Fast Optimal Transport Averaging of Neuroimaging Data

Alexandre Gramfort, Gabriel Peyré, Marco Cuturi

Fast Optimal Transport Averaging of Neuroimaging Data

A. Gramfort[1,2], G. Peyré[3], M. Cuturi[4]

[1] Institut Mines-Télécom, Telecom ParisTech, CNRS LTCI
[2] NeuroSpin, CEA Saclay, Bat. 145, 91191 Gif-sur-Yvette, cedex France
[3] CNRS and CEREMADE, Université Paris-Dauphine
[4] Graduate School of Informatics, Kyoto University

Abstract. Knowing how the Human brain is anatomically and functionally organized at the level of a group of healthy individuals or patients is the primary goal of neuroimaging research. Yet computing an average of brain imaging data defined over a voxel grid or a triangulation remains a challenge. Data are large, the geometry of the brain is complex and the between subjects variability leads to spatially or temporally non-overlapping effects of interest. To address the problem of variability, data are commonly smoothed before performing a linear group averaging. In this work we build on ideas originally introduced by Kantorovich [18] to propose a new algorithm that can average efficiently non-normalized data defined over arbitrary discrete domains using transportation metrics. We show how Kantorovich means can be linked to Wasserstein barycenters in order to take advantage of the entropic smoothing approach used by [7]. It leads to a smooth convex optimization problem and an algorithm with strong convergence guarantees. We illustrate the versatility of this tool and its empirical behavior on functional neuroimaging data, functional MRI and magnetoencephalography (MEG) source estimates, defined on voxel grids and triangulations of the folded cortical surface.

1 Introduction

Computing the average of some observations may seem like a trivial problem, yet it remains an active topic of research in mathematics, statistics and applications such as medical imaging. The problem of atlas computation from images [17], or meshes [11], or the problem of group analysis from functional imaging data [25] are particularly relevant for this field. The challenge is that natural phenomena are usually described in terms of physical and temporal event locations, along with their intensity. While Euclidean averaging is standard and has some benefits such as low computation time, this procedure ignores the geometry of the space the observations belong to; the image of an average brain image obtained by Euclidean averaging of individual voxels does not yield the image of the brain of an average individual.

Starting from observations defined on a regular or irregular grid, our aim is to provide a *model-free* approach to *average* them that only builds upon geometric

arguments. An example of such data are functional MRI (fMRI) data defined on a voxel grid or a triangulated cortical surface. The approach aims to be intrinsically geometric in the sense that it *only* requires the prior knowledge of a metric between the locations on the grid. The technique aims to be versatile in the sense that it can be applied to weighted samples taking values on a discretized space with no assumptions on the regularity of the metric.

The approach we propose is inspired by optimal transport theory [26] and can be seen as an extension of the Wasserstein barycenter problem [1, 22, 7], which aims at estimating a probability measure which bests approximates a family of probability measures in the Wasserstein metric sense. The challenge of using optimal transport in the setting we consider comes from the fact that Wasserstein distances (and their barycenters) are defined for *probability* measures only. While some data are normalized such the Orientation Diffusion Function (ODF) in diffusion MRI [10], a number of medical imaging data are non-normalized. Here we bypass this limitation using a generalization of optimal transport distances proposed by [18]. This extension comes at a price, since it introduces an additional parameter (the cost of adding or removing mass) which can be difficult to tune. We propose a simple way to mitigate this problem by introducing a natural constraint on the overall mass of the barycenter, which we set to be equal to the average of the masses of all samples. We provide an efficient method to compute Kantorovich means by building upon the first algorithm of [7]. We provide intuitions on the behavior of our method and demonstrate its relevance with simulations and experimental results obtained with fMRI and MEG data which are neuroimaging data defined on a voxel grid or a triangulation of the folded cortical mantle.

This paper is organized as follows. We start in Section 2 with a reminder on optimal transport and the Kantorovich metric for non-normalized measures. We introduce Kantorovich means in Section 3 and describe efficient algorithms to compute them. Section 4 contains simulations and results on publicly available fMRI data with 20 subjects and MEG data with 16 subjects.

2 Optimal Transport Between (Un)Normalized Measures

We introduce in this section Wasserstein distances for non-negative measures on a finite metric space (Ω, D). Simply put, we consider normalized histograms on a grid of locations, and assume the distance between any two locations is provided. Next, we extend Wasserstein distances to non-negative vectors of *bounded mass*.

