NeuroImage 84 (2014) 476–487

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Technical Note

Algorithmic procedures for Bayesian MEG/EEG source reconstruction in SPM☆

J.D. López [a,*], V. Litvak [b], J.J. Espinosa [c], K. Friston [b], G.R. Barnes [b]

[a] Departamento de Ingeniería Electrónica, Universidad de Antioquia, Medellín, Colombia
[b] Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, UK
[c] Universidad Nacional de Colombia, Medellín, Colombia

ARTICLE INFO

Article history:
Accepted 3 September 2013
Available online 13 September 2013

Keywords:
MEG/EEG inverse problem
Multiple Sparse Priors
Free energy
Bayesian model selection

ABSTRACT

The MEG/EEG inverse problem is ill-posed, giving different source reconstructions depending on the initial assumption sets. Parametric Empirical Bayes allows one to implement most popular MEG/EEG inversion schemes (Minimum Norm, LORETA, etc.) within the same generic Bayesian framework. It also provides a cost-function in terms of the variational Free energy—an approximation to the marginal likelihood or evidence of the solution. In this manuscript, we revisit the algorithm for MEG/EEG source reconstruction with a view to providing a didactic and practical guide. The aim is to promote and help standardise the development and consolidation of other schemes within the same framework. We describe the implementation in the Statistical Parametric Mapping (SPM) software package, carefully explaining each of its stages with the help of a simple simulated data example. We focus on the Multiple Sparse Priors (MSP) model, which we compare with the well-known Minimum Norm and LORETA models, using the negative variational Free energy for model comparison. The manuscript is accompanied by Matlab scripts to allow the reader to test and explore the underlying algorithm.

Introduction

In this technical note, we revisit the algorithmic procedures required for source reconstruction of EEG and MEG. We will take a pragmatic and algorithmic approach and focus on a particular function within the SPM software (Litvak et al., 2011) that implements Bayesian model inversion and transformations necessary for source reconstruction. This is a multi-function routine that performs: (*i*) a projection on spatial and temporal subspaces, (*ii*) model inversion and (*iii*) ensuing estimation of the maximum a posteriori cortical source estimates. We focus on single subject source reconstruction (although the routine can handle multi-modal data from multiple subjects), as a vehicle to link the mathematical concepts with their implementation and show how they can be unpacked in terms of Matlab pseudo-code. Our aim is to provide an operational understanding of the reconstruction scheme so that the reader can change various parameters and examine the resulting effects on source reconstruction results. Our focus is more on the algorithmic architecture and implementation, rather than on providing a comprehensive survey of the underlying theory. Having said this, we make every effort to motivate each step of the scheme in terms of its theoretical principles.

In MEG/EEG source reconstruction based on the Bayesian framework, we briefly review the basic linear model upon which source reconstruction is based. We will see that the key ingredient is the specification of the prior covariance of source activity. This prior covariance accommodates the basic distinctions between commonly employed regularisation schemes in the source reconstruction literature, and is generalised by the use of multiple and sparse spatial priors. In Pre-processing stage, we start with a brief review of the necessary pre-processing components required for the inversion. We will cover the specification of both spatial and temporal modes in channel space and how these are used to finesse the estimation of parameters controlling the prior covariance above. In Inversion scheme we turn to the inversion routine itself. We survey the different approaches to optimising the prior covariance parameters and how their form relates to different prior assumptions about the distribution of cortical activity.

Having established the overall structure of the scheme, in Simulation example we present some illustrative examples showing how one can manipulate various parameters to change the data features that are reconstructed. Some of these parameters can be specified as arguments or inputs to the routine, whereas others (such as the number of sparse prior components) can be changed by modifying the code. We first demonstrate how to produce synthetic data with a known source geometry and then illustrate the differences between solutions based on the prior covariance models implicit in the Minimum Norm, LORETA-like, and sparse priors inversion schemes.

* Corresponding author.
E-mail address: josedavid@udea.edu.co (J.D. López).

1053-8119/$ – see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2013.09.002

These simulations are not meant to be exhaustive explorations of prior models but are used to illustrate how the readers can reproduce the results that follow to optimise the scheme for their own work. The simulated dataset and results presented in this paper are available in the Supplementary material or via the SPM website.

MEG/EEG source reconstruction based on the Bayesian framework

The raw MEG and EEG data can be regarded as sets of waveforms or scalp topographies that change over time. When attempting to reconstruct the neural activity producing these topographies, one usually assumes that they are generated by a set of discrete brain sources. Each of these sources is formed by a group of neurons (around 10^4) whose membrane potentials fluctuate synchronously over a certain time scale. The signal generated by such a group of neurons can be represented by a current dipole that gives rise to an electrical potential difference on the scalp or generates magnetic fields measurable outside the head.

There are two dominant frameworks for M/EEG neural source activity reconstruction: (*i*) Assume a small number of active sources and fit to the data using a non-linear search throughout the brain (Supek and Aine, 1993). (*ii*) Use a large number of fixed dipoles that fill the search space (the grey matter surface for example) and estimate their amplitude (Dale and Sereno, 1993; Hämäläinen and Ilmoniemi, 1984). The first (few sources) approach is powerful but severely compromised by the emergence of local extrema in the objective function as the number of sources increases. The second (distributed) approach has the advantage that the model is linear with respect to neuronal currents, but the large number of unknowns creates an ill-posed problem that can only be solved by including prior information (see Baillet et al., 2001; Grech et al., 2008; Michel et al., 2004; Pascual-Marqui, 1999 for reviews on the field). In the recent years, major efforts have been devoted to distributed solutions because they are linear and independent of the number and characteristics of activated regions. Furthermore, using strategies to reduce the noise and search space size, distributed solutions have become robust and computationally feasible.

Distributed neural source activity reconstruction

The distributed solution is based on the linear mapping between the dipole moments for a fixed set of dipoles distributed inside the brain and a set of signals recorded by electrodes/gradiometers placed outside the head. This relation is given by Dale and Sereno (1993):

$$Y = LJ + \epsilon \tag{1}$$

where the MEG/EEG dataset $Y \in \Re^{N_c \times N_n}$ is formed by N_c sensors and N_n time samples, and the neural source activity $J \in \Re^{N_d \times N_n}$ is represented by the amplitude of N_d current dipoles distributed through the cortical surface — generally, with fixed orientations perpendicular to the surface. The data and sources are related through the gain matrix L (also known as the lead field matrix), and the measurements are affected by zero mean Gaussian noise ϵ with covariance: $\mathrm{cov}(\epsilon) = Q_\epsilon$.

The selection of a distributed approach ($N_d \gg N_c$) means that the lead field matrix L is non-invertible, and that the source estimates $\hat{J}$ cannot be recovered directly. This problem can be solved within the Bayesian framework by assuming a priori that J is a zero mean Gaussian process with covariance: $\mathrm{cov}(J) = Q$. The representation of the MEG/EEG inverse problem within the Bayesian framework has been widely studied (Auranen et al., 2005; Baillet and Garnero, 1997; Phillips et al., 1997; Sato et al., 2004; Schmidt et al., 1999; Trujillo-Barreto et al., 2004; Wipf and Nagarajan, 2009). Within this framework, source estimates can be expressed as the expected value of the posterior distribution of the source activity given the data: $\hat{J} = E[p(J|Y)]$. This estimate can be computed using Bayes' theorem to define $p(J|Y)$ in terms of known distributions:

$$p(J|Y) = \frac{p(Y|J)p(J)}{p(Y)}. \tag{2}$$

Here the evidence $p(Y)$ can be neglected, because it is a constant value for a given dataset:

$$p(J|Y) \propto p(Y|J)p(J) \tag{3}$$

In other words, the objective is to obtain the current source distribution J based on the dataset Y, where the prior probability of the source activity $p(J)$, is what we expect before observing the data. The likelihood $p(Y|J)$, gives us the probability of the data for a given source activity $p(Y|J) = \mathcal{N}(LJ, Q_\epsilon)$, with $\mathcal{N}(\cdot)$ the multivariate Gaussian probability distribution. Given that the prior and likelihood are Gaussian, the right hand side of Eq. (3) can be expressed as:

$$p(Y|J)p(J) \propto \Theta = \exp\left(-\frac{1}{2}(LJ-Y)^T Q_\epsilon^{-1}(LJ-Y) - \frac{1}{2}J^T Q^{-1} J\right) \tag{4}$$

where $(\cdot)^T$ denotes the transpose operator. The optimal value of source activity is the value that minimises Θ, which is equivalent to finding the source activity where the gradient of $\log(\Theta)$ is zero:

$$\frac{d(\log\Theta)}{dJ}\Big|_{J=\hat{J}} = 0 = -L^T Q_\epsilon^{-1}\left(L\hat{J} - Y\right) - Q^{-1}\hat{J} \tag{5}$$

gives $\hat{J}$ (Dale and Sereno, 1993):

$$\hat{J} = QL^T\left(Q_\epsilon + LQL^T\right)^{-1} Y. \tag{6}$$

This is the canonical equation used in all distributed source reconstruction algorithms based on Gaussian assumptions (see Liu et al., 2002, Appendix, for other approaches to obtain this equation). Since the data Y are known – and the lead fields can be computed based on a physical model of the head – one only requires estimates of the sensor and source level covariances to compute the source currents J with a single algebraic step.

So the problem of finding a distributed solution reduces to finding a good estimate of the two covariance matrices Q, and Q_ϵ (Baillet and Garnero, 1997; Phillips et al., 1997, 2005). This is the main objective of the steps described below. To render the estimation of the source covariance matrix more computationally efficient, this estimation is preceded by several data reduction steps.

