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

Variational Bayesian localization of EEG sources with generalized Gaussian priors

J.M. Cortes[1,2,3], A. Lopez[4], R. Molina[1], and A.K. Katsaggelos[5]

[1] Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada, E-18071 Granada, Spain. Emails: jesus.cortesdiaz@osakidetza.net,rms@decsai.ugr.es
[2] Current Address: IKERBASQUE, The Basque Foundation for Science, E-48011, Bilbao, Spain
[3] Current Address: Biocruces Health Research Institute. Hospital Universitario de Cruces. Plaza de Cruces s/n, E-48903. Barakaldo, Spain
[4] Departamento de Lenguajes y Sistemas Informáticos. Universidad de Granada, E-18071 Granada, Spain. Email: alopez@ugr.es
[5] Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, Illinois 60208-3118, USA. Email: aggk@eecs.northwestern.edu

the date of receipt and acceptance should be inserted later

Abstract. Although in the last decades the use of Magnetic Resonance Imaging has grown in popularity as a tool for the structural analysis of the brain, including MRI, fMRI and recently DTI, the ElectroEncephaloGraphy (EEG) is still-today an interesting technique for the understanding of brain organization and function. The main reason for this is that the EEG is a direct measure of brain bioelectrical activity, and such activity can be monitorized in the millisecond time-window. For some situations and cognitive scenarios, such fine-temporal resolution might suffice for some aspects of brain function; however, the EEG spatial resolution is very poor since it is based on a small number of scalp recordings, thus turning the source localization problem into an ill-posed in which infinite possibilities exist for the localization of the neuronal generators. This is an old problem in computational neuroimaging; indeed, many methods have been proposed to overcome this localization. Here, by performing a Variational Bayesian Inference procedure with a generalized Gaussian prior, we come out with an algorithm that performs simultaneously the estimation of both sources and model parameters. The novelty for the inclusion of the generalized Gaussian prior allows to control the smoothness degree of the estimated sources. Finally, the suggested algorithm is validated on simulated data.

PACS. 02.50.-r: Statistics – 02.50.Tt: Inference methods – 87.19.le: EEG, in neuroscience

1 Introduction

Electroencephalography (EEG) is a widely used technique to look into the brain [1–3]; compared to other neuroimaging modalities, its main advantage is that EEG allows for a direct measurement of neuronal activity with a time-resolution as fine as to account for variations of neuronal activity in the order of few milliseconds. The main EEG disadvantage is that its spatial resolution is very poor. In fact, a big challenge for the EEG is to achieve a good physiological solution to the (so called) ill-posed localization problem; namely, based on the scalp recordings to approach an estimation for the cortical generators of the scalp signals. Bayesian Inference has a long tradition in the EEG source localization problem, see for instance [4–12], in which by adding some prior information into the sources space it is possible to find (under those assumptions) the localization of the sources which is consistent with the observed scalp recordings.

Following previous studies, here we present a Bayesian Inference formalism with two important features: 1) it utilizes a generalized Gaussian prior for the sources and 2) uses a Variational Bayesian approach for the estimation of both hyperparameters and sources. The combined estimation for hyperparameters and sources has been addressed before by applying an Expectation-Maximization algorithm [5,6], see also [9] for a discussion of empirical Bayesian approaches to the source localization problem. Estimation of both sources and hyperparameters have been also approached before by Variational Bayesian methods with Gaussian priors [10]. In this paper, the use of a generalized Gaussian prior distribution for the sources introduces a control (*shape*) parameter $1 \leq p \leq 2$ which allows to control the smoothness

Send offprint requests to:

degree of the estimated sources. Interestingly, the generalized Gaussian prior is equivalent to the use of the *Lp*-norm to define the energy of the prior distribution [13]. Thus, the use of the generalized Gaussian prior allows to recover in an elegant form previously reported source localization models; for instance, a Gaussian prior (achieved for $p = 2$) coincides with the so-celebrated algorithm LORETA [4] if no hyperparameters estimation is performed; the case for $p = 2$ and hyperparameters estimation by the Variational Bayesian approach was addressed in [10]; the case $p = 1$ produces a Laplacian prior, which was shown to approximate Total-Variations priors [14]. For other values of p new solutions can be explored. Once the modelling is completed, the use of the Variational Bayesian inference [15,16] comes-out with an iterative algorithm to estimate, given the scalp observations, the posterior distribution of sources and hyperparameters. This inference procedure is, as we will explain later, much more powerful than the use of solely point estimates since it allows to examine, for instance, the variance of the estimated unknown variables.

Interestingly Babacan et al. [17] showed in image processing that the use of Gaussian priors ($p = 2$) worked well for reconstruction of smooth images whilst Laplacian priors ($p = 1$) worked better for edge reconstruction; they also showed the possibility of improving the recostruction of some images for intermediate values of $1 < p < 2$ compared to the extremal situations of $p = 1$ and $p = 2$. Here, we bring these results to the problem of EEG source localization and similarly conclude that the abrupt distributed sources are better localized for $p = 1$ and this happens independently on the noise level. In contrast, the localization of smooth profiles depends critically on noise: for high noise, the localization is practically independent on p but for low noise the localization is better for $p = 2$. Intermediate situations can occur for a different p.

It is important to note that, while smooth sources are physiologically plausible to describe most of brain activity situations, in pathological conditions such as localized epilepsy, in which the electrical activity becomes strong, highly synchronous, in a small well-localized region (even that small as an individual point-source), smooth source estimation is unrealistic and it is necessary the use of models that account for abrupt source-patterns. So, we need workable localization methods as the one presented here which can localize both smooth or abrupt sources.

To summarize, we present here a generalized Gaussian prior, which, by simply varying a parameter p, allows the use of the same algorithm to localize sources with very different profiles of electrical activity, and also a Variational Bayesian inference method (which avoids hyperparameter hand-tuning) and provides information on the uncertainty of the estimated parameters and sources.

The paper is organized as follows. In section 2 we detail the methods, including: 1) the Modelization, 2) the Bayesian formulation, 3) the Variational Bayesian approach, 4) the iterative algorithm for the EEG source localization, 5) a comparison to other similar existing methods, 6) the calculation of the Lead Field matrix, 7) the explanation of the two different simulated conditions, 8) the generation of the observations and finally 9) the initial conditions used for simulations. In section 3 we apply our algorithm to simulated data for two different conditions, one in which strong sources are initially localized in a narrow region, named *abrupt* sources, and one in which weak sources are widespread throughout the whole cortical surface, named *smooth* sources. For both situations, we utilize different values of the p parameter, with $1 \leq p \leq 2$, quantify performance and numerically validate the hyperparameters estimation provided by our algorithm. Finally, conclusions are presented in section 4.

2 Methods

2.1 Modelization

We do not assume here to have a given numbers of dipoles at a specific location and responsible for the scalp signals; in contrast, we use distributed source modeling, in which it is possible to have source dipoles at all possible locations within the brain space (representing for instance source-points in grey matter). We assume there are n_d dipoles generating activity represented by the vector $\mathbf{j}$ and responsible of the observations, which represented by the vector $\mathbf{v}$ correspond to the electrical potentials registered by $n_e \ll n_d$ sensors on the scalp. In general, for each dipole its sources activity is represented by a 3D vector (its electrical field), but hereon we consider only one component, corresponding to the perpendicular direction to the cortical surface. At each time instant, observations and sources are linearly related, and we can write

$$\mathbf{v} = \mathcal{L}\mathbf{j} + \epsilon, \tag{1}$$

where the noise ϵ, hereon, is assumed to be Gaussian with a covariance matrix in which all off-diagonal terms are equal to zero, i.e., the noise at each sensor is independent from the noise at other sensors. In addition to this, it is assumed that the noise at each sensor has zero mean and unknown variance β^{-1}.