Notations. Let d be the cardinal of Ω. Relabeling arbitrarily all the elements of Ω as $\{1, \cdots, d\}$ we represent the set of non-negative measures on Ω by the non-negative orthant $\mathbb{R}^d_+$. Let $\mathbf{1}_d$ be the d-dimensional vector of ones. For any vector $u \in \mathbb{R}^d$ we write $|u|_1$ for the l_1 norm of u, $\sum_{i=1}^d |u_i|$. When $u \in \mathbb{R}^d_+$ we also call $|u|_1$ the (total) *mass* of vector u. Let $M = [m_{ij}] \in \mathbb{R}^{d \times d}_+$ be the matrix that describes the metric between all locations in Ω, namely $m_{ij} = D(i, j), i, j \leq d$.

Wasserstein Distance for Normalized Histograms. Consider two vectors $a, b \in \mathbb{R}_+^d$ such that $|a|_1 = |b|_1$. Both can be interpreted as histograms on Ω of the same mass. A non-trivial example of such normalized data in medical imaging is the discretized ODF used for diffusion imaging data [10]. For $p \geq 1$, the p-Wasserstein distance $W_p(a, b)$ between a and b is the p^{th} root of the optimum of a linear program, known as a *transportation problem* [4, §7.2]. A transport problem is a network flow problem on a bipartite graph with cost M^p (the pairwise distance matrix M raised element-wise to the power p), and feasible set of flows $U(a, b)$ (known as the transportation polytope of a and b), where $U(a, b)$ is the set of $d \times d$ nonnegative matrices such that their row and column marginals are equal to a and b respectively:

$$U(a, b) \stackrel{\text{def}}{=} \{T \in \mathbb{R}_+^{d \times d} \mid T\mathbf{1}_d = a,\, T^T\mathbf{1}_d = b\}. \tag{1}$$

Given the constraints induced by a and b, one naturally has that $U(a, b)$ is empty when $|a|_1 \neq |b|_1$ and non-empty when $|a|_1 = |b|_1$ (in which case one can easily check that the matrix $ab^T/|a|_1$ belongs to that set). The p-Wasserstein distance $W_p(a, b)$ raised to the power p (written $W_p^p(a, b)$ below) is equal to the optimum of a parametric Optimal Transport (**OT**) problem on d^2 variables,

$$W_p^p(a, b) = \mathbf{OT}(a, b, M^p) \stackrel{\text{def}}{=} \min_{T \in U(a,b)} \langle T, M^p \rangle, \tag{2}$$

parameterized by the marginals a, b and matrix M^p.

Optimal Transport for Unnormalized Measures. If the total masses of a and b differ, namely $|a|_1 \neq |b|_1$, the definition provided above is not useful because $U(a, b) = \emptyset$. Several extensions of the OT problem have been proposed in that setting; we recall them here for the sake of completeness. In the computer vision literature, [23] proposed to handle that case by: *(i) relaxing* the equality constraints of $U(a, b)$ to inequality constraints $T\mathbf{1}_d \leq a$, $T^T\mathbf{1}_d \leq b$ in Equation (1); *(ii) adding an equality constraint* on the total mass of the solution $\mathbf{1}_d^T T \mathbf{1}_d = \min(|a|_1, |b|_1)$; *(iii) dividing* the minimum of $\langle T, M \rangle$ under constraints *(i,ii)* by $\min(|a|_1, |b|_1)$. This modification does not, however, result in a metric. [20] proposed later a variant of this approach called EMD-hat that incorporates constraints *(i,ii)* but *(iii')* adds to the optimal cost $\langle T^\star, M \rangle$ a constant times $\min(|a|_1, |b|_1)$. When that constant is large enough M, [20] claim that EMD-hat is a metric. We also note that [2] proposed a quadratic penalty between the differences of masses and made use of a dynamic formulation of the transportation problem.

Kantorovich Norms for Signed Measures. We propose to build on early contributions by Kantorovich to define a generalization of optimal transport distance for unnormalized measures, making optimal transport applicable to a wider class of problems, such as the averaging of functional imaging data. [18] proposed such a generalization as an intermediary result of a more general definition, the *Kantorovich norm* for signed measures on a compact metric space, which was itself

