+1} - \beta^t\|_2^2 - \nabla_\beta F(u^{t+1}, v^{t+1}, \beta^t) \cdot (\beta^t - \beta^{t+1})$$

but thanks to (3.23) we also have

$$F(u^{t+1}, v^{t+1}, \beta^t) \geq \widetilde{F}(u^{t+1}, v^{t+1}, \beta^{t+1})$$
$$+ \nabla_\beta F(u^{t+1}, v^{t+1}, \beta^t) \cdot (\beta^t - \beta^{t+1}) - \frac{\alpha}{2}\|\beta^{t+1} - \beta^t\|_2^2.$$

Summing this inequality with (3.24) and using (3.22), we thus get

$$(3.25) \qquad \Phi(u^{t+1}, v^{t+1}, \beta^t) \geq \widetilde{F}(u^{t+1}, v^{t+1}, \beta^{t+1}) + \gamma|\beta^{t+1}|_1 + \frac{\alpha}{2}\|\beta^{t+1} - \beta^t\|_2^2.$$

If $F(u^{t+1}, v^{t+1}, \beta^{t+1}) \leq 2A$ then by (3.21) we have $\widetilde{F}(u^{t+1}, v^{t+1}, \beta^{t+1}) \geq F(u^{t+1}, v^{t+1}, \beta^{t+1})$ and therefore $\Phi(u^{t+1}, v^{t+1}, \beta^{t+1}) \leq C$. If, on the contrary, $F(u^{t+1}, v^{t+1}, \beta^{t+1}) > 2A$

then (3.25) would imply that $2A \leq F(u^{t+1}, v^{t+1}, \beta^t) + \gamma|\beta^t|_1 \leq A + \gamma R$ which contradicts our choice of A in (3.20). This shows that

$$(3.26) \qquad \Phi(u^{t+1}, v^{t+1}, \beta^t) \geq \Phi(u^{t+1}, v^{t+1}, \beta^{t+1}) + \frac{\alpha}{2}\|\beta^{t+1} - \beta^t\|_2^2$$

and in particular this proves the desired conclusion that $\Phi(u^{t+1}, v^{t+1}, \beta^{t+1}) \leq C$. The iterates of SISTA therefore remain bounded and the value of Φ along these iterates converges monotonically. In particular, these iterates remain in a ball where both the uniform ellipticity condition (3.12) and the Lipschitz estimate (3.13) on ∇F are satisfied. Using (3.12) together with $\nabla_u F(u^{t+1}, v^t, \beta^t) = 0$, $\nabla_v F(u^{t+1}, v^{t+1}, \beta^t) = 0$, we get

$$F(u^t, v^t, \beta^t) - F(u^{t+1}, v^t, \beta^t) \geq \frac{v}{2}\|u^{t+1} - u^t\|_2^2$$

$$F(u^{t+1}, v^t, \beta^t) - F(u^{t+1}, v^{t+1}, \beta^t) \geq \frac{v}{2}\|v^{t+1} - v^t\|_2^2.$$

Summing and using $F(u^t, v^t, \beta^t) - F(u^{t+1}, v^{t+1}, \beta^t) = \Phi(u^t, v^t, \beta^t) - \Phi(u^{t+1}, v^{t+1}, \beta^t)$, we deduce

$$(3.27) \qquad \Phi(u^t, v^t, \beta^t) - \Phi(u^{t+1}, v^{t+1}, \beta^t) \geq \frac{v}{2}(\|u^{t+1} - u^t\|_2^2 + \|v^{t+1} - v^t\|_2^2).$$

Together with (3.26), this gives

$$\Phi(u^t, v^t, \beta^t) - \Phi(u^{t+1}, v^{t+1}, \beta^{t+1}) \geq \frac{v}{2}(\|u^{t+1} - u^t\|_2^2 + \|v^{t+1} - v^t\|_2^2) + \frac{\alpha}{2}\|\beta^{t+1} - \beta^t\|_2^2.$$