Selection of the prior covariance components

The accuracy of the reconstructed image of source activity is highly dependent on the constraints implicit in the form of Q and Q_ϵ used in Eq. (6). In absence of information about noise over sensors, one generally assumes a sensor noise covariance matrix of the form: $Q_\epsilon = h_0 I_{N_c}$, where $I_{N_c} \in \Re^{N_c \times N_c}$ is an identity matrix, and h_0 is the sensor noise variance. That is, the amount of noise variance is the same on all sensors (uniformity). This covariance parameter can also be viewed as a regularisation parameter (Golub et al., 1979; Hansen, 2000) or hyperparameter (Phillips et al., 2002b). Prior information about sensor noise can also be based on empty room recordings — and some estimate of empirical noise covariance can enter as an additional covariance component at the sensor level (Henson et al., 2011).

Single covariance matrix based approaches

There are multiple constraints that can be used as prior source covariance matrix Q. The simplest (Minimum Norm) assumption about

the sources is that all dipoles have approximately the same prior variance and no covariance (Hämäläinen and Ilmoniemi, 1984):

$$Q = h_0 I_{N_d}. \tag{7}$$

Another assumption is to consider that the sources vary smoothly over space — as assumed in the LORETA model (Pascual-Marqui, 1999; Pascual-Marqui et al., 1994). One such smoothing function was proposed in Harrison et al. (2007): a Green's function based on a graph Laplacian was computed using the vertices and faces provided from a cortical surface mesh (derived from a structural MRI). The graph Laplacian $G_L \in \Re^{N_d \times N_d}$ is based on an adjacency matrix $\mathcal{A} \in \Re^{N_d \times N_d}$, with $\mathcal{A}_{ij} = 1$ if there is face connectivity (maximum six neighbours for each voxel), and zero otherwise. The graph Laplacian is then defined as:

$$G_{L_{ij}} = \begin{cases} -\sum_{k=1}^{N_d} \mathcal{A}_{ik} & , \text{for } i = j, \text{with } \mathcal{A}_{i\cdot} \text{ the } i\text{-th row of } \mathcal{A} \\ \mathcal{A}_{ij} & , \text{for } i \neq j \end{cases}. \tag{8}$$

Note that the sum of each column of G_L is zero. Finally, Green's function $Q_G \in \Re^{N_d \times N_d}$ is defined as:

$$Q_G = e^{\sigma G_L} \tag{9}$$

with σ a positive constant value that determines the smoothness of the current distribution or spatial extent of the activated regions. A LORETA-like solution can be obtained by using Green's function $Q = h_0 Q_G$ In other words, replacing the identity matrix of the Minimum Norm solution with a smooth prior covariance component.

As superficial sources in M/EEG have a much larger impact on the sensors than deeper ones, both Minimum Norm and LORETA tend to produce solutions with a superficial bias (as these solutions can explain most of the data with the least source power). There have been several modifications to these algorithms to correct for this bias by means of column weighting (Fuchs et al., 1999; Hauk, 2004; Ioannides et al., 1990; Lin et al., 2006) or normalisation by noise (Dale et al., 2000; Pascual-Marqui, 2002) but we do not consider them here.

The assumption that all the dipoles are active at the same time tends to make the final solution smooth but also renders it sensitive to external artefacts (i.e., there will be a tendency to explain artefacts in the source space). An alternative approach known as beamforming (Hillebrand et al., 2005; Sekihara et al., 1999; Van Veen et al., 1997) actively attempts to remove smoothness (or covariance) from the solution; these algorithms have excellent robustness to noise but suffer when there is true source covariance. In this case, Q is a diagonal matrix formed from a direct projection of the data (covariance matrix) into the source space (Belardinelli et al., 2012).

Multiple Sparse Priors

The classical approaches above can be generalised within the Bayesian framework, by considering the prior source covariance as the weighted sum of multiple prior components: $C = \{C_1, \ldots, C_{N_q}\}$, commonly known as Empirical Bayes (see Wipf and Nagarajan, 2009 for a review on its treatment in source reconstruction):

$$Q = \sum_{i=1}^{N_q} h_i C_i. \tag{10}$$

Here, each $C_i \in \Re^{N_d \times N_d}$ is a prior source covariance matrix, and can take any form. For simplicity we consider the case where prior component corresponds to a single potentially activated region of cortex. The hyperparameters $h = \{h_1, \ldots, h_{N_q}\}$ weight these covariance components. Regions with large hyperparameters will have large prior variances. Note that these components may embody different types of informative priors, e.g., different smoothing functions, medical knowledge, fMRI priors (Henson et al., 2011). The choice of the set of prior components C used in Eq. (10) determines the sets of prior assumptions that define the model; and specific forms of C can be used to emulate standard source reconstruction approaches. For the Minimum Norm solution, for example, the set is just one identity matrix: $C = I_{N_d}$, and for the LORETA-like solution it will be a smoothed version $C = Q_G$.

In the absence of prior information, the most inclusive set C should have the same number of components as there are dipoles distributed through the source space (around 8000). However, this (over-complete) set precludes beliefs or constraints on source activity: the number of components usually considered is of the same order as the number of channels (<500). As we know a-priori that neuronal current flow has some local coherence, we model the basic unit of current flow as a spatially smooth impulse (Green's) function at selected vertices on the cortical surface. The size and the number of the ensuing patches can be defined based on prior knowledge (López et al., 2012a). Current implementations are based on fixed sets of patches. For example, the SPM software package uses a set of $N_q = 512$ patches covering the entire cortical surface (Fig. 1(b)), the centres of these patches are a sparse sample of the original set of dipoles used to form the lead field matrix (Fig. 1(a) shows a set of $N_d = 8196$ dipoles for the "normal" grid in SPM). Fig. 1(c) shows different sizes of patches obtained with different values of σ in Eq. (9); i.e., they can be modified if there is prior knowledge about the size or location of the region of neural activity.

Rather than each covariance component corresponding to a single patch, this set can then be supplemented by a further N_d / 2 covariance components; in which patches in opposite hemispheres are correlated.

The priors for different inversion schemes are summarised in Fig. 2. The Minimum Norm prior is an identity matrix (Fig. 2(a)), while the LORETA prior is based on a fixed smoothing function that couples nearby sources (Fig. 2(b)). Finally, MSP is based on a library of hundreds of covariance components, each corresponding to a different locally smooth focal region (or patch) of cortex. Figs. 2(c) and (d) show two possible covariance components, corresponding to two distinct cortical patches. After the optimisation process (reviewed in the following section) the prior covariance matrix Q of the MSP will be formed by a linear mixture of covariance components from this library of priors.

Exogenous source priors

If there is prior knowledge that activity is restricted to a volume of interest, the dipoles outside this volume can be masked and the solution will be forced to be inside the specified volume. However this procedure can lead to errors as all data, regardless of origin, will be explained by activity in this volume. An alternative and preferred approach is to use a soft constraint by creating extra components in the set C that specify sources inside the volume of interest. For example, when functional MRI (fMRI) data are available, regional activations can be included as extra components in C, translating the fMRI information into candidate MSP patches within the library (see Henson et al., 2010). These must be soft constraints as one cannot assume that volume showing fMRI responses will necessarily contribute to MEG/EEG data. Note that incorporating prior knowledge in this way does not bias the estimate of source activity — rather it allows the estimate to take non-zero values.

Pre-processing stage

The source reconstruction scheme implemented in SPM allows group-based inversions with multiple modalities (MEG and/or EEG, see Litvak and Friston, 2008), but to simplify things, we will restrict ourselves to a single subject and a single modality (MEG).

Before source estimation, several stages are required to prepare the data for the inversion. Principally, this involves data reduction to increase effective SNR and decrease the computational burden on the subsequent optimisation.

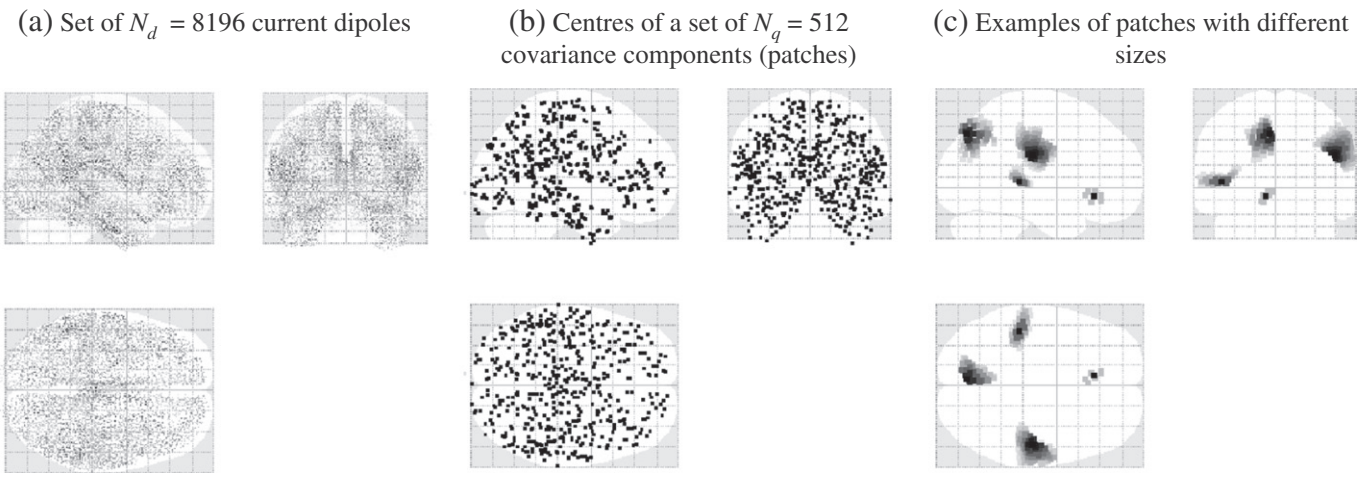

Fig. 1. Glass brain showing sagittal, axial and coronal views of the vertex and patch centres. (a) The sources of neural activity are limited to this set of current dipoles distributed over the cortical surface. (b) Each dot represents the centre of an MSP patch. (c) The parameter of the Green's function controls the size of the focal regions.

