ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 39 (2008) 1104–1120

Multiple sparse priors for the M/EEG inverse problem

Karl Friston,[a,*] Lee Harrison,[a] Jean Daunizeau,[a] Stefan Kiebel,[a] Christophe Phillips,[b] Nelson Trujillo-Barreto,[c] Richard Henson,[d] Guillaume Flandin,[e] and Jérémie Mattout[f]

[a] *The Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, 12 Queen Square, London, WC1N 3BG, UK*
[b] *Centre de Recherches du Cyclotron, Université de Liège, Belgium*
[c] *Cuban Neuroscience Centre, Havana, Cuba*
[d] *MRC Cognition & Brain Sciences Unit, Cambridge, UK*
[e] *Service Hospitalier Frédéric Joliot, CEA-SHFJ, Orsay, France*
[f] *INSERM U821-Dynamique Cérébrale et Cognition, Lyon, France*

Received 15 May 2007; revised 19 September 2007; accepted 22 September 2007
Available online 10 October 2007

This paper describes an application of hierarchical or empirical Bayes to the distributed source reconstruction problem in electro- and magnetoencephalography (EEG and MEG). The key contribution is the automatic selection of multiple cortical sources with compact spatial support that are specified in terms of empirical priors. This obviates the need to use priors with a specific form (e.g., smoothness or minimum norm) or with spatial structure (e.g., priors based on depth constraints or functional magnetic resonance imaging results). Furthermore, the inversion scheme allows for a sparse solution for distributed sources, of the sort enforced by equivalent current dipole (ECD) models. This means the approach automatically selects either a sparse or a distributed model, depending on the data. The scheme is compared with conventional applications of Bayesian solutions to quantify the improvement in performance.
© 2007 Elsevier Inc. All rights reserved.

Keywords: Variational Bayes; Free energy; Expectation maximization; Restricted maximum likelihood; Model selection; Automatic relevance determination; Sparse priors

Introduction

Bayesian approaches to the inverse problem in EEG represent an exciting development over past years (see Baillet and Garnero, 1997; Russell et al., 1998; Sato et al., 2004; Jun et al., 2006; Nagarajan et al., 2006; Daunizeau et al., 2007; Nummenmaa et al., 2007 for some important developments). A special instance of Bayesian analysis rests on empirical Bayes in which spatial priors are estimated from the data. Parametric empirical Bayesian (PEB) models are simple hierarchical linear models under parametric assumptions (i.e., additive Gaussian random effects at each level). Their hierarchical form enables one level to constrain the parameters of the level below and therefore act as empirical priors (Efron and Morris, 1973; Kass and Steffey, 1989). In the context of the EEG inverse problem, the parameters correspond to unknown source activity and the priors represent spatially varying constraints on the values the parameters can take. PEB models furnish priors on the parameters through hyperparameters encoding the covariance components of random effects at each level. However, these models can also be extended hierarchically by inducing hyperpriors on the hyperparameters themselves (see Trujillo-Barreto et al., 2004; Sato et al., 2004; Daunizeau and Friston, 2007). This is the hierarchical extension considered in Sato et al. (2004) and evaluated using sampling techniques in Nummenmaa et al. (2007). Under these models, it is possible to estimate the inverse variance (i.e., precision) of each prior, even when the number of hyperparameters exceeds the number of observations. Sato et al. used this to estimate an empirical prior precision on a large number of sources on the cortical mesh. This estimation used standard variational techniques to estimate the conditional density of the parameters and precision hyperparameters. In this context, non-informative gamma hyperpriors on the precision of random effects are also known as automatic relevance determination or ARD priors (Neal, 1998; Tipping, 2001). This approach gives better results, in terms of location and resolution, compared to standard minimum norm estimators.

The approach taken here uses covariance as opposed to precision hyperparameters (see also Wipf et al., 2006). This has two advantages: first the fixed-form variational scheme used for estimation reduces to a very simple and efficient classical covariance component estimation based on ReML (Patterson and Thompson, 1971, Harville, 1977; Friston et al., 2007). This means one can consider a large range of models with additive covariance components in source space (e.g., different source configurations) using exactly the same variational scheme (i.e., there is no need to

* Corresponding author. Fax: +44 207 813 1445.
E-mail address: k.friston@fil.ion.ucl.ac.uk (K. Friston).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2007.09.048

derive special update rules for different components). Second, one avoids the improper densities associated with non-informative ARD priors based on the gamma density (see Gelman, 2006).