extended to separable metric spaces by [15]. We summarize this idea here by simplifying it to the case of interest in this paper where Ω is a finite (of size d) probability space, in which case signed measures are equivalent to vectors in $\mathbb{R}^d$. [18] propose first a norm for vectors z in the orthogonal of $\mathbf{1}_d$ (vectors z such that $z^T\mathbf{1}_d = 0$), by considering the 1-Wasserstein distance between the positive and negative parts of z, $\|z\|_K = W_1(z_+, z_-)$. A penalty vector $\Delta \in \mathbb{R}^d_+$ is then introduced to define the norm $\|x\|_K$ of *any* vector x as the minimal value of $\|z\|_K + \Delta^T|z - x|$ when z is taken in the space of all vectors z with zero sum, and $|z - x|$ is the element-wise absolute value of the difference of vectors z and x. For this to define a true norm in $\mathbb{R}^d$, Δ must be such that $\Delta_i \geq \max_j m_{ij}$ and $|\Delta_i - \Delta_j| \leq m_{ij}$. The distance between two arbitrary non-negative vectors a, b of different mass is then defined as $\|a - b\|_K$. As highlighted by [26, p.108], and if we write e_i for the i^{th} vector of the canonical basis of $\mathbb{R}^d$, this norm is the maximal norm in $\mathbb{R}^d$ such that for any $i, j \leq d$, $\|e_i - e_j\|_K = m_{ij}$, namely the maximal norm in the space of signed measures on Ω such that the norm between two Dirac measures coincides with Ω's metric between these points.

Kantorovich Distances for Unnormalized Nonnegative Measures. [14] noticed that Kantorovich's distance between unnormalized measures can be cast as a regular optimal transport problem. Indeed, one simply needs to: *(i)* add a *virtual point* ω to the set $\Omega = \{1, \cdots, d\}$ whose distance $D(i, \omega) = D(\omega, i)$ to any element i in Ω is set to Δ_i ; *(ii)* use that point ω as a buffer when comparing two measures of different mass. The appeal of Kantorovich's formulation in the context of this work is that it boils down to a classic optimal transport problem, which can be approximated efficiently using the smoothing approach of [6] as discussed in Section 3. To simplify our analysis in the next section, we only consider non-negative vectors (histograms) $a \in \mathbb{R}^d_+$ such that their total mass is upper bounded by a known positive constant. This assumption alleviates the definition of our distance below, since it does not require to treat separately the cases where either $|a|_1 \geq |b|_1$ or $|a|_1 < |b|_1$ when comparing $a, b \in \mathbb{R}^d_+$. Note also that this assumption always holds when dealing with finite collections of data. Without loss of generality, this is equivalent to considering vectors a in $\mathbb{R}^d_+$ such that $|a|_1 \leq 1$ with a simple rescaling of all vectors by that constant. We define next the Kantorovich metric on S_d, where $S_d = \{u \in \mathbb{R}^d_+, |u|_1 \leq 1\}$.

Definition 1 (Kantorovich Distances on S_d). *Let $\Delta \in \mathbb{R}^d_+$ such that $\Delta_i \geq \max_j m_{ij}$ and $|\Delta_i - \Delta_j| \leq m_{ij}$. Let $p \geq 0$. For two elements a and b of S_d, we write $\alpha = 1 - |a|_1 \geq 0$ and $\beta = 1 - |b|_1 \geq 0$. Their p-Kantorovich distance raised to the power p is*

$$K^p_{p\Delta}(a, b) = \mathbf{OT}(\left[\begin{smallmatrix} a \\ \alpha \end{smallmatrix}\right], \left[\begin{smallmatrix} b \\ \beta \end{smallmatrix}\right], \hat{M}^p), \text{ where } \hat{M} = \begin{bmatrix} M & \Delta \\ \Delta^T & 0 \end{bmatrix} \in \mathbb{R}^{d+1 \times d+1}_+. \quad (3)$$

The Kantorovich distance inherits all metric properties of Wasserstein distances: the mapping which to a vector a associates a vector $[a; 1 - |a|_1] \in \Sigma_{d+1}$ can be regarded as a feature map, to which the standard Wasserstein distance using $\hat{M}$ (which is itself a metric matrix) is applied.

3 Kantorovich Mean of Unnormalized Measures

Consider now a collection $\{b^1, \cdots, b^N\}$ of N non-negative measures on (Ω, D) with mass upper-bounded by 1, namely N vectors in S_d. Let $\beta^j = 1 - \left|b^j\right|$ be the deficient mass of b^j. Our goal in this section is to find, given a vector of virtual costs Δ and an exponent p, a vector a in S_d which minimizes the sum of its p-Kantorovich distances $K^p_{p\Delta}$ to all the b^j,

$$a \in \operatorname*{argmin}_{u \in S_d} \frac{1}{N} \sum_{j=1}^{N} K^p_{p\Delta}(u, b^j) = \operatorname*{argmin}_{u \in S_d} \frac{1}{N} \sum_{j=1}^{N} \mathbf{OT}(\begin{bmatrix} u \\ 1-|u|_1 \end{bmatrix}, \begin{bmatrix} b^j \\ \beta^j \end{bmatrix}, \hat{M}^p). \quad \text{(P1)}$$

Because of the equivalence between Kantorovich distances for points in S_d and Wasserstein distances in the $d+1$ simplex, this problem can be naturally cast as a Wasserstein barycenter problem [1] with metric $\hat{M}$. Problem (P1) can be cast as a linear program with $N(d+1)^2$ variables. For the applications we have in mind, where d is of the order or larger than 10^4, solving that program is not tractable. We discuss next computational approaches to solve it efficiently.