Spatial projector

Here, the sensor space is transformed to a subset of orthogonal sensors (or spatial modes) with a singular value decomposition (SVD) over the lead field matrix (Gener and Williamson, 1998; Phillips et al., 2002a). The problem with using the original number of sensors is that there is some redundancy of information due to the high correlations between nearby sensors; this redundancy adds unnecessarily to the computational load. The use of a spatial projection over sensor space finesses these problems by generating a new set of orthogonal sensors. For example, Fig. 5(a) shows the singular values of a lead field matrix with $N_m = 274$ sensors, illustrating that only the few spatial modes could actually be generated by the lead field.

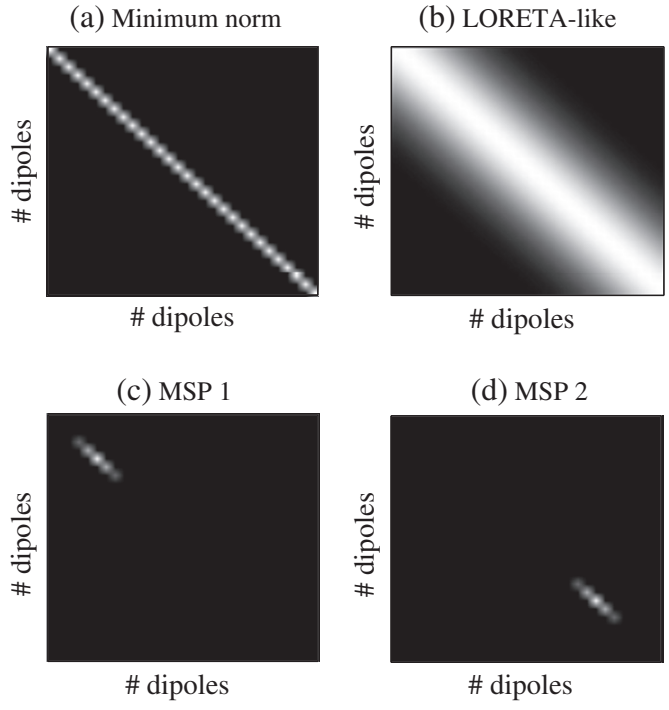

Fig. 2. Priors for different inversion schemes. (a) Minimum Norm solution does not include spatially structured prior information. (b) LORETA-like is based on a smoother, coupling each dipole with its nearest neighbours. (c) and (d) MSP is based on a set of covariance components, each with a different possible location (here two examples).

This procedure starts by selecting the lead field matrix and reducing it into the space of singular values: $USV^T = LL^T > e^{-16}$, where $U \in \Re^{N_c \times N_m}$ is the transformation matrix from the sensor space to the space of the $N_m \leq N_c$ largest singular values (spatial modes larger than e^{-16}), that forms the spatial projector, so that the new data becomes AY, with $A = U^T$. Finally the gain matrix is projected into the new sensor sub-space:

$$\mathbb{L} = AL. \tag{11}$$

Typically this orthogonal set comprises the first 100 eigenmodes for 274 sensors.

Temporal projector

The inclusion of temporal data reduction helps to reduce noise, and guarantees a continuous temporal evolution of the estimated brain activity. Again the temporal domain data is transformed into a sub-space of its principal singular components or temporal modes (Phillips et al., 2002a). Being orthogonal, each component can be regarded as an independent waveform. Effectively, this allows the joint inversion of a small number of "instantaneous" forward problems, where the data are summarised by the spatial patterns generating the temporal modes.

Applying the spatial projector

EEG/MEG data typically contains several minutes of recordings, separated into trials and conditions that we will assume have been averaged to produce the dataset Y. This averaging by itself results in significant noise reduction, but the sensors still share a large amount of information. This redundancy can be reduced by projecting the dataset into N_m orthogonal virtual sensors using the spatial projector:

$$\widetilde{Y} = AY. \tag{12}$$

In this case we define our source space based on a cortical manifold (and its constituent lead fields) but other elegant approaches (Taulu and Kajola, 2005) have used Maxwells equations to define the signal subspace within the head volume.

Computation of the temporal projector

In computing the temporal projector one can suppress signals at early and late peri-stimulus time to accentuate event related

(or induced) responses, using a Hanning window $W \in \Re^{N_n \times N_n}$ to compute the data covariance over time bins:

$$\widetilde{W} = W^T\left(\widetilde{Y}^T\widetilde{Y}\right)W. \tag{13}$$

Then, a discrete cosine transform (DCT) is used to filter the data using a transformation matrix $K \in \Re^{N_n \times N_f}$ where K contains DCT coefficients corresponding to the frequency window of interest:

$$\widetilde{K} = K^T\widetilde{W}K. \tag{14}$$

If there are multiple modalities a normalisation is performed to remove scaling differences (see Henson et al., 2009), and the filtered datasets of all modalities are averaged into a single $\widetilde{K} \in \Re^{N_f \times N_f}$ matrix. At this point the main diagonal of $\widetilde{K}$ contains the main frequencies present in the data as shown in Fig. 5(b), where a pure sinusoid signal of 22 Hz can be observed.

Applying a SVD over the filtered data covariance: $\widetilde{U}\widetilde{S}\widetilde{V}^T = \widetilde{K} > e^{-8}$, gives N_r orthogonal temporal modes. The temporal projector $P \in \Re^{N_n \times N_r}$ is then obtained with:

$$P = W\widetilde{U}. \tag{15}$$

As with the spatial projector, the temporal projector is an orthogonal basis set that is used to perform a linear transformation. In this case the model reduction is considerably higher (from thousands of time bins to around $N_r = 16$ temporal modes in SPM).

Data (and model) reduction

Both spatial and temporal data reductions are simple to implement. For the spatial projector it is only necessary to change the lead field matrix with $\mathbb{L}$ from Eq. (11). The temporal projector is applied directly to the data, which is in turn reduced to the space of orthogonal sensors:

$$\mathbb{Y} = AYP \tag{16}$$

with $\mathbb{Y} \in \Re^{N_m \times N_r}$. Note that all the mathematical formalism in this paper can be equally portrayed using $\mathbb{L}$ and $\mathbb{Y}$ instead of L and Y. For simplicity, the original notation will be retained; however, all the simulation examples were performed using the reduced forms.

Inversion scheme

In this section, the model specific (Minimum Norm—IID, LORETA, LOR, and Multiple Sparse Priors—MSP) covariance components are used to provide an empirically optimised weighted mixture Q matrix. This estimated source covariance matrix enters the final optimisation stage (along with the sensor level covariance matrix) to estimate the current density $\hat{J}$, using Eq. (6).

Optimisation over averaged data

For any set of covariance components C, the optimisation is based on a standard "Variational Laplace" scheme with the negative variational Free energy (henceforth "Free energy") as the objective function (Friston et al., 2008). Variational Laplace is an approximate Bayesian inference scheme that assumes that the posterior is Gaussian (the Laplace assumption). The objective is to obtain the set of hyperparameters that maximise the evidence for the data (or Free energy). The resulting hyperparameters will be used to form the prior source covariance matrix Q in a second inversion stage.

This implementation is made computationally feasible by using the (reduced) sensor rather than source level covariance matrix (see Appendix A for a full derivation of this equation):

$$\Sigma = Q_\epsilon + LQL^T \tag{17}$$

where Q_ϵ is the sensor noise covariance. This definition allows us to project the source covariance components into the (typically more compact) sensor space. Now, given a set of arbitrary source covariance components C_i and sensor noise covariance Q_ϵ, the sensor covariance can be modelled as:

$$\Sigma = e^{\lambda_0}Q_\epsilon + \sum_{i=1}^{N_q} e^{\lambda_i}LC_iL^T, \tag{18}$$

the change of variable $h_i = e^{\lambda_i}$ guarantees positive values, a convex optimisation, and Gaussian assumptions on the prior of hyperparameters (Friston et al., 2008; Wipf and Nagarajan, 2009).

There are two advantages of using Σ instead of Q for optimising the hyperparameters: the size of the matrices is significantly reduced due to the projection into the sensor space, and the inclusion of the noise variance into the equation allows the regularisation parameter to be treated as another hyperparameter.

Free energy as an objective or cost function

For the linear Gaussian models underlying source reconstruction, the model evidence $p(Y)$ (see Eq. (2)) is well approximated by the variational Free energy (Friston et al., 2007b; Penny, 2012; Wipf and Nagarajan, 2009). The Free energy is used as the cost function to fit the modelled covariance (determined by the hyperparameters λ in Eq. (18)) to the data covariance: $\Sigma_Y = \frac{1}{N_t}YY^T$. The Free energy can be expressed as Friston et al. (2007b):

$$\begin{aligned} F = &-\frac{N_n}{2}\mathrm{tr}\left(\Sigma_Y\Sigma^{-1}\right) - \frac{N_n}{2}\log|\Sigma| - \frac{N_nN_c}{2}log2\pi \\ &-\frac{1}{2}\left(\hat{\lambda}-\nu\right)^T\Pi\left(\hat{\lambda}-\nu\right) + \frac{1}{2}\log|\Sigma_\lambda\Pi| \end{aligned} \tag{19}$$

where $|\cdot|$ is the matrix determinant operator. Here, we consider the prior: $q(\lambda)$, and approximate posterior: $p(\lambda)$, densities of the hyperparameters as Gaussian:

$$q(\lambda) = \mathcal{N}\left(\lambda;\nu,\Pi^{-1}\right) \qquad p(\lambda) = \mathcal{N}\left(\lambda;\hat{\lambda},\Sigma_\lambda\right). \tag{20}$$

Each term of the Free energy can be expressed in words as follows:

$$\begin{aligned} F = &-\begin{bmatrix}\text{Model}\\ \text{error}\end{bmatrix} - \begin{bmatrix}\text{Size of model}\\ \text{covariance}\end{bmatrix} - \begin{bmatrix}\text{Num of data}\\ \text{samples}\end{bmatrix} \\ &-\begin{bmatrix}\text{Error in}\\ \text{hyperparameters}\end{bmatrix} + \begin{bmatrix}\text{Error in covariance}\\ \text{of hyperparameters}\end{bmatrix}. \end{aligned}$$

The Free energy can be divided into accuracy and complexity (Penny, 2012). The accuracy is given by the model error, the size of the model based covariance Σ, and the number of data samples. The complexity is the key difference between the Free energy and other Bayesian approaches (Wipf and Nagarajan, 2009). The complexity acts as a penalty term and defines the "distance" between the prior and posterior hyperparameter means and covariances.