The matrix $\mathcal{L}$ relating observations and sources is of size $n_e \times n_d$ and is named the *Lead Field* matrix. To calculate $\mathcal{L}$, one needs the positions of both sensors and sources and knowledge on the specific head model. In section 2.6, we provide details on its calculation.

2.2 Bayesian Formulation

The Bayesian formulation of the EEG source localization problem needs to define and manipulate the joint distribution to estimate, consistent with the observations, both sources and hyperparameters. Firstly, we will consider that the observations probability distribution (i.e. the likelihood) is Gaussian, i.e.,

$$\text{prob}(\mathbf{v}|\mathbf{j},\beta) \propto \beta^{\frac{n_e}{2}} \exp\left[-\frac{\beta}{2} \parallel \mathbf{v} - \mathcal{L}\mathbf{j} \parallel^2\right], \tag{2}$$

where the hyperparameter β controls the sensor noise variability. For the prior, we will take

$$\text{prob}(\mathbf{j}|\alpha) \propto \alpha^{\frac{n_d}{p}} \exp\left[-\alpha \sum_i^{n_d} \sum_{l=1}^{\eta_i} |j_i - j_{i:l}|^p\right], \tag{3}$$

in which the hyperparameter α, named the *scale* hyperparameter, controls the contribution that the prior has on the source estimation compared to the contribution coming solely from the observations. This prior is named a *generalized* Gaussian prior as the parameter p, with $0 \leq p \leq 2$, allows for localization in different scenarios: $p = 2$ recovers the standard Gaussian prior [4]; $p = 1$ the Laplacian prior; and otherwise it accounts for intermediate situations. For p values smaller than 1, the generalized Gaussian prior allows for localization of sparser sources (this possibility has not been explored in this paper). The index l in Eq. (3) labels the different nearest neighbors at site i; the total number of nearest neighbors (per site) is represented by η_i. In constrast to regular (cubic) lattices, η_i in a realistic cortical mesh is not always the same but varies depending on i, and this is why we are denoting it by η_i.

To model the joint distribution we will make use of the hierarchical Bayesian paradigm in which the estimation is performed in two stages: first, over the two distributions, observations and prior, and second, over the hyperparameters. The joint global distribution is then conveniently written as

$$\text{prob}(\alpha,\beta,\mathbf{j},\mathbf{v}) = \text{prob}(\alpha)\text{prob}(\beta)\text{prob}(\mathbf{j} \mid \alpha)\text{prob}(\mathbf{v} \mid \mathbf{j},\beta). \tag{4}$$

The Gaussian distribution for the likelihood, Eq. (2), is known to be conjugated to Gamma distributed priors. This means that when a Gaussian likelihood is multiplied by a Gamma prior the resulting posterior distribution is Gamma distributed as well [18]. We use for this reason Gamma distributions to model the hyperpriors prob(β) and prob(α), given by

$$\text{prob}(\omega) = \Gamma(\omega|a_\omega^o, b_\omega^o) \propto \omega^{a_\omega^o - 1} \exp\left[-b_\omega^o \omega\right], \tag{5}$$

where $\omega > 0$ denotes either hyperparameter α or β, and $a_\omega^o > 0$ and $b_\omega^o > 0$ are the parameters defining the Gamma distribution, which are assumed to be known (see Appendix for further details about the Gamma distribution).

Now, with Eqs. (2), (3) and (5) the joint distribution Eq. (4) can be explicitly written and ready to perform Bayesian inference, which is based on calculating (or approximating)

$$\text{prob}(\alpha,\beta,\mathbf{j}|\mathbf{v}) = \text{prob}(\alpha,\beta,\mathbf{j},\mathbf{v})/\text{prob}(\mathbf{v}). \tag{6}$$

Note that if α and β were known we could use the so called Maximum a Posteriori (MAP) approach and estimate $\mathbf{j}$ by solving

$$\hat{\mathbf{j}} = \arg\max_{\mathbf{v}} \text{prob}(\mathbf{j}|\mathbf{v},\alpha,\beta) = \arg\min_{\mathbf{j}} \; \alpha \sum_i^{n_d} \sum_{l=1}^{\eta_i} |j_i - j_{i:l}|^p + \frac{\beta}{2} \parallel \mathbf{v} - \mathcal{L}\mathbf{j} \parallel^2 . \tag{7}$$

However, this procedure requires knowledge on model parameters that we may not have; furthermore, even if the model parameters are known it is of great interest to estimate the posterior distribution (for instance, by using its variance it allows to address how confident is our method on the provided estimates). In next subsection, we derive a Variational Bayesian approach for the estimation of the posterior distribution.

2.3 The Variational Bayesian approach

We present here an approximation to estimate the posterior distribution, Eq. (6). More concretely, the posterior distribution will be approximated by q($\alpha,\beta,\mathbf{j}$) , a distribution that factorizes over the hyperparameters, i.e.,

$$\text{q}(\alpha,\beta,\mathbf{j}) = \text{q}(\alpha,\beta)\text{q}(\mathbf{j}) \tag{8}$$

and which is found by minimizing the Kullback-Leibler (KL) distance [19] between $\mathrm{q}(\alpha,\beta,\mathbf{j})$ and $\mathrm{prob}(\alpha,\beta,\mathbf{j}\mid\mathbf{v})$, i.e.,

$$\begin{aligned} C_{KL}(\mathrm{q}(\alpha,\beta,\mathbf{j})\parallel\mathrm{prob}(\alpha,\beta,\mathbf{j}\mid\mathbf{v})) &= \int_\alpha\int_\beta\int_{\mathbf{j}}\mathrm{q}(\alpha,\beta,\mathbf{j})\log\left(\frac{\mathrm{q}(\alpha,\beta,\mathbf{j})}{\mathrm{prob}(\alpha,\beta,\mathbf{j}\mid\mathbf{v})}\right)d\alpha d\beta d\mathbf{j} \\ &= \int_\alpha\int_\beta\int_{\mathbf{j}}\mathrm{q}(\alpha,\beta,\mathbf{j})\log\left(\frac{\mathrm{q}(\alpha,\beta,\mathbf{j})}{\mathrm{prob}(\alpha,\beta,\mathbf{j},\mathbf{v})}\right)d\alpha d\beta d\mathbf{j}+\log\mathrm{prob}(\mathbf{v}), \end{aligned} \tag{9}$$

This quantity is always non negative and equal to zero if, and only if, $\mathrm{q}(\alpha,\beta,\mathbf{j})=\mathrm{prob}(\alpha,\beta,\mathbf{j}\mid\mathbf{v})$. Notice that in the last equality in Eq. (9) we are using $\mathrm{prob}(\alpha,\beta,\mathbf{j},\mathbf{v})$ which we know and not $\mathrm{prob}(\alpha,\beta,\mathbf{j}|\mathbf{v})$ which is unknown. Notice also that explicit knowledge of $\mathrm{prob}(\mathbf{v})$ is not needed in the estimation of $\mathrm{q}(\alpha,\beta,\mathbf{j})$.