Smooth Optimal Transport. [22] and [5] have proposed efficient algorithms to solve the Wasserstein barycenter problem in low dimensional Euclidean spaces. These approaches are not, however, suitable when one considers observations on the cortex, for which *all pairs shortest path* metrics (inferred from a graph structure connecting all voxels) are preferred over Euclidean metrics. To solve Problem (P1) we turn instead to a recent series of algorithms proposed in [7], [3] and [8] that all exploit the regularized OT approach suggested in [6]. Among these recent approaches, we propose to build in this work upon the first algorithm in [7], which can be easily modified to incorporate constraints on a. This flexibility will prove useful in the next section.

The strategy of [7] is to regularize directly the optimal transport problem by an entropic penalty, whose weight is parameterized by a parameter $\lambda > 0$,

$$\mathbf{OT}_\lambda(a, b, M^p) \stackrel{\text{def}}{=} \min_{T \in U(a,b)} \langle T, M^p \rangle - \frac{1}{\lambda} H(T),$$

where $H(T)$ stands for the entropy of the matrix T seen as an element of the simplex of size d^2, $H(T) \stackrel{\text{def}}{=} -\sum_{ij} t_{ij} \log(t_{ij})$. As shown by [7], the regularized transport problem $\mathbf{OT}_\lambda$ admits a unique optimal solution. As such, $\mathbf{OT}_\lambda(a, b, M^p)$ is a differentiable function of a whose gradient can be recovered through the solution of the corresponding smoothed dual optimal transport. Without elaborating further on this approach, we propose to simply replace all expressions that involve an optimal transport problem $\mathbf{OT}$ in our formulations by their smoothed counterpart $\mathbf{OT}_\lambda$.

Sensitivity of Kantorovich Means to the Parameter Δ. The magnitude of the solution a to Problem (P1) depends directly on the virtual distance Δ. Suppose,

for instance, that $\Delta = \varepsilon \mathbf{1}_d$ with ε arbitrarily small. In that case a should converge to a unit mass on the last (virtual) bin and would therefore be equal to the null histogram $\mathbf{0}_d$ on the d other bins. If, on the contrary, $\Delta = \gamma \mathbf{1}_d$ and γ is large, we obtain that $K^p_{p\Delta}(a,b)/\gamma$ grows as $||a|_1 - |b|_1|$. Therefore a minimum of Problem (P1) would necessarily need to have a total mass that minimizes $\sum_j \big| |a|_1 - |b^j|_1 \big|$, namely a total mass equal to the median mass of all b^j. This sensitivity of the solution a to the magnitude of Δ may be difficult to control. Choosing adequate values for Δ, namely setting the distance of the virtual point to the d other points, may also be a difficult parameter choice. To address this issue we propose to simplify our framework by introducing an equality constraint on the mass of the barycenter a in our definition, and let Δ be any non-negative vector, typically set to a large quantile of the distribution of all pairwise distances M^p_{ij} times the vector of ones $\mathbf{1}_d$. Under these assumptions, we can now propose p-Kantorovich means with a constraint on the total mass of a. Remaining parameters in our approach are therefore only p and λ. In practice we will fix $p = 1$, which corresponds to the Earth Mover's Distance [23], and use a high λ, namely a small entropic regularization of order $1/\lambda$, which has also the merit of making Problem (P1) strongly convex. λ is set in our experiments to $100/\text{median}(M)$, where $\text{median}(M)$ is the median of all pairwise distances $\{M_{ij}\}_{ij}$.

Definition 2 (p-Kantorovich Means with Constrained Mass). *Let $\Delta \in \mathbb{R}^d_+$ and $p \geq 0$. A Kantorovich mean with a target mass $\rho \leq 1$ of a set of N histograms $\{b^1, \cdots, b^N\}$ in S_d is the unique vector a in S_d such that:*

$$a \in \operatorname*{argmin}_{\substack{a \in S_d \\ |a|_1 = \rho}} \frac{1}{N} \sum_j \mathbf{OT}_\lambda(a, b^j, \hat{M}^p).$$

We provide in Algorithm 1 an implementation of [7, Alg.1]. Unlike their version, we only consider a fixed step-length exponentiated gradient descent, and add a mass renormalization step. We set the default mass of the barycenter to be the mean of the masses of all histograms. We use the notation $\circ$ for the elementwise (Schur) product of vectors. Note that the computations of N dual optima in line 7 of Algorithm 1 below can be vectorized and computed using only matrix-matrix products. We use GPGPUs to carry out these computations.