The optimal combination of hyperparameters is achieved for the maximum Free energy value: $\hat{\lambda} = arg\,max_\lambda F$, where the Free energy approximates the log evidence. The maximum of this function can be located with a gradient ascent, which is based on the gradient and Hessian of the Free energy (Friston et al., 2008). The gradient is calculated as the derivative of Eq. (19) with respect to the hyperparameters:

$$\frac{\partial F}{\partial \lambda_i} = -\frac{N_n}{2}\mathrm{tr}(D_i(\Sigma_Y - \Sigma)) - \Pi_{ii}(\lambda - \nu) \tag{21}$$

with

$$D_i = \frac{\partial \Sigma^{-1}}{\partial \lambda_i} = e^{\lambda_i}\Sigma^{-1}C_i\Sigma^{-1} \tag{22}$$

and its curvature is obtained with the derivative of the gradient:

$$\frac{\partial^2 F}{\partial\lambda_i\partial\lambda_j} = -\frac{N_n}{2}\mathrm{tr}\left(D_i C_i D_i C_j\right) - \Pi_{ii}. \tag{23}$$

This gradient ascent is known as the Newton non-linear search algorithm (see Grippo et al., 1989 and the references therein) — an efficient minimisation approach for high dimensional problems.

Variational Laplace

Variational Laplace (VL) is an iterative optimisation process based on variational Bayes and generalises things like Restricted Maximum Likelihood (ReML) and Expectation–Maximization (by including hyperpriors on the hyperparameters). Its objective is to obtain the combination of hyperparameters λ that maximise Free energy, by following its gradient at a rate that is determined by its Hessian — with Eqs. (21) and (23) respectively. The VL optimisation proceeds as follows:

1. For the *k*-th iteration, compute the model based sample covariance matrix $\Sigma^{(k)}$ with Eq. (18). The hyperparameters can be initialized with zero values for the first iteration — if there are no informative hyperpriors.
2. Compute the gradient of the Free energy with Eq. (21) for each hyperparameter. In absence of informative hyperpriors use: $v = 0$, $\Pi \approx 0 I_{N_q}$, with I_{N_q} a $N_q \times N_q$ identity matrix.
3. Compute the curvature of the Free energy with Eq. (23) for each hyperparameter.
4. Update the hyperparameters:

$$\lambda_i^{(k)} = \lambda_i^{(k-1)} + \Delta\lambda_i \tag{24}$$

where the variation on each parameter $\Delta\lambda_i$ is computed with a Fisher scoring over the Free energy variation

$$\Delta\lambda_i = -\left(\frac{\partial^2 F}{\partial\lambda_i\partial\lambda_j}\right)^{-1}\frac{\partial F}{\partial\lambda_i}. \tag{25}$$

5. Eliminate those hyperparameters near to zero, and (implicitly) their corresponding covariance component.
6. Update the Free energy variation

$$\Delta F = \frac{\partial F}{\partial\lambda}\Delta\lambda. \tag{26}$$

Finish if the variation is less than a given tolerance (here $\Delta F < 0.01$). Otherwise go back to step 1.

Variational Laplace enables us to estimate the most likely value of the hyperparameters associated with each of the multiple prior covariance components (or patches), but this does not afford a sparse solution. In other words, we have not yet implemented the prior belief that only a small number of "patches" will be active at any one time. It is at this point that the full variational scheme comes into play. This is because we can implement the sparsity assumption by eliminating silent patches by optimising hyperpriors — namely, priors that can shrink the covariance hyperparameters to zero. If a hyperparameter is zero the patch can have no variance and is effectively eliminated. The problem now is to find the hyperpriors that maximise model evidence or Free energy. This is essentially a model selection problem, because each set of hyperpriors (combination of patches with zero and nonzero hyperpriors) represents a different model. This model selection can also be viewed as optimisation of the hyperpriors, because both maximise variational Free energy. SPM model selection uses two schemes: An Automatic Relevance Determination (ARD) and a Greedy Search (GS) over the Multiple Sparse Priors, optimised for sparse patterns.

Finally, the source estimation is obtained by using another round of VL to weight the source covariance estimates produced by these two subsidiary optimisation stages. We will now look at this part of the optimisation more closely:

The search for optimal priors

The goal now is to find the optimal mixture of prior covariance components *C*, that optimises the model evidence (or Free energy). There are many possible schemes to do this, the most computationally intensive (and impractical) being the sequential testing of all possible combinations of prior covariance components. The two schemes, ARD and GS, used in SPM are both deterministic schemes that use different computational strategies to simplify this high dimensional problem. In brief, both schemes use informative hyperpriors that ensure that most of the hyperparameters shrink to zero — thereby producing a sparse solution in source space. The ARD scheme does this by iteratively optimising Free energy in a bottom up fashion (removing redundant hyperparameters), while the GS uses a top-down strategy (creating new hyperparameters by partitioning the covariance component set). Instead of working with the original covariance component matrices, the computations are made more efficient by encoding covariance components in terms of their eigenvectors, that can be stacked into a single large matrix $\mathbb{Q} \in \Re^{N_d \times N_q}$ (as opposed to an array or list of matrices), with each column of $\mathbb{Q}$ being the main diagonal of its corresponding covariance component C_i (in these approaches only diagonal covariance components are allowed).

The ARD scheme exploits efficient matrix computation in Matlab to optimise all N_q hyperparameters (wrt Free energy) simultaneously through a gradient descent — removing hyperparameters that fall below some (small) threshold. In contrast the GS algorithm reduces the dimensionality of the problem by optimising mixtures of covariance components and then splitting these mixtures until the Free energy ceases to increase.

Automatic relevance determination. The main objective of the ARD approach is to avoid the computation of the gradient and curvature of the Free energy for each hyperparameter (Eqs. (21) and (23)), which accounts for the greatest computational cost. To achieve this, ARD performs a projection that allows matrix computations to optimise the hyperparameters. This is implemented using a stacked matrix $\mathbb{Q}$. With this stacked matrix $\mathbb{Q}$, it is possible to obtain the gradient of the Free energy with respect to all hyperparameters with a single computation:

$$\frac{dF}{d\lambda} = -\frac{1}{2}\mathrm{diag}\left(e^{\lambda}\right)\mathbb{Q}^T\left(\Sigma^{-1}\Sigma_Y\Sigma^{-1} - \Sigma^{-1}\right)\mathbb{Q} \tag{27}$$

with diag(*x*) a diagonal matrix with the vector *x* on its main diagonal. In similar way the curvature of the Free energy can be computed:

$$\frac{d^2 F}{d\lambda^2} = \mathrm{diag}\left(e^{\lambda}\right)\left(\left(\mathbb{Q}\Sigma^{-1}\mathbb{Q}\right)^T \cdot \left(\mathbb{Q}\Sigma^{-1}\mathbb{Q}\right)\right)\mathrm{diag}\left(e^{\lambda}\right). \tag{28}$$

Just by replacing Eqs. (21) and (23) with Eqs. (27) and (28), the ARD approach allows one to optimise large sets of hyperparameters without the use of "for" loops, considerably reducing the computation time.

Greedy search

In contrast to ARD, the Greedy Search performs a single-to-many optimisation of hyperparameters (Friston et al., 2007a). It is initialized by including all covariance components in the stacked matrix $\mathbb{Q}$. It then prunes this matrix iteratively by removing columns of $\mathbb{Q}$ that do not contribute to the solution.

Before going into the GS algorithm in depth, let us define the set of up to N_g diagonal matrices $G = \{G_1, G_2, \ldots, G_{N_g}\}$ that will be used to switch on or off the columns of $\mathbb{Q}$ that model the original covariance components. Each $G_i \in \Re^{N_q \times N_q}$ is generated with ones on the diagonal values corresponding to active components. This set will grow with iterations, but it will be initialised with the identity matrix: $G_1 = I_{N_q}$, indicating that all the components are equally feasible at the beginning. The main idea of the GS algorithm is to apply VL with relatively few hyperparameters ($N_g \ll N_q$) while allowing for a larger number of covariance components C.

The following is the iterative algorithm performed for the GS optimisation:

1. For the k-th iteration solve the inverse problem in the space of N_q covariance components. First compute the reduced source covariance matrix:

$$\Sigma_{GS} = Q_\epsilon + \sum_{i=1}^{k} e^{\beta_i} L\mathbb{Q}G_i\mathbb{Q}^T L^T \tag{29}$$

with β the new set of hyperparameters computed with VL. Then obtain the source reconstruction in the space of covariance components:

$$\hat{J}_Q = \left(\sum_{i=1}^{k} e^{\beta_i} G_i\right)\mathbb{Q}^T L^T \Sigma_{GS}^{-1} Y \tag{30}$$

where the set of neural sources in the space of covariance components is $\hat{J}_Q \in \Re^{N_q \times N_r}$.