The Generalized Gaussian prior considered here and given in Eq. (3) makes hard to perform the integral appearing in Eq. (9). To overcome this difficulty, we will approximate the original prior $\mathrm{prob}(\mathbf{j}|\alpha)$ by a lower bound that makes plausible the integration. In concrete, and similar to the strategy followed in [17], we will make use of the following quadratic function

$$M(\alpha,\mathbf{j},\mathbf{u})=\alpha^{N/p}exp\left[-\frac{\alpha p}{2}\sum_{i=1}^{n_d}\sum_{l=1}^{\eta_i}\left[\frac{(\Delta_i^l(\mathbf{j}))^2+\frac{2-p}{p}u_{i,l}}{u_{i,l}^{1-p/2}}\right]\right], \tag{10}$$

which satisfying that $\mathrm{prob}(\mathbf{j}|\alpha)\geq cM(\alpha,\mathbf{j},\mathbf{u})$ (with c a constant), makes easier the integration of the KL distance. The vector $\mathbf{u}$ has elements $u_{i,l}$ with values, which as we will see later, depend on the first-order spatial differences of the sources $\mathbf{j}$ with respect the distribution $\mathrm{q}(\mathbf{j})$, c.f. Eq. (16); in such a way the vector $\mathbf{u}$ is encoding the local-spatial sources activity. In other words, we are using a spatially varying Gaussian distribution to provide a majorization of the original prior distribution.

The operator Δ_i^l in Eq. (10) denotes the first-order difference with respect to the l neighbor of site i. By these considerations, it can be shown that

$$\mathrm{prob}(\alpha,\beta,\mathbf{j},\mathbf{v})\geq c\ \mathrm{prob}(\alpha)\,\mathrm{prob}(\beta)M(\alpha,\mathbf{j},\mathbf{u})\mathrm{prob}(\mathbf{v}|\mathbf{j},\beta)\equiv F(\alpha,\beta,\mathbf{j},\mathbf{u},\mathbf{v}), \tag{11}$$

where c is a constant. Thus the estimation of the posterior distribution is finally mapped to the minimization of an upper bound of the KL distance, i.e.,

$$C_{KL}(\mathrm{q}(\alpha,\beta,\mathbf{j})\parallel\mathrm{prob}(\alpha,\beta,\mathbf{j}\mid\mathbf{v}))\leq\int_\alpha\int_\beta\int_{\mathbf{j}}\mathrm{q}(\alpha,\beta)\mathrm{q}(\mathbf{j})\log\left(\frac{\mathrm{q}(\alpha,\beta)\mathrm{q}(\mathbf{j})}{F(\alpha,\beta,\mathbf{j},\mathbf{u},\mathbf{v})}\right)d\alpha d\beta d\mathbf{j}. \tag{12}$$

To summarize, using the Variational Bayesian approach and the lower bound in Eq. (11), the original problem, the estimation of the posterior distribution with all the unknowns has been formulated as finding the two distributions $\mathrm{q}(\mathbf{j})$ and $\mathrm{q}(\alpha,\beta)$ and the vector $\mathbf{u}$ which minimizes the right-hand side of Eq. (12). In next subsection, we provide details for an iterative algorithm to find such distributions.

2.4 The iterative Algorithm

1. Give the initial estimates of the distribution $\mathrm{q}(\alpha,\beta)$ and $\mathbf{u}$, represented by $\mathrm{q}^1(\alpha,\beta)$ and $\mathbf{u}^1$ respectively. Remark that for the calculation of $\mathrm{q}^1(\alpha,\beta)$ only the mean values of α and β are needed.
2. Do for $k=1,\ldots,$ (until convergence)
   2.1 Find the solution of

$$\mathrm{q}^k(\mathbf{j})=\arg\min_{\mathrm{q}(\mathbf{j})}\Big(\int_{\mathbf{j}}\int_\alpha\int_\beta\mathrm{q}^k(\alpha,\beta)\mathrm{q}(\mathbf{j})\times\log\left(\frac{\mathrm{q}^k(\alpha,\beta)\mathrm{q}(\mathbf{j})}{F(\alpha,\beta,\mathbf{j},\mathbf{u}^k,\mathbf{v})}\right)d\alpha d\beta d\mathbf{j}\Big), \tag{13}$$

   which is given by the probability distribution

$$\mathrm{q}^k(\mathbf{j})\propto\exp\{E_{\mathrm{q}^k(\alpha,\beta)}\left[\ln F(\alpha,\beta,\mathbf{j},\mathbf{u}^k)\right]\}. \tag{14}$$

   Notice that at iteration k the estimation of the distribution of $\mathbf{j}$ is Gaussian, whose mean is given by $E_{\mathrm{q}^k(\mathbf{j})}[\mathbf{j}]=cov_{\mathrm{q}^k(\mathbf{j})}[\mathbf{j}]E_{\mathrm{q}^k(\beta)}[\beta]\mathcal{L}^T\mathbf{v}$. If a point estimate of $\mathbf{j}$ is needed, this mean can be used. However, a key advantage of the Variational Bayesian approach is that it provides not only the point estimate but the whole probability distribution over $\mathbf{j}$. The covariance of the Gaussian distribution of $\mathbf{j}$ is given by

$$cov_{\mathrm{q}^k(\mathbf{j})}[\mathbf{j}]=\left(E_{\mathrm{q}^k(\beta)}[\beta]\mathcal{L}^T\mathcal{L}+pE_{\mathrm{q}^k(\alpha)}[\alpha]\sum_{i=1}^{n_d}\Delta_{\eta_i}^TW_{\eta_i}\left(\mathbf{u}^k\right)\Delta_{\eta_i}\right)^{-1},$$

where the matrix Δ_{η_i} have elements which are all zeros at the l-th row except for two positions: value of 1 at column i and of -1 at column $l \in \eta_i$. Thus, Δ_{η_i} has dimensions of $\eta_i \times n_d$. We have also defined that $W_{\eta_i}\left(\mathbf{u}^k\right) \equiv diag\left(\frac{1}{(u_{i,l}^k)^{1-p/2}}\right)$.

2.2 Find the solution of

$$\mathbf{u}^{k+1} = \arg\min_{\mathbf{u}}\Big(\int_\alpha\int_\beta\int_{\mathbf{j}} \mathrm{q}^k(\alpha,\beta)q^k(\mathbf{j}) \times \log\left(\frac{\mathrm{q}^k(\alpha,\beta)\mathrm{q}^k(\mathbf{j})}{F(\alpha,\beta,\mathbf{j},\mathbf{u},\mathbf{v})}\right) d\alpha d\beta d\mathbf{j}\Big), \tag{15}$$

which is given by the vector

$$u_{i,l}^{k+1} = \left(\Delta_i^l\left[E_{\mathrm{q}^k(\mathbf{j})}[\mathbf{j}]\right]\right)^2 + \frac{1}{n_d\ \eta_i}\sum_{j=1}^{n_d} trace\left[cov_{\mathrm{q}^k(\mathbf{j})}[\mathbf{j}] \times \Delta_{\eta_j}^T\Delta_{\eta_j}\right]. \tag{16}$$

Notice that the sum appearing in the second term in the right hand-side is constant and independent of i, thus the only dependence of $u_{i,l}^{k+1}$ in that second term comes from η_i.