4 Application to the Averaging of Neuroimaging Data

Neuromaging data are defined on a grid of voxels, eventually restricted to the brain volume, or on a triangulation of the cortical mantle obtained by segmentation of MRI data. Examples of data most commonly analyzed on a grid are fMRI data, while neural activity estimates derived from MEG/EEG data are often restricted to the cortical surface [9]. Anatomical data such as cortical thickness, which is a biomarker of certain neurodegenerative pathologies, is also defined on the surface. In all cases the data are defined on a discrete set of points (voxels or vertices) which have a natural distance given by the geometry of the brain. The

Algorithm 1 p-Kantorovich Barycenter with Constrained Mass

Input: $\{b^1, \cdots, b^N\} \subset S_d$, metric M, quantile q, $p \geq 0$, entropic regularizer $\lambda > 0$, step size c.
Compute mean mass $\rho = \frac{1}{N} \sum_i |b^j|_1$.
Form virtual cost vector $\Delta = \text{quantile}(M, q\%)\mathbf{1}_d$.
Form augmented $d+1 \times d+1$ ground metric $\hat{M}$ as in Equation (3)
Set $a = \mathbf{1}_{d+1}/(d+1)$.
while a changes **do**
  Compute all dual optima α^j of $\mathbf{OT}_\lambda(a, b^j, \hat{M})$ using [7, Alg.3]
  $a \leftarrow a \circ \exp(-c\frac{1}{N}\sum_{j=1}^N \alpha^j)$; (gradient update)
  $a_i \leftarrow \begin{cases} \rho a_i / \sum_{l=1}^d a_l & \text{if } i \leq d, \\ 1-\rho & \text{if } i = d+1. \end{cases}$ (projection on the simplex/mass constraint)
end while
Output $a_{1:d} \in S_d$.

data are also non-normalized as they represent physical or statistical quantities, such as thickness in millimeters or F statistics. Such data are therefore particularly well adapted to the algorithm proposed in this paper: they are defined on discrete space, are non-normalized and there exists a natural ground metric.

The difficulty when averaging neuroimaging data is the anatomo-functional variability: every brain is different. The standard approach to compensate for this variability across subjects is to smooth the data to favor the overlap of signal of interest after the individual data have been ported to a common space using anatomical landmarks (anatomical registration). Volume data are typically redefined in MNI space while surface data are transferred to an average cortical surface, using for instance the FreeSurfer software[5]. When working with fMRI the spatial variability is commonly compensated by smoothing the data with an isotropic Gaussian kernel of Full Width at Half Maximum (FWHM) between 6 and 8 mm. MEG and EEG suffer from the same spatial variability, but also from the temporal variability of neural responses which is compensated by low pass filtering the data. By employing a transportation metric informed by the geometry of the domain, this smoothing procedure as well as the setting of the kernel bandwidth are not needed.

In the following experiments we focus on spatial averaging, although extension to spatiotemporal data is straightforward provided the metric is defined along the time axis. When working with a voxel grid the distance is the Euclidian distance taking into account the voxel size in millimeters, and when working on a cortical triangulation, the distance used is the geodesic distance computed on the folded cortical mantle. We now present results of a simulation study where standard averaging with Gaussian smoothing is compared to Kantorovich means. Simulation results are followed by experimental results obtained with fMRI data from 20 subjects and MEG data on a population of 16 subjects.

[5] https://surfer.nmr.mgh.harvard.edu/

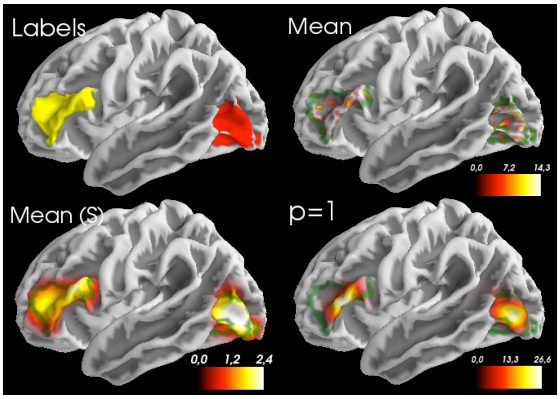

Fig. 1. Simulation results with focal random signals generated in areas/labels BA45 (yellow) and MT (red) in a group of 100 subjects. Data are defined on a surface with 10,024 vertices (FreeSurfer fsaverage 5). One shows the standard averaging referred to as *Mean*, the averaging after Gaussian smoothing is referred to as *Mean (S)* (mean after Gaussian smoothing with FWHM=8 mm), and the Kantorovich mean (p=1). The result *Mean* shows the focal signals with random positions in the labels delineated in green. The Kantorovich mean highlights clear foci of activations in the ROIs without smearing the activation as with Gaussian smoothing which furthermore significantly dampens the amplitudes.

