Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods

Alexandre Gramfort, Matthieu Kowalski, Matti Hämäläinen

Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods

Alexandre Gramfort[12], **Matthieu Kowalski**[3], **Matti Hämäläinen**[2]

[1] Parietal Project Team, INRIA Saclay-Ile de France, France.
[2] Department of Radiology, Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, MA.
[3] Laboratoire des Signaux et Systèmes
Univ. Paris-Sud
SUPELEC (C-4-20), Plateau de Moulon, France.

E-mail: alexandre.gramfort@inria.fr

Abstract.

Magneto- and electroencephalography (M/EEG) measure the electromagnetic fields produced by the neural electrical currents. Given a conductor model for the head, and the distribution of source currents in the brain, Maxwell's equations allow one to compute the ensuing M/EEG signals. Given the actual M/EEG measurements and the solution of this forward problem, one can localize, in space and in time, the brain regions than have produced the recorded data. However, due to the physics of the problem, the limited number of sensors compared to the number of possible source locations, and measurement noise, this inverse problem is ill-posed. Consequently, additional constraints are needed. Classical inverse solvers, often called Minimum Norm Estimates (MNE), promote source estimates with a small ℓ_2 norm. Here, we consider a more general class of priors based on *mixed-norms*. Such norms have the ability to *structure* the prior in order to incorporate some additional assumptions about the sources. We refer to such solvers as Mixed-Norm Estimates (MxNE). In the context of M/EEG, MxNE can promote spatially focal sources with smooth temporal estimates with a two-level ℓ_1/ℓ_2 mixed-norm, while a three-level mixed-norm can be used to promote spatially non-overlapping sources between different experimental conditions. In order to efficiently solve the optimization problems of MxNE, we introduce fast first-order iterative schemes that for the ℓ_1/ℓ_2 norm give solutions in a few seconds making such a prior as convenient as the simple MNE. Furhermore, thanks to the convexity of the optimization problem, we can provide optimality conditions that guarantee global convergence. The utility of the methods is demonstrated both with simulations and experimental MEG data.

Keywords: Magnetoencephalography, Electroencephalography, inverse problem, convex optimization, mixed-norms, structured sparsity, functional brain imaging

1. Introduction

Inverse problems are common in applied physics. They consist of estimating the parameters of a model from incomplete and noisy measurements. Examples are tomography which is a key technology in the field of medical imaging, or identifying the targets using sonars and radars. Blind source separation, which is an active topic of research in audio-processing, also falls in this category. In this contribution, we target the localization of the brain regions whose neural activations produce electromagnetic fields measured by Magnetoencephalography (MEG) and Electroencephalography (EEG), which we will refer to collectively as M/EEG. The *sources* of M/EEG are current generators classically modeled by current dipoles. Given a limited number of noisy measurements of the electromagnetic fields associated to neural activity, the task is to estimate the positions and amplitudes of the sources that have generated the signals. By solving this problem, M/EEG become noninvasive methods for functional brain imaging with a high temporal resolution.

Finding a solution to an inverse problem requires finding a good model for the observed data given the model parameters: this is called the forward problem. The task in the inverse problem is to infer the model parameters given the measurements. This is particularly challenging for an under-determined problem where the number of parameters to estimate is greater than the number of measurements. In such settings, several different source configurations can explain the experimental data and additional constraints are needed to provide a sound solution. In addition, the solution may be highly sensitive to noise in the measurements. Such problems are said to be *ill-posed*. Note that even over-determined problems can be ill-posed.

The linearity of Maxwell's equations implies that the signals measured by M/EEG sensors are linear combinations of the electromagnetic fields produced by all current generators. The linear operator, called *gain matrix* in the context of M/EEG, predicts the fields measured by the sensors due to a configuration of sources (Mosher et al. 1999). Computing the gain matrix accurately is particularly crucial for EEG, and involves complex numerical solvers (Kybic et al. 2005, Gramfort et al. 2010). In the M/EEG literature, solvers known as *distributed inverse solvers* essentially seek to invert the gain matrix. In practice, the *distribution* of estimated currents is defined over a discrete set of locations where are positioned current dipoles. The distribution is scalar valued when only their amplitudes are unknown, and vector valued when both amplitudes and orientations of the dipoles need to be estimated. The current generators are commonly assumed to lie on the cortex and are in practice fixed at the locations of the vertices of a cortical mesh (Dale & Sereno 1993). However, the number of generators largely exceeds the number of M/EEG sensors. To tackle this problem, one needs to use *a priori* knowledge on the characteristics of a realistic source configuration.

The priors most commonly used in the M/EEG community are based on the ℓ_2 norm, leading to what is known as the *Minimum Norm* (MN) inverse solver (Wang et al. 1992, Dale & Sereno 1993, Hämäläinen & Ilmoniemi 1994, Pascual-Marqui et al. 1994). This MN inverse solver leads to a linear solution: *i.e.,* Minimum Norm Estimates (MNE) are obtained by simple matrix multiplication (Tikhonov & Arsenin 1977). This makes the estimation extremely fast. However, ℓ_2-based solvers suffer from several limitations. Among which is the smearing of the even focal activations, often leading to overestimation of the extents of the activated areas. Also, they require to use a *two-step* approach where the MNE are post-processed to

obtain an interpretable picture of the spatio-temporal activation patterns (Pantazis et al. 2003, Gramfort et al. 2011). To address these limitations many alternatives to MNE have been proposed.

In the mid 90's, Matsuura *et al.* (Matsuura & Okabe 1995) proposed to regularize the amplitudes of the estimated sources with an ℓ_1 prior using an optimization procedure based on the simplex method. This approach was then later slightly modified by Uutela *et al.* (Uutela et al. 1999), who called the ℓ_1 penalized solutions minimum-current estimates (MCEs). About the same time, Gorodnitsky *et al.* proposed to use Iterative Reweighted Least Squares (IRLS) with the FOCUSS algorithm (Gorodnitsky et al. 1995) to approximate the solution that would be obtained with an ℓ_0 prior. Subsequently, (Phillips et al. 1997) proposed a Bayesian approach based of Markov random fields (MRF) and solved with mean field annealing. All these approaches are motivated by the fact that realistic source configurations are likely to have only a limited number of active sites. For example, only a few brain regions are typically significantly activated by a given cognitive task. The source configuration is said to be spatially *sparse*. This assumption has proved to be relevant for clinical applications (Huppertz et al. 2001) and also justifies dipole fitting which is currently the most widely used method in clinical settings.

However, the above approaches suffer from significant limitations. As they promote a sparse solution independently at each time instant, they fail to recover the time courses of cortical sources. In order to go beyond these limitations, there has been a growing interest for methods that promote spatially sparse solutions while taking into the temporal dynamics of the data (Phillips et al. 2005, Friston et al. 2008, Wipf & Nagarajan 2009, Valdés-Sosa et al. 2009, Haufe et al. 2008, Ou et al. 2009). While the methods proposed in (Phillips et al. 2005, Friston et al. 2008, Wipf & Nagarajan 2009) are related to sparse Bayesian learning where the problem boils down to the maximization of a non-convex cost function called the *model evidence*, (Haufe et al. 2008, Ou et al. 2009) address the problem using a sparsity-inducing prior that mixes both ℓ_1 and ℓ_2 norms. A ℓ_1 prior is used to promote a spatially sparse solution and a smooth ℓ_2 prior is used either for orientations (Haufe et al. 2008) or both time and orientations (Ou et al. 2009), leading to a convex optimization problem. A problem is convex when it consists in minimizing a convex function over a convex set (Boyd & Vandenberghe 2004). The main reason for the success of these solvers is the *structured sparsity* induced by the ℓ_{21} mixed-norm. Figure 1 illustrates source estimates with a simple ℓ_1 norm compared to a structured prior with a ℓ_{21} mixed-norm. The latter leads to a structured sparsity pattern while a simple ℓ_1 norm provides a scattered pattern that is not consistent with what is known about the sources. Here, the ℓ_{21} prior guarantees that the active source sites will be the same over the entire time interval of interest. Furthermore, grouping the temporal coefficients with an ℓ_{21} norm is a natural way to take into account the smooth temporal dynamics of M/EEG data. More generally, mixed-norm based priors offer a general way to take the structure of a problem into consideration. We call solutions obtained with such priors Mixed-Norm Estimates (MxNE). For an application to other brain imaging methods, see, for example (Varoquaux et al. 2010), where a two-level mixed norm was employed for the identification of brain networks using functional Magnetic Resonance Images (MRI) data.

Despite this growing interest, the use of sparsity-inducing priors is still limited to a small group of researchers. One possible reason is that solvers proposed so far are slow when applied to the analysis of real datasets. Another explanation is that

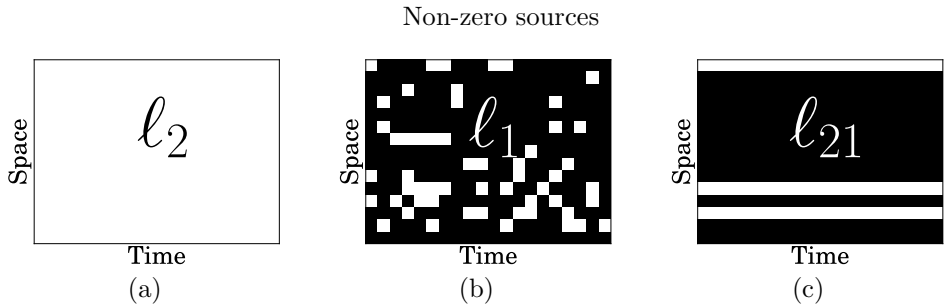


Figure 1. (a), (b) and (c) show in white the non-zero in the estimated source amplitudes obtained with the three norms. The non-zero coefficients are shown in white. While ℓ_2 yields only non-zero coefficients (all sources have a non-zero amplitude), ℓ_{21} promotes non-zero coefficients with a row structure (only a few sources have non-zero amplitude over the entire time interval of interest).

algorithms proposed so far are complex and difficult to implement. Indeed, while a basic minimum norm can be computed in a few hundreds of milliseconds, sparse inverse solvers as proposed in (Haufe et al. 2008, Ou et al. 2009) can take an hour to converge when realistic dimensions are used. A challenge is therefore to develop efficient optimization strategies that can solve the M/EEG inverse with such priors in a few seconds.

In the last few years, the machine learning and signal processing communities have devoted a lot of efforts into the improvement of the optimization methods that help to solve non-differentiable problems arising when considering sparse priors. One reason is that, under certain conditions, it has been proved that *sparsity* could enable the perfect resolution of ill-posed problems (Donoho 2006, Candès & Tao 2005). Among the list of algorithms that have been proposed are IRLS methods, similar to the FOCUSS algorithm, that consist in iteratively computing weighted MN solutions with weights updated after each iteration (Li 1993, Daubechies et al. 2008). The LARS-LASSO algorithm (Tibshirani 1996, Efron et al. 2004), which is a variant of the homotopy method from Osborne (Osborne et al. 2000), is an extremely powerful method for solving the ℓ_1 problem. Simple coordinate descent methods (Friedman et al. 2007) or blockwise coordinate descent, also called Block Coordinate Relaxation (BCR) (Bruce et al. 1998), are also possible strategies. Alternatively, methods based on projected gradients and proximity operators have been proposed (Daubechies et al. 2004, Combettes & Wajs 2005, Nesterov 2007*a*, Beck & Teboulle 2009). Even if some MxNE can be obtained efficiently, *e.g.*, with coordinate descent, the algorithms proposed in this contribution rely on proximal operators and gradient based methods as they provide a generic approach for all MxNE. They are also grounded on the current mathematical understanding and convergence properties of these solvers.

