Group level MEG/EEG source imaging via optimal transport: minimum Wasserstein estimates

Hicham Janati, Thomas Bazeille, Bertrand Thirion, Marco Cuturi, Alexandre Gramfort

Group level MEG/EEG source imaging via optimal transport: minimum Wasserstein estimates

H. Janati[1], T. Bazeille[1], B. Thirion[1], M. Cuturi[2], A. Gramfort[1]

[1] INRIA, CEA Neurospin, Université Paris-Saclay, France
[2] Google and CREST ENSAE

Abstract. Magnetoencephalography (MEG) and electroencephalography (EEG) are non-invasive modalities that measure the weak electromagnetic fields generated by neural activity. Inferring the location of the current sources that generated these magnetic fields is an ill-posed inverse problem known as source imaging. When considering a group study, a baseline approach consists in carrying out the estimation of these sources independently for each subject. The ill-posedness of each problem is typically addressed using sparsity promoting regularizations. A straightforward way to define a common pattern for these sources is then to average them. A more advanced alternative relies on a joint localization of sources for all subjects taken together, by enforcing some similarity across all estimated sources. An important advantage of this approach is that it consists in a single estimation in which all measurements are pooled together, making the inverse problem better posed. Such a joint estimation poses however a few challenges, notably the selection of a valid regularizer that can quantify such spatial similarities. We propose in this work a new procedure that can do so while taking into account the geometrical structure of the cortex. We call this procedure Minimum Wasserstein Estimates (MWE). The benefits of this model are twofold. First, joint inference allows to pool together the data of different brain geometries, accumulating more spatial information. Second, MWE are defined through Optimal Transport (OT) metrics which provide a tool to model spatial proximity between cortical sources of different subjects, hence not enforcing identical source location in the group. These benefits allow MWE to be more accurate than standard MEG source localization techniques. To support these claims, we perform source localization on realistic MEG simulations based on forward operators derived from MRI scans. On a visual task dataset, we demonstrate how MWE infer neural patterns similar to functional Magnetic Resonance Imaging (fMRI) maps.

Keywords: Brain · Inverse modeling · EEG / MEG source imaging

1 Introduction

Magnetoencephalography (MEG) measures the components of the magnetic field surrounding the head, while Electroencephalography (EEG) measures the electric

potential at the surface of the scalp. Both can do so with a temporal resolution of less than a millisecond. Localizing the underlying neural activity on a high resolution grid of the cortex, a problem known as source imaging, is inherently an "ill-posed" linear inverse problem: Indeed, the number of potential sources is larger than the number of MEG and EEG sensors, which implies that, even in the absence of noise, different neural activity patterns could result in the same electromagnetic field measurements.

To limit the set of possible solutions, prior hypotheses on the nature of the source distributions are necessary. The minimum-norm estimates (MNE) for instance are based on ℓ_2 Tikhonov regularization which leads to a linear solution [11]. An ℓ_1 norm penalty was also proposed by [34], modeling the underlying neural pattern as a sparse collection of focal dipolar sources, hence their name "Minimum Current Estimates" (MCE). These methods have inspired a series of contributions in source localization techniques relying on noise normalization [6,28] to correct for the depth bias [1] or block-sparse norms [30,10] to leverage the spatio-temporal dynamics of MEG signals. While such techniques have had some success, source estimation in the presence of complex multi-dipole configurations remains a challenge. In this work we aim to leverage the anatomical and functional diversity of multi-subject datasets to improve localization results.

Related work. This idea of using multi-subject information to improve statistical estimation has been proposed before in the neuroimaging literature. In [20] it is showed that different anatomies across subjects allow for point spread functions that agree on a main activation source but differ elsewhere. Averaging across subjects thereby increases the accuracy of source localization. On fMRI data, [35] proposed a probabilistic dictionary learning model to infer activation maps jointly across a cohort of subjects. A similar idea led [19] to introduce a Bayesian framework to account for functional intersubject variability. To our knowledge, the only contribution formulating the problem as a multi-task regression model employs a Group Lasso with an ℓ_{21} block sparse norm [21]. Yet this forces every potential neural source to be either active for all subjects or for none of them.

Contribution. The assumption of identical functional activity across subjects is clearly not realistic. Here we investigate several multi-task regression models that relax this assumption. One of them is the multi-task Wasserstein (MTW) model [15]. MTW is defined through an Unbalanced Optimal Transport (UOT) metric that promotes support proximity across regression coefficients. However, applying MTW to group level data assumes that the signal-to-noise ratio is the same for all subjects. We propose to build upon MTW and alleviate this problem by inferring estimates of both sources and noise variance for each subject. To do so, we follow similar ideas that lead to the concomitant Lasso [27,31,25] or the multi-task Lasso [24].

This paper is organized as follows. Section 2 introduces the multi-task regression source imaging problem. Section 3 presents some background on UOT metrics and explains how MWE are carried out. Section 4 presents the results of our experiments on both simulated and MEG datasets.