Simulation setup. In this experiment using on a triangulation of the cortex, we simulated signals of interest in two brain regions using the functional parcellation provided by the FreeSurfer software. We used regions Broadman area 45 (BA45) and the visual area MT. We simulated for a group of 100 subjects random positive signals in these two regions. For each subject and each region, the signal is focal at a random location with a random amplitude generated with a truncated Gaussian distribution (mean 5, std. dev. 1.). We use here focal signals to exemplify the effect of optimal transport. Such signals could correspond to dipolar activations derived from MEG/EEG using dipole fitting methods [24] or sparse regression techniques [27, 13].

Figure 1 presents the locations of the two regions (labels), the averages with and without Gaussian smoothing and the Kantorovich average. Gaussian smoothing leads to a highly blurred average which exceeds the extent of the regions of interest, while it also strongly reduces the amplitudes of the signals, potentially washing out the statistical effects. The peak amplitudes obtained with optimal transport are also higher and closer to the individual peak amplitudes. One can clearly observe the limitations of Gaussian smoothing, which furthermore requires to set the bandwidth of the kernel. The Kantorovich average nicely highlights two foci of signals at the group level.

Results on fMRI data. We used here fMRI data analyzed on a voxel grid. It corresponds to 20 subjects from the database described in [21]. We average here

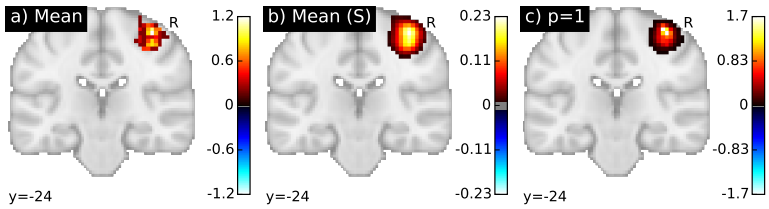

Fig. 2. Averaging of the standardized effect of interest on fMRI data. From left to right, a) the Euclidian mean without smoothing, b) the Euclidian mean with smoothing (FWHM=8 mm), c) the Kantorovich mean with Euclidian ground metric and p=1. The later result highlights a clear foci of activations in the ROI without smearing the activation nor damping the amplitudes as much as the kernel smoothing.

the standardized effect of interest induced by left hand button press. In Fig. 2-a we show the Euclidian average without smoothing. In Fig.2-b we report results obtained by classical averaging following Gaussian smoothing with FWHM of 8 mm. Fig.2-c shows the Kantorovich mean with constrained mass. One can observe that this barycenter highlights a clear active region without requiring any kernel smoothing. It also leads to a amplitude in the average standardized effect around 1.7 which is much higher than the 0.23 obtained when smoothing.

Results on MEG data. We now evaluate the benefit of the proposed approach on experimental data. These data were acquired with a Neuromag VectorView system (Elekta Oy, Helsinki, Finland) with 306 sensors arranged in 102 triplets, each comprising two orthogonal planar gradiometers and one magnetometer. Subjects are presented with images containing faces of familiar (famous) or unfamiliar persons and so called "scrambled" faces. See [16] for more details. Dataset contains 16 subjects. For each one, event related fields (ERF) were obtained by averaging about 200 repetitions of recordings following stimuli presentations. Data were band pass filtered between 1 and 40 Hz. Following standard MEG source localization pipelines [12], a noise covariance was estimated from prestimulus time intervals and used for source reconstruction with the cortically constrained dSPM method [9]. The values obtained with dSPM can be considered as F statistics, where high values are located in active regions.

In Figure 3, we present results at a single time point, 190 ms after stimulus onset, which corresponds to the time instant where the dSPM amplitudes are maximum. Data correspond the visual presentation of *famous faces*. In green, is the border of the primary visual cortex (V1) provided by the FreeSurfer functional atlas. One can observe that the Kantorovich barycenter yields a more focal average nicely positioned in the middle of the calcarine fissure where V1 is located. Such a strong activation in V1 is expected in such an experiment consisting of visual stimuli. To investigate more subtle cognitive effects, such as the response of the fusiform face area (FFA) reported about 170 ms after stimulation in the literature [19, 16], we report results obtained on contrasts of ERFs measured after famous faces presentations *vs.* scrambled faces. As illustrated in

Figure 4, Kantorovich mean nicely delineates a focal source of activity in the ventral part of the cortex known as the fusiform gyrus.