2.3 Find the solution of

$$\mathrm{q}^{k+1}(\alpha,\beta) = \arg\min_{\mathrm{q}(\alpha,\beta)}\Big(\int_\alpha\int_\beta\int_{\mathbf{j}} \mathrm{q}(\alpha,\beta)q^k(\mathbf{j}) \times \log\left(\frac{\mathrm{q}(\alpha,\beta)\mathrm{q}^k(\mathbf{j})}{F(\alpha,\beta,\mathbf{j},\mathbf{u}^{k+1},\mathbf{v})}\right) d\alpha d\beta d\mathbf{j}\Big), \tag{17}$$

which is given by the probability distribution

$$\mathrm{q}^{k+1}(\alpha,\beta) = \mathrm{q}^{k+1}(\alpha)\mathrm{q}^{k+1}(\beta) \propto \exp\{E_{\mathrm{q}^k(\mathbf{j})}\left[\ln F(\alpha,\beta,\mathbf{j},\mathbf{u}^{k+1})\right]\}. \tag{18}$$

This equation ensures that $\mathrm{q}^{k+1}(\alpha)$ and $\mathrm{q}^{k+1}(\beta)$ are Gamma distributions. At the k-th iteration the estimation of the two hyperparameters α and β are given respectively by the expectations of $\mathrm{q}^{k+1}(\alpha)$ and $\mathrm{q}^{k+1}(\beta)$. The inverse of the expectations are given by:

$$\left(E_{\mathrm{q}^{k+1}(\alpha)}[\alpha]\right)^{-1} = \gamma_\alpha\frac{1}{\bar{\alpha}^o} + (1-\gamma_\alpha)\frac{p\sum_{i=1}^{n_d}\sum_{l=1}^{\eta_i}\left(u_{i,l}^{k+1}\right)^{\frac{p}{2}}}{n_d}, \tag{19}$$

and

$$\left(E_{\mathrm{q}^{k+1}(\beta)}[\beta]\right)^{-1} = \gamma_\beta\frac{1}{\bar{\beta}^o} + (1-\gamma_\beta)\frac{E_{\mathrm{q}^k(\mathbf{j})}\left[\|\ \mathbf{v}-\mathcal{L}\mathbf{j}\ \|^2\right]}{n_e}, \tag{20}$$

where we have defined the ratios $\bar{\alpha}^o \equiv a_\alpha^o/b_\alpha^o$ and $\bar{\beta}^o \equiv a_\beta^o/b_\beta^o$, where a_ω^o and b_ω^o were introduced in Eq. (5). The parameters $0 \le \gamma_\alpha \le 1$ and $0 \le \gamma_\beta \le 1$ are modeling our confidence on the parameters of the Gamma priors. When this confidence is zero, only the data are responsible for the estimation of α and β. When they are equal to 1, α and β are specified in advance by the user and fixed in the whole iterative process. The term $E_{\mathrm{q}^k(\mathbf{j})}\left[\|\ \mathbf{v}-\mathcal{L}\mathbf{j}\ \|^2\right]$ in Eq. (20) is calculated by using

$$E_{\mathrm{q}^k(\mathbf{j})}\left[\|\,\mathbf{v}-\mathcal{L}\mathbf{j}\,\|\right]^2 = \|\ \mathbf{v}-\mathcal{L}E_{\mathrm{q}^k(\mathbf{j})}[\mathbf{j}]\ \|^2 + trace\left[cov_{\mathrm{q}^k(\mathbf{j})}[\mathbf{j}]\mathcal{L}^T\mathcal{L}\right]. \tag{21}$$

Finally, it is important to remark that the algorithm presented here has an interesting limit: when $p = 2$ and without hyperparameter estimation, the method coincides with the so-celebrated LORETA algorithm [4]. Thus, fixing the values of α and β in the calculation of $E_{\mathrm{q}^k(\mathbf{j})}[\mathbf{j}]$ and $cov_{\mathrm{q}^k(\mathbf{j})}[\mathbf{j}]$ (that is, giving full credibility to the values provided by the user), needed in the algorithm stage 2.1, and observing that the W matrix (appearing into the covariance $cov_{\mathrm{q}^k(\mathbf{j})}[\mathbf{j}]$) is the identity matrix, one can write for the mean value that $\mathbf{j} = \mathcal{T}\mathbf{v}$ with $\mathcal{T} = \left(\mathcal{L}^T\mathcal{L} + \mu\sum_{i=1}^{n_d}\Delta_{\eta_i}^T\Delta_{\eta_i}\right)^{-1}\mathcal{L}^T$ and $\mu = \frac{2\alpha}{\beta}$, which is the LORETA solution [4]. Notice that this solution does not require any iteration, so a point estimation of the localization problem can be achieved by only inverting once the matrix $\mathcal{T}$.

2.5 Comparison to other similar existing methods

Together with the localization method presented in this paper, there are other methods that estimate both sources and hyperparameters. In [5,6,9] the authors used hierarchical linear priors incorporating a linear combination of quadratic priors. The hierarchical linear prior belongs to the family of parametric empirical Bayes approaches [5].

In ths family the prior covariance of the sources is defined as a linear mixture of covariance components. In this context, the model parameters at a given level can be treated as a prior for the level one-step below. Interestingly, this approach allows to introduce in the prior covariance different terms, such as a smoothness prior (3D Laplacian prior), a compensation term for the estimation bias of deep sources (a weighting prior), and even spatio-temporal location priors accounting for a set of activated regions at a given time-instant and/or location (by a diagonal matrix whose elements are equal to 1 if the source is active), see [6,9] for further details. This approach needs a careful definition of the activated regions to provide an appropriate localization; to overcome this problem, the authors suggested the use of anatomical information and fMRI-thresholded statistical maps. The sources and the hyperparameters are estimated via ReML (Restricted Maximum Likelihood) method, a variant of EM (Expectation Maximization). There are two main differences between the methods presented in [5,6,9] and our method here: first, we use (they did not) a Variational Bayesian approach which minimizes the KL-distance; second, our generalized Gaussian prior allows not only for the consideration of Gaussian priors (as they did) but for other classes of Lp-norm based priors.

In [5,6,9] the authors performed EM-based hyperparameters estimation, but Variational Bayesian localization was used as well in [10]. Here, the authors proposed a Gaussian prior where the covariance was a non-negative weighted sum of positive semidefinite matrices. Each positive semidefinite matrix defined a localization prior. A localization prior corresponds, for instance, to a single dipolar source at a given location or to sources with some spatial extent. Then the authors performed Automatic Relevance Determination principles to propose three different methods to estimate the weights, (in other words) the active positive semidefinite matrices. Our approach differs from the one in [10] in the use of a source prior with non quadratic energy which is able to estimate the location of punctual as well as reduced size areas of activity of the brain without specifying all their possible locations.

Despite the similarity of our approach to the methods mentioned here, the LORETA algorithm [4] has been by-far much more widely-used for the EEG source localization problem and it is for this reason why in section "Results" we choose the LORETA method to be the control for testing the performance of the our localization method. Further research will elaborate a more systematic (simulations-based) comparison with the other methods, the ones acknowledged here that use Gaussian priors, and other methods specifically desgined for focused localization, see for instance, [20,21].

2.6 Calculation of $\mathcal{L}$ in Eq. (1), the Lead Field matrix

Similarly to the approach utilized in [14], here the matrix $\mathcal{L}$ was calculated based on the template for the cortical-mesh which is included in SPM8 [22]. The initial 8196 vertices were downsampled to $n_d = 1200$ vertices. While the coarser mesh provided a less accurate cortex geometry, however, it signicantly reduced the computational cost. The matrix $\mathcal{L}$ was then computed using the BEM method from FieldTrip [23] in which sensors are located according to a 64 channel montage, canonical scalp and meshes for the outer and inner skull, all of them included in SPM8.