Notation. We denote by $\mathbb{1}_p$ the vector of ones in $\mathbb{R}^p$ and by $[\![q]\!]$ the set $\{1, \dots, q\}$ for any integer $q \in \mathbb{N}$. The set of vectors in $\mathbb{R}^p$ with non-negative (resp. positive) entries is denoted by $\mathbb{R}^p_+$ (resp. $\mathbb{R}^p_{++}$). On matrices, log, exp and the division operator are applied elementwise. We use $\odot$ for the elementwise multiplication between matrices or vectors. If $\mathbf{X}$ is a matrix, $\mathbf{X}_{i\cdot}$ denotes its i^{th} row and $\mathbf{X}_{\cdot j}$ its j^{th} column. We define the Kullback-Leibler (KL) divergence between two positive vectors by $\mathrm{KL}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \log(\mathbf{x}/\mathbf{y}) \rangle + \langle \mathbf{y} - \mathbf{x}, \mathbb{1}_p \rangle$ with the continuous extensions $0 \log(0/0) = 0$ and $0 \log(0) = 0$. We also make the convention $\mathbf{x} \neq 0 \Rightarrow \mathrm{KL}(\mathbf{x}|0) = +\infty$. The entropy of $\mathbf{x} \in \mathbb{R}^n$ is defined as $H(\mathbf{x}) = -\langle \mathbf{x}, \log(\mathbf{x}) - \mathbb{1}_p \rangle$. The same definition applies for matrices with an element-wise double sum.

2 Source imaging as a multi-task regression problem

We formulate in this section the inverse problem of interest in this paper, and recall how a multi-task formulation can be useful to carry out a joint estimation of all these parameters through regularization.

Source modeling. Using a volume segmentation of the MRI scan of each subject, the positions of potential sources are constructed as a set of coordinates uniformly distributed on the cortical surface of the gray matter. Moreover, synchronized currents in the apical dendrites of cortical pyramidal neurons are thought to be mostly responsible for MEG signals [26]. Therefore, the dipole orientations are usually constrained to be normal to the cortical surface. We model the current density as a set of focal current dipoles with fixed positions and orientations. The purpose of source localization is to infer their amplitudes. The ensemble of possible candidate dipoles forms the *source space*.

Forward modeling. Let n denote the number of sensors (EEG and/or MEG) and p the number of sources. Following Maxwell's equations, at each time instant, the measurements $\mathbf{B} \in \mathbb{R}^n$ are a linear combination of the current density $\mathbf{x} \in \mathbb{R}^p : \mathbf{B} = \mathbf{L}\mathbf{x}$. However, we observe noisy measurements $\mathbf{Y} \in \mathbb{R}^n$ given by:

$$\mathbf{Y} = \mathbf{B} + \varepsilon = \mathbf{L}\mathbf{x} + \varepsilon \ , \tag{1}$$

where ε is the noise vector. The linear forward operator $\mathbf{L} \in \mathbb{R}^{n \times p}$ is called the *leadfield* or *gain matrix*, and can be computed by solving Maxwell's equations using the Boundary element method [12]. Up to a whitening pre-processing step, ε can be assumed Gaussian distributed $\mathcal{N}(0, \sigma I_n)$.

Source localization. Source localization consists in solving in $\mathbf{x}$ the inverse problem (1) which can be cast as a least squares problem:

$$\mathbf{x}^\star = \operatorname*{arg\,min}_{\mathbf{x} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{L}\mathbf{x}\|_2^2 \ . \tag{2}$$

Since $n \ll p$, problem (2) is ill-posed and additional constraints on the solution $\mathbf{x}^\star$ are necessary. When analyzing evoked responses, one can promote source

configurations made of a few focal sources, e.g. using the ℓ_1 norm. This regularization leads to problem (3) called minimum current estimates (MCE), also known in the machine learning community as the Lasso [32].

$$\mathbf{x}^\star = \underset{\mathbf{x}\in\mathbb{R}^p}{\arg\min}\ \frac{1}{2n}\|\mathbf{Y}-\mathbf{L}\mathbf{x}\|_2^2+\lambda\|\mathbf{x}\|_1\ , \tag{3}$$

where $\lambda > 0$ is a tuning hyperparameter.

Common source space. Here we propose to go beyond the classical pipeline and carry out source localization jointly for S subjects. First, dipole positions (features) must correspond to each other across subjects. To do so, the source space of each subject is mapped to a high resolution average brain using morphing where the sulci and gyri patterns are matched in an auxiliary spherical inflating of each brain surface [8]. The resulting leadfields $\mathbf{L}^{(1)}, \ldots, \mathbf{L}^{(S)}$ have therefore the same shape $(n \times p)$ with aligned columns.

Multi-task framework. Jointly estimating the current density $\mathbf{x}^{(s)}$ of each subject s can be expressed as a multi-task regression problem where some coupling prior is assumed on $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)}$ through a penalty Ω:

$$\min_{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(S)}\in\mathbb{R}^p}\ \frac{1}{2n}\sum_{s=1}^{S}\|\mathbf{Y}^{(s)}-\mathbf{L}^{(s)}\mathbf{x}^{(s)}\|_2^2+\Omega(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(S)})\ . \tag{4}$$

Following the work of [15], we propose to define Ω using an UOT metric.