In this paper, we introduce efficient methods to compute mixed-norm estimates from M/EEG data. The three main contributions of this article are:

(i) We present the M/EEG inverse problem as a convex optimization problem and we explain how structured solutions can be promoted via appropriate priors based on mixed-norms.

(ii) We present in detail optimization methods that outperform in terms of convergence speed previously proposed algorithms and derive optimality conditions to control the convergence of the algorithm.

(iii) We then give two examples of MxNE that are relevant for M/EEG using two and three-level mixed-norms, including application to real data.

The first section of the paper provides the necessary background and notation. Mixed-norm estimates with two or three-level mixed-norms are introduced. The second contains the algorithmic and mathematical details of the optimization methods. The third section provides experimental results on real MEG data, demonstrating the efficiency and relevance of the proposed methods.

2. Mixed-norm estimates (MxNE)

In this section, we introduce inverse problems with linear forward models and more specifically the M/EEG inverse problem. We then define formally the one, two and three-level mixed-norms, explaining their influence on the solutions when used as priors. We explain how a three-level mixed-norm can be used for functional mapping and detail how the ℓ_{21} norm can be combined to obtain focal source estimates while promoting smooth time courses.

2.1. Framework and notation

Solving an inverse problem consists of estimating one or more unknown signals from observations, typically incomplete and noisy. When considering linear models, the observations, also called *measurements*, are linear combinations of the signals, also called *sources*. The linear relationship between the sources and the measurements, of this model, also called the *forward model*, is commonly derived from the physics of the problem.

Distributed source models in M/EEG use the individual anatomical information derived from MRI (Dale & Sereno 1993). The putative source locations can be then restricted to the brain volume or to the cortical mantle. Due to the linearity of the forward problem, each source adds its contribution independently to the measured signal. We focus here on source models where one dipole with a known orientation is positioned at each location. Source estimates are the amplitudes of the dipoles. Such models are known as *fixed orientation*. The framework however holds also in the *free orientation* case where three dipoles share a same spatial location. In this case both amplitudes and orientations need to be estimated.

The measurements $M \in \mathbb{R}^{N\times T}$ (N number of sensors and T number of time instants) are obtained by multiplying the source amplitudes $X \in \mathbb{R}^{S\times T}$ (S number of dipoles) by a forward operator $G \in \mathbb{R}^{N\times S}$, *i.e.*, $M = GX$. In addition, the measurements are corrupted by an additive noise E:

$$M = GX + E \ .$$

In the context of M/EEG, N lies between 50 for EEG only and 400 for M/EEG combined measurements, while S lies between 5000 and 50000 depending on the precision of the source model considered.

A classical approach to estimate X given M consists in introducing a cost function $\mathcal{F}$ whose minimum provides the solution:

$$X^* = \arg\min_X \mathcal{F}(X) = \arg\min_X \left(f_1(X) + \lambda f_2(X)\right) \ . \tag{1}$$

The cost function is composed of two terms:

- A data-fit term, f_1, that quantifies how well the estimated sources match the measured data. This term takes into account the characteristics of the measurement noise.
- A regularization term, *a.k.a.*, penalty term or prior, denoted f_2, that is used to introduce *a priori* knowledge on the solution. This term is mandatory to render the solution unique when considering ill-posed problems.

These two terms are balanced by the regularization parameter $\lambda > 0$. In the context of M/EEG, f_2 can be directly a function of the source amplitudes or introduce a regularization matrix like a spatial Laplacian D leading to a regularization term of the form $f_2(DX)$ (Pascual-Marqui et al. 1994).

This contribution focuses on cases where f_1 and f_2 are convex functions (Boyd & Vandenberghe 2004). As will be detailed later, the convexity of $\mathcal{F}$ is a key assumption that allows to obtain globally optimal solutions which are independent of the initialization of the solver. As will be discussed in Section 3, this assumption allows us to employ very efficient optimization procedures whose mathematical properties in terms of complexity and convergence rate are fully understood. Another benefit of convexity observed in practice is the increased stability of the solutions in the presence of noise.

In M/EEG, $f_1(X)$ is usually the squared ℓ_2 norm of the residual $R = M - GX$:

$$f_1(X) = \frac{1}{2}\|M - GX\|_2^2 = \frac{1}{2}\|R\|_2^2 = \frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{T} R_{n,t}^2 \ . \tag{2}$$

The smaller is the residual, the better the sources explain the data. The minimization (2) is equivalent to finding a maximum likelihood estimate under the assumption that the additive noise E is Gaussian, *i.e.*, $E \sim \mathcal{N}(0, I)$. In practice, the M/EEG noise is not white but one can estimate the noise-covariance matrix, which can be employed in whitening, either from empty-room data or from periods of actual brain signals void of data of interest (Hansen et al. 2010). Note that what is called here ℓ_2 norm is in fact the Frobenius matrix norm, since, generally, $T > 1$. Note also that the factor $\frac{1}{2}$ is included for convenience in the derivation of the optimization methods.

The problem therefore reads:

$$X^* = \arg\min_X \left(\frac{1}{2}\|M - GX\|_2^2 + \lambda f_2(X)\right) \ , \lambda \in \mathbb{R}_+ \ . \tag{3}$$

The variance of each of the uncorrelated whitened signals is unity, which can help to set the λ parameter. Indeed, assuming that E is Gaussian, the expected value of $\|E\|_2^2$ is NT. It suggests that λ should be chosen such that $\|M - G\hat{X}\|_2^2 \approx NT$. This is known as the *discrepancy principle* (Morozov 1966). It will be used in Section 4.

We now proceed to discuss suitable priors f_2 for M/EEG.

2.2. The $\ell_{\mathbf{w};p}$ norm

The most common choice for $f_2(X)$ is the squared ℓ_2 (Frobenius) norm of the sources amplitudes X. This lies in the category of ℓ_p norms (Matsuura & Okabe 1995, Wagner et al. 1996, Uutela et al. 1999, Gorodnitsky et al. 1995) that work on a time-by-time basis. Throughout this paper, we are often interested in estimating $X \in \mathbb{R}^{S\times T}$ or $X \in \mathbb{R}^{S\times KT}$ when considering $K > 1$ datasets with a joint estimation. When

considering $\ell_{\mathbf{w};p}$ norms applied to such X, we consider X as a set of coefficients that can be seen as a vector $\mathbf{x} \in \mathbb{R}^P$ where $P = ST$ or $P = SKT$.

Definition 1 ($\ell_{\mathbf{w};p}$ norm) *Let $\mathbf{x} = (x_1, \ldots, x_P) \in \mathbb{R}^P$ and $\mathbf{w} = (w_1, \ldots, w_P) \in \mathbb{R}_+^P$, some positive weights. Let $p \geq 1$. Then the $\ell_{\mathbf{w};p}$ norm of $\mathbf{x}$ is:*

$$\|\mathbf{x}\|_{\mathbf{w};p} = \left(\sum_{s=1}^{P} w_s |x_s|^p \right)^{1/p} ,$$

which is known to be convex for $p \geq 1$ and strictly convex for $p > 1$ (Boyd & Vandenberghe 2004).

The reason for introducing weights in the ℓ_p norms is due to the fact the columns $(G_{\cdot s})_s$ of M/EEG forward operators are not normalized. The closer the dipole s is to the head surface, the bigger $\|G_{\cdot s}\|_2$. This implies that a naïve inverse procedure would favor dipoles close to the head surface. In the M/EEG literature, this is known as the "depth bias" (Lin et al. 2006). Using a weighted norm is a way to address this problem.

2.2.1. The $\ell_{\mathbf{w};2}$ norm The squared ℓ_2 norm when used both for the data-fit and the penalty term f_2 is known as MNE in the M/EEG literature. The optimization problem reads:

$$\hat{X} = \arg\min_X \left(\frac{1}{2}\|M - GX\|_2^2 + \frac{\lambda}{2}\|X\|_{\mathbf{w};2}^2 \right) . \tag{4}$$

This corresponds to a penalized maximum likelihood estimate assuming the sources are Gaussian and normally distributed, with a diagonal covariance matrix (Wipf & Nagarajan 2009). By using such a prior, one spreads the energy of the solution over all the sources. In the context of M/EEG source localization, it leads to activation maps where every brain region has a non-zero amplitude (see Fig. 1-a) and where the extent of active regions is often over-estimated. Solvers based on $\ell_{\mathbf{w};2}$ penalty fail to recover high spatial frequencies. In order to avoid this, we can employ a prior that promotes spatially sparse solutions where the data will be explained by a few sources. Keeping f_2 convex, this is can be done with an ℓ_1 norm.

2.2.2. The $\ell_{\mathbf{w};1}$ norm The ℓ_1 norm promotes sparse solutions, which is a strong hypothesis: the solution should only have a small number of non-zero coefficients.

While sparsity can be a valuable assumption in some applications, *e.g.,* denoising (Kowalski & Torrésani 2008, Févotte et al. 2008, Dupé et al. 2009), it can also lead to unrealistic solutions in other applications, *e.g.,* blind source separation (Bobin et al. 2008), coding (Daudet et al. 2004), and for M/EEG. Indeed, as illustrated in Fig. 1, an ℓ_1 prior should be used with some caution when performing M/EEG source imaging with temporally correlated data. Such a prior, which estimates the active sources independently at each time instant, will very likely fail to recover the smooth temporal dynamics of a realistic source. To address this limitation a solution recently proposed in the literature estimates the sources for all time instants jointly after introducing a coupling between the estimates (Ou et al. 2009). This is achieved using a penalty based on a two-level mixed-norm.

2.3. Two-level mixed-norms

In order to define the two-level mixed-norm, we must consider a sequence indexed by a double index $(g, m) \in \mathbb{N}^2$ such that $(\mathbf{x}) = (x_{g,m})_{(g,m)\in\mathbb{N}^2}$. One can then consider the two canonical subsequences $(x_{g,.}) = (x_{g,1}, x_{g,2}, \ldots)$ for a fixed g, and $(x_{.,m}) = (x_{1,m}, x_{2,m}, \ldots)$ for a fixed m. This labeling convention introduces a grouping of the coefficients and will be utilized below.

Definition 2 (Two-level mixed-norms) *Let $\mathbf{x} \in \mathbb{R}^P$ be indexed by a double index (g, m) such that $\mathbf{x} = (x_{g,m})$.*

Let $p, q \geq 1$, and $\mathbf{w} \in \mathbb{R}^P_{+,*}$ be a sequence of strictly positive weights labeled by a double index $(g, m) \in \mathbb{N}^2$. We call mixed-norm of $\mathbf{x} \in \mathbb{R}^P$, the norm $\ell_{\mathbf{w};p,q}$ defined by

$$\|\mathbf{x}\|_{\mathbf{w};p,q} = \left(\sum_g \left(\sum_m w_{g,m} |x_{g,m}|^p \right)^{q/p} \right)^{1/q} .$$

Cases $p = +\infty$ and $q = +\infty$ are obtained by replacing the corresponding norm by the supremum.

The two indices g and m can be interpreted as a hierarchy of the coefficients. The double indexing needed by the definition of mixed-norms allows to consider coefficients by groups. Coefficients are indeed distinguished between groups which are *blind* to each other, and the coefficients that belong to a same group are correlated. With the notation above, the g index can be seen as the "group index" and the m index as the "membership" index. Mixed-norms are then a practical way to induce explicitly a coupling between coefficients, instead of the independence hypothesis behind the ℓ_p norms. Hence, mixed-norms allow to promote some structures that have been observed in real signals. Properties of such norms, convexity in particular, enable the use of efficient optimization strategies.