2.7 The two simulated scenarios: abrupt and smooth profiles of electrical activity

We have simulated two plausible scenarios. The first one is generated by a strong circular-area of electrical activity, 20mm radius and strength of 30 (adimensional). In this case, there are in total a number of 26 active dipoles of the possible 1200 existing in the cortical mesh, which can be easily eye-counted from Fig. 1A, considering the fact that each dipole is located at a vertex of the cortical mesh. This scenario is named the *abrupt* case.

The second scenario considers a similar but weaker circular-area (with the same radius of 20mm but with a weaker strength of 15) which smoothly is attenuating up to a distance of 150mm from the center of the circular area. In this case, there are 1039 active dipoles (of a total number of 1200 dipoles), see Fig. 2 A. This scenario is named the *smooth* case.

These two scenarios have been designed based on previous studies: first, the authors in [6] worked with a realistic head model (T1- weighted MRI based) of 12300 dipoles uniformly distributed throughout the entire brain volume and 61 EEG sensors to define priors locations (fMRI-based) with a spheric shape of 7 mm far from the central dipole. Only dipoles within this sphere were defined as active. Then, this mask was introduced into the prior to perform localization. The use of this spherical geometry has inspired the design of the two scenarios presented here. In another paper, for a very simplified head model (a three-spheres shell model) with 1716 dipoles and only 27 EEG electrodes, the authors in [9] simulated a single source as the one we have considered, cf. Fig .2 (top left) in [9]. Finally, in another paper and for a realistic mesh of 6004 dipoles the authors in [12] considered spatial profiles for the neuronal generators consisting in Gaussian blobs (in concrete two blobs) whose locations change with time, i.e. the activated sources have a spatio-temporal dynamics. Notice that, although we have not considered any dynamics in the generators, for a given snapshot the profile of the neural generators for a given blob in [12] is quite similar in comparison with our case.

2.8 Generation of observation vector v

The observation vector $\mathbf{v}$ is generated according to Eq. (1) for different levels of noise: 0dB, 10dB, 20dB, 30dB. The calculation of the Lead Field matrix is explained in section 2.6. The neuronal generators were simulated in two possible scenarios (c.f. section 2.7); thus Eq. (1) comes-out with two $\mathbf{v}$ vectors, each one for each of the situations. For each of the described scenarios, the values of $p = 1.00, 1.25, 1.50, 1.75, 2.00$ defining the prior and the two observed $\mathbf{v}$ are used to perform localization. In such a way, the algorithm performance is studied over *quenched*-noise observations rather than varying the noise and computing means over multiple noise-realizations.

2.9 Initial conditions used in the iterative Algorithm

The initial condition of the source activity was $\mathbf{j}^1 = \mathcal{L}^T\mathbf{v}$, which, according to Eq. (1), gives a *naive* backprojection estimation of the sources, see Figs. 1 and 2, "initial" label.

For the vector $\mathbf{u}^1$ we utilized $u_i^1 = \left(\Delta_i^x \mathbf{j}^1\right)^2 + \left(\Delta_i^y \mathbf{j}^1\right)^2 + \left(\Delta_i^z \mathbf{j}^1\right)^2$, neglecting the second term in the right-hand side of Eq. (16).

For the two hyperparameters α and β their initialization is flexible, and different values did perform equally well simply by keeping one constraint, the hyperparameter α has to be initialized about 10 orders (or more) of magnitude smaller than β. Otherwise, the algorithm resulted in a very high MSE. For example, if $\alpha^1 = 1$ then $\beta^1 = 10^{10}$ (or larger) did work well. This issue is related to the magnitude values of the Lead Field matrix; if the values are small as it occurs in our case with the calculation performed with FieldTrip (section 2.6), then β has to be initialized to a very large value in comparison to α. Another alternative would have been to rescale β to the values of the Lead Field matrix, but this possibility has not been explored here.

Without loss of generality, we use $\gamma_\alpha = 0$ and $\gamma_\beta = 0$. Thus, according to Eqs. (19) and (20), solely the data are responsible for the estimation of α and β, and no prior information on them is included.

3 Results

In this section, we applied the algorithm for source localization presented in section 2.4 to two possible scenarios: one in which the cortical sources are abruptly distributed and other in which sources are smoothly expanded throughout the entire cerebral cortex. The electrical activity in each scenario is plotted in Fig. 1A and Fig. 2A, label "original".

3.1 Algorithm performance

The algorithm performed source localization for different values of the shape parameter p. Fig. 1 shows the results for the abrupt scenario. The original sources are represented in Fig. 1A; next, the localization was performed in presence of two levels of noise, SNR=30dB (Fig. 1B) and SNR=10dB (Fig. 1C). In both Figs. 1B-C, the label "initial" with the initial condition in the iterative algorithm for the estimation (explained in section 2.9). After algorithm convergence, the final estimation is represented for values of $p = 1.0, 1.5, 2.0$. Looking at the "final" estimation, one can see by eye-inspection that for $p = 2$ the localization is more spread compared to the case of $p = 1$, and this happens for both situations high (10dB) and low noise (30dB). The having a *more spread* estimation means that some of the dipoles have a higher activity compared with the original sources, what will produce a larger Mean Squared Error (MSE) for the case of $p = 2$ compared to $p = 1$.

In a similar manner, Fig. 2 shows for smooth sources, the "original", "initial" and "final" estimates of source activity. Now in Fig. 2, in contrast with what happened for abrupt sources, the final estimate hardly depended on the noise level. Thus, for high noise of SNR= 10dB the estimation is p-independent (Fig. 2C). For low noise of SNR= 30dB, Fig. 2B shows that the estimation of $p = 1$ is worse compared to $p = 2$, and this can be observed by looking at the central area of the brain activity, the region in which sources have larger values. Here, the final estimation for $p = 1$ has more active sources within the central area, which will increase the MSE for $p = 1$ compared to $p = 2$.

The performance visualization shown in Figs. 1 and 2 was also quantified by the MSE between the final estimate and the original sources. The MSE versus the parameter p is illustrated in Fig. 3 for different values of noise (SNR=0dB, 10dB , 20dB and 30dB) and the two scenarios: abrupt (red lines) and smooth (blue lines). The high noise of 0dB shows that in this case neither for the abrupt nor for the smooth scenario there is a p-dependence for the estimation performance. At 10dB the MSE is not p-dependent for the smooth situation but it do depends in the abrupt scenario; indeed, one can see how the estimation for $p = 1$ performs with less MSE compared to $p = 2$, with a relative MSE improvement of the 167% (MSE of 32.90 for $p = 1$ vs 87.96 for $p = 2$). A similar tendency occured for 20dB noise. The MSE was quite-independent on p for smooth sources and better for $p = 1$ compared with $p = 2$ for the abrupt scenario (with a relative MSE improvement of the 317%; MSE=10.60 for $p = 1$ vs 44.23 for $p = 2$).

The MSE obtained at the very low noise of 30dB shows a peculiarity compared to higher noise situations: for smooth sources the case of $p = 2$ performed better than $p = 1$. The relative MSE improvement was for this case of the 100%, MSE=58.53 for $p = 1$ and MSE=29.26 for $p = 2$. When sources were abruptly generated, $p = 1$ performed better than $p = 2$, similar to the other situations in presence of a higher noise; the relative MSE improvement was of the 199% (MSE=8.73 for $p = 1$ and 26.11 for $p = 2$).