Setting $A_t := \Phi(x^t) - \Phi(x^*)$, using that $\alpha \geq v$, we thus have:

$$(3.28) \qquad A_{t-1} - A_t \geq \frac{v}{2}\|x^t - x^{t-1}\|_2^2, \ \forall t \geq 1.$$

Let us now bound A_t from above. By construction of β^t by proximal gradient descent, we have

$$q_t := \frac{\beta^{t-1} - \beta^t}{\rho} - \nabla_\beta F(u^t, v^t, \beta^{t-1}) \in \gamma\partial(|.|_1)(\beta^t).$$

Using (3.12) and the fact that $q_t \in \gamma\partial(|.|_1)(\beta^t)$, we thus obtain

$$\Phi(x^*) \geq \Phi(x^t) + \nabla_u F(x^t) \cdot (u^* - u^t) + \frac{v}{2}\|u^t - u^*\|_2^2$$

$$+ \nabla_v F(x^t) \cdot (v^* - v^t) + \frac{v}{2}\|v^t - v^*\|_2^2$$

$$+ (\nabla_\beta F(x^t) + q_t) \cdot (\beta^* - \beta^t) + \frac{v}{2}\|\beta^t - \beta^*\|_2^2.$$

Using Young's inequality: $q \cdot z + \frac{v}{2}\|z\|_2^2 \geq -\frac{1}{2v}\|q\|_2^2$, this yields

$$(3.29) \qquad A_t \leq \frac{1}{2v}\left(\|\nabla_u F(x^t)\|_2^2 + \|\nabla_v F(x^t)\|_2^2 + \|\nabla_\beta F(x^t) + q_t\|_2^2\right).$$

Now since $\nabla_u F(u^t, v^{t-1}, \beta^{t-1}) = 0$ and $\nabla_v F(u^t, v^t, \beta^{t-1}) = 0$, thanks to (3.13), we have

$$\|\nabla_u F(x^t)\|_2^2 + \|\nabla_v F(x^t)\|_2^2 \leq 2\alpha^2 \|x^t - x^{t-1}\|_2^2.$$

Thanks to (3.13), the monotonicity of $\nabla_\beta F(u^t, v^t, .)$ and the definition of q_t, we have

$$\begin{aligned}
\|\nabla_\beta F(x^t) + q_t\|_2^2 &= \left\| \frac{\beta^{t-1} - \beta^t}{\rho} + \nabla_\beta F(x^t) - \nabla_\beta F(u^t, v^t, \beta^{t-1}) \right\|_2^2 \\
&\leq (\rho^{-2} + \alpha^2) \|\beta^{t-1} - \beta^t\|_2^2 \\
&\quad + \frac{2}{\rho} (\beta^{t-1} - \beta^t) \cdot (\nabla_\beta F(x^t) - \nabla_\beta F(u^t, v^t, \beta^{t-1})) \\
&\leq (\rho^{-2} + \alpha^2) \|\beta^{t-1} - \beta^t\|_2^2.
\end{aligned}$$

Using these inequalities in (3.29), we deduce that

$$(3.30) \qquad A_t \leq \frac{3\alpha^2 + \rho^{-2}}{2\nu} \|x^t - x^{t-1}\|_2^2,$$

which, combined with (3.28), gives the desired linear convergence rate:

$$A_t \leq \frac{A_{t-1}}{1+\delta} \text{ with } \delta = \frac{\nu^2 \rho^2}{3\alpha^2 \rho^2 + 1}.$$

Remark 3.4. Since $\|x^t - x^{t+1}\|_2 \leq \sqrt{\frac{2(\Phi(x^t) - \Phi(x^*))}{\nu}}$, linear convergence of $\Phi(x^t) - \Phi(x^*)$ together with the triangle inequality straightforwardly gives $\|x^t - x^*\|_2 = O((1+\delta)^{-\frac{t}{2}})$.

Discussion

We point out a few remarks that are useful in applying SISTA in practice.