3 Minimum Wassertein Estimates

We start this section with background material on UOT. Consider the finite metric space (E, d) where each element of $E = \{1, \ldots, p\}$ corresponds to a vertex of the source space. Let $\mathbf{M}$ be the matrix where $\mathbf{M}_{ij}$ corresponds to the geodesic distance between vertices i and j. Kantorovich [16] defined a distance for normalized histograms (probability measures) on E. However, it can easily be extended to non-normalized measures by relaxing marginal constraints [4].

Marginal relaxation. Let $\mathbf{a}, \mathbf{b}$ be two normalized histograms on E. Assuming that transporting a fraction of mass $\mathbf{P}_{ij}$ from i to j is given by $\mathbf{P}_{ij}\mathbf{M}_{ij}$, the total cost of transport is given by $\langle \mathbf{P}, \mathbf{M}\rangle = \sum_{ij}\mathbf{P}_{ij}\mathbf{M}_{ij}$. Minimizing this total cost with respect to $\mathbf{P}$ must be carried out on the set of feasible transport plans with marginals $\mathbf{a}$ and $\mathbf{b}$. The (normalized) Wasserstein-Kantorovich distance reads:

$$\mathrm{WK}(\mathbf{a},\mathbf{b}) = \min_{\substack{\mathbf{P}\in{\mathbb{R}_+}^{p\times p}\\ \mathbf{P}\mathbb{1}=\mathbf{a},\mathbf{P}^\top\mathbb{1}=\mathbf{b}}} \langle \mathbf{P}, \mathbf{M}\rangle\ . \tag{5}$$

In practice, if $\mathbf{a}$ and $\mathbf{b}$ are positive and normalized current densities, $\mathrm{WK}(\mathbf{a}, \mathbf{b})$ will quantify the geodesic distance between their supports along the curved

geometry of the cortex. This property makes OT metrics adequate for assessing the proximity of functional patterns across subjects. To allow $\mathbf{a}, \mathbf{b}$ to be non-normalized, the marginal constraints in (5) can be relaxed using a KL divergence:

$$\min_{\mathbf{P} \in {\mathbb{R}_+}^{p \times p}} \langle \mathbf{P}, \mathbf{M} \rangle + \gamma \mathrm{KL}(\mathbf{P}\mathbb{1}|\mathbf{a}) + \gamma \mathrm{KL}(\mathbf{P}^\top \mathbb{1}|\mathbf{b}) \ , \tag{6}$$

where $\gamma > 0$ is a hyperparameter that enforces a fit to the marginals.

Entropy regularization. Entropy regularization was introduced by [5] to propose a faster and more robust alternative to the direct resolution of the linear programming problem (5). Formally, this amounts to minimizing the loss $\langle \mathbf{P}, \mathbf{M} \rangle - \varepsilon H(\mathbf{P})$ where $\varepsilon > 0$ is a tuning hyperparameter. This penalized loss function can be written: $\varepsilon \mathrm{KL}(\mathbf{P}, e^{-\frac{\mathbf{M}}{\varepsilon}})$ up to a constant [3]. Combining entropy regularization with marginal relaxation in (6), we get the unbalanced Wasserstein distance as introduced by [4]:

$$W(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in {\mathbb{R}_+}^{p \times p}} \varepsilon \mathrm{KL}(\mathbf{P}|e^{-\frac{\mathbf{M}}{\varepsilon}}) + \gamma \mathrm{KL}(\mathbf{P}\mathbb{1}|\mathbf{a}) + \gamma \mathrm{KL}(\mathbf{P}^\top \mathbb{1}|\mathbf{b}) \ , \tag{7}$$

Generalized Sinkhorn. Problem (7) can be solved as follows. Let $\mathbf{K} = e^{-\frac{\mathbf{M}}{\varepsilon}}$ and $\psi = \gamma/(\gamma + \epsilon)$. Starting from two vectors $\mathbf{u}, \mathbf{v}$ set to $\mathbb{1}$ and iterating the scaling operations $\mathbf{u} \leftarrow (\mathbf{a}/\mathbf{K}\mathbf{v})^\psi$, $\mathbf{v} \leftarrow \left(\mathbf{b}/\mathbf{K}^\top \mathbf{u}\right)^\psi$ until convergence, the minimizer of (7) can be computed as $P^\star = (\mathbf{u}_i \mathbf{K}_{ij} \mathbf{v}_j)_{i,j \in [\![p]\!]}$. This algorithm is a generalization of the Sinkhorn algorithm [18]. Since it involves matrix-matrix operations, it benefits from parallel hardware, such as GPUs.

Extension to $\mathbb{R}^p$. We extend next the Wasserstein distance to signed measures. We adopt a similar idea to what was suggested in [23,29,15] using a decomposition into positive and negative parts, $\mathbf{x}^{(s)} = \mathbf{x}^{(s)+} - \mathbf{x}^{(s)-}$ where $\mathbf{x}^{(s)+} = \max(\mathbf{x}^{(s)+}, 0)$ and $\mathbf{x}^{(s)-} = \max(-\mathbf{x}^{(s)+}, 0)$. For any vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^p$, we define the generalized Wasserstein distance as:

$$\widetilde{W}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def}}{=} W(\mathbf{a}^+, \mathbf{b}^+) + W(\mathbf{a}^-, \mathbf{b}^-) \ . \tag{8}$$

Note that $W(\mathbf{0}, \mathbf{0}) = 0$ (see [15] for a proof), thus on positive measures $\widetilde{W} = W$. For the sake of convenience, we refer to $\widetilde{W}$ in (8) by the Wasserstein distance, even though it does not verify indiscernability. In practice, this extension allows to compare current dipoles across subjects according to their polarity which could be either towards the deep or superficial layers of the cortex.