Finally, it is important to remark that in Fig. 3, the values of MSE versus p allows to monitor not only the extremal values of $p = 1$ and $p = 2$ but its tendency; thus, when $p = 1$ performed better than $p = 2$ (for abrupt sources at 10dB, 20dB and 30dB) the improvement occurred gradually modulated along the different values of p. And similarly for the smooth scenario at 30dB; the performance monotonously increased (the MSE decreased) in the direction from $p = 1$ to $p = 2$.

3.2 Validation of the hyperparameter estimation

We have shown in section 2.3 how the Variational Bayesian approach produces the Eqs. (19) and (20) to perform hyperparameter estimation. Herein, we have examined the quality of the estimated parameters α and β, and this is illustrated in Fig. 4 in panels A and C for abrupt sources and $p = 1$, and in panels B and D for smooth sources and $p = 2$.

To explore appropriately the 2D space (α, β), we have defined the hyperparameters ratio $\lambda \equiv \alpha/\beta$ to be the control parameter. If $\hat{\lambda}$ denotes the estimation achieved by the iterative algorithm, then values of $\lambda < \hat{\lambda}$ consider for the localization more dominant the observations than the prior, and viceversa, the values satisfying that $\lambda > \hat{\lambda}$ weight more the prior than the observations. This can be eye-observed for instance in Fig.4A. The brain sources corresponding to $\lambda = 10^{-1}\hat{\lambda}$ shows that the estimated sources are localized within a smaller region compared to the estimated solution as observations are more dominant than the prior. For values of $\lambda = 10\hat{\lambda}$, the localization spreads-out to the whole left hemisphere (the anterior brain –the front– coincides with the upper brain). A similar interpretation can be made to Fig. 4B.

The localization with no-hyperparameters estimation and $p = 2$ coincides with the localization performed by LORETA [4], what is represented along the MSE curve in Fig. 4D (all the points with $\lambda \neq \hat{\lambda}$).

For both plots, Figs. 4C-D the algorithm estimation gives an excellent hyperparameters estimation $\hat{\lambda}$, as the localization performed at that point falls very close to the minimum of the MSE curve.

4 Discussion

We have presented a new Bayesian method for localization of the EEG sources. The method uses a Varional Bayesian approach to derive an algorithm which allows for the simultaneous estimation of both hyperparameters and sources. In addition, the method incorporates a generalized Gaussian prior, in which the shape parameter $1 \leq p \leq 2$ operates as a control parameter in the model, allowing the transition from the standard Gaussian prior ($p = 2$) as in LORETA [4] to the Laplacian prior ($p = 1$). In between the two cases, other values of p can be explored by our algorithm.

We have performed the localization based on simulated EEG data corresponding to two different situations, one in which strong sources are originally compacted in a narrow region (named the abrupt situation) and a smooth situation in which the sources are widely spread throughout (almost) the entire cortical surface.

We have provided quantitative evidence that for abrupt sources, the case of $p = 1$ performs better (with smaller MSE) than $p = 2$, and this occurred for a wide regime of noise values (SNR=10dB,20dB, 30dB). For smooth sources, the estimation was p-independent for most of the noise simulated scenarios (0dB, 10dB and 20dB). However, for very low noise such as SNR=30dB, $p = 2$ performed with lower MSE than $p = 1$. This tendency was preserved from intermediate values of p, i.e., from $p = 1$ to $p = 2$ the MSE increased for abrupt sources and decreased for smooth ones.

We have quantitatively validated the hyperparameters estimation achieved by our algorithm. We have simulated different fixed values of hyperparameters and computed the MSE curve as a function of the hyperparameter ratio $\lambda \equiv \alpha/\beta$. For both abrupt and smooth sources, the values of the estimated hyperparameters are close to the minimum value of the MSE curve.

We have simulated different noise levels ranging from SNR=0dB to 30dB. Although in practical EEG source localization the SNR is generally below 0 dB, there are different possibilities to increase the SNR after processing the EEG data; one, for instance, is the performed during Event-Related Potentials (ERP) in which by averaging over multiple EEG time-series of same-conditions experiments, it is possible to increase the SNR up to values of about 10dB (this detail was acknowledged by one of the reviewers). Notice, however, that the ERP procedure is strongly

limited by the possibility of performing multiple repetitions of same-conditions experiments, useful for instance to study sensory, cognitive and behavioural neuroscience in which many pairs of stimulation-response can be recorded. However, this possibility does not exist in other situations such as during epileptic seizures, in which only a small number of recordings sessions are available (very often there is only one), so the ERP data manipulation is meaningless. Thus, even if the exploration of this possibility is beyond the scope of the present work, we believe that for these situations in which ERP manipulation is not possible to be performed, an alternative procedure to increase the SNR of the EEG time-series might be the consideration of spatio-temporal priors in which if two EEG time-events are close each other, it is more likely to have a higher temporal correlation between those two events compared with events which are far-separated in time-distance (the closer, the more correlated), similar to the modelization presented in [24,12]. For these reasons it is important to remark that the practical limitation of having the SNR of the EEG time series very low does not mean that the method presented here is not valid. Simply that in order to have a meaningful localization, one needs to perform inference over time-series in which the signal is enough to be decoded; otherwise the localization on purely noise is meaningless.

Last but not least, we want to comment on some possible extensions to the modelization we have presented here; our algorithm fixes p and then performs localization. But real physiological conditions might require a more self-adapting and dynamical mechanism for the parameter p. Unlike to our modelization, sources which are smooth or abrupt at one time-instant do not necessary remain invariant but may vary with the time-course of neural activity. Thus, for instance, during an epileptic seizure neural sources can be more abrupt distributed at the seizure onset (then better localized by $p = 1$) but later, after the seizure propagation, sources can become more smooth, thus $p = 2$ can be more appropriate for localization. Another interesting issue is related to the fact that in our modelization the parameter p is global and the same for all the possible cortical dipoles. But a more flexible and adaptable localization might require to have different values of p for different dipoles, which might have important consequences for real localization of cortical sources.

Acknowledgments

Work supported by MICINN, Ministerio de Ciencia e Innovacion, Ref. TIN2010-15137. JMC thanks previous funding from MICINN (Programa Ramón y Cajal) and currently from Ikerbasque: The Basque Foundation for Science. Special thanks to Martin Luessi, who performed the calculation of the *Lead Field* matrix used in this paper.

Appendix: Simple statistics of the Gamma distribution

Here, we mention some of the properties used in the text related with the Gamma distribution, c.f. Eq. (5). Its mean, mode and variance are given by:

$$\begin{aligned} \text{mean}[\omega] &= \frac{a_\omega^o}{b_\omega^o}, \\ \text{mode}[\omega] &= \frac{a_\omega^o - 1}{b_\omega^o}, \\ \text{variance}[\omega] &= \frac{a_\omega^o}{(b_\omega^o)^2}. \end{aligned} \tag{22}$$

Notice that the gamma distribution has a very interesting property, if we multiply a_ω and b_ω by $\lambda > 0$, the mean of the new Gamma distribution does not change but its variance is divided by λ. This allows the introduction of prior information on ω, very vague when λ is small or very precise when λ is large.