In order to illustrate this, let us consider the use of the $\ell_{\mathbf{w};21}$ norm in the problem (3). The $\ell_{\mathbf{w};21}$ norm defined on a matrix $X \in \mathbb{R}^{S\times T}$, and with weights $\mathbf{w}$ depending only on the space index s (t indexes time), is given by:

$$\|X\|_{\mathbf{w};21} = \sum_s \sqrt{\sum_t w_s X_{s,t}^2} .$$

This corresponds to the sum of the ℓ_2 norm of the lines. As a consequence, an estimation of X given by the minimization of Eq. (3) is sparse through the lines, *i.e.,* all the coefficients of a line of X are either jointly nonzero, or all set to zero (see Fig. 1-c). Such a behavior will become more explicit with the definition of the so called *proximity operator*, see Section 3.2. This approach, proposed earlier for M/EEG (Ou et al. 2009), avoids the irregular time series obtained with a simple ℓ_1 norm. Note that the general formulation in Definition 2 using (g, m) covers the case with sources having unconstrained orientations. In this case g indexes each spatial location which contains three dipoles (Haufe et al. 2008, Ou et al. 2009).

The two-level mixed-norms were introduced during the 60's in (Benedek & Panzone 1961). These norms were then studied more formally in the context of Besov and modulation spaces (Samarah & Salman 2006, Feichtinger 2006, Rychkov 1999, Grochenig & Samarah 2000). Also see (Kowalski 2009) and (Kowalski & Torrésani 2009), who introduced the use of the ℓ_{12} norm under the name *Elitist-Lasso*.

2.4. Three-level mixed-norms

In this section, we are interested by models where sources X can be indexed by three indices. In the context of M/EEG, these three indices can correspond to the spatial location, the experimental conditions, and the time. For example, for somatosensory data of Section 4, an experimental condition corresponds to the finger that is stimulated. Let us denote this new index by k. The sources, with elements indexed by (s, k, t), are denoted by $X \in \mathbb{R}^{S \times KT}$ (K concatenated datasets) or simply $\mathbf{x} \in \mathbb{R}^P$ with $P = SKT$. Using this notation we can define a three-level mixed norm.:

Definition 3 (Three-level mixed-norms) *Let $\mathbf{x} \in \mathbb{R}^P$ be indexed by a triple index (s, k, t) such that $\mathbf{x} = (x_{s,k,t})$. Let $p, q, r \geq 1$ and $\mathbf{w} \in \mathbb{R}^P_{+,*}$ a sequence of strictly positive weights. We call mixed norm of $\mathbf{x}$ the norm $\ell_{\mathbf{w};p,q,r}$ defined by*

$$\|\mathbf{x}\|_{\mathbf{w};pqr} = \left(\sum_{s=1}^{S} \left(\sum_{k=1}^{K} \left(\sum_{t=1}^{T} w_{s,k,t} |x_{s,k,t}|^p \right)^{q/p} \right)^{r/q} \right)^{1/r} .$$

Cases $p = +\infty$, $q = +\infty$ and $r = +\infty$ are obtained by replacing the corresponding norm by the supremum.

The inverse problem then reads:

$$X^* = \arg\min_X \left(\frac{1}{2} \|M - GX\|_2^2 + \frac{\lambda}{r} \|X\|^r_{\mathbf{w};pqr} \right) , \lambda \in \mathbb{R}_+ . \tag{5}$$

For our application, we will use the $\ell_{\mathbf{w};212}$ mixed-norm. Note that the ℓ_2, ℓ_{12} and ℓ_{21} norms are special cases of the latter norm. Indeed ℓ_2 is obtained by setting $K = 1$, ℓ_{21} by setting $S = 1$ and ℓ_{12} by setting $T = 1$. This suggests that an optimization procedure for the ℓ_{212} norm readily works for both ℓ_{12} and ℓ_{21} norms. Also, it can be observed that $\ell_{\mathbf{w};221}$ is equivalent to $\ell_{\mathbf{w};21}$ after grouping conditions as well as time instants. By doing so one imposes the active sources to be common between all experimental conditions.

With the ℓ_1 norm to penalize the experimental conditions, while keeping the ℓ_2 norm on other indices, source estimates with non-zero activations for few conditions are promoted. By doing so, one penalizes the overlap between the active regions for the different conditions. With the somatosensory example, such a mixed-norm promotes activations where a given spatial location is active only for one, or at least few, experimental conditions. By definition, this norm corresponds to the *a priori* information that the stimulation of the different fingers leads to brain activations at different cortical locations, see Section 4.

3. Algorithms

The algorithms we employ are first-order methods that fit in the same category as the iterative thresholding procedures proposed in (Daubechies et al. 2004) for the ℓ_1 penalty. We extend them to problems where f_2 is a convex mixed-norm (Kowalski 2009), which, as explained in Section 2, can take into account the specific characteristics of, e.g., M/EEG source localization. The properties of such algorithms are based on recent mathematical results (Combettes & Wajs 2005).

Let us first introduce the notion of proximity operator, *a.k.a.*, proximal operator (Moreau 1965):

Definition 4 (Proximity operator) *Let $\phi : \mathbb{R}^P \to \mathbb{R}$ be a proper convex function. The proximity operator associated to ϕ and $\lambda > 0$, denoted by $\text{prox}_{\lambda\phi} : \mathbb{R}^P \to \mathbb{R}^P$ reads:*

$$\text{prox}_{\lambda\phi}(\mathbf{y}) = \underset{\mathbf{x}\in\mathbb{R}^P}{\arg\min}\ \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \lambda\phi(\mathbf{x})\ . \tag{6}$$

3.1. Iterative proximal gradient methods

In a nutshell, proximal gradient methods are a natural extension of gradient-based techniques when the objective function to minimize has an amenable non-smooth part. Such procedures based on iterative thresholding and more generally on projected gradients require that the cost function (1) meets the following hypothesis (Combettes & Wajs 2005):

- f_1 is a proper convex function whose gradient is Lipschitz continuous: $\exists L \in \mathbb{R}_+$ such that $\|\nabla f_1(\mathbf{x}) - \nabla f_1(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|$ for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^P$. L is called the Lipschitz constant.
- f_2 is a proper convex function (not necessarily differentiable).

  In our case, the gradient of the data-fit (2) is Lipschitz continuous. It reads:

$$\nabla f_1(X) = -G^T(M - GX)\ ,$$

and its Lipschitz constant is given by $L = \|G^TG\|$ (spectral norm which corresponds to the largest singular value).

The "simplest" iterative scheme to minimize (1) given in Algorithm 1 is called Iterative soft-thresholding and sometimes referred to as Forward-Backward or Landweber iterations (Combettes & Wajs 2005).

Algorithm 1: ISTA (Iterative shrinkage/thresholding algorithm)

Initialization: Let $X^{(0)} \in \mathbb{R}^{S\times KT}$ (for example $\mathbf{0}$).
repeat
  $X^{(k+1)} = \text{prox}_{\mu\lambda\ f_2}\left(X^{(k)} + \mu G^T(M - GX^{(k)})\right)$, with $0 < \mu < \frac{2}{L}$.
until *convergence*;

The idea is to alternate the minimization over f_1 using a small gradient step and the computation of the proximal operator associated with f_2. As the proximal operator can be seen as a generalized projection, this algorithm is a generalized iterative projected gradient method (see (Combettes & Wajs 2005) for a proof of convergence).

Unfortunately, this algorithm may converge rather slowly. It has been proved that its convergence rate is $\mathcal{O}(1/k)$, where k is the number of iterations

$$\exists C > 0, \forall k \text{ such that } \|\mathcal{F}(X^{(k)}) - \mathcal{F}(X^*)\| \leq \frac{C}{k}\ .$$

To improve the convergence speed, at least two accelerated projected gradient schemes whose convergence speed is $\mathcal{O}(1/k^2)$ have been proposed (Nesterov 2007*b*, Weiss 2008, Beck & Teboulle 2009). The FISTA (Fast Iterative shrinkage/thresholding algorithm) algorithm (Beck & Teboulle 2009) is one of them. It's a small modification of ISTA that takes into account the previous descent direction. It is a two-step approach. Note

that a classical example of a multi-step approach is the conjugate gradient algorithm used to solve positive definite linear systems. More details on these approaches can be found in (Tseng 2010).

Algorithm 2: FISTA

Initialization: $X^{(0)} \in \mathbb{R}^{S\times KT}$, $Z^{(1)} = X^{(0)}$, $\tau^{(1)} = 1$, $k = 1$, $0 < \mu < \frac{1}{L}$ **repeat**

$X^{(k)} = \text{prox}_{\mu\lambda f_2}\left(Z^{(k)} + \mu G^T(M - GZ^{(k)})\right)$

$\tau^{(k+1)} = \frac{1+\sqrt{1+4\tau^{(k)2}}}{2}$

$Z^{(k+1)} = X^{(k)} + \frac{\tau^{(k)}-1}{\tau^{(k)}}(X^{(k)} - X^{(k-1)})$

until *convergence*;

In order to tackle the optimization problem (3), one just needs to know how to compute $\text{prox}_{\mu\lambda f_2}$ where f_2 is a mixed-norm presented in Section 1.

3.2. Proximity operators corresponding to mixed-norms

The following proposition gives details of the proximity operators associated with the mixed-norms presented in Section 1. It corresponds to the solutions of (6) when ϕ is a mixed-norm.

Proposition 1 (Proximal operators for MxNE) *Let* $\mathbf{x} \in \mathbb{R}^P$ *and* $\mathbf{y} \in \mathbb{R}^P$. *Let* $\mathbf{w} \in \mathbb{R}_+^{*P}$ *be a vector of weights.*

ℓ_2 ***norm*** *Let* $\mathbf{x}$ *be indexed by* s. *The proximity operator associated to the squared* ℓ_2 *norm is given by* $\mathbf{x} = \text{prox}_{\lambda\|.\|^2_{\mathbf{w},2}}(\mathbf{y})$ *where* $\mathbf{x}$ *reads coordinate by coordinate:*

$$x_s = \frac{y_s}{1 + \lambda w_s} \ .$$

ℓ_1 ***norm*** *Let* $\mathbf{x}$ *be indexed by* s. *The proximity operator associated to the* ℓ_1 *norm is given by* $\mathbf{x} = \text{prox}_{\lambda\|.\|_{\mathbf{w},1}}(\mathbf{y})$ *where* $\mathbf{x}$ *reads coordinate by coordinate:*

$$x_s = \frac{y_s}{|y_s|}\left(|y_s| - \lambda w_s\right)^+ \ .$$

The function $(\cdot)^+$ *is defined as* $(a)^+ = \max(a, 0)$ *and we use the convention* $0/0 = 0$. *The proximity operator for the* ℓ_1 *norm is known as "soft-thresholding".*

ℓ_{21} ***norm*** *Let* $\mathbf{x}$ *be indexed by* (s, t). *Let us consider a vector of weights used to weight each group. The proximity operator associated to the* ℓ_{21} *norm is given by* $\mathbf{x} = \text{prox}_{\lambda\|.\|_{\mathbf{w},21}}(\mathbf{y})$ *where* $\mathbf{x}$ *reads for each coordinate:*

$$x_{s,t} = y_{s,t}\left(1 - \frac{\lambda\sqrt{w_s}}{\|\mathbf{y}_s\|_2}\right)^+ \ ,$$

where $\mathbf{y}_s$ *is the vector formed by the coefficients indexed by* s.