The MTW model. The multi-task Wasserstein model is the specific case of (4) with a penalty Ω promoting both sparsity and supports' proximity:

$$\Omega_{\mathrm{MTW}}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(S)}) \stackrel{\text{def}}{=} \mu \min_{\bar{\boldsymbol{x}} \in \mathbb{R}^p} \frac{1}{S} \sum_{s=1}^{S} \widetilde{W}(\mathbf{x}^{(s)}, \bar{\boldsymbol{x}}) + \lambda \|\mathbf{x}^{(s)}\|_1 \ , \tag{9}$$

where $\mu, \lambda \geq 0$ are tuning hyperparameters. The OT term in (9) can be seen as a spatial variance. Indeed, the minimizer $\bar{\boldsymbol{x}}$ corresponds to the Wasserstein barycenter with respect to the distance $\widetilde{W}$.

Algorithm 1 MWE algorithm

Input: σ_0, $\mu, \epsilon, \gamma, \lambda$ and cost matrix $\mathbf{M}$. data $(\mathbf{Y}^{(s)})_s (\mathbf{L}^{(s)})_s$.
Output: MWE: $(\mathbf{x}^{(s)})$, minimizers of (10).
repeat
  for $s = 1$ **to** S **do**
    Update $\mathbf{x}^{(s)+}$ with proximal coordinate descent to solve (12).
    Update $\mathbf{x}^{(s)-}$ with proximal coordinate descent to solve (12).
    Update $\sigma^{(s)}$ with (11).
  end for
  Update left marginals $\mathbf{m}^{(1)+}, \ldots, \mathbf{m}^{(S)+}$ and $\bar{\boldsymbol{x}}^{+}$ with generalized Sinkhorn.
  Update left marginals $\mathbf{m}^{(1)-} \ldots, \mathbf{m}^{(S)-}$ and $\bar{\boldsymbol{x}}^{-}$ with generalized Sinkhorn.
until convergence

Minimum Wasserstein Estimates. One of the drawbacks of MTW is that λ is common to all subjects. Indeed, the loss considered in MTW implicitly assumes that the level of noise is the same across subjects. Following the work of [25] on the smoothed concomitant Lasso, we propose to extend MTW by inferring the specific noise standard deviation $\sigma^{(s)}$ along with the regression coefficient $\mathbf{x}^{(s)}$ of each subject. This allows to scale the weight of the ℓ_1 according to the level of noise. The Minimum Wasserstein Estimates (MWE) model reads:

$$\min_{\substack{\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(S)} \in \mathbb{R}^p \\ \sigma^{(1)},\ldots,\sigma^{(S)} \in [\sigma_0, +\infty]}} \sum_{s=1}^{S} \frac{1}{2n\sigma^{(s)}} \|\mathbf{Y}^{(s)} - \mathbf{L}^{(s)}\mathbf{x}^{(s)}\|_2^2 + \frac{\sigma^{(s)}}{2} + \Omega_{\mathrm{MTW}}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(S)}) \ , \tag{10}$$

where σ_0 is a pre-defined constant. This lower bound constraint avoids numerical issues when $\lambda \to 0$ and therefore the standard deviation estimate also tends to 0. In practice σ_0 can be set for example using prior knowledge on the variance of the data or as a small fraction of the initial estimate of the standard deviation $\sigma_0 = \alpha \min_s \frac{\|\mathbf{Y}^{(s)}\|}{\sqrt{n}}$. In practice we adopt the second option and set $\alpha = 0.01$.

Algorithm. By combining (7), (8) and (10), we obtain an objective function taking as arguments $\big((\mathbf{x}^{(s)+})_s, (\mathbf{x}^{(s)-})_s, (\mathbf{P}^{(s)+})_s, (\mathbf{P}^{(s)-})_s, \bar{\boldsymbol{x}}^{+}, \bar{\boldsymbol{x}}^{-}, (\sigma^{(s)})_s\big)$. This function restricted to all parameters except $(\sigma^{(s)})_s$ is jointly convex [15]. Moreover, each $\sigma^{(s)}$ is only coupled with the variable $\mathbf{x}^{(s)}$. The restriction on every pair $(\mathbf{x}^{(s)}, \sigma^{(s)})$ is also jointly convex [25]. Thus the problem is jointly convex in all its variables. We minimize it by alternating optimization. To justify the convergence of such an algorithm, one needs to notice that the non-smooth ℓ_1 norms in the objective are separable [33]. The update with respect to each $\sigma^{(s)}$ is given by solving the first order optimality condition (Fermat's rule):

$$\sigma^{(s)} \leftarrow \frac{\|\mathbf{Y}^{(s)} - \mathbf{L}^{(s)}\mathbf{x}^{(s)}\|_2}{\sqrt{n}} \wedge \sigma_0 \ , \tag{11}$$