2. Select the most active dipoles in $\hat{J}_Q$ and create a new G_{k+1} matrix with ones on the corresponding diagonal elements; by switching off the least active dipoles. Eliminating one half of the dipoles seems to provide a reasonable trade-off between speed and efficiency, although other proportions could be entertained.
3. Go back to Step 1 until the log evidence converges.

When the GS optimisation is complete, it is possible to recover the source space estimates with:

$$\hat{J} = \mathbb{Q}\hat{J}_Q. \tag{31}$$

In summary, ARD takes the set of all possible priors and prunes them until convergence, while GS splits and prunes mixtures of these priors. Both ARD and GS are based on the same prior information and ideally should provide the solutions; however, given the dimensionality of the problem, the two searches may get stuck in local maxima. For this reason, solving the problem with two search schemes provides a more robust solution. An example of how ARD and GS prune the covariance components over iterations is shown in Fig. 3.

Fig. 3(a) shows the evolution of the hyperparameters across the iterations of the ARD reconstruction of two sources (this example is reduced to 30 components). In ARD there is one hyperparameter per covariance component. Initially (iteration 1) all components are equally likely. After each iteration, those hyperparameters (one per component) close to zero are removed. At the end of the iterative process only those covariance components corresponding to active hyperparameters (in this case components 21 and 24) are used to reconstruct the sources.

Fig. 3(b) shows the corresponding evolution of the selected components over 4 iterations of the GS source reconstruction for the same data. After each iteration, a new hyperparameter is added and the new set of components is scored and pruned. The first set (encoded by the G matrix) just contains the identity matrix, indicating that at the beginning, all covariance components are equally probable. In this example hyperparameters 1 and 2 were eliminated by the algorithm and so, the final solution was produced using combinations of priors determined by hyperparameters 3 and 4 (i.e., prior components 21 and 24).

Final optimisation

In the SPM implementation of MSP, a second inversion is performed using the prior covariance matrices produced by the GS and ARD searches. This adds some robustness in the event that either deterministic search got stuck in a local maximum. In this second inversion the ARD, GS and sensor noise covariances are mixed using VL, and the resulting single covariance matrix is used to get the posterior mean and variance of the current density:

$$\widetilde{J} = QL^T\left(Q_\epsilon + LQL^T\right)^{-1} Y \tag{32}$$

$$\widetilde{Q} = \left(L^T Q_\epsilon^{-1} L + Q^{-1}\right)^{-1} \tag{33}$$

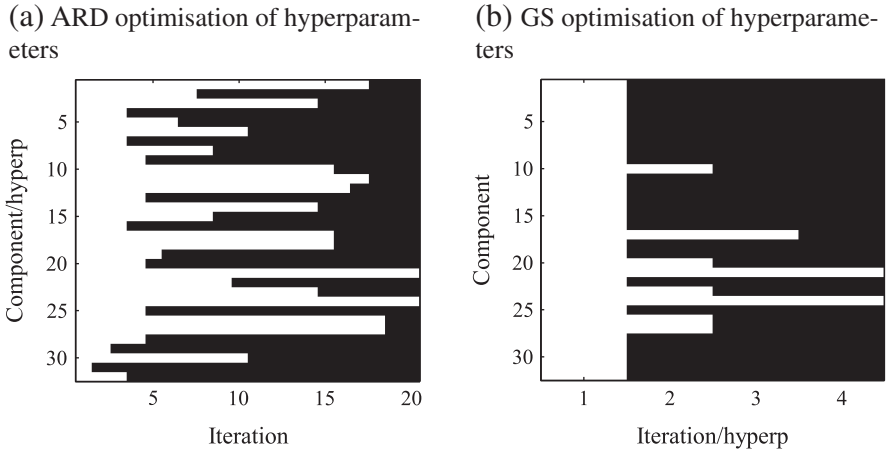

Fig. 3. ARD and GS optimisations for a set of 30 dipoles or patches. Active components are coloured in white: (a) For ARD, each patch has a covariance component has an associated hyperparameter and these are pruned as the search evolves. Note that at iteration 1 all patches are considered but by iteration 20 all but two have been eliminated. (b) The Greedy search is based on a single matrix in which each column defines a covariance component in terms of combinations or mixtures of patches. At each iteration, a new component is created (with the most active patches) and this component has its own associated hyperparameter. That is, in contrast to panel (a) (ARD) in which there is a hyperparameter for every patch, in panel (b) (GS) each iteration produces a different combination of patches and it is these combinations that are weighted by the hyperparameters.

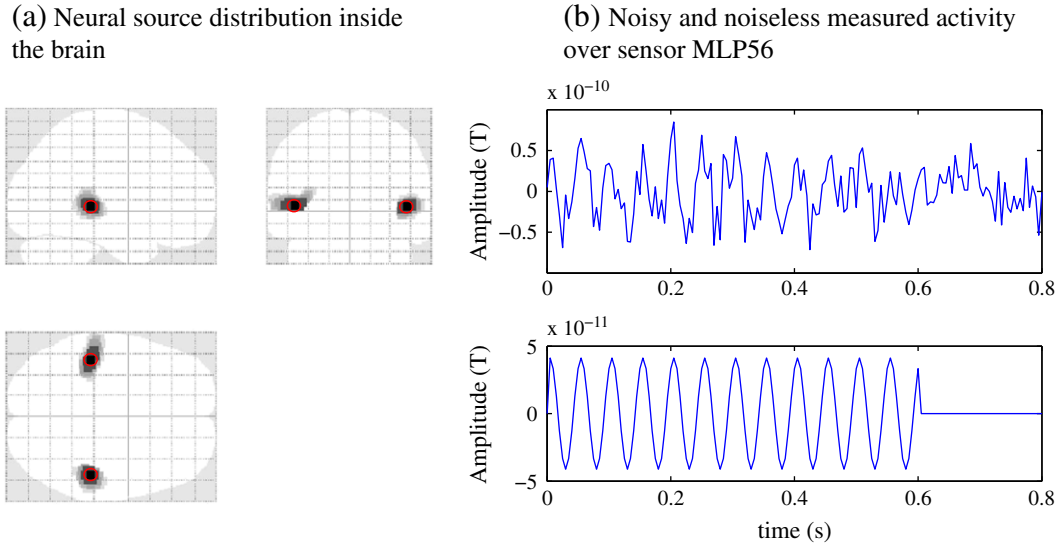

Fig. 4. Glass brain with simulated activity. (a) The locations of the simulated sources on a glass brain. (b) Measured activity over sensor MLP56, the original source activity (bottom) cannot be clearly observed from the noisy measured data (top).

where $\widetilde{Q}$ is the estimated posterior covariance over source space. The temporal responses in source space can be recovered with the temporal projector:

$$\hat{J} = \widetilde{J}P^T. \tag{34}$$

Simulation example

To help understand the different inversion approaches implemented in the SPM software,[1] we have developed a simulation example that can be downloaded from the SPM web page — and run to obtain the results presented below.

A single trial dataset of $N_n = 161$ samples over $N_c = 274$ MEG sensors was generated from the neural source distribution shown in Fig. 4(a). These sources consisted of two synchronous lateral sinusoidal signals of 20 Hz. Noise was added to the data using white random noise, where the signal-to-noise ratio was: $SNR = 10log_{10}|\text{var}(Y)/\text{var}(noise)|$. Both sources were focal Gaussian sources (on a cortical mesh), with a spatial extent of approximately 10 mm. Fig. 4(b) shows the data collected by sensor MLP56 with and without noise. The head model used for simulations is the canonical model provided with the SPM software package, it consists in a Single shell head model (Nolte, 2003). The source model consisted of a canonical cortical mesh (Mattout et al., 2007) of $N_d = 8196$ dipoles distributed over the cortical surface (see Fig. 1(a)), each with fixed perpendicular orientation; computed following the procedure described in Phillips et al. (2002a). The same head and source models were used to simulate data and solve the inverse problem (although this needs not be the case, see López et al., 2012b). The glass brains (maximum intensity projections) in Fig. 4 show the frontal, lateral and superior views of the 512 sources with the highest variance during the time window of interest.

Pre-processing stage

The synthetic dataset generated for this example is based on a single subject with a single (averaged) trial. In the spatial projector stage, the sensor space was reduced to $N_m = 103$ spatial modes (Fig. 5(a)).

In this case $N_r = 16$ temporal modes were selected, accounting 84.56% of the total variance present in the data. Fig. 5(b) shows an image of $\widetilde{K}$ computed with Eq. (14), which is the basis set of the temporal projector.

Inversion scheme

This reduced dataset $\mathbb{Y} \in \Re^{103\times16}$ was generated and used as a benchmark to compare a number of inversion schemes:

Using suboptimal priors

In order to demonstrate the sensitivity of the inversion to the correct prior assumptions, we first consider the case of reconstructing data simulated with MSP priors using Minimum Norm or LORETA assumptions. Note that this is not a comparison of different algorithms, simply a demonstration that using sub-optimal prior assumptions will result in sub-optimal source reconstruction. In this case we simulate sources consistent with MSP assumptions (above) and try and reconstruct using alternative (Minimum Norm and LORETA) prior covariance matrices. Fig. 6(a) shows the Minimum Norm and LORETA reconstructions of the source distributions shown in Fig. 4(a): as expected (consistent with the prior assumptions) both source estimates were superficial and extended relative to the true simulated source. Note that ad-hoc solutions exist to compensate for this superficial bias (Dale et al., 2000; Fuchs et al., 1999; Hauk, 2004; Ioannides et al., 1990; Lin et al., 2006; Pascual-Marqui, 2002), however, the aim here is to illustrate how this bias arises from a suboptimal prior covariance model.