ℓ_{12} ***norm*** *Let* $\mathbf{x}$ *be indexed by* (s, k). *Let us consider* $\mathbf{w}$ *a vector of weights used to weight each group. Let* $r_{s,k}$ *be defined such that* $r_{s,k} \stackrel{\text{def}}{=} y_{s,k}/w_{s,k}$. *For each* s, *let the*

indexing denoted by k'_s *be defined such that* $\forall k'_s, r_{s,k'_s+1} \leq r_{s,k'_s}$. *Let the index* K_s *be such that:*

$$\lambda \sum_{k'_s=1}^{K_s} w_{s,k'_s}^2 \left(r_{s,k'_s} - r_{s,K_s}\right) < r_{s,K_s} \leq \lambda \sum_{k'_s=1}^{K_s+1} w_{s,k'_s}^2 \left(r_{s,k'_s} - r_{s,K_s}\right)$$

The proximal operator $\mathbf{x} = \text{prox}_{\lambda\|.\|_{\mathbf{w},12}}(\mathbf{y})$ *is given coordinate by coordinate:*

$$x_{s,k} = \frac{y_{s,k}}{|y_{s,k}|} \left(|y_{s,k}| - \frac{\lambda}{1 + \lambda K_{\mathbf{w}_s}} \sum_{k'_s=1}^{K_s} y_{s,k'_s} \right)^+ ,$$

where $K_{\mathbf{w}_s} = \sum_{k'_s=1}^{K_s} w_{s,k'_s}^2$.

This proposition shows the effect of such proximal operators on their inputs. For ℓ_2, it is a simple weighting. The associated cost function being differentiable, a proof is obtained simply by computing the derivative with respect to each x_s. For ℓ_1 it is a thresholding of all the coefficients independently (see (Donoho 1995) for a proof). As a result, some coefficients are set to zero which reflects the sparsity obtained with such a penalty. With the ℓ_{21} norm, a group is globally set to zero depending on its norm. A coefficient is non-zero only if the norm of the group it belongs to is large enough. If groups are formed by rows then the ℓ_{21} prior promotes a row structured sparsity pattern as illustrated in Fig. 1-c. For completeness, a derivation of this proximal operator is given in Appendix A. Note that such results have been previously obtained like in (Kowalski 2009). However, the later work does not address the weighted case.

From Proposition 1, the proximal operators associated to any mixed-norm combining ℓ_1 and ℓ_2 norms can be derived. This is in particular the case of the ℓ_{212} norm for which the proximal operator is given by the following proposition.

Proposition 2 (Proximal operator associated to the $\ell_{\mathbf{w};212}$ norm) *Let* $\mathbf{y} \in \mathbb{R}^P$ *be indexed by* (s,k,t). *Let* $\mathbf{w} \in \mathbb{R}^P$ *be a vector of positive weights such that* $\forall t,\ w_{s,k,t} = w_{s,k}$. *Let us define* $[\mathbf{y}_{s,k}] \stackrel{\text{def}}{=} \sqrt{w_{s,k} \sum_t y_{s,k,t}^2}$ *and* $r_{s,k} \stackrel{\text{def}}{=} [\mathbf{y}_{s,k}]/w_{s,k}$. *For each* s, *let the indexing denoted by* k'_s *be defined such that* $\forall k'_s, r_{s,k'_s+1} \leq r_{s,k'_s}$. *Let the index* K_s *be such that:*

$$\lambda \sum_{k'_s=1}^{K_s} w_{s,k'_s} \left(r_{s,k'_s} - r_{s,K_s}\right) < r_{s,K_s} \leq \lambda \sum_{k'_s=1}^{K_s+1} w_{s,k'_s} \left(r_{s,k'_s} - r_{s,K_s}\right) \ .$$

Then, $\mathbf{x} = \text{prox}_{\lambda\|.\|^2_{\mathbf{w};212}}(\mathbf{y})$ *is given, for each coordinate* (s,k,t), *by:*

$$x_{s,k,t} = y_{s,k,t} \left(1 - \frac{\lambda\sqrt{w_{s,k}}}{1 + \lambda K_{\mathbf{w}_s}} \frac{\sum_{k'_s=1}^{K_s} [\mathbf{y}_{s,k'_s}]}{\|\mathbf{y}_{s,k}\|_2} \right)^+ ,$$

where $K_{\mathbf{w}_s} = \sum_{k_s=1}^{K_s} w_{s,k_s}$.

A proof of this proposition is given in Appendix B.

Having established the minimization procedures for MxNE, we need next to test for the convergence and the optimality of the current iterate $X^{(k)}$.

3.3. Optimality conditions and stopping criterion

When the cost function $\mathcal{F}$ to be optimized is smooth, a natural optimality criterion is obtained by checking whether the gradient is small: $\|\nabla\mathcal{F}(X^{(k)})\| < \varepsilon$. Unfortunately, this approach does not apply to the non-differentiable cost-functions involving ℓ_1 norms.

An answer for convex problems more generally consists of computing, if possible, a *duality gap*. For a subset of these problems the Slater's conditions apply and, consequently, the gap at the optimum proves to be zero (Boyd & Vandenberghe 2004). Computing the gap starts by deriving a dual formulation of the original problem, also called the *primal* problem. The duality gap is defined as the difference between the minimum of the primal cost function $\mathcal{F}_p$ and the maximum of the dual cost function $\mathcal{F}_d$. For a value of $X^{(k)}$ of the primal variable at iteration k, if one can exhibit a dual variable $Y^{(k)}$, the duality gap $\eta^{(k)}$ is defined as:

$$\eta^{(k)} = \mathcal{F}_p(X^{(k)}) - \mathcal{F}_d(Y^{(k)}) \ \geq 0$$

At the optimum (corresponding to X^*), if the $Y^{(k)}$ associated to $X^{(k)}$ is properly chosen, $\eta^{(k)}$ is 0. By exhibiting a pair $(X^{(k)}, Y^{(k)})$ one can guarantee that: $\|\mathcal{F}_p(X^{(k)}) - \mathcal{F}_p(X^*)\| \leq \|\mathcal{F}_p(X^{(k)}) - \mathcal{F}_d(Y^{(k)})\|$. A good stopping criteria is therefore given by $\eta^{(k)} < \varepsilon$. The solution meeting this condition is said to be ε-optimal. The challenge in practice is to find an expression for $\mathcal{F}_d$ and to be able to associate a "good" Y to a given X for the problems of the form (3).

We now give a general answer for the class of problems detailed in this contribution. The solution is derived from the Fenchel-Rockafellar duality theorem (Rockafellar 1972) which leads to the following dual cost function:

$$\mathcal{F}_d(Y) = -\frac{1}{2}\|Y\|_2^2 + \mathrm{Tr}(Y^T M) - \lambda f_2^*(G^T Y/\lambda) \tag{7}$$

where Tr stands for the trace of a matrix and f_2^* is the Fenchel conjugate of f_2 defined by:

$$f_2^*(Z) \stackrel{\text{def}}{=} \sup_{X\in\mathbb{R}^P} \mathrm{Tr}(Z^T X) - f_2(X) \ .$$

In Appendix C we provide the Fenchel-Rockafellar duality theorem in order to obtain this result.

The Fenchel conjugates of mixed-norms and squared mixed-norms, which remain to be given in (7), can be computed thanks to the following proposition:

Proposition 3 (Fenchel conjugate of a mixed-norm) *(i) The Fenchel conjugate of norm $\|u\|_{p_1,\ldots,p_n}$ is the indicator function of the dual norm:*

$$v \mapsto i_{\|v\|^*_{p_1,\ldots,p_n}} = \begin{cases} 1 & \text{if } \|v\|_{p'_1,\ldots,p'_n} \leq 1 \\ +\infty & \text{if not} \end{cases}$$

where $\forall i, p'_i$ is such that $\frac{1}{p_i} + \frac{1}{p'_i} = 1$

(ii) The Fenchel conjugate of the function $u \mapsto \frac{1}{2}\|u\|^2_{p_1,\ldots,p_n}$ is the function:

$$v \mapsto \frac{1}{2}\|v\|^2_{p'_1,\ldots,p'_n}$$

Moreover, the Karush-Khun-Tucker (KKT) conditions of the Fenchel-Rockafellar duality theorem (see Appendix C) give a natural mapping from the primal space to the dual space:

$$Y^{(k)} = M - GX^{(k)} \ ,$$

which then needs to be modified in order to satisfy the constraint of f_2^*. When f_2^* is an indicator function, it corresponds to a projection on the associated convex set. It is then possible to use the dual gap as stopping criterion. Algorithm 3 summarizes the computation of the dual gap, in cases of ℓ_{21} and ℓ_{212} penalty function. An 1D illustration with the ℓ_1 is provided in Fig. 2.

Algorithm 3: Duality gap for ℓ_{212} or ℓ_{21}

Entries: $X^{(k)}$
Mapping to the dual space: $Y^{(k)} = M - GX^{(k)}$
Compute $f_2^*(G^T Y^{(k)})$:
if $f_2 = \ell_{21}$ **then**
  Project dual variable on the constraint if necessary:
  $Y^{(k)} = Y^{(k)} / \max(\|G^T Y^{(k)}\|_{2\infty}/\lambda, 1)$
  $f_2^*(G^T Y^{(k)}/\lambda) = 0$
if $f_2 = \ell_{212}$ **then**
  $f_2^*(G^T Y^{(k)}/\lambda) = \frac{1}{2}\|G^T Y^{(k)}/\lambda\|_{2,\infty,2}^2$
Compute duality gap:
$\eta^{(k)} = \frac{1}{2}\|M - GX^k\|_2^2 + \lambda f_2(X^k) + \frac{1}{2}\|Y^{(k)}\|_2^2 - \mathrm{Tr}(Y^T M) + \lambda f_2^*(G^T Y/\lambda)$

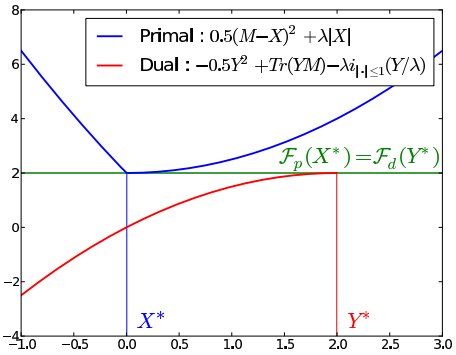


Figure 2. Duality gap illustration with a 1D ℓ_1 problem ($\lambda = 2$ and $M = 2$). It can be observed that the minimum of $\mathcal{F}_p$ is equal to the maximum of $\mathcal{F}_d$, i.e., that the duality gap is 0.

From a numerical point of view, every solution is ε-optimal for a particular value of ε. The duality gap observed at the end of the computation is for example limited by machine precision. Also, Algorithm 3 shows that $\eta^{(k)}$ depends on the scaling of the data. Therefore, in practice the duality gap is meaningful if the input data have been scaled or normalized in a certain way. This is guaranteed with M/EEG data by the

pre-whitening step. Our experimental results show that for whitened data a duality gap lower than 10^{-5} does not produce distinguishable solutions.