References

1. P. Nunez, *Electric Fields of the Brain: The Neurophysics of EEG* (Oxford University Press, 1981)
2. M. Hamalainen, R. Hari, R. Ilmoniemi, J. Knuutila, O. Lounasmaa, Review of Modern Physiscs **65**, 413 (1993)
3. S. Baillet, J. Mosher, R. Leahy, IEEE Signal Processing Magazine **18**, 14 (2001)
4. R. Pascual-Marqui, C. Michel, D. Lehmann, International Journal of Psychophysiologyl **18**, 49 (1994)
5. K. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, J. Ashburner, Neuroimage **16**, 465 (2002)
6. C. Phillips, M. Rugg, K. Friston, Neuroimage **17**, 287 (2002)
7. C. Phillips, M. Rugg, K. Friston, Neuroimage **16**, 678 (2002)
8. N. Trujillo-Barreto, E. Aubert-Vazquez, P. Valdes-Sosa, Neuroimage **21**, 1300 (2004)

9. C. Phillips, J. Mattout, M. Rugg, P. Maquet, K. Friston, Neuroimage **24**, 997 (2005)
10. D. Wipf, R. Ramirez, J. Palmer, S. Makeig, B. Rao, Advances in Neural Information Processing Systems **19** (2007)
11. K. Friston, L. Harrison, J. Daunizeau, S. Kiebel, C. Phillips, N. Trujillo-Barreto, R. Henson, G. Flandin, J. Mattout, Neuroimage **39**, 1104 (2008)
12. N. Trujillo-Barreto, E. Aubert-Vazquez, W. Penny, Neuroimage **39**, 318 (2008)
13. C. Bouman, K. Sauer, IEEE Transactions on Image Processing **2**, 296310 (1993)
14. M. Luessi, S. Babacan, R. Molina, J. Booth, A. Katsaggelos, Neuroimage **55**, 113 (2011)
15. C.M. Bishop, *Pattern recognition and Machine learning* (Springer, 2006)
16. S. Babacan, R. Molina, A. Katsaggelos, IEEE Transactions on Image Processing **17**, 326 (2008)
17. D. Babacan, R. Molina, A. Katsaggelos, *Generalized Gaussian Markov Random Field Image Restoration using Variational Distribution Approximation*, in *IEEE International Conference on Audio, Speech and Signal Processing* (2008), pp. 1265–1268
18. J. Berger, *Statistical Decision Theory and Bayesian Analysis* (Berlin, Springer-Verlag, 1985)
19. S. Kullback, *Information Theory and Statistics* (New York, Dover Publications, 1959)
20. J. Mosher, R. Leah, IEEE Transactions on Biomedical Engineering **45**, 1342 (1998)
21. R.G. de Peralta Menendez, S.G. Andino, G. Lantz, C. Michel, T. Landis, Brain Topography **14**, 131 (2001)
22. K. Friston, J. Ashburner, S. Kiebel, T. Nichols, W. Penny, *Statistical Parametric Mapping: The Analysis of Functional Brain Images* (Academic Press, 2006)
23. R. Oostenveld, P. Fries, E. Maris, J. Schoffelen, Computational Intelligence and Neuroscience **2011**, 156869 (2011)
24. J. Daunizeau, J. Mattout, D. Clonda, B. Goulard, H. Benali, J. Lina, IEEE Transactions on Biomedical Engineering **53**, 503 (2006)

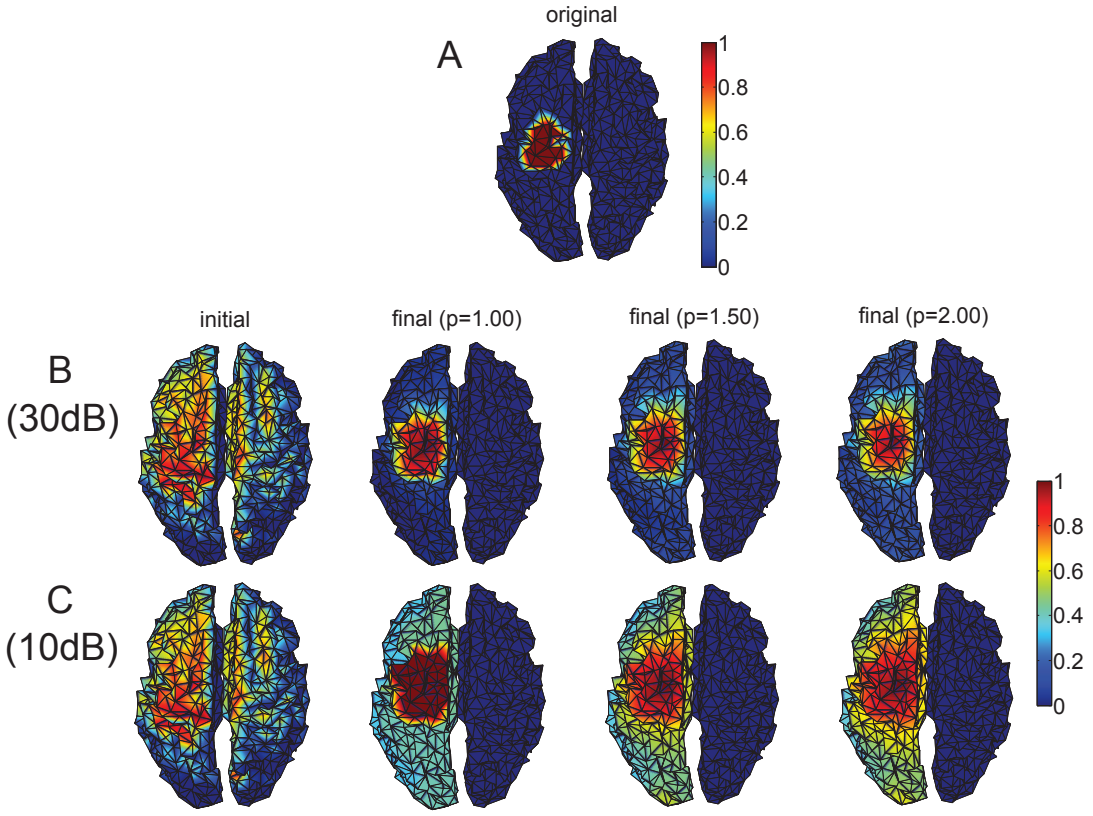

Fig. 1. Visualization of the algorithm performance for abrupt sources. A: Original sources. B: at SNR=30dB the initial and final algorithm estimates for $p = 1.00, 1.50, 2.00$. C: similar to B but for SNR=10dB. In all figures the neural activity has been rescaled to its maximum value, thus normalizing activities between the values 0 and 1.