These results show that Kantorovich means provides focal activation at the population level despite the challenging problem of inter-subject anatomo-functional variability. They avoid the smearing of the signal or statistical effects of interests which naturally occur when data are spatially smoothed before standard averaging. Note again that here no smoothing parameter with FWHM in millimeters is manually specified. Their solution only depends on the metric naturally derived from the geometry of the cortical surface. With a cortical triangulation containing 10,024 vertices and 16 subjects the computation on a Tesla K40 GPU of one barycenter takes less than 1 min.

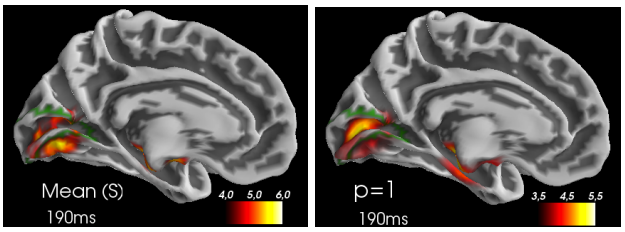

Fig. 3. Average of dSPM estimates derived from MEG ERF data on a group of 16 subjects stimulated with pictures of famous faces. From left to right: standard mean and Kantorovich mean. The left hemisphere is displayed in medial view. In green is the border of the primary visual cortex (V1) provided by FreeSurfer. One can observe that the Kantorovich mean has its peak amplitude within V1.

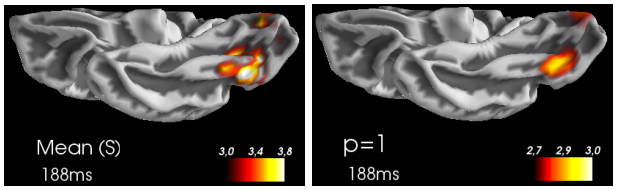

Fig. 4. Group averages (16 subjects) of dSPM estimates derived from MEG ERF data obtained by contrasting the famous faces stimulation with the scrambled faces. From left to right: standard mean and the Kantorovich mean. The right hemisphere is displayed in ventral view. Optimal transport results highlight a focal activity in the Fusiform gyrus known to be implicated in face processing [19].

5 Conclusion

The contributions of this work are two-fold. First, by considering non-normalized measures particularly relevant for medical imaging data we extend the current state of the art in barycenter estimation using transportation metrics. Following recent contributions on discrete optimal transport we propose a smoothed version of the transport problem that leads us to an efficient optimization algo-

rithm. While many contributions on optimal transport work only in one or two dimensions on a regular grid, our approach can cope with the complex geometry of the brain (irregular grids and surfaces). Only the definition of a ground metric is here required. The algorithm proposed involves simple operations that are particularly adapted to modern GPU hardware and allows us to compute barycenters on full brain data in a few minutes.

Second, with simulations defined on the cortex triangulation, a publicly available fMRI dataset with 20 subjects and an MEG dataset processed with a standard analysis pipeline with 16 subjects we demonstrated the ability of the method to clearly highlight activation foci while avoiding the need to smooth the data. The fMRI data showed a clear activation in the right motor cortex and on the MEG data we showed that the proposed approach better identified activation foci in the primary visual cortex and the fusiform gyrus. Both findings, that are consistent with previous neuroscience literature, show that method proposed yields more accurate results than the current pipelines which furthermore requires to set a kernel bandwidth parameter. The removal of any free parameter in the pipeline is a way towards more reproducible neuroimaging results.

Due to the non-linearity of the approach the estimation of statistical threshold shall be performed with non-parametric permutation tests. When thresholding barycenters as presented in Section 4 it is expected that one will obtain clear clusters.

Acknowledgements A. Gramfort was supported by the ANR grant THALAMEEG, ANR-14-NEUC-0002-01. M. Cuturi gratefully acknowledges the support of JSPS young researcher A grant 26700002, the gift of a K40 card from NVIDIA and fruitful discussions with K.R. Müller. The work of G. Peyré has been supported by the European Research Council (ERC project SIGMA-Vision).