3.4. An active set method to improve the convergence speed using the ℓ_{21} norm

Like the ℓ_1 norm, the ℓ_{21} norm leads to sparse solutions; only a few sources will have non-zero activations. Knowing this, we can accelerate the computation with an active set strategy which will restrict computations to sources likely to have non-zero activations. This amounts to solving problems with only a subset of columns of G. Let us call Γ the set of sources considered in the sub-problem and X^*_Γ the associated estimated sources. By computing the duality gap associated to the full problem for a value of X, such that X restricted to Γ is equal to X^*_Γ and where the rest of X is filled with 0, one can test the quality of the solution for the full problem. If this solution is not good enough according to the stopping criteria, one needs to add to Γ sources that are likely to be active. Such sources violate the KKT optimality conditions (Boyd & Vandenberghe 2004). These conditions are specific to the penalty considered. With a $\ell_{\mathbf{w};21}$ penalty, denoting by W the diagonal matrix of weights, the KKT optimality conditions impose the following constraint on X:

$$\|W^{-1}G^T(M - GX)\|_{2\infty} \leq \lambda \ . \tag{8}$$

The indices of the sources that need to be added to the active set at a next iteration, are given by the indices of the rows of $W^{-1}G^T(M - GX)$ that do not meet this constraint. Intuitively it says that the sources that should be added to the active set are the ones whose forward fields correlate the most with the current residual. Such an active set strategy is known as *forward* as the size of the problem keeps increasing.

In practice, one can simply add to Γ the source that violates the most the latter constraint. This can however be rather slow if the active set contains hundreds of variables. That would mean running FISTA hundreds of times. A natural idea consists in adding groups of sources, *i.e.*, the set of sources that violate the most the constraint. When no more source violates the KKT constraint, the solver has converged to an optimal solution. The number of sources that should be added to the active set at each iteration is however application specific. For M/EEG, we have found that adding blocks of 10 new variables is a good trade-off. For an optimal solution containing at the most about a hundred active sources a solution is obtained in practice by running the solver ten times at the most on very small problems. Note that the procedure do guarantee the optimality of the solution at the end as the active set can only grow meaning that the solver will end up solving the original full problem (Roth & Fischer 2008).

Using the active set strategy, the solution corresponding to an experimental data set (with about 300 channels, 200 time samples and 10000 sources) can be obtained in a few seconds. This means that the proposed optimization strategy makes the use of the ℓ_{21} penalty computationally trackable in practical M/EEG applications, which was not the case using the methods relying on second order cone solvers proposed in (Haufe et al. 2008, Ou et al. 2009).

3.5. Convergence results on simulated data

In order to illustrate the convergence rate of the algorithms detailed above in a realistic experimental setting, we performed a simulation using a real MEG gain matrix (151 sensors and 5,000 sources). The implementation used is written in Matlab and involves

only linear algebra and standard vectorized operations. Fig. 3 shows the size of the duality gap as a function of computation time using ISTA, FISTA, or FISTA with the active set (AS-FISTA) approach on a problem with an ℓ_{21} prior. One can observe that FISTA actually converges much faster than ISTA and that AS-FISTA clearly outperforms both of them. For comparison we also ran on the same configuration of dipoles a SOCP (Second Order Cone Program) solver using the CVX Matlab toolbox (http://cvxr.com/cvx/) as in (Ou et al. 2009). Employing only 1,000 dipoles out of 5,000, the computation took 86 s. Such a solver relies of the inversion of a Hessian matrix whose size in $S \times S$. It's cost per iteration is therefore $\mathcal{O}(S^3)$, *i.e.*, cubic in the number of dipoles, and it also requires to store in memory of matrix of size $S \times S$ which may be prohibitive. To tackle the realistic problem used for Fig. 3, it suggests that besides the problem of storing a 5000×5000 matrix in memory, computation should approximately be multiplied by 125 which corresponds to more than 2 hours of computation.

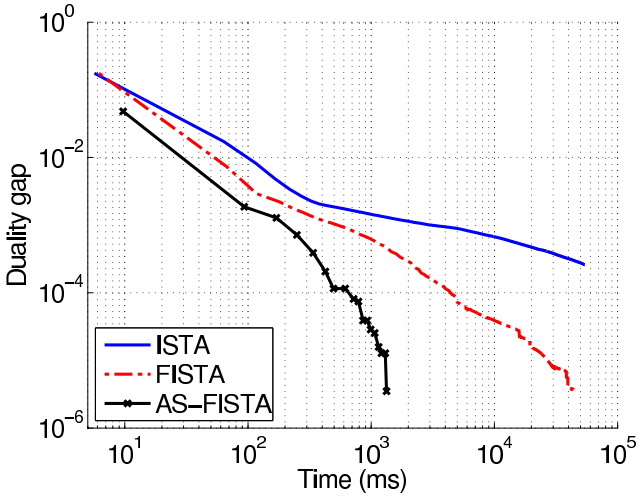


Figure 3. Convergence of ISTA, FISTA and AS-FISTA using a $\|.\|_{\mathbf{w};21}$ penalty with a real MEG lead field (151 sensors and 5,000 sources) and synthetic measurements. It can be observed that ISTA can be slow to converge compared to FISTA and that the active set strategy speeds significantly the convergence.

4. Simulations and MEG results

The following section first presents results with the ℓ_{21} norm applied to M/EEG data, and then some simulation and experimental results obtained with the ℓ_{212} prior. We show that our solver applied with an ℓ_{21} norm provides accurate results in a few seconds on a real auditory M/EEG dataset and that the ℓ_{212} norm can improve the accuracy of reconstructions when performing functional mapping of the somatosensory cortex.

4.1. MxNE with the ℓ_{21} norm

The data used to illustrate the performance of the ℓ_{21} MxNE consists of MEG and EEG combined recordings of the evoked response to left-ear auditory pure-tone stimulus‡. Data were acquired using a 306-channel MEG Neuromag Vectorview system with 60 EEG electrodes, sampled at 600 samples/s. The signals were low-pass filtered at 40 Hz. The noise covariance matrix was estimated from the 200 ms of recordings preceding each stimulus. The source space consisted of 8192 dipoles on the cortex with orientations constrained to be normal to the cortical mantle. Two channels with technical artifacts were ignored from the analysis resulting in a lead field matrix of size 364×8192. We averaged 55 epochs and the sources were estimated between 0 and 400 ms resulting in 241 time samples, hence $M \in \mathbb{R}^{364\times241}$.

The results are presented in Fig. 4. The computation time using AS-FISTA for the entire source estimation was 20 s on a laptop (4 GB of RAM and 2.8 GHz CPU). At the optimum the cost function values was 2073.2 and the duality gap about 10^{-5} corresponding to a change in fifth significant digit of the cost function. The estimated sources are located in both contralateral and ipsilateral primary auditory cortices (cA and iA). A first early component is observed in cA between 30 and 50 ms with a peak around 90 s in cA and later at 100 ms in iA.

In many respects the source estimates look similar to standard multi-dipole fittings results. However, a few remarks should be made about the present results. First, when working with constrained orientations, the signs of the estimated wave forms are dependent on the orientation used for the dipole, *i.e.*, the wall of a sulcus on which the dipoles are located. This is a fundamental problem of M/EEG source imaging well known from the classical MNE. Also, the cluster of 3 active dipoles in iA illustrates a natural behavior of convex priors. These 3 dipoles have very similar forward fields making them very hard to disambiguate with M/EEG. The stability of the ℓ_{21} convex prior produces this cluster of dipoles while a non-convex prior, *e.g.*, ℓ_{2p} with $p < 1$, would certainly pick any one of these dipoles.

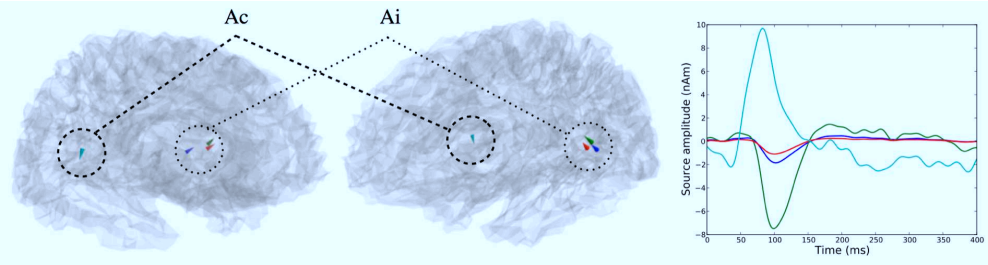


Figure 4. $\ell_{\mathbf{w};21}$ estimates on auditory M/EEG data. Estimation leads to 4 active dipoles in both contralateral and ipsilateral auditory cortices (cA and iA).

4.2. MxNE with the ℓ_{212} prior: Functional mapping

4.2.1. Motivation During an M/EEG experiment, a subject is generally asked to perform different cognitive tasks or to respond to various stimuli. Without an adapted prior, it may occur that the estimated active cortical region in experimental condition

‡ Condition 1 of the sample data provided by the MNE software.

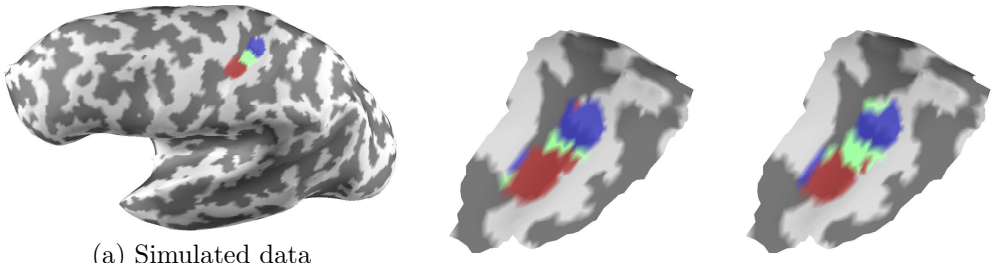
1 overlaps with the active region of experimental condition 2, which may in practice be unrealistic considering the underlying physiology. The primary somatosensory cortex (S1) (Penfield & Rasmussen 1950) and the primary visual cortex (V1) (Wandell et al. 2007) are examples of brain regions with such known organizations. The different body parts are mapped to locations at S1 and a location in the visual field maps to an area at V1 This later is known as the retinotopic organization of V1 (Wandell et al. 2007). Recent work, such as (Sharon et al. 2007, Hagler et al. 2009), emphasize the interest for advanced methods able to reveal such functional organizations non-invasively. However, while (Hagler et al. 2009) propose to use functional MRI data to improve MNE results, we propose the ℓ_{212} prior which is a principled way of taking into account the spatial properties of such regions in order to obtain better functional mappings of the brain using only M/EEG data.

To illustrate the fact that an ℓ_2 prior tends to smear the activations and therefore to overestimate the extent of the active regions, we first performed a simulation study.

4.2.2. Simulation study We generated a synthetic dataset that mimics the organization of the primary somatosensory cortex (S1) (Penfield & Rasmussen 1950). Three non-overlapping cortical regions with a similar area (cf. Fig. 5a), that could correspond to the localization of three right hand fingers were assumed and were used to generate synthetic measurements corrupted with an additive Gaussian noise. The amplitude of activation for the most lateral region (colored in red in Fig. 5), that could correspond to the thumb, was set to be two times as big as the amplitudes of the two other regions. An inverse source estimate was then computed with a standard $\ell_{\mathbf{w};2}$ (4) norm, an $\ell_{\mathbf{w};1}$ norm, and the $\ell_{\mathbf{w};212}$ mixed-norm (5). Each source in the three simulated active regions was then assigned a label corresponding to the condition for which its estimated amplitude was the largest. Quantification of performance was done for multiple values of signal-to-noise ratio (SNR) by counting the percentage of dipoles that have been incorrectly labeled. The SNR is defined here as 20 times the log of the ratio between the norm of the signal and the norm of the added noise. The results are also presented in Fig. 6. It can be observed that the $\ell_{\mathbf{w};212}$ always produces the best result and that the ℓ_1 norm is more strongly affected by the decrease in SNR. In order to have a fair comparison between all methods, the λ was set in each case to have $\|M - GX^*\|_2$ equal to the norm of the added noise, always known in simulations. The depth bias was corrected in the $\ell_{\mathbf{w};212}$ norm case by setting $w_{s,k} = w_s = \|G_{\cdot s}\|_2^2$. This amounts to scaling the columns of G. The depth bias correction was applied the same way for ℓ_1 and ℓ_2 solutions.