Multiple Sparse Priors

The MSP algorithm returns a Variational Laplace optimisation of the candidate covariance estimates from GS and ARD optimisations. Fig. 7(a) shows the source reconstruction given by the GS: both dominant sources perfectly matched the original sources, with some spurious activity attributable to noise. This algorithm is well suited to deal with bilateral synchronous sources because it can accommodate them within a single column of G (controlled with a single hyperparameter). The ARD starts with a large number of hyperparameters and prunes the patches independently. In this case (Fig. 7(b)) the covariance prior associated with the smaller amplitude correlated source was erroneously rejected by the algorithm. Both candidate covariance models are then further mixed with a final VL stage to give a more robust estimate of source covariance weighted by model evidence. Given that the MSP is an optimal mixture of the GS and ARD solutions, we would expect that it should rely more on the GS reconstruction due to its higher

[1] Available for free download from http://www.fil.ion.ucl.ac.uk/spm/.

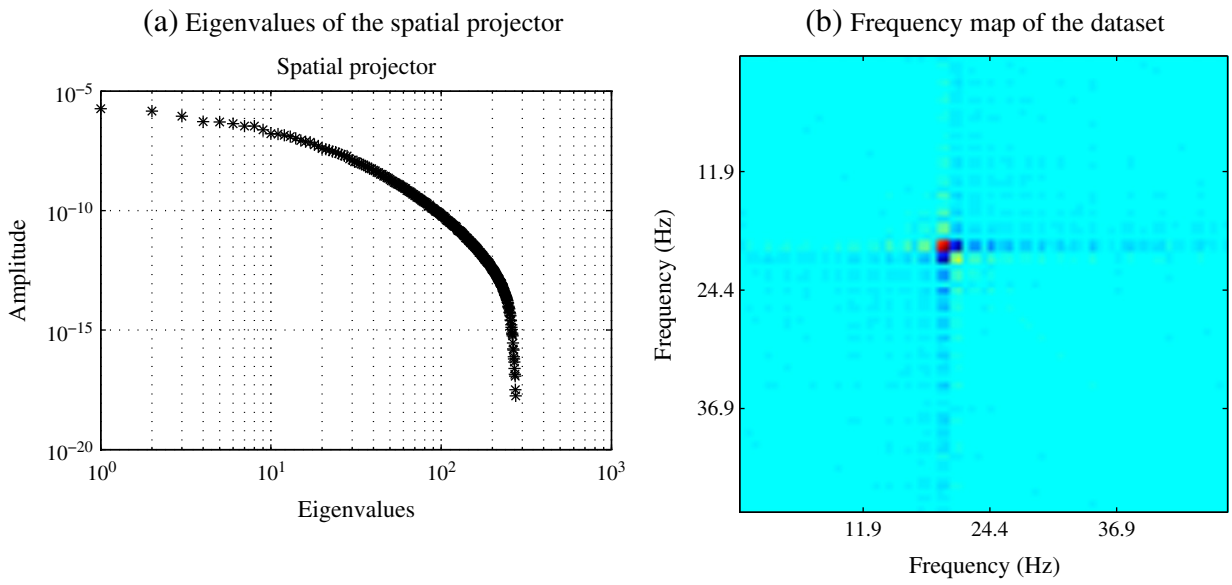

Fig. 5. Spatial and temporal projectors. (a) The first 103 eigenvalues of the spatial projector A were selected. (b) The matrix K whose eigenvectors provide the temporal projector.

Free energy (see below). This is indeed the case, as can be seen in Fig. 7(c).

In order to illustrate the link between better models (as quantified through Free energy) and more traditional measures of algorithm performance through localisation error, we simulated (100 times) a single smooth patch of active cortex (i.e., a source matching MSP prior assumptions) randomly located on the cortical surface, and reconstructed using the different modelling assumptions. Fig. 8 shows the distance of the simulated source to the peak in the source reconstruction for each assumption set (blue bars) alongside the corresponding relative (to the minimum) Free energy or log model evidence (green bars). Note that the prior assumptions with the highest model evidence also have the smallest localisation error. Note that this is not a demonstration that MSP works better than other algorithms; rather it shows that model evidence that allows us to score different prior source covariance matrices and reconstructions based on an optimal model will generally minimise localisation error. In this case we set the prior covariance matrix to be consistent with MSP but we could have just as easily simulated data with LORETA like assumptions.

Discussion and conclusions

In this tutorial paper, we have tried to unpack and explain the inversion scheme used in the SPM software. This manuscript and accompanying software provides examples of the pre-processing and inversion procedure involved in the classical (Minimum Norm, LORETA etc.).

Importantly all algorithms can be seen within the same mathematical framework and differing only in their prior assumptions about the structure of the source covariance matrix (Mosher et al., 2003; Wipf and Nagarajan, 2009). In turn, the fitness of any prior covariance matrix to explain measured data can be quantified in terms of Free energy or model evidence. Here, we have focused on describing the MSP framework and have consequently simulated (sparse focal) data on the cortical surface — consistent with the sparse priors algorithm. It is important to note that had we simulated data based on LORETA-like assumptions, then the LORETA priors would have had maximum evidence and least localisation error. The question of which priors work best in practice is an empirical one, but hopefully we have been able to describe some of the tools that can be used to establish this optimal set. Note that because

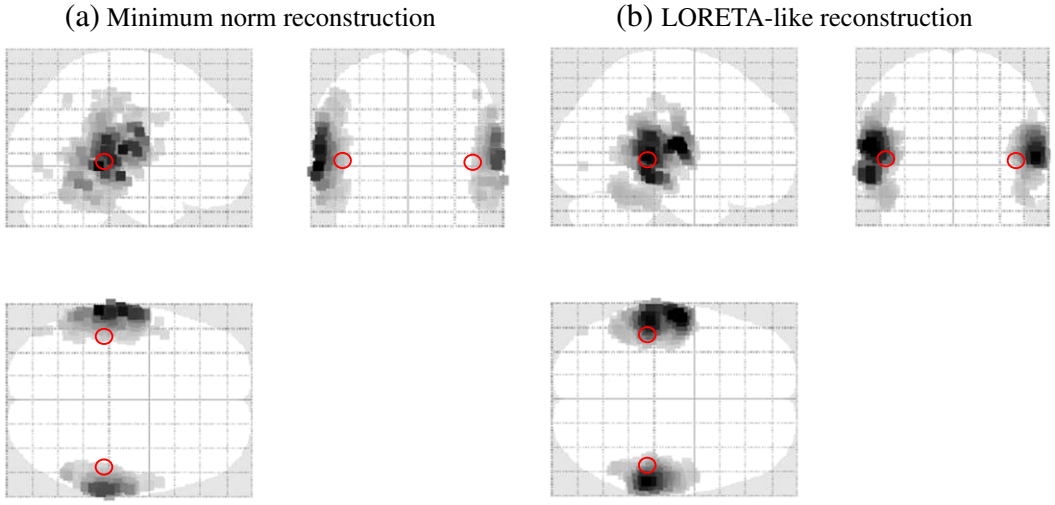

Fig. 6. RMS current estimates for reconstructions of the focal sources simulated in Fig. 4(a) (red circles) based on (a) Minimum Norm and (b) LORETA prior covariance matrices. Note that both these assumption sets lead to localisation error, as both algorithms tend to project sources superficially and increase the spatial extent of the estimated current distribution (158 and 226 sources exceed half the maximum amplitude).

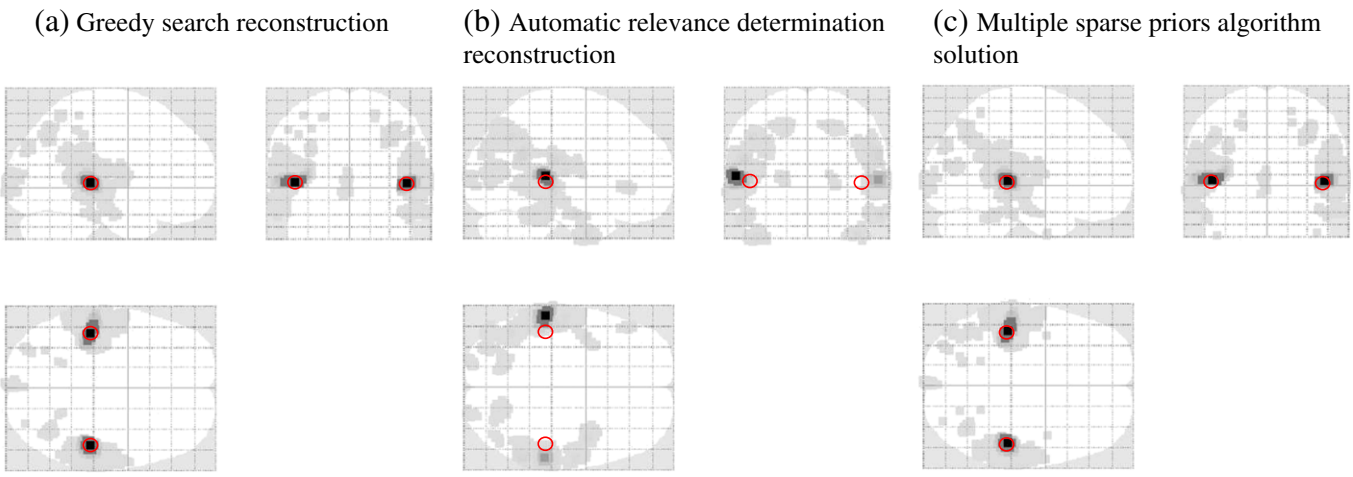

Fig. 7. Using the variational Laplace schemes to select the optimal prior set from the library. (a) Greedy search reconstruction with zero localisation error. (b) Automatic relevance determination performs poorly in this case. (c) Multiple Sparse Priors are based on the weighted mixture of GS and ARD estimates (in this case the localisation error is zero and there are 8 sources that exceed half the maximum amplitude).

a combination of MSPs can emulate the LORETA covariance assumptions, in principle, it should never be necessary to actually use the LORETA covariance component – because this will be selected automatically – if it provides the best explanation for the data (c.f., the first component considered by the Greedy search above).