References

1. Agueh, M., Carlier, G.: Barycenters in the Wasserstein space. SIAM Journal on Mathematical Analysis 43(2), 904–924 (2011)
2. Benamou, J.D.: Numerical resolution of an unbalanced mass transport problem. ESAIM: Mathematical Modelling and Numerical Analysis 37(5), 851–868 (2003)
3. Benamou, J.D., Carlier, G., Cuturi, M., Nenna, L., Peyré, G.: Iterative bregman projections for regularized transportation problems. arXiv preprint arXiv:1412.5154 (2014)
4. Bertsimas, D., Tsitsiklis, J.: Introduction to linear optimization. Athena Scientific Belmont, MA (1997)
5. Bonneel, N., Rabin, J., Peyré, G., Pfister, H.: Sliced and radon wasserstein barycenters of measures. Journal of Mathematical Imaging and Vision pp. 1–24 (2014)
6. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: Advances in Neural Information Processing Systems 26. pp. 2292–2300 (2013)
7. Cuturi, M., Doucet, A.: Fast computation of wasserstein barycenters. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14) (2014)
8. Cuturi, M., Peyré, G., Rolet, A.: A smoothed dual approach for variational wasserstein problems. arXiv preprint arXiv:1503.02533 (2015)

9. Dale, A., Liu, A., Fischl, B., Buckner, R.: Dynamic statistical parametric neurotechnique mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. Neuron 26, 55–67 (2000)
10. Descoteaux, M., Deriche, R., Knosche, T., Anwander, A.: Deterministic and probabilistic tractography based on complex fibre orientation distributions. Medical Imaging, IEEE Transactions on 28(2), 269–286 (Feb 2009)
11. Durrleman, S., Prastawa, M., Charon, N., Korenberg, J.R., Joshi, S., Gerig, G., Trouvé, A.: Morphometry of anatomical shape complexes with dense deformations and sparse parameters. NeuroImage 101(0), 35 – 49 (2014)
12. Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Parkkonen, L., Hämäläinen, M.: MNE software for processing MEG and EEG data. NeuroImage 86(0), 446 – 460 (2014)
13. Gramfort, A., Strohmeier, D., Haueisen, J., Hämäläinen, M., Kowalski, M.: Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations. NeuroImage 70(0), 410 – 422 (2013)
14. Guittet, K.: Extended kantorovich norms: a tool for optimization. Tech. Rep. 4402, INRIA (2002)
15. Hanin, L.: An extension of the kantorovich norm. Contemporary Mathematics 226, 113–130 (1999)
16. Henson, R.N., Wakeman, D.G., Litvak, V., Friston, K.J.: A parametric empirical bayesian framework for the EEG/MEG inverse problem: generative models for multisubject and multimodal integration. Front. in Human Neuro. 5(76) (2011)
17. Joshi, S., Davis, B., Jomier, M., Gerig, G.: Unbiased diffeomorphic atlas construction for computational anatomy. NeuroImage 23(0), 151–160 (2004)
18. Kantorovich, L., Rubinshtein, G.: On a space of totally additive functions, vestn. Vestn Lening. Univ. 13, 52–59 (1958)
19. Kanwisher, N., Mcdermott, J., Chun, M.M.: The fusiform face area: A module in human extrastriate cortex specialized for face perception. Journal of Neuroscience 17, 4302–4311 (1997)
20. Pele, O., Werman, M.: A linear time histogram metric for improved sift matching. In: European Conference on Computer Vision, ECCV 2008, pp. 495–508 (2008)
21. Pinel, P., Thirion, B., Meriaux, S., Jobert, A., Serres, J., Le Bihan, D., Poline, J., Dehaene, S.: Fast reproducible identification and large-scale databasing of individual functional cognitive networks. BMC neuroscience 8, 91 (2007)
22. Rabin, J., Peyré, G., Delon, J., Bernot, M.: Wasserstein barycenter and its application to texture mixing. In: Scale Space and Variational Methods in Computer Vision, Lecture Notes in Computer Science, vol. 6667, pp. 435–446. Springer (2012)
23. Rubner, Y., Guibas, L., Tomasi, C.: The earth movers distance, multi-dimensional scaling, and color-based image retrieval. In: Proceedings of the ARPA Image Understanding Workshop. pp. 661–668 (1997)
24. Scherg, M., Von Cramon, D.: Two bilateral sources of the late AEP as identified by a spatio-temporal dipole model. Electroencephalogr Clin Neurophysiol 62(1), 32–44 (Jan 1985)
25. Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., Poline, J.B.: Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. NeuroImage 35(1), 105 – 120 (2007)
26. Villani, C.: Optimal transport: old and new, vol. 338. Springer Verlag (2009)
27. Wipf, D., Ramirez, R., Palmer, J., Makeig, S., Rao, B.: Analysis of empirical bayesian methods for neuroelectromagnetic source localization. In: Proc. Neural Information Processing Systems (NIPS) (2007)