- It is clear from the proof above that using varying steps ρ_t that are bounded away from 0 and satisfy (3.22) leads to the same convergence results, which also allows in practice to use line search methods.
- Assumption (3.3) may look restrictive in some applications. However, all of our analysis can be generalized to cases where some of the $\hat{\pi}_{ij}$ are zero. In those cases, setting $I^+ := \{(i, j) : \hat{\pi}_{ij} > 0\}$, F should just be replaced by the corresponding sum over I^+. In the context of migration we study in our application, for example, the number of migrants from country i to country j is only defined if $i \neq j$.
- Our method also applies to the case where one imposes sign constraints on some of the β_k's. Indeed, the corresponding proximal operator is explicit as well. More generally, SISTA can be used for any *simple* (i.e. having a closed-form prox) penalization, for instance, the group lasso.
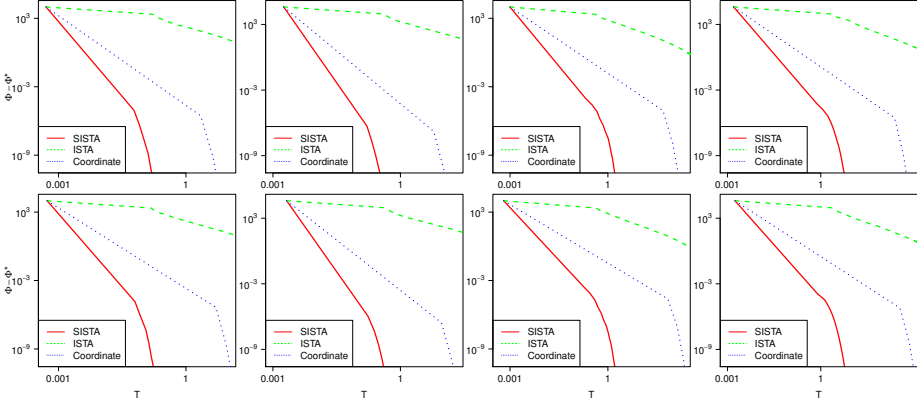
(A) $K = 100, N = 100$ (B) $K = 100, N = 200$ (C) $K = 500, N = 100$ (D) $K = 500, N = 200$

FIGURE 3.1. Comparison of SISTA, coordinate descent, and ISTA for the specified problem sizes. The top row has sparsity level $= 0.05$ and the bottom row has sparsity level $= 0.1$.

3.4 Numerical simulations

We demonstrate the fast convergence of SISTA in practice on simulated examples varying K and N, the dimensionality of β and size of the margins of π, respectively. We compare against two other algorithms: ISTA and coordinate descent. In ISTA, we perform gradient descent on u, v (as the objective function is smooth for these parameters), and ISTA on β, which is the only part of the variables subject to a nonsmooth penalization. For coordinate descent, we apply Sinkhorn minimization to u, v and univariate coordinate descent (using the bisection method) to each component of β.

We generate each d_{ij}^k from an i.i.d. standard normal distribution. We additionally draw each $\hat{\pi}_{ij}$ from an i.i.d. standard log normal distribution. For each problem size, we choose γ such that the sparsity level, defined as the number of non-zero β components divided by K, is at 0.05 or 0.1. We run SISTA on the simulated data with high precision to obtain (u^*, v^*, β^*) and then plot $|\Phi(u^t, v^t, \beta^t) - \Phi(u^*, v^*, \beta^*)|$ against $T(t)$, the computation time at iteration t in seconds, in a log-log plot, as illustrated in figure 3.1.

For comparability, the three methods are run using the same initial estimate of the variables, in practice by setting them all to be zero. We can see from figure 3.1 that SISTA runs an order of magnitude faster than the two pure methods it is a hybrid of. The ISTA method has two major drawbacks compared to the other two. First, a proper initialization of u, v is crucial to ensure a quick convergence of gradient based approaches. Second, a step size has to be set for u, v in addition to β, which adds extra complexities to the algorithm as they do not necessarily have the same

scale as β. The limitations of coordinate descent are also straightforward to see. As the dimensionality of β increases, which is common in practical applications, the cost of coordinate descent increases dramatically, since the algorithm needs to evaluate the gradient after each component's update, which is time-intensive.