which also corresponds to the empirical estimator of the standard deviation when the constraint is not active. To update the remaining variables, we follow the same optimization procedure laid out in [15] and adapted to MWE in Algorithm 1. Briefly, let $\mathbf{m}^{(s)+} \stackrel{\text{def}}{=} \mathbf{P}^{(s)+}\mathbb{1}$ (resp. $\mathbf{m}^{(s)+} \stackrel{\text{def}}{=} \mathbf{P}^{(s)+}\mathbb{1}$), when minimizing with respect to one $\mathbf{x}^{(s)+}$ (resp. $\mathbf{x}^{(s)-}$), the resulting problem can be written (dropping the exponents for simplicity):

$$\min_{\mathbf{x}\in\mathbb{R}_+^p} \frac{1}{2n}\|\mathbf{Y}-\mathbf{L}\mathbf{x}\|_2^2 + \frac{\mu\gamma}{S}(\langle \mathbf{x}, \mathbb{1}\rangle - \langle \log(\mathbf{x}), \mathbf{m}\rangle) + \lambda\sigma\|\mathbf{x}\|_1 \ , \tag{12}$$

which can be solved using proximal coordinate descent [7]. Note that the additional inference of a specific $\sigma^{(s)}$ for each subject allows to scale the Lasso penalty depending on their particular level of noise. The final update with respect to $((\mathbf{P}^{(s)+})_s, (\mathbf{P}^{(s)-})_s, \bar{\boldsymbol{x}}^+, \bar{\boldsymbol{x}}^-)$ can be cast as two Wasserstein barycenter problems, carried out using generalized Sinkhorn iterations [4]. Note that we do not need to compute the transport plans $P^{(s)}$ since inferring every source estimate $\mathbf{x}$ only requires the knowledge of the left marginal $\mathbf{m} = \mathbf{P}\mathbb{1}$ which does not require storing $\mathbf{P}$ in memory.

4 Experiments

Benchmarks: Dirty models and Multi-level Lasso. As discussed in introduction, standard sparse source localization solvers are based on an ℓ_1 norm regularization, applied to the data of each subject independently. We use the independent Lasso estimator as a baseline. We compare MWE to the Group-Lasso estimator [37,2] which was proposed in this context to promote functional consistency across subjects [21]. It falls in the multi-task framework of (4) where the joint penalty is defined through an ℓ_{21} mixed norm $\|\mathbf{X}\|_{21} = \sum_{j=1}^p \sqrt{\sum_{s=1}^S \mathbf{x}_j^{(s)^2}}$ where $\mathbf{X} = (\mathbf{x}_j^{(s)})_{(j,s)} \in \mathbb{R}^{p\times S}$. We also evaluate the performance of more flexible block sparse models where only a fraction of the source estimates are shared across all tasks: Dirty models [14] and the multivel lasso [22]. In Dirty models source estimates are written as a sum of two parts which are penalized with different norms. One common to all subjects (penalty ℓ_{21}) and one specific for each subject (penalty ℓ_1). The Multi-level Lasso (MLL) [22] applies the same idea using instead a product decomposition and a Lasso penalty on both parts. We also compare MWE with MTW to evaluate the benefits of inferring noise levels adaptively.

Simulation data and MEG/fMRI datasets. We use the public dataset DS117 [36] which provides MEG, EEG and fMRI data of 16 healthy subjects to whom were presented images of famous, unfamiliar and scrambled faces. Using the MRI scan of each subject, we compute a source space and its associated leadfield comprising around 2500 sources per hemisphere [9]. Keeping only MEG gradiometer channels, we have $n = 204$ observations per subject.

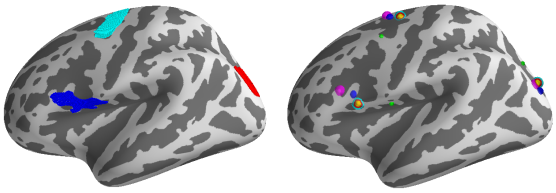

Fig. 1. Left: 3 labels from the aparc.a2009s parcellation. **Right:** Simulated activations for $S = 6$ subjects. Each color corresponds to a subject. Different radiuses are used to distinguish overlapping sources.

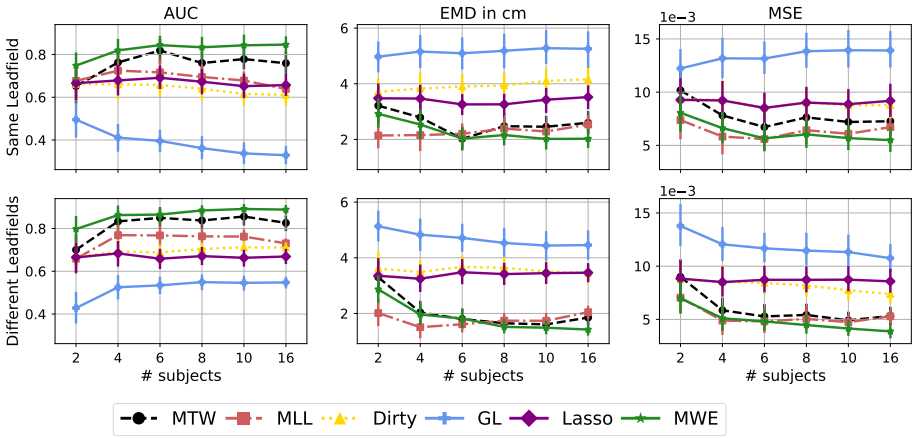

Fig. 2. Performance of different models over 30 trials in terms of AUC, EMD and MSE using the same leadfield for all subjects (randomly selected in each trial) (**top**) and specific leadfields (**bottom**).