For real data, no ground truth is available, and although one solution might appear more focal than another, there is no reason why it should be more accurate. Given certain caveats (see below), it is clear that the model evidence is a useful and objective test of plausible prior covariance models, and these should provide the most accurate estimates of neuronal current distribution.

It is also important to note that in all of these examples the propagation model (the lead-field matrix) has been considered as ground truth. This is rarely the case in practice, several errors such as co-registration error, head movement, MRI distortion, poor cortical segmentation, amongst others, add extra uncertainty to the problem that should ideally be accounted for in the confidence interval on the final solution (see Chung et al., 2008; López et al., 2012b, 2013; Troebinger et al., 2013 for some recent work in this field). One important drawback of methods with higher resolution (MSP, beamformers, etc.) is that they are also the most sensitive to errors in the forward model (Hillebrand and Barnes, 2002; López et al., 2012b). For example, in (López et al., 2012b) we showed that co-registration errors of the order of 4 mm and 4° were enough to compromise the MSP inversion. It is also the case that methods which require a non-linear search over the space of priors, although more flexible, may also be less robust to situations in which there are large numbers of independent sources (due to local maxima in the cost function) or non-Gaussian noise (such as unmodelled artefacts in the data). In Appendix B we show the relative performances of the different algorithms for different numbers of simulated sources in both ideal Gaussian and real noise conditions.

There are many unsolved problems and many possible improvements to the schemes we have considered. These include the optimal tree size for the Greedy search, the optimal number of patches, patch spacing and patch smoothness. We hope that the technical details presented here will be sufficient for others to familiarise themselves with this software and address these and other outstanding issues.

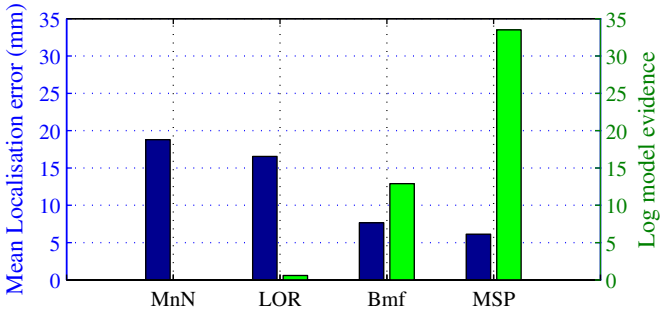

Fig. 8. Mean source localisation error (blue bars) and log model evidence (green bars) relative to the poorest model. In this case, the underlying source model is consistent with MSP and so MSP has the highest model evidence and smallest localisation error.

Acknowledgments

J.D. López and J.J. Espinosa are supported by the ARTICA Research Center for Excellence, Ministerio de Educación Nacional Colombiano and Colciencias, projects 1115-489-25190 and 1115-545-31374. The Wellcome Trust Centre for Neuroimaging is supported by a strategic award from the Wellcome Trust, grant number 091593/Z/10/Z.

Conflict of interest

There is no conflict of interest.

Appendix A. Model based sample covariance matrix

The definition of the prior source covariance matrix Q with Eq. (10) implies a redefinition of the joint probability distribution, due to the inclusion of hyperparameters: $p(Y, J, h)$. A priori, the weights of Q (hyperparameters) are independent, the parameters J are fully dependent on them, and the model based data is also strictly dependent on J, allowing us to define:

$$p(Y,J,h) = p(Y|J)p(J|h)p(h) \tag{A.1}$$

The prior distribution of the parameters now depends on h: $p(J|h)$. It is now necessary to assume a prior on h (such as the general probability distribution proposed in Wipf and Nagarajan (2009):

$$p(h) \propto \prod_{i=1}^{N_q} e^{f_i(h_i)} \tag{A.2}$$

where each $f_i(\cdot)$ is a known unspecified function (preferably convex). With a known distribution on h it can be integrated out (marginalised) on what is known as a Gaussian scale mixture:

$$p(J) = \int p(J,h)dh = \int p(J|h)p(h)dh \tag{A.3}$$

and the prior on h is again independent. Now the problem is how to obtain these h values. Rather than estimate a complete posterior distribution of the hyperparameters, it should be possible to obtain their expected value $\hat{h}$.

Initially let us assume that h is known, then Q is known and the conditional distribution $p(J|Y, h)$ can be expressed as a fully specified Gaussian distribution. However, since h is not known, a suitable approximation $h \approx \hat{h}$ must be computed:

$$p\left(J|Y, h = \hat{h}\right) = p(J|Y) \tag{A.4}$$

to solve the problem with Eq. (6). This approximation can be optimised with Empirical Bayes (Berger, 1985), where the prior $p(J|h)$ can be empirically learned from the data using the evidence $p(Y)$ as a cost function. This approach is based on the fact that each set of hyperparameters will approximate the solution to the evidence, and that the optimal set $\hat{h}$ is the one that provides the highest evidence.

Given that the parameters J are fully dependent on h, they can be marginalised out of the optimisation problem by integrating them out of the joint probability distribution $p(Y, J, h)$:

$$p(Y,h) = \int p(Y,J,h)dJ = p(Y|h)p(h) \tag{A.5}$$

where $p(Y|h)$ can be derived from Eq. (A.1):

$$p(Y,h) = \int p(Y|J)p(J|h)p(h)dJ \tag{A.6}$$

because $p(h)$ is independent of J, it can be extracted from the integral and by comparison with Eq. (A.5):

$$p(Y|h) = \int p(Y|J)p(J|h)dJ. \tag{A.7}$$

To which solution is a Gaussian distribution:

$$p(Y|h) \propto exp\left(-\frac{1}{2}\mathrm{tr}\left(Y^T\Sigma_Y^{-1}Y\right)\right) \tag{A.8}$$

where $\Sigma_Y = \Sigma_\epsilon + LQL^T$ is the "model based sample covariance matrix" given the set of hyperparameters h. This result is important because it obviates the use of J in the optimisation problem, and allows us to formulate a cost function for h exclusively in terms of the data. The ensuing evidence, computed with the optimal set of hyperparameters: $p(Y) = p\left(Y|h = \hat{h}\right)$, is a rigorous upper bound that can be used for model selection (Friston et al., 2008; López et al., 2012b).

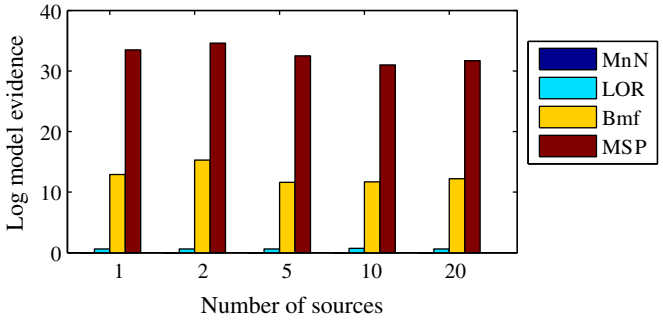

Fig. B.9. Average log evidence of different approaches for 100 random locations of 1, 2, 5, 10 and 20 simultaneous active sources of neural activity.

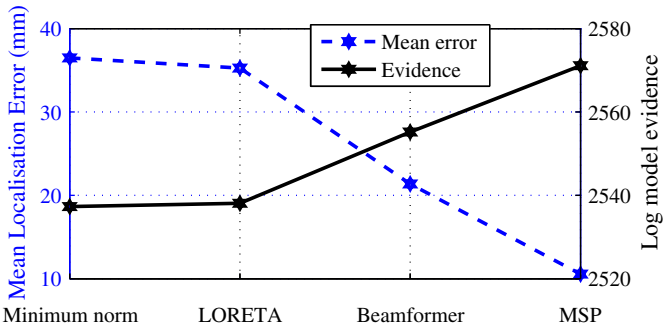

Fig. B.10. Average localisation error and average Free energy for 100 simulations of each tested approach with a single simulated source under real noise conditions. The tendency is the same observed with Gaussian noise.

Appendix B. Performance analysis of different algorithms under Gaussian and real noise conditions

Fig. B.9 shows the relative log model evidence (as approximated through Free energy) of 100 simulations with 1, 2, 5, 10, and 20 simultaneous active sources (consisting of sinusoids of random frequency) randomly located on the cortical surface. All the data were simulated using MSP priors and so it is no surprise that MSP models have the highest evidence. One might expect changes in differential evidence amongst solutions that involve a non-linear search over multiple priors (like MSP) and single component models (like Minimum Norm, LORETA, and Beamformer) as the complexity of the solution increases, but we did not observe that here.

Fig. B.10 shows the results of the same example of Fig. 8 but with real noise acquired from resting state activity (the complete experiment is presented in Sedley et al. (2011). Note that the average localisation error increased in all approaches, but trends were unaffected. This figure also shows how the increase in Free energy correlates with lower average localisation error.

Appendix C. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2013.09.002.