4 Application to migration flows

4.1 Literature

Our goal in this section is to predict migration flows between countries. In this application, $\hat{\pi}_{ij}$ is the probability that a migrant drawn from the overall population of migrants has origin country i and destination country j.

The recent compilation of country-to-country migration data (see [18] for instance) has initiated a fast growing literature in economics whose aim is to predict bilateral migration flows using measures of dissimilarity between an origin and a destination country (see [3] for a recent critical review). The dissimilarity measures are built to capture various dimensions of attractiveness of a destination country for migrants from a given origin country, which can be classified into two broad classes: those that are constructed using country-specific characteristics such as Examples (a) and (b) in section 2.2 and pairwise measures such as Example (c).

The current literature has only considered a limited number of dissimilarity measures so far – a minimum of five in [2] and up to a maximum of eight in [20]. The considered ones generally include geographic distances, economic gaps, differences in immigration policy, and cultural differences (see [2, 4, 5, 13, 20]). The parameters are estimated using classical non-linear panel data techniques, where the dependent variable is the logarithm of the bilateral migration flow, and the dissimilarity measures act as independent variables. Countries of origin and destination fixed-effects are usually included in the regression to account for the potential bias due to the fact that the attractiveness of a destination country is relative to that of alternative destination countries, the so-called "multilateral resistance."

However, the relatively small number of dissimilarity measures studied in this literature is in sharp contrast with the fact that countries differ along many dimensions. A recent effort from various organizations including the World Bank Group, the Centre d'Études Prospectives et d'Informations Internationales (CEPII), Freedom House, has greatly increased the range of measures available for countries. Combining these various sources of data enables one to distinguish more than 200 countries along over 100 dimensions, significantly augmenting the size of dissimilarity measures from five or eight.

While this offers scope for improvement in the prediction of bilateral migration flows, it also raises the questions of how to select the most relevant measures of dissimilarity and how to estimate their corresponding parameters. The methodology outlined in the previous section contributes to this literature by providing a solution to them.

4.2 Data

Our application requires access to a working dataset containing information about bilateral migration flows. Each unit in our dataset is an origin-destination country pair. We compile the dataset using five different sources, which gives us over 40,000 observations of bilateral migration flows and more than 100 measures of dissimilarity, including nine pairwise measures.

We use data about bilateral migration flows between 1990 and 2000. These data are obtained in two steps. First, we obtain data about migration stocks in 1990 and 2000 from the dataset compiled by [18], which were collected using census records from 226 countries of origin and destination. Second, we follow the literature (e.g. [4]) and calculate bilateral migration flows between 1990 and 2000 as the change in bilateral stocks of migrants between the two decades.

[4] argues that the existence of a social network at destination, defined as the stock of migrants from the same origin, explains the size and the patterns of international migration flows. We follow [4] and use the stock of migrants from origin country i in destination country j in 1990 as a proxy measure of the network of migrants from i in j in 2000.

We obtain the second set of dissimilarity measures using the data compiled by CEPII and documented in [17]. These data cover 223 countries and contain eight pairwise measures of dissimilarity, including the geographic distance between two countries, whether they are contiguous, and whether they have had a colonial link after 1945.

We incorporate three additional sources to append information about country-specific characteristics covering a vast range of topics.

From the World Development Indicators (WDI), compiled by the World Bank, we obtain 73 variables containing information about economic development (GDP, income, unemployment, inflation, human capital), demography, health (life expectancy, access to improved sanitation facilities), and environment (climate, pollution) for 217 countries.

From the Worldwide Governance Indicators (WGI), also compiled by the World Bank, we obtain six variables about the quality of governance: voice and accountability, political stability and absence of violence, government effectiveness, regulatory quality, rule of law, and control of corruption for 213 countries.