For realistic data simulation, we use the actual leadfields from all subjects, yet restricted to the left hemisphere. We thus have 16 leadfields with $p = 2500$. We simulate an inverse solution $\mathbf{x}^s$ with q sources (q-sparse vector) by randomly selecting one source per label among q pre-defined labels using the *aparc.a2009s* parcellation of the Destrieux atlas. To model functional consistency, 50% of the subjects share sources at the same locations, the remaining 50% have sources randomly generated in the same labels (see Figure 1). Their amplitudes are taken uniformly between 20 and 30 nAm. Their sign is taken at random with a Bernoulli distribution (0.5) for each label. We simulate $\mathbf{Y}$ using the forward model with a variance matrix σI_n. We set σ so as to have an average signal-to-noise ratio across subjects equal to 4 ($\mathrm{SNR} \overset{\text{def}}{=} \sum_{s=1}^{S} \frac{\|\mathbf{L}^{(s)}\mathbf{x}^{(s)}\|}{S\sigma}$). We evaluate the performance of all models knowing the ground truth by comparing the best estimates in terms of three metrics: the mean squared error (MSE) to quantify accuracy in amplitude estimation, AUC and a generalized Earth mover distance (EMD) to assess supports estimation. We generalize the PR-AUC (Area under the curve Precision-recall) by defining $\mathrm{AUC}(\hat{\mathbf{x}}, \mathbf{x}^\star) = \frac{1}{2}\text{PR-AUC}(\hat{\mathbf{x}}^+, \mathbf{x}^{\star+}) + \frac{1}{2}\text{PR-AUC}(\hat{\mathbf{x}}^-, \mathbf{x}^{\star-})$ where PR-AUC is computed between the estimated coefficients and the true

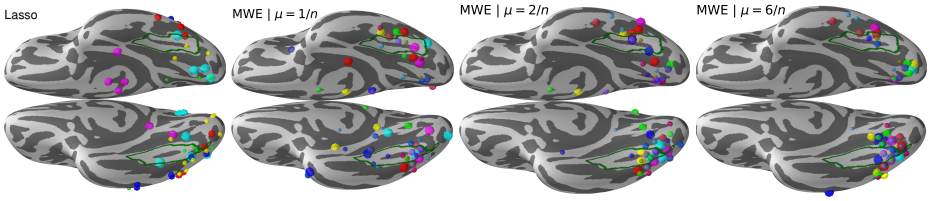

Fig. 3. Support of source estimates. Each color corresponds to a subject. Different radiuses are displayed for a better distinction of sources. The fusiform gyrus is highlighted in green. Increasing μ promotes functional consistency across subjects.

supports. We compute EMD between normalized values of sources: $\mathrm{EMD}(\hat{\mathbf{x}}, \mathbf{x}^\star) = \frac{1}{2}\mathrm{WK}(\frac{\hat{\mathbf{x}}^+}{\hat{\mathbf{x}}^+\mathbb{1}}, \frac{\mathbf{x}^{\star+}}{\mathbf{x}^{\star+}\mathbb{1}}) + \frac{1}{2}\mathrm{WK}(\frac{\hat{\mathbf{x}}^-}{\hat{\mathbf{x}}^-\mathbb{1}}, \frac{\mathbf{x}^{\star-}}{\mathbf{x}^{\star-}\mathbb{1}})$. Since $\mathbf{M}$ is expressed in centimeters, WK can be seen as an expectation of the geodesic distance between sources. The mean across subjects is reported for all metrics.

Simulation results. We set the number of sources to 3 and vary the number of subjects under two conditions: (1) using one leadfield for all subjects, (2) using individual leadfields. Each model is fitted on a grid of hyperparameters and the best AUC/MSE/EMD scores are reported. We perform 30 different trials (with different true activations and noise, different common leadfield for condition (1)) and report the mean within a 95% confidence interval in Figure 2.

Various observations can be made. The Group Lasso performs poorly – even compared to independent Lasso – which is expected since sources are not common for all subjects. Non-convexity allows MLL to be very effective with less than 2-4 subjects. Its performance yet degrades with more subjects. OT-based models (MWE and MTW) however benefit from the presence of more subjects by leveraging spatial proximity. They reduce the average error distance from 4 cm (Lasso) to less than 1 cm and reach an AUC of 0.9. One can also observe that the estimation of the noise standard deviation in the MTW model does improve performance. Finally, we can appreciate the improvement of multi-task models when increasing the number subjects, especially when using different leadfield matrices. We argue that the different folding patterns of the cortex across subjects lead to different dipole orientations thereby increasing the chances of source identification.