The results are illustrated in Fig. 5b and 5c on a region of interest (ROI) around the left primary somatosensory cortex. It can be observed that in the $\ell_{\mathbf{w};2}$ norm result the extend of the most lateral region is overestimated while the result obtained with the $\ell_{\mathbf{w};212}$ mixed-norm is clearly more accurate.

4.2.3. MEG data We also analyzed somatosensory data acquired using a CTF Systems Inc. Omega 151 system at a 1250 Hz sampling rate. The somatosensory stimulus was an electrical square-wave pulse delivered randomly to the thumb, index, middle, and little finger of the right hand of a healthy right-handed subject. The evoked response was computed by averaging 400 repetitions of the stimulation of each finger. The triangulation over which cortical activations have been estimated contained a high number of vertices (about 55,000). The forward solution was



(a) Simulated data (b) $\ell_{\mathbf{w};2}$ results (ROI) (c) $\ell_{\mathbf{w};212}$ results (ROI)

Figure 5. Simulation results on the primary somatosensory cortex (S1) (SNR = 20dB). Neighboring active regions reproduce the organization of S1. It illustrates that the $\ell_{\mathbf{w};212}$ prior improves over a simple $\ell_{\mathbf{w};2}$ prior.

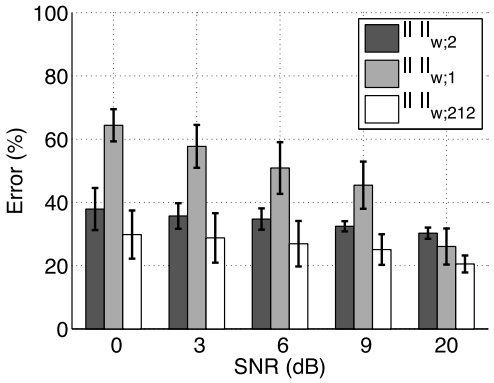


Figure 6. Evaluation of $\ell_{\mathbf{w};2}$, $\ell_{\mathbf{w};212}$ and $\ell_{\mathbf{w};1}$ estimates on synthetic somatosensory data. The error represents the percentage of wrongly labeled dipoles.

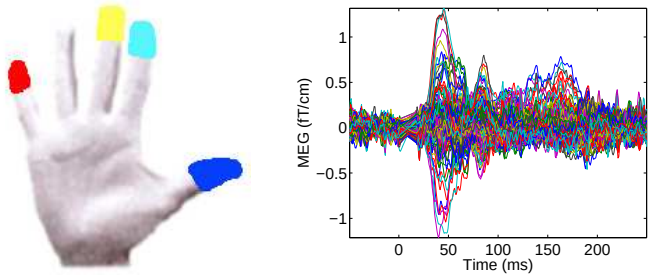
computed using the spherically symmetric head model (Sarvas 1987). The sphere was positioned to match the shape on the inner surface of the skull near the primary somatosensory cortex (central sulcus). Prior to source estimation, data were whitened using a noise covariance matrix estimated during baseline periods.

The source estimates during the period between 42 and 46 ms are presented in Fig. 7. For each type of prior, the regularization parameter λ was set in order for X^* to satisfy $\|M - GX^*\|_2^2 \approx NKT$, knowing that after whitening the data, NKT is a good estimate of the noise variance.

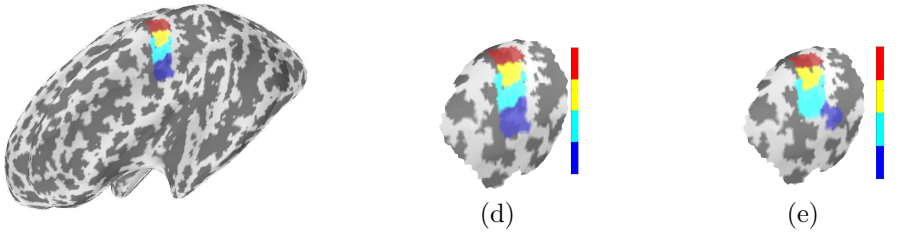
During the time interval of interest the measured magnetic fields indicate the currents are directed into the cortex at S1 which is known to lie on the posterior bank of the central sulcus. Therefore, we next ignored the positive activations located on the anterior bank. Each dipole with negative amplitude was then assigned a label between 1 and 4 based on its maximum amplitude across the 4 conditions. For each condition, equivalently each label, the biggest connected component of dipoles with the same label was kept. Each of the 4 estimated components, corresponding to the 4 right hand fingers are presented in Fig. 7. Solutions using both $\ell_{\mathbf{w};2}$ and $\ell_{\mathbf{w};212}$ norms

are presented. The solution with $\ell_{\mathbf{w};1}$ is not represented as the sparsity promoted do not produce a continuous mapping which make results difficult to compare.

With $\ell_{\mathbf{w};212}$ the well known organization of the primary somatosensory cortex (Penfield & Rasmussen 1950) is successfully recovered with regions of similar size for each finger. While with $\ell_{\mathbf{w};2}$, the component corresponding to the index finger is overestimated leading to an incorrect localization of the area corresponding to the thumb. Note that some activation does remain at the right location for the thumb using the $\ell_{\mathbf{w};2}$ norm. However, it is not strong enough to match with the biggest connected component represented here. These results demonstrate that an alternative



(a) Fingers color coded (b) MEG data for the index finger



(c) $\ell_{\mathbf{w},212}$ results (d) $\ell_{\mathbf{w},212}$ results (zoom) (e) $\ell_{\mathbf{w},2}$ results (zoom)

Figure 7. Labeling results of the left primary somatosensory (S1) cortex in MEG using both $\ell_{\mathbf{w};2}$ and $\ell_{\mathbf{w};212}$ priors. Source estimated during the period between 42 and 46 ms. The $\ell_{\mathbf{w};212}$ leads to a more coherent estimate of the functional organization of S1.

to the standard ℓ_2 priors, can improve the localization of cortical activations by offering the possibility to use a prior between different conditions. By solving the inverse problem of multiple conditions simultaneously and by using a mixed-norm that sets an ℓ_1 prior between each condition, our method penalizes current estimates with an overlap between the corresponding active regions. When such a hypothesis holds, the localization of the neural activity becomes more accurate with increasing number of conditions recorded and included in the analysis.

5. Discussion

In the present paper, we capitalize on advanced numerical methods to tackle multiple convex optimization problems present in many applications such as functional brain imaging using M/EEG. Our paper provides a unifying view of many solvers previously proposed in the M/EEG literature and is to our knowledge the first demonstration that the M/EEG inverse problem can be solved in a few seconds using non-ℓ_2 priors. Rapid

computations are essential in functional brain imaging, since interactive exploratory analysis is often needed.

This work relates to the distributed solvers based on sparse Bayesian regression that have been recently proposed (Nummenmaa et al. 2007, Friston et al. 2008, Wipf & Nagarajan 2009). These solvers are not explicitly derived from cost-functions like (3) and they lead to non-convex optimization problems not covered by the algorithms detailed above. Note also that formulating the inverse problem as the minimization of a cost function does not guarantee convexity. For example, Valdes-Sosa *et al.* propose in (Valdés-Sosa et al. 2009) to estimate the source activations as a linear combination of a small number of spatiotemporal maps. Here again, sparsity is a key assumption of the method, however, the minimized cost function is not jointly convex in space and time. The consequence of the non-convexity for all these methods, is that the optimality of the solutions cannot be guaranteed and that solutions depend on the initialization of the algorithm. Our formulation of MxNE does not suffer from these shortcomings.

Another benefit of MxNE is the diversity of *a priori* knowledge that can be taken into account. With the same mathematical foundations and very similar implementations, the ℓ_{21} norm can be used to promote sparse source estimates with smooth temporal activations, an ℓ_{221} norm can furthermore impose a common set of active dipoles between conditions, and the ℓ_{212} norm can be used for more accurate functional mapping. From the neuroscience perspective, the ℓ_{21} prior models the *a priori* assumption that active brain regions should be consistent during a time interval. This assumption is adapted to some datasets like the auditory data presented here, however it is likely to be wrong for a long time interval during which active sources are changing. The ℓ_{212} prior is motivated by its ability to explicitly model the functional specificity of brain regions thus leading to more insights on neural circuitry (Chklovskii & Koulakov 2004). The latter application is also to our knowledge the first attempt to improve the M/EEG inverse problem by using multiple datasets jointly.

Finally, an important point is that there is no prior that fits all needs. That is why MxNE does not refer to a particular prior but to a class of solvers that use mixed-norms to better constrain the M/EEG inverse problem. This implies that depending on the assumptions made about the underlying sources, a particular mixed-norm can be more relevant than others.

Conclusion

In this article, we have shown how mixed-norms can be used to promote structure on inverse estimates in order to take into account some a priori knowledge. In the case of M/EEG, the *a priori* knowledge is based on the understanding of both neurophysiology and biophysics.

This contribution provides principled first-order optimization strategies which are simple to implement and fast, especially when an active set approach is used. Furthermore, the speed of convergence of these algorithms is well understood thanks to a theoretical analysis. All these algorithms rely on proximity operators, which are in practice special thresholding operators. Moreover, we were able to construct practical criteria to stop the optimization process while guaranteeing the optimality of the solutions obtained.

The utility of mixed-norms is demonstrated with both synthetic and experimental MEG data. Our results match existing knowledge about auditory evoked responses

and lead to a more accurate mapping of the somatosensory homunculus than the unstructured standard methods.

Acknowledgments

The authors acknowledge support from the ANR ViMAGINE ANR-08-BLAN-0250-02, the National Center for Research Resources P41 RR014075-11, and the NIH National Institute of Biomedical Imaging and Bioengineering 5R01EB009048. The authors also wish to thank the anonymous reviewers for their careful comments which helped improve the presentation and the readability of this article.

Appendix A. Derivation of the (ℓ_{21} proximal operator)

We are looking for:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_{\mathbf{w};21} \ ,$$

with

$$\|\mathbf{x}\|_{\mathbf{w};21} = \sum_k \|\mathbf{x}_k\|_{\mathbf{w};2} = \sum_k \sqrt{\sum_t w_k x_{k,t}^2} \ .$$

We can differentiate the functional, when $\|\mathbf{x}_k\|_{\mathbf{w};2} \neq 0$, to obtain the variational system:

$$\left\{ \begin{array}{rl} |x_{k,t}| & = |y_{k,t}| - \lambda\sqrt{w_k}|x_{k,t}|\|\mathbf{x}_k\|_2^{-1} \\ \operatorname{sign}(x_{k,t}) & = \operatorname{sign}(y_{k,t}) \end{array} \right.$$

which gives:

$$|x_{k,t}|\left(1+\lambda\sqrt{w_k}\|\mathbf{x}_k\|_2^{-1}\right) = |y_{k,t}|$$

As $1+\lambda\sqrt{w_k}\|\mathbf{x}_k\|_2^{-1}$ does not depend on t, it implies that for all t and ν:

$$|x_{k,t}| = \frac{|y_{k,t}|}{|y_{k,\nu}|}|x_{k,\nu}| \ .$$

By injecting this last result in Eq. (A.1) we obtain

$$\begin{aligned} |x_{k,t}| &= \left(|y_{k,t}| - \frac{\lambda\sqrt{w_k}|x_{k,t}|}{\sqrt{\sum_\nu x_{k,\nu}^2}}\right)^+ \\ |x_{k,t}| &= \left(|y_{k,t}| - \frac{\lambda\sqrt{w_k}|x_{k,t}|}{\sqrt{\sum_\nu \left(y_{k,\nu}^2 \frac{x_{k,t}^2}{y_{k,t}^2}\right)}}\right)^+ \\ &= \left(|y_{k,t}| - \frac{\lambda\sqrt{w_k}|x_{k,t}|}{\frac{|x_{k,t}|}{|y_{k,t}|}\|\mathbf{y}_k\|_2}\right)^+ \\ &= |y_{k,t}|\left(1-\frac{\lambda\sqrt{w_k}}{\|\mathbf{y}_k\|_2}\right)^+ \ , \end{aligned}$$

which is the desired result.