Finally, from Freedom in the World Indicators, compiled by the Freedom House organization, we obtain another 15 variables providing information about political rights and civil liberties for 210 countries.

Combining all these variables, we end up with a vector of 94 characteristics for each country. We then construct 94 measures of dissimilarity as in Example (a) in section 2.2.

4.3 Results

We apply SISTA to our working dataset complied in the previous section. Since the current literature uses between five and eight dissimilarity measures, we perform two sets of experiments, where we conduct a grid search of γ so that five and eight components are non-zero in the learned β.

TABLE 4.1. Learned β with (top) five (bottom) eight non-zero components.

Contiguity	Common language	Colonial link after WWII	(log) Geo. distance	Network	Improved sanitation facilities, urban Ori.*Dest.	Area Ori.*Dest.	Life expect. female Ori.*Dest.
1.797 (0.057)	0	1.224 (0.194)	-1.540 (0.020)	1.309 (0.178)	0	-0.630 (0.043)	0
2.438 (0.058)	4.412 (0.108)	2.118 (0.431)	-1.992 (0.048)	1.633 (0.211)	2.801 (0.318)	-0.927 (0.062)	4.202 (0.535)

Note: Standard errors, calculated using 1,000 bootstrapped samples, are in parentheses.

Table 4.1 shows the learned β from the two experiments. The top row indicates that if only five measures of dissimilarity are allowed in the prediction of migration flows, three of the selected ones are used in most papers in this literature: whether two countries are contiguous, logarithm of geographic distance between them, and whether they have had a colonial link after World War II. The fourth one is the network measure introduced in [2]. Moreover, the signs of β_k's corresponding to these four measures match precisely with previous results. Interestingly, the fifth selected one has never been used. It corresponds to the interaction between the areas of the origin and destination countries. The negative sign of the parameter indicates that migration flows are larger between small (vast) origin countries and vast (resp. small) destination countries compared to origin-destination countries with similar areas.

The second row indicates that if only eight measures of dissimilarity are to be selected, besides the five mentioned above, the three additional ones are: whether two countries share a language spoken by at least nine percent of their respective population, the interaction between the share of urban population that has access to improved sanitation facilities, and the interaction between the female life expectancy at birth in the origin and destination countries. While the first measure has already been used in [20], with its parameter having the same sign as ours, the other two have not appeared in the literature. The parameters for these new measures are positive, indicating that migration flows tend to be larger between countries with similar urban access to improved sanitation facilities and female life expectancy at birth.

5 Conclusion

This paper introduces a new algorithm, which we called SISTA, to learn the transport cost in optimal transport problems. In this type of problems, one needs to optimize simultaneously over the potentials and over the parameters of the cost. As

the parameter vector is sparse, we add an l_1 penalization over the parameters. SISTA alternates between a phase of exact minimization over the transport potentials and a phase of proximal gradient descent over the parameters of the transport cost. We prove its linear convergence and illustrate through numerical experiments its rapid convergence compared to coordinate descent and ISTA. In our application of predicting bilateral migration flows, SISTA allows us to learn which measures of dissimilarity between an origin and a destination country are the most important ones. Our approach reveals that dissimilarities between origin and destination countries in terms of area, female life expectancy, and urban access to improved sanitation facilities, are critical predictors of bilateral migrations flows that have been absent from this literature.

Our method applies to a broad range of problems in quantitative social sciences. We present an application to the prediction of bilateral migration flows using country-specific characteristics and pairwise measures of dissimilarity between countries. The same technique could be applied to predicting bilateral trade flows, the matching of workers to jobs, men to women, children to schools, etc. In all these applications, there exists a long list of attributes/characteristics upon which measures of dissimilarity between countries of origin and countries of destination, men and women, workers and jobs can be constructed and could explain flows or matches. Our approach allows one to select those that matter the most.