References

Auranen, T., Nummenmaa, A., Hämäläinen, M., Jääskeläinen, I., Lampinen, J., Vehtari, A., Sams, M., 2005. Bayesian analysis of the neuromagnetic inverse problem with lp-norm priors. NeuroImage 26, 870–884.
Baillet, S., Garnero, L., 1997. A Bayesian approach to introducing anatomo-functional priors in the EEG/MEG inverse problem. IEEE Trans. Biomed. Eng. 44 (5), 374–385 (May).
Baillet, S., Mosher, J., Leahy, R., 2001. Electromagnetic brain mapping. IEEE Signal Process. Mag. 18 (6), 14–30.
Belardinelli, P., Ortiz, E., Barnes, G., Noppeney, U., Preissl, H., 2012. Source reconstruction accuracy of MEG and EEG Bayesian inversion approaches. PLoS One 7, 12.
Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis, 2nd edition. Springer Verlag, New York.
Chung, M.K., Dalton, K.M., Davidson, R.J., 2008. Tensor-based cortical surface morphometry via weighted spherical harmonic representation. IEEE Trans. Med. Imaging 27 (8), 1143–1151 (august).
Dale, A.M., Sereno, M., 1993. Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach. J. Cogn. Neurosci. 5, 162–176.
Dale, A., Liu, A., Fischl, B., Buckner, R., Belliveau, J., Lewine, J., Halgren, E., 2000. Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. Neuron 26, 55–67.
Friston, K., Chu, C., Mourao-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J., 2007a. Bayesian decoding of brain images. NeuroImage 39 (1), 181–205.
Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007b. Variational free energy and the Laplace approximation. NeuroImage 34, 220–234.

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, 1104–1120.
Fuchs, M., Wagner, M., Kohler, T., Wischman, H., 1999. Linear and nonlinear current density reconstructions. J. Clin. Neurophysiol. 16, 267–295.
Gener, N., Williamson, S., July 1998. Differential characterization of neural sources with the bimodal truncated SVD pseudo-inverse for EEG and MEG measurements. IEEE Trans. Biomed. Eng. 45 (7), 827–838.
Golub, G.H., Heath, M., Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21 (2), 215–223 (May).
Grech, R., Cassar, T., Muscat, J., Camilleri, K., Fabri, S., Zervakis, M., Xanthopoulos, P., Sakkalis, V., Vanrumste, B., 2008. Review on solving the inverse problem in EEG source analysis. J. Neuro Eng. Rehabil. 5 (1), 25.
Grippo, L., Lampariello, F., Lucidi, S., March 1989. A truncated Newton method with nonmonotone line search for unconstrained optimization. J. Optim. Theory Appl. 60 (3), 401–419.
Hämäläinen, M.S., Ilmoniemi, R.J., 1984. Interpreting measured magnetic fields of the brain: estimates of current distributions. Tech. rep. Helsinki University of Technology.
Hansen, P.C., 2000. The L-Curve and its use in the numerical treatment of, inverse problems. Computational Inverse Problems in Electrocardiology.WIT Press 119–142 (Ch.).
Harrison, L., Penny, W., Ashburner, J., Trujillo-Barreto, N., Friston, K., 2007. Diffusion-based spatial priors for imaging. NeuroImage 38, 677–695.
Hauk, O., 2004. Keep it simple: a case for using classical minimum norm estimation in the analysis of EEG and MEG data. NeuroImage 21, 1612–1621.
Henson, R., Mouchlianitis, E., Friston, K., 2009. MEG and EEG data fusion: simultaneous localisation of face-evoked responses. NeuroImage 47 (2), 581–589.
Henson, R., Flandin, G., Friston, K., Mattout, J., 2010. A parametric empirical Bayesian framework for fMRI-constrained MEG/EEG source reconstruction. Hum. Brain Mapp. 31 (10), 1512–1531.
Henson, R.N., Wakeman, D.G., Litvak, V., Friston, K.J., 2011. A parametric empirical Bayesian framework for the EEG/MEG inverse problem: generative models for multi-subject and multi-modal integration. Front. Hum. Neurosci. 5, 16.
Hillebrand, A., Barnes, G., 2002. A quantitative assessment of the sensitivity of whole-head MEG to activity in the adult human cortex. NeuroImage 16, 638–650.
Hillebrand, A., Singh, K.D., Holliday, I.E., Furlong, P.L., Barnes, G.R., 2005. A new approach to neuroimaging with magnetoencephalography. Hum. Brain Mapp. 25 (2), 199–211.
Ioannides, A., Bolton, J., Clarke, C., 1990. Continuous probabilistic solutions to the biomagnetic inverse problem. Inverse Prob. 6, 523–542.
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), 160–171.
Litvak, V., Friston, K., 2008. Electromagnetic source reconstruction for group studies. NeuroImage 42 (4–24), 1490–1498.
Litvak, V., Mattout, J., Kiebel, S., Phillips, C., Henson, R., Kilner, J., Barnes, G., Oostenveld, R., Daunizeau, J., Flandin, G., Penny, W., Friston, K., 2011. EEG and MEG data analysis in SPM8. Comput. Intell. Neurosci. 2011, 32.
Liu, A.K., Dale, A.M., Belliveau, J.W., 2002. Monte Carlo simulation studies of EEG and MEG localization accuracy. Hum. Brain Mapp. 16, 47–62.
López, J.D., Barnes, G., Espinosa, J., 2012a. Single MEG/EEG source reconstruction with multiple sparse priors and variable patches. DYNA 174, 136–144.
López, J.D., Penny, W.D., Espinosa, J.J., Barnes, G.R., 2012b. A general Bayesian treatment for MEG source reconstruction incorporating lead field uncertainty. NeuroImage 60, 1194–1204.
López, J.D., Troebinger, L., Penny, W., Espinosa, J.J., Barnes, G.R., 2013. Cortical surface reconstruction based on MEG data and spherical harmonics. 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 6449–6452.
Mattout, J., Henson, R.N., Friston, K.J., 2007. Canonical source reconstruction for MEG. Comput. Intell. Neurosci. 10 (article ID 67613).
Michel, C., Murray, M., Lantz, G., Gonzalez, S., Spinelli, L., Grave de Peralta, R., 2004. EEG source imaging. Clin. Neurophysiol. 115, 2195–2222.
Mosher, J., Baillet, S., Leahi, R.M., 2003. Equivalence of linear approaches in bioelectromagnetic inverse solutions. IEEE Workshop on Statistical Signal Processing, 294–297.
Nolte, G., 2003. The magnetic lead field theorem in the quasi-static approximation and its use for magnetoencephalography forward calculation in realistic volume conductors. Phys. Med. Biol. 48 (22), 3637–3652.
Pascual-Marqui, R., 1999. Review of methods for solving the EEG inverse problem. Int. J. Bioelectromagn. 1 (1), 75–86.
Pascual-Marqui, R., 2002. Standardized low resolution brain electromagnetic tomography (sLORETA): technical details. Methods Find. Exp. Clin. Pharmacol. 24, 5–24.
Pascual-Marqui, R., Michel, C.M., Lehmann, D., 1994. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. Int. J. Psychophysiol. 18, 49–65.
Penny, W., 2012. Comparing dynamic causal models using AIC, BIC and free energy. NeuroImage 59 (1), 319–330.
Phillips, J., Leahy, R., Mosher, J., 1997. MEG-based imaging of focal neuronal current sources. IEEE Trans. Med. Imaging 16 (3), 338–348.
Phillips, C., Rugg, M., Friston, K., 2002a. Anatomically informed basis functions for EEG source localization: combining functional and anatomical constraints. NeuroImage 16, 678–695.
Phillips, C., Rugg, M., Friston, K., 2002b. Systematic regularization of linear inverse solutions of the EEG source localization problem. NeuroImage 17, 287–301.
Phillips, C., Mattout, J., Rugg, M., Maquet, P., Friston, K., 2005. An empirical Bayesian solution to the source reconstruction problem in EEG. NeuroImage 24 (4), 997–1011.
Sato, M., Yoshioka, T., Kajihara, S., Toyama, K., Goda, N., Doya, K., Kawato, M., 2004. Hierarchical Bayesian estimation for MEG inverse problem. NeuroImage 23 (3), 806–826.
Schmidt, D.M., George, J.S., Wood, C., 1999. Bayesian inference applied to the electromagnetic inverse problem. Hum. Brain Mapp. 7 (3), 195–212.
Sedley, W., Teki, S., Kumar, S., Overath, T., Barnes, G., Griffiths, T., 2011. Gamma band pitch responses in human auditory cortex measured with magnetoencephalography. NeuroImage 59 (2), 1904–1911.
Sekihara, K., Poeppel, D., Marantz, A., Koizumi, H., Miyashita, Y., 1999. MEG spatio-temporal analysis using a covariance matrix calculated from nonaveraged multiple-epoch data. IEEE Trans. Biomed. Eng. 46 (5), 515–521.
Supek, S., Aine, C.J., 1993. Simulation studies of multiple dipole neuromagnetic source localization: model order and limits of source resolution. IEEE Trans. Biomed. Eng. 40, 529–540.
Taulu, S., Kajola, M., 2005. Presentation of electromagnetic multichannel data: the signal space separation method. J. Appl. Phys. 97, 10.
Troebinger, L., López, J.D., Lutti, A., Bradbury, D., Bestmann, S., Barnes, G., 2013. High precision anatomy for MEG. NeuroImage. http://dx.doi.org/10.1016/j.neuroimage.2013.07.065.
Trujillo-Barreto, N.J., Aubert-Vazquez, E., Valdes-Sosa, P.A., 2004. Bayesian model averaging in EEG/MEG imaging. NeuroImage 21, 1300–1319.
Van Veen, B.D., van Drongelen, W., Yuchtman, M., Suzuki, A., 1997. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans. Biomed. Eng. 44 (9), 867–880 (September).
Wipf, D., Nagarajan, S., 2009. A unified Bayesian framework for MEG/EEG source imaging. NeuroImage 44, 947–966.