Results on MEG/fMRI data The fusiform face area specializes in facial recognition and activates around 170ms after stimulus [17,13]. To study this response, we perform MEG source localization using Lasso and MWE. We pick the time point with the peak response for each subject within the interval 150-200 ms after visual presentation of famous faces. For both models, we select the smallest ℓ_1 tuning parameter λ for which less than 10 sources are active for each subject. Figure 3 shows how UOT regularization favors activation in the ventral pathway

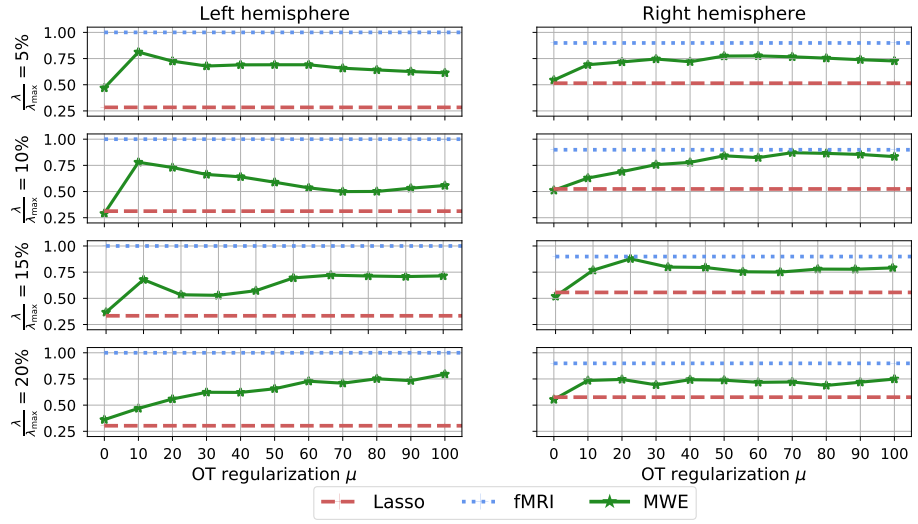

Fig. 4. Ratio of maximum absolute amplitude in the fusiform gyrus over maximum absolute amplitude in the hemisphere. The mean across the 16 subjects is reported for different ℓ_1 norm regularization weights λ.

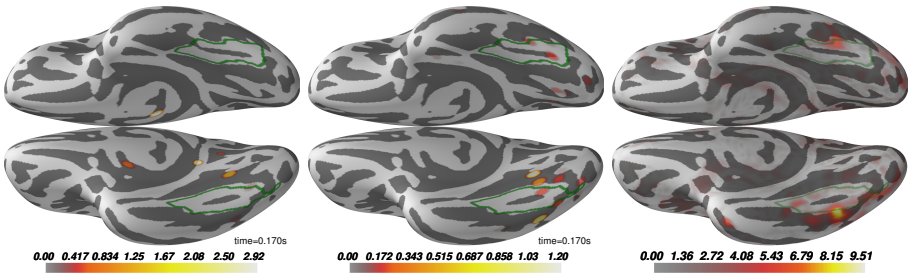

Fig. 5. Neural patterns of subject 2. Absolute amplitudes of MEG source estimates (in nAm) given by Lasso (**Left**) and MWE (**Middle**). Absolute values of fMRI Z-scores. (**Right**). The fusiform gyrus is highlighted in green.

of the visual cortex. The Lasso solutions in Figure 3 show significant differences between subjects. Since no ground truth exists, one could argue that MWE promotes consistency at the expense of individual signatures. To address this concern we compute the standardized fMRI Z-score of the conditions *famous vs scrambled faces*. We compare Lasso, MWE and fMRI by computing for each subject the ratio *largest value in fusiform gyrus / largest absolute value*. We report the mean across all subjects in Figure 4. Note that for all subjects, the fMRI Z-score reaches its maximum in the fusiform gyrus, and that MWE regularization leads to more agreement between MEG and fMRI. Figure 5 shows MEG with MWE and fMRI results for subject 2.

Conclusion

We proposed in this work a novel approach to promote functional consistency through a convex model defined using an Unbalanced Optimal Transport regularization. Using a public MEG and fMRI dataset, we presented experiments demonstrating that MWE outperform multi-task sparse models in both amplitude and support estimation. We have shown in these experiments that MWE can close the gap between MEG and fMRI source imaging by gathering data from different subjects.

Acknowledgements

This work was funded by the ERC Starting Grant SLAB ERC-YStG-676943 and a chaire d'excellence de l'IDEX Paris Saclay.