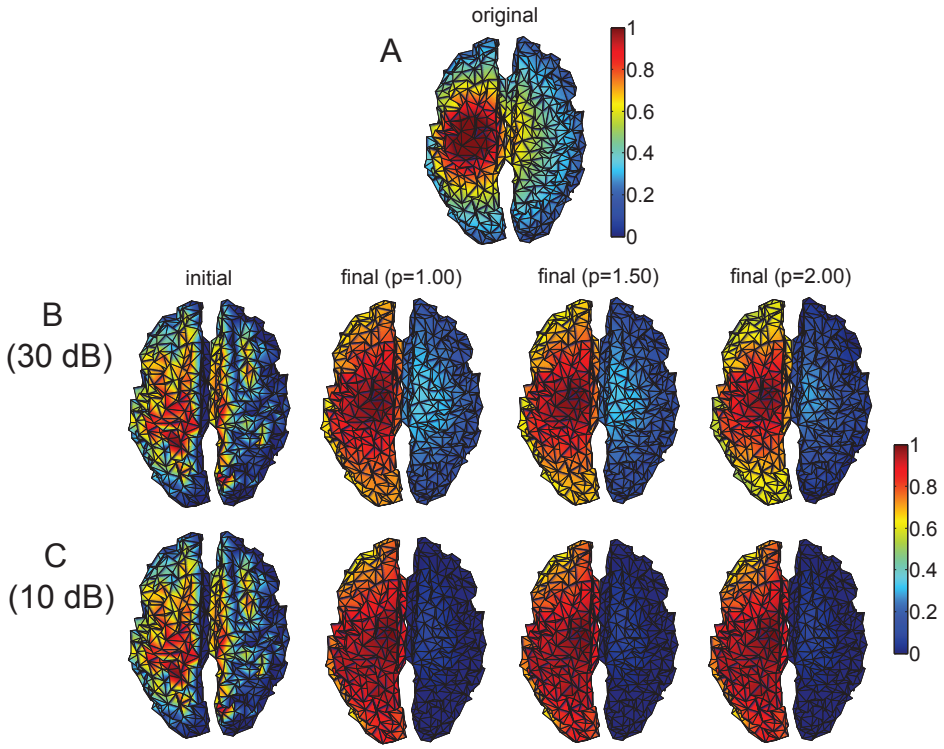

Fig. 2. Visualization of the algorithm performance for smooth sources. Similar to Fig.1 but for smooth sources (see Methods). Activity has been normalized between values 0 and 1 for all figures.

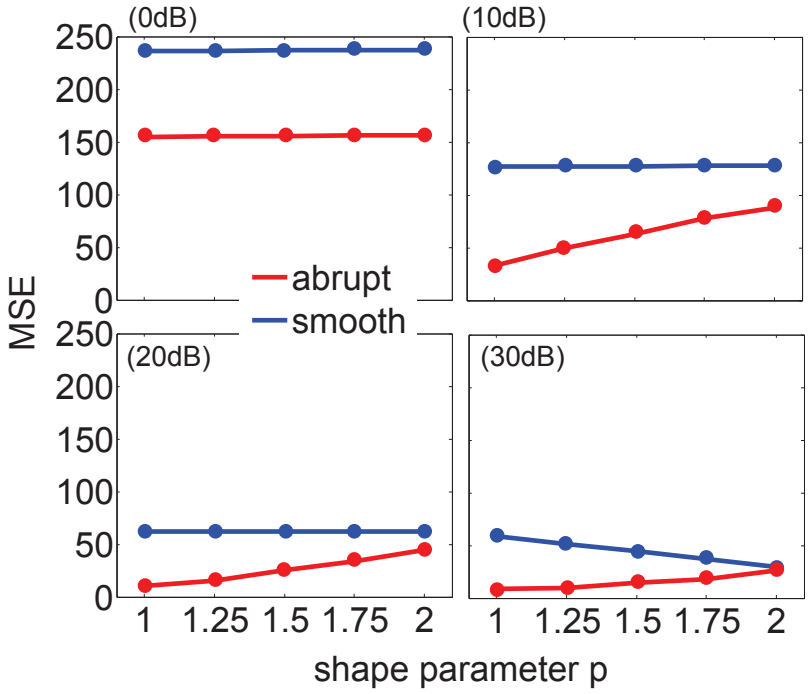

Fig. 3. Quantification of the algorithm performance. From very high (SNR=0dB) to very low noise (30dB), the MSE between the original sources and the final estimates is plotted versus the prior shape parameter p. Concretely we have simulated $p = 1.00, 1.25, 1.50, 1.75, 2.00$. Red lines are corresponding to the MSE obtained for abrupt sources and the blue lines are for the smooth situation.

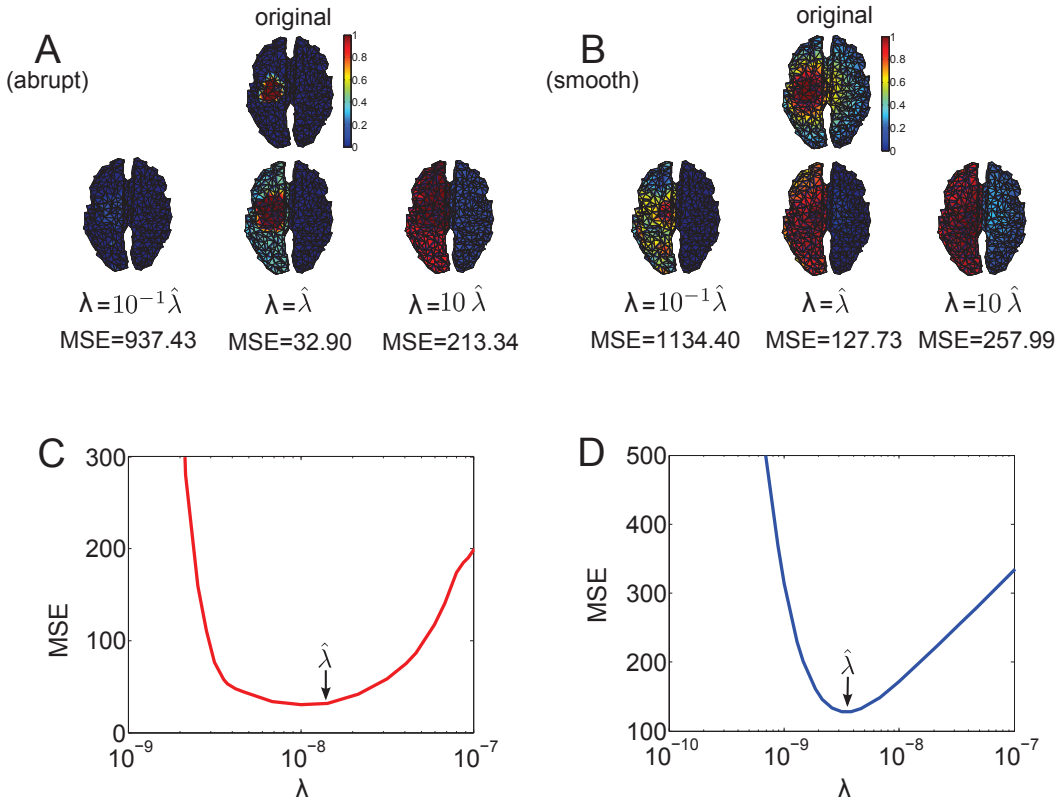

Fig. 4. Validation of the hyperparameter estimation. A,C: Abrupt sources and $p = 1$. B,D: Smooth sources and $p = 2$. A,B: Brain activity visualization for the original and the final estimates at different values of the hyperparameters ratio $\lambda \equiv \alpha/\beta$ (details in the text). The value of $\lambda = \hat{\lambda}$ is the one corresponding to the algorithm estimation, marked with an arrow in panels C (abrupt) and D (smooth). For illustration purposes, we also showed the final estimates and their corresponding MSE localization values for values of $\lambda = 10\hat{\lambda}$ and $\lambda = 10^{-1}\hat{\lambda}$. C,D: For both panels, the algorithm estimation performs close to the minimum of the MSE curve. A,B,C,D: Noise was fixed to SNR=10dB.