More generally, the idea underlying the SISTA algorithm is broadly applicable. In many other optimization problems, it may be worthwile using hybrid methods combining the strengths of several existing descent methods. As exemplified by SISTA, the hybrid version can be much more efficient than the methods it combines.

Bibliography

[1] Beck, A.; Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** (2009), no. 1, 183–202.

[2] Beine, M.; Parsons, C. Climatic factors as determinants of international migration. *Scand. J. Econ.* **117** (2015)., no. 2, 723–767.

[3] Beine, M.; Bertoli, S.; Fernández-Huertas Moraga, J. A practitioners' guide to gravity models of international migration. *World Econ.*, **39** (2016), no. 4, 496–512.

[4] Beine, M. The role of networks for migration flows: an update. *Int. J. Manpow.* (2016).

[5] Belot, M.; Ederveen, S. Cultural barriers in migration between OECD countries. *J. Popul. Econ.* **25** (2012), no. 3, 1077–1105.

[6] Benamou, J. D.; Carlier, G.; Cuturi, M.; Nenna, L.; Peyré, G. Iterative Bregman projections for regularized transportation problems. *SIAM J. Sci. Comput.* **37** (2015), no. 2, A1111–A1138.

[7] Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS* **26** (2013), 2292–2300.

[8] Cuturi, M.; Avis, D. Ground metric learning. *J. Mach. Learn. Res.* **15** (2014), no. 1, 533–564.

[9] Dupuy, A.; Galichon, A. Personality traits and the marriage market. *J. Political Econ.*, **122** (2014), no. 6, 1271–1319.

[10] Dupuy, A.; Galichon, A.; Sun, Y. Estimating matching affinity matrices under low-rank constraints. *Inf. Inference* **8** (2019), no. 4, 677–689.

[11] Galichon, A. *Optimal Transport Methods in Economics*. Princeton University Press, Princeton, N.J., 2016.

[12] Galichon, A.; Salanié, B. Cupid invisible hand: Social surplus and identification in matching models. *Available at SSRN 1804623* (2015).

[13] Grogger, J.; Hanson, G. H. Income maximization and the selection and sorting of international migrants. *J. Dev. Econ.* **95** (2011), no. 1, 42–57.

[14] Huang, G.; Guo, C.; Kusner, M. J.; Sun, Y.; Sha, F.; Weinberger, K. Q. Supervised word mover's distance. *NeurIPS* **29** (2016), 4862–4870.

[15] Kulis, B. Metric learning: A survey. *Found. Trends Mach. Learn.* **5** (2013), no. 4, 287–364.

[16] Idel, M. A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps. *arXiv preprint arXiv:1609.06349* (2016).

[17] Mayer, T.; Zignago, S. Notes on CEPII distances measures: The GeoDist database (2011).

[18] Özden, Ç.; Parsons, C. R.; Schiff, M.; Walmsley, T. L. Where on earth is everybody? The evolution of global bilateral migration 1960-2000. *World Bank Econ. Rev.* **25** (2011), no. 1, 12–56.

[19] Peyré, G.; Cuturi, M. Computational optimal transport. *Found. Trends Mach. Learn.* **11** (2019), no. 5–6, 355–607.

[20] Ramos, R.; Suriñach, J. A gravity model of migration between the ENC and the EU. *Tijdschr. Econ. Soc. Geogr.* **108** (2017), no. 1, 21–35.

[21] Santambrogio, F. *Optimal Transport for Applied Mathematicians*. Progress in Nonlinear Differential Equations and their Applications, Vol. 87. Birkhäuser, New York, 2015.

[22] Villani, C. *Topics in Optimal Transportation*. Graduate Studies in Mathematics, Vol. 58. American Mathematical Society, Providence, R.I., 2003.

[23] Villani, C. *Optimal Transport: Old and New*. Grundlehren der Mathematischen Wissenschaften [Comprehensive Studies in Mathematics], Vol. 338. Springer-Verlag, Berlin, 2009.