Appendix B. Proof of Proposition 2 (ℓ_{212} proximal operator)

Proof We are looking for:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \|\mathbf{y}-\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_{\mathbf{w};212}^2 \ ,$$

with

$$\|\mathbf{x}\|_{\mathbf{w};212}^2 = \sum_s \left(\sum_k \left(\sum_t w_{s,k} x_{s,k,t}^2 \right)^{1/2} \right)^2 \ .$$

We can differentiate the functional, when $x_{s,k,t} \neq 0$, to obtain the variational system:

$$\left\{ \begin{array}{rl} |x_{s,k,t}| & = |y_{s,k,t}| - \lambda \sqrt{w_{s,k}} |x_{s,k,t}| \|\mathbf{x}_{s,k}\|_2^{-1} \|\mathbf{x}_s\|_{\mathbf{w};21} \\ \mathrm{sign}(x_{s,k,t}) & = \mathrm{sign}(y_{s,k,t}) \end{array} \right. \tag{B.1}$$

which gives:

$$\begin{aligned} |x_{s,k,t}| \left(1 + \lambda \sqrt{w_{s,k}} \|\mathbf{x}_{s,k}\|_2^{-1} \|\mathbf{x}_s\|_{\mathbf{w};21}\right) &= |y_{s,k,t}| \qquad (B.2) \\ \Rightarrow x_{s,k,t}^2 \left(1 + \lambda \sqrt{w_{s,k}} \|\mathbf{x}_{s,k}\|_2^{-1} \|\mathbf{x}_s\|_{\mathbf{w};21}\right)^2 &= y_{s,k,t}^2 \end{aligned}$$

By summing over t, we get:

$$\begin{aligned} \|\mathbf{x}_{s,k}\|_2 \left(1 + \lambda \sqrt{w_{s,k}} \|\mathbf{x}_{s,k}\|_2^{-1} \|\mathbf{x}_s\|_{\mathbf{w};21}\right) &= \|\mathbf{y}_{s,k}\|_2 \\ \Rightarrow \|\mathbf{x}_{s,k}\|_2 + \lambda \sqrt{w_{s,k}} \|\mathbf{x}_s\|_{\mathbf{w};21} &= \|\mathbf{y}_{s,k}\|_2 \ . \qquad (B.3) \end{aligned}$$

This last equality is true for all k. Then for k and l satisfying $\|\mathbf{x}_{s,k}\|_2 > 0$ and $\|\mathbf{x}_{s,l}\|_2 > 0$, we have:

$$\frac{\|\mathbf{x}_{s,k}\|_2}{\sqrt{w_{s,k}}} = \frac{\|\mathbf{x}_{s,l}\|_2}{\sqrt{w_{s,l}}} + \frac{\|\mathbf{y}_{s,k}\|_2}{\sqrt{w_{s,k}}} - \frac{\|\mathbf{y}_{s,l}\|_2}{\sqrt{w_{s,l}}} \ . \tag{B.4}$$

Furthermore, we have that:

$$\|\mathbf{x}_s\|_{\mathbf{w};21} = \sum_{l/\|\mathbf{x}_{s,l}\|_2 > 0} \sqrt{w_{s,l}} \|\mathbf{x}_{s,l}\|_2 \ .$$

which implies that:

$$\|\mathbf{x}_{s,k}\|_2 = \|\mathbf{y}_{s,k}\|_2 - \lambda \sqrt{w_{s,k}} \sum_{l/\|\mathbf{x}_{s,l}\|_2 > 0} \sqrt{w_{s,l}} \|\mathbf{x}_{s,l}\|_2 \ . \tag{B.5}$$

By injecting (B.4) in (B.5) we get:

$$\begin{array}{rl} \|\mathbf{x}_{s,k}\|_2 & = \|\mathbf{y}_{s,k}\|_2 - \lambda \sqrt{w_{s,k}} \sum_{l/\|\mathbf{x}_{s,l}\|_2>0} w_{s,l} \left(\frac{\|\mathbf{x}_{s,k}\|_2}{\sqrt{w_{s,k}}} + \frac{\|\mathbf{y}_{s,l}\|_2}{\sqrt{w_{s,l}}} - \frac{\|\mathbf{y}_{s,k}\|_2}{\sqrt{w_{s,k}}} \right) \\ \Leftrightarrow \|\mathbf{x}_{s,k}\|_2 & = \|\mathbf{y}_{s,k}\|_2 - \lambda K_{\mathbf{w}_s} (\|\mathbf{x}_{s,k}\|_2 - \|\mathbf{y}_{s,k}\|_2) - \lambda \sqrt{w_{s,k}} \sum_{l/\|\mathbf{x}_{s,l}\|_2>0} \sqrt{w_{s,l}} \|\mathbf{y}_{s,l}\|_2 \quad , \\ \Leftrightarrow \|\mathbf{x}_{s,k}\|_2 & = \|\mathbf{y}_{s,k}\|_2 - \frac{\lambda \sqrt{w_{s,k}}}{1+\lambda K_{\mathbf{w}_s}} \sum_{l/\|\mathbf{x}_{s,l}\|_2>0} \sqrt{w_{s,l}} \|\mathbf{y}_{s,l}\|_2 \end{array}$$

where $K_{\mathbf{w}_s} = \sum_{k/\|\mathbf{x}_{s,k}\|_2>0} w_{s,k}$. We therefore have for all s:

$$\|\mathbf{x}_{s,k}\|_2 = \left(\|\mathbf{y}_{s,k}\|_2 - \frac{\lambda \sqrt{w_{s,k}}}{1+\lambda K_{\mathbf{w}_s}} \sum_{k/\|\mathbf{x}_{s,k}\|_2>0} w_{s,k} \|\mathbf{y}_{s,k}\|_2 \right)^+ \ .$$

If we reorder the $\sqrt{w_{s,k}}\|\mathbf{y}_{s,k}\|_2$ and introduce K_s defined in the proposition, it leads to:

$$\|\mathbf{x}_{s,k}\|_2 = \left(\|\mathbf{y}_{s,k}\|_2 - \frac{\lambda\sqrt{w_{s,k}}}{1+\lambda K_{\mathbf{w}_s}}[\mathbf{y}_s]\right)^+ . \tag{B.6}$$

with $[\mathbf{y}_s] = \sum_{k=1}^{K_s}\sqrt{w_{s,k}}\|\mathbf{y}_{s,k}\|_2$. Let us rewrite (B.2):

$$|x_{s,k,t}| = \frac{|y_{s,k,t}|}{1+\lambda\sqrt{w_{s,k}}\|\mathbf{x}_{s,k}\|_2^{-1}\|\mathbf{x}_s\|_{\mathbf{w};21}} = \frac{|y_{s,k,t}|\|\mathbf{x}_{s,k}\|_2}{\|\mathbf{x}_{s,k}\|_2+\lambda\sqrt{w_{s,k}}\|\mathbf{x}_s\|_{\mathbf{w};21}}$$

Using (B.3), we get:

$$|x_{s,k,t}| = \frac{|y_{s,k,t}|\|\mathbf{x}_{s,k}\|_2}{\|\mathbf{y}_{s,k}\|_2} .$$

By injecting the result (B.6) in this equation we get:

$$|x^*_{s,k,t}| = \frac{|y_{s,k,t}|\left(\|\mathbf{y}_{s,k}\|_2 - \frac{\lambda\sqrt{w_{s,k}}}{1+\lambda K_{\mathbf{w}_s}}[\mathbf{y}_s]\right)^+}{\|\mathbf{y}_{s,k}\|_2} = |y_{s,k,t}|\left(1-\frac{\lambda\sqrt{w_{s,k}}}{1+\lambda K_{\mathbf{w}_s}}\frac{[\mathbf{y}_s]}{\|\mathbf{y}_{s,k}\|_2}\right)^+ .$$

Note that this result gives also the proof of the proximal operator associated to the Elitist-Lasso in Proposition 1.

Appendix C. Proof of equation (7)

Theorem 1 (Fenchel-Rockafellar duality (Rockafellar 1972)) *Let $f : \mathbb{R}^M \cup \{+\infty\} \to \mathbb{R}$ be a convex function and $g : \mathbb{R}^N \cup \{+\infty\} \to \mathbb{R}$ a concave function. Let G be a linear operator mapping vectors of $\mathbb{R}^M$ to $\mathbb{R}^N$. Then*

$$\inf_{X\in\mathbb{R}^M}\{f(X)-g(GX)\} = \sup_{Y\in\mathbb{R}^N}\{g^*(Y)-f^*(G^TY)\}$$

where f^* (resp. g^*) is the Fenchel conjugate associated to f (resp. g), and G^T the adjoint operator of G.

Moreover, the Karush-Kuhn-Tucker (KKT) conditions read:

$$f(X)+f^*(G^Tu) = \langle X, G^TY\rangle , \quad g(GX)+g^*(Y) = \langle GX, Y\rangle .$$

We can apply this Theorem to the functional (3) with $g(X) = -\frac{1}{2}\|M-X\|_2^2$ and $f(X) = \lambda f_2(X)$. Then, one just have to compute the conjugate of f. By definition of the dual, given here for a concave function, we have

$$g^*(Y) \stackrel{\text{def}}{=} \inf_X \mathrm{Tr}(Y^TX) + \frac{1}{2}\|M-X\|_2^2 = -\sup_X -\mathrm{Tr}(Y^TX) - \frac{1}{2}\|M-X\|_2^2$$

then, by the change of variable $Y = M-X$ we have

$$g^*(Y) = \mathrm{Tr}(Y^TM) - \sup_Z\left(\mathrm{Tr}(Y^TZ) - \frac{1}{2}\|Z\|_2^2\right) .$$

Moreover, we know by Proposition 3 that the Fenchel conjugate of a squared norm is the squared dual norm. Then, we have

$$g^*(Y) = \mathrm{Tr}(Y^TM) - \frac{1}{2}\|Y\|_2^2 .$$

For the Fenchel conjugate of $f(X) = \lambda f_2(X)$, one can apply simple calculus rules given in (Boyd & Vandenberghe 2004), which give $f^*(Y) = \lambda f_2^*(Y/\lambda)$. Finally, the Fenchel-Rockafellar dual function of (3) is given by

$$\mathrm{Tr}(Y^T M) - \frac{1}{2}\|Y\|_2^2 - \lambda f_2^*(G^T Y/\lambda) \ .$$

Furthermore, one can check that the dual variable $Y = M - GX$ verifies the KKT condition

$$-\frac{1}{2}\|M - GX\|_2^2 + \mathrm{Tr}(Y^T M) - \frac{1}{2}\|Y\|_2^2 = \langle GX, Y\rangle \ .$$

References

Beck A & Teboulle M 2009 'A fast iterative shrinkage-thresholding algorithm for linear inverse problems' *SIAM Journal on Imaging Sciences* **2**(1), 183–202.