References

1. Ahlfors, S.P., Ilmoniemi, R.J., Hämäläinen, M.S.: Estimates of visually evoked cortical currents. Electroencephalography and Clinical Neurophysiology **82**(3), 225–236 (2018/11/20 1992)
2. Argyriou, A., Evgeniou, T., Pontil, M.: Multi-task feature learning. In: NIPS (2007)
3. Benamou, J., Carlier, G., Cuturi, M., Nenna, L., Peyré, G.: Iterative Bregman Projections For Regularized Transportation Problems. Society for Industrial and Applied Mathematics (2015)
4. Chizat, L., Peyré, G., Schmitzer, B., Vialard, F.X.: Scaling Algorithms for Unbalanced Transport Problems. arXiv:1607.05816 [math.OC] (2017)
5. Cuturi, M.: Sinkhorn Distances: Lightspeed Computation of Optimal Transport. NIPS (2013)
6. Dale, A.M., Liu, A.K., Fischl, B.R., Buckner, R.L., Belliveau, J.W., Lewine, J.D., Halgren, E.: Dynamic statistical parametric mapping. Neuron **26**(1), 55–67 (2000)
7. Fercoq, O., Richtárik, P.: Accelerated, parallel and proximal coordinate descent. SIAM Journal on Optimization **25**, 1997–2023 (2015)
8. Fischl, B., Sereno, M.I., Dale, A.M.: Cortical surface-based analysis: Ii: Inflation, flattening, and a surface-based coordinate system. NeuroImage **9**, 195 – 207 (1999), mathematics in Brain Imaging
9. Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Parkkonen, L., Hämäläinen, M.: MNE software for processing MEG and EEG data. NeuroImage **86** (10 2013)
10. 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)
11. Hämäläinen, M.S., Ilmoniemi, R.J.: Interpreting magnetic fields of the brain: minimum norm estimates. Medical & Biological Engineering & Computing **32**(1), 35–42 (Jan 1994)
12. Hämäläinen, M.S., Sarvas, J.: Feasibility of the homogeneous head model in the interpretation of neuromagnetic fields. Physics in Medicine and Biology **32**(1), 91 (1987)

13. 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 multi-subject and multi-modal integration. Frontiers in human neuroscience **5**, 76; 76–76 (08 2011)
14. Jalali, A., Ravikumar, P., Sanghavi, S., Ruan, C.: A Dirty Model for Multi-task Learning. NIPS (2010)
15. Janati, H., Cuturi, M., Gramfort, A.: Wasserstein regularization for sparse multi-task regression. Arxiv. preprint (2018)
16. Kantorovic, L.: On the translocation of masses. C.R. Acad. Sci. URSS (1942)
17. 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**(11), 4302–4311 (1997)
18. Knopp, P., Sinkhorn, R.: Concerning nonnegative matrices and doubly stochastic matrices. . Pacific Journal of Mathematics **1**(2), 343–348 (1967)
19. Kozunov, V.V., Ossadtchi, A.: Gala: group analysis leads to accuracy, a novel approach for solving the inverse problem in exploratory analysis of group MEG recordings. Frontiers in Neuroscience **9**, 107 (2015)
20. Larson, E., Maddox, R.K., Lee, A.K.C.: Improving spatial localization in meg inverse imaging by leveraging intersubject anatomical differences. Frontiers in Neuroscience **8**, 330 (2014)
21. Lim, M., Ales, J., Cottereau, B.M., Hastie, T., Norcia, A.M.: Sparse eeg/meg source estimation via a group lasso. PLOS (2017)
22. Lozano, A., Swirszcz, G.: Multi-level Lasso for Sparse Multi-task Regression. ICML (2012)
23. Mainini, E.: A description of transport cost for signed measures. Journal of Mathematical Sciences **181**(6), 837–855 (Mar 2012)
24. Massias, M., Fercoq, O., Gramfort, A., Salmon, J.: Generalized concomitant multi-task lasso for sparse multimodal regression. Proceedings of Machine Learning Research, vol. 84, pp. 998–1007. PMLR (09–11 Apr 2018)
25. Ndiaye, E., Fercoq, O., Gramfort, A., Leclère, V., Salmon, J.: Efficient smoothed concomitant lasso estimation for high dimensional regression. Journal of Physics: Conference Series **904**(1), 012006 (2017)
26. Okada, Y.: Empirical bases for constraints in current-imaging algorithms. Brain Topography p. 373–377
27. Owen, A.B.: A robust hybrid of lasso and ridge regression. Contemporary Mathematics **443**, 59–72 (2007)
28. Pascual-Marqui, R.: Standardized low-resolution brain electromagnetic tomography (sloreta): technical details. Methods Find Exp Clin Pharmacol **24**, D:5–12 (2002)
29. Profeta, A., Sturm, K.T.: Heat flow with dirichlet boundary conditions via optimal transport and gluing of metric measure spaces (2018)
30. Strohmeier, D., Bekhti, Y., Haueisen, J., Gramfort, A.: The iterative reweighted mixed-norm estimate for spatio-temporal MEG/EEG source reconstruction. IEEE Transactions on Medical Imaging **35**(10), 2218–2228 (Oct 2016)
31. Sun, T., Zhang, C.H.: Scaled sparse linear regression. Biometrika **99**, 879–898 (2012)
32. Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society **58**(1), 267–288 (1996)
33. Tseng, P.: Convergence of a block coordinate descent method for nondifferentiable minimization. J. Optim. Theory Appl. **109**(3), 475–494 (2001)
34. Uutela, K., Hämäläinen, M.S., Somersalo, E.: Visualization of magnetoencephalographic data using minimum current estimates. NeuroImage **10**(2), 173–180 (1999)

35. Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., Thirion, B.: Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In: Information Processing in Medical Imaging. vol. 6801, pp. 562–573. Springer (2011)
36. Wakeman, D., Henson, R.: A multi-subject, multi-modal human neuroimaging dataset. Scientific Data **2**(150001) (2015)
37. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society **68**(1), 49–67 (2006)