Benedek A & Panzone R 1961 'The space l^p with mixed norm' *Duke Mathematical Journal* **28**, 301–324.

Bobin J, Starck J L, Moudden Y & Fadili M 2008 *in* P Hawkes, ed., 'Advances in Imaging and Electron Physics' Academic Press, Elsevier pp. 221–298.

Boyd S & Vandenberghe L 2004 *Convex Optimization* Cambridge University Press.

Bruce A, Sardy S & Tseng P 1998 'Block coordinate relaxation methods for nonparamatric signal denoising' *Proceedings of SPIE* **3391**(75).

Candès E J & Tao T 2005 'The dantzig selector: statistical estimation when p is much larger than n' *Annals of Statistics* **35**, 2313–2351.

Chklovskii D B & Koulakov A A 2004 'Maps in the brain: what can we learn from them?' *Annu Rev Neurosci* **27**, 369–92.

Combettes P L & Wajs V R 2005 'Signal recovery by proximal forward-backward splitting' *Multiscale Modeling and Simulation* **4**(4), 1168–1200.

Dale A & Sereno M 1993 'Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach' *Journal of Cognitive Neuroscience* **5**(2), 162–176.

Daubechies I, Defrise M & De Mol C 2004 'An iterative thresholding algorithm for linear inverse problems with a sparsity constraint' *Commun. Pure Appl. Math.* **57**(11), 1413 – 1457.

Daubechies I, DeVore R, Fornasier M & Gunturk S 2008 'Iteratively re-weighted least squares minimization: Proof of faster than linear rate for sparse recovery' *Information Sciences and Systems* .

Daudet L, Molla S & Torrésani B 2004 *in* J. P Li, ed., 'International Conference Wavelet analysis and Applications' Chongqing, China pp. 13–24.

Donoho D L 1995 'De-noising by soft-thresholding' *IEEE Transactions on Information Theory* **41**(3), 613–627.

Donoho D L 2006 'Compressed sensing' *IEEE Transactions on Information Theory* **52**(4), 1289–1306.

Dupé F X, Fadili M & Starck J L 2009 'A proximal iteration for deconvolving poisson noisy images using sparse representations' *IEEE Transactions on Image Processing* **18**(2), 310–321.

Efron B, Hastie T, Johnstone L & Tibshirani R 2004 'Least angle regression' *Annals of Statistics* **32**, 407–499.

Feichtinger H G 2006 'Modulation spaces: Looking back and ahead' *Sampling Theory in Signal and Image Processing* **5**(3), 109–140.

Févotte C, Torrésani B, Daudet L & Godsill S J 2008 'Sparse linear regression with structured priors and application to denoising of musical audio' *IEEE Transactions on Audio, Speech and Language Processing* **16**(1), 174–185.

Friedman J, Hastie T, Höfling H & Tibshirani R 2007 'Pathwise coordinate optimization' *Annals of Applied Statistics* **1**(2), 302–332.

Friston K, Harrison L, Daunizeau J, Kiebel S, Phillips C, Trujillo-Barreto N, Henson R, Flandin G & Mattout J 2008 'Multiple sparse priors for the M/EEG inverse problem' *Neuroimage* **39**(3), 1104–20.

Gorodnitsky I, George J & Rao B 1995 'Neuromagnetic source imaging with focuss: a recursive weighted minimum norm algorithm' *Electroencephalography and clinical Neurophysiology* .

Gramfort A, Papadopoulo T, Baillet S & Clerc M 2011 'Tracking cortical activity from M/EEG using graph-cuts with spatiotemporal constraints' *NeuroImage* **54**(3), 1930–1941.

Gramfort A, Papadopoulo T, Olivi E & Clerc M 2010 'OpenMEEG: opensource software for quasistatic bioelectromagnetics' *BioMedical Engineering OnLine* **9**(1), 45.

Grochenig K & Samarah S 2000 'Non-linear approximation with local fourier bases' *Constr. Approx.* **16**, 317–332.

Hagler D J, Halgren E, Martinez A, Huang M, Hillyard S A & Dale A M 2009 'Source estimates for MEG/EEG visual evoked responses constrained by multiple, retinotopically-mapped stimulus locations' *Human Brain Mapping* **30**, 1290–1309.

Hämäläinen M & Ilmoniemi R 1994 'Interpreting magnetic fields of the brain: minimum norm estimates' *Medical and Biological Engineering and Computing* **32**(1), 35–42.

Hansen P C, Kringelbach M L & Salmelin R 2010 *MEG: An Introduction to Methods* Oxford University Press US.

Haufe S, Nikulin V V, Ziehe A, Müller K R & Nolte G 2008 'Combining sparsity and rotational invariance in EEG/MEG source reconstruction' *NeuroImage* **42**(2), 726–38.

Huppertz H J, Hoegg S, Sick C, Lücking C H, Zentner J, Schulze-Bonhage A & Kristeva-Feige R 2001 'Cortical current density reconstruction of interictal epileptiform activity in temporal lobe epilepsy' *Clinical Neurophysiology* **112**(9), 1761–72.

Kowalski M 2009 'Sparse regression using mixed norms' *Appl. Comput. Harmon. Anal.* **27**(3), 303–324.

Kowalski M & Torrésani B 2008 'Random models for sparse signals expansion on unions of basis with application to audio signals' *IEEE Transactions on Signal Processing* **56**(8).

Kowalski M & Torrésani B 2009 'Sparsity and persistence: mixed norms provide simple signals models with dependent coefficients' *Sig Imag Video Process* **3**(3), 251–264.

Kybic J, Clerc M, Abboud T, Faugeras O, Keriven R & Papadopoulo T 2005 'A common formalism for the integral formulations of the forward EEG problem' *IEEE Transactions on Medical Imaging* **24**(1), 12–28.

Li Y 1993 'A globally convergent method for l_p problems' *SIAM Journal on Optimization* **3**(3), 609–629.

Lin F H, Witzel T, Ahlfors S P, Stufflebeam S M, Belliveau J W & Hämäläinen M S 2006 'Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates' *NeuroImage* **31**(1), 160–71.

Matsuura K & Okabe Y 1995 'Selective minimum-norm solution of the biomagnetic inverse problem.' *IEEE Trans Biomed Eng* **42**(6), 608–615.

Moreau J J 1965 'Proximité et dualité dans un espace hilbertien' *Bull. Soc. Math. France* **93**, 273–299.

Morozov V A 1966 'On the solution of functional equations by the method of regularization' *Soviet Math. Dokl.* **7**, 414–417.

Mosher J, Leahy R & Lewis P 1999 'EEG and MEG: Forward solutions for inverse methods' *IEEE Transactions on Biomedical Engineering* **46**(3), 245–259.

Nesterov Y 2007*a* Gradient methods for minimizing composite objective function CORE Discussion Papers 2007076 Universit catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Nesterov Y E 2007*b* Gradient methods for minimizing composite objective function Technical report CORE discussion paper – Université Catholique de Louvain.

Nummenmaa A, Auranen T, Hamalainen M, Jaaskelainen I, Sams M, Vehtari A & Lampinen J 2007 'Automatic relevance determination based hierarchical Bayesian MEG inversion in practice' *Neuroimage* **37**(3), 876–889.

Osborne M R, Presnell B & Turlach B A 2000 'A new approach to variable selection in least squares problems' *IMA J Numer Anal* **20**(3), 389–403.

Ou W, Hämaläinen M & Golland P 2009 'A distributed spatio-temporal EEG/MEG inverse solver' *NeuroImage* **44**(3), 932–946.

Pantazis D, Nichols T E, Baillet S & Leahy R 2003 'Spatiotemporal localization of significant activation in MEG using permutation tests.' *Inf Process Med Imaging* **18**, 512–523.

Pascual-Marqui R D, Michel C M & Lehman D 1994 'Low resolution electromagnetic tomography: A new method for localizing electrical activity of the brain' *Psychophysiology* **18**, 49–65.

Penfield W & Rasmussen T 1950 *The Cerebral Cortex of Man: A Clinical Study of Localization of Function* Macmillan.

Phillips C, Mattout J, Rugg M & Maquet P 2005 'An empirical bayesian solution to the source reconstruction problem in eeg' *Neuroimage* .

Phillips J, Leahy R & Mosher J 1997 'MEG-based imaging of focal neuronal current sources' *Medical Imaging, IEEE Transactions on* **16**(3), 338 – 348.

Rockafellar R T 1972 *Convex Analysis* Princeton University Press.

Roth V & Fischer B 2008 *in* 'ICML '08: Proceedings of the 25th international conference on Machine

learning' pp. 848–855.
Rychkov V S 1999 'On restrictions and extensions of the besov and triebel–lizorkin spaces with respect to lipschitz domains' *Journal of London Mathematical Society* **60**(1), 237–257.
Samarah S & Salman R 2006 'Local Fourier bases and modulation spaces' *Turkish Journal of Mathematics* **30**(4), 447–462.
Sarvas J 1987 'Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem' *Phys. Med. Biol.* **32**(1), 11–22.
Sharon D, Hämäläinen M, Tootell R & Halgren E 2007 'The advantage of combining MEG and EEG: Comparison to fMRI in focally stimulated visual cortex' *Neuroimage* **36**, 1225–1235.
Tibshirani R 1996 'Regression shrinkage and selection via the Lasso' *J.R. Statist. Soc.* **58**(1), 267–288.
Tikhonov A & Arsenin V 1977 *Solutions of Ill-Posed Problems* Winston & Sons, Washington.
Tseng P 2010 'Approximation accuracy, gradient methods, and error bound for structured convex optimization' *Mathematical Programming* **125**, 263–295. 10.1007/s10107-010-0394-2.
Uutela K, Hämäläinen M & Somersalo E 1999 'Visualization of magnetoencephalographic data using minimum current estimates' *Neuroimage* **10**, 173–180.
Valdés-Sosa P A, Vega-Hernández M, Sánchez-Bornot J M, Martínez-Montes E & Bobes M A 2009 'EEG source imaging with spatio-temporal tomographic nonnegative independent component analysis' *Human Brain mapping* **30**(6), 1898–910.
Varoquaux G, Gramfort A, Poline J B & Thirion B 2010 *in* Richard Zemel & John Shawe Taylor, eds, 'Advances in Neural Information Processing Systems Advances in Neural Information Processing Systems' John Lafferty Vancouver Canada.
Wagner M, Wischmann H, Fuchs M, Kohler T & Drenckhahn R 1996 *in* C Aine, ed., 'Adv Biomagn Res (BIOMAG)' Santa Fe, NM, USA pp. 393–396.
Wandell B, Dumoulin S & Brewer A 2007 'Visual field maps in human cortex' *Neuron* **56**(2), 366–383.
Wang J Z, Williamson S J & Kaufman L 1992 'Magnetic source images determined by a lead-field analysis: the unique minimum-norm least-squares estimation' *Biomedical Engineering, IEEE Transactions on* **39**(7), 665–675.
Weiss P 2008 Algorithmes rapides d'optimisation convexe. Applications à la reconstruction d'images et à la détection de changements. PhD thesis Université de Nice Sophia-Antipolis.
Wipf D & Nagarajan S 2009 'A unified bayesian framework for MEG/EEG source imaging' *Neuroimage* **44**(3), 947–966.