Empirical Bayes

Previously, we have described the use of parametric empirical Bayes (PEB) to invert electromagnetic models and localize distributed sources in EEG and MEG (Phillips et al., 2002a,b, 2005). Empirical Bayes provides a principled way of quantifying the relative importance of spatial priors that replaces heuristics like L-curve analysis. Furthermore, PEB can accommodate multiple priors and provides more accurate and efficient source reconstruction than its precedents (Phillips et al., 2002a,b). After this, we explored the use of PEB to identify the most likely combination of priors using model selection, where each model comprises a different set of priors (Mattout et al., 2006). This was based on the fact that the restricted maximum likelihood (ReML) objective function used in the optimization of the model parameters is the log-likelihood, $\ln p(y|\lambda,m)$, of the covariance hyperparameters, λ for a model m and data y; a model is defined by its covariance components associated with activity over sources. We have since applied the ensuing inversion schemes to evoked and induced responses in both EEG and MEG (see Friston et al., 2006).

Finally, we showed that adding the entropy of the conditional density on the hyperparameters to the ReML objective function provides a free energy bound on the log-evidence or marginal likelihood $\ln p(y|m)$ of the model itself (Friston et al., 2007). Although this result is well-known to the machine learning community, it is particularly important here because it means one can use ReML within an evidence (i.e., free energy) maximization framework to optimize the parameters and hyperparameters of electromagnetic forward models. The key advantage of ReML is that optimization can proceed using the sample covariance of the data in measurement or channel space, which does not increase in size with the number of sources. The result is an efficient optimization, which uses classical methods, designed originally to estimate Gaussian covariance components (Patterson and Thompson, 1971). The ensuing approach is related formally to Gaussian process modeling (Ripley, 1994; Rasmussen, 1996; Kim and Ghahramani, 2006), where empirical Gaussian process priors are furnished by a hierarchical (PEB; Kass and Steffey, 1989) model.

Model selection and ARD

The fact that ReML can be used to optimize a bound on the marginal likelihood or evidence means that it can be used for model selection, specifically to select or compare models with different Gaussian process priors. Furthermore, under simple hyperpriors, ReML selects the best model automatically. This is because the hyperpriors force the conditional variance of hyperparameters to zero, when their conditional mean is zero. This means the free energy is the same that one would obtain with formal model comparison. In short, ReML can be used to estimate the hyperparameters controlling mixtures of covariance components in both measurement and source space that generate data. If there are redundant components, ReML will automatically switch them off, or suppress them, to provide a forward model with the greatest evidence or marginal likelihood. This is an example of automatic relevance determination (ARD). ARD refers to a general phenomenon, in hierarchical Bayesian models, where maximizing the evidence (often through EM-like algorithms) leads to pruning away of unnecessary model components (see Neal, 1996, 1998).

Recently, Wipf et al. (2006) provided an extremely useful formulation of empirical Bayesian approaches to the electromagnetic inverse problem and show how existing schemes "can be related via the notion of automatic relevance determination (Neal, 1996) and evidence maximization (MacKay, 1992)". The approach adopted here conforms exactly to the principles articulated in Wipf et al. and re-iterates the generality of free energy or evidence maximization. Wipf et al. (2006) also consider particular maximization schemes, based on standard variational updates, under inverse gamma hyperpriors. We use an ReML scheme, which is much simpler and uses log-normal hyperpriors. This allows us to use the Laplace approximation to the curvatures of the log-evidence during optimization (Friston et al., 2007).

In summary, this paper takes the application of ReML to the EEG inverse problem to its natural conclusion;[1] instead of using a small number of carefully specified prior covariance components (e.g., Laplace, minimum norm, depth constraints etc.) we use a large number of putative sources with compact (but not necessarily continuous) support on the cortical surface. The inversion scheme automatically selects which priors are needed, furnishing sparse or distributed solutions, depending on the data. This provides a graceful balance between the two extremes offered by sparse ECD models and the distributed source priors implicit in weighted minimum norm solutions (see also Daunizeau and Friston, 2007). Critically, the inversion scheme is fast, principled and uses a linear model, even when sparse ECD-like solutions are selected.

Overview

This paper comprises three sections. In the first, we present the theory and operational details of the inversion scheme. We then compare its performance to existing applications using distributed constraints and simulated EEG data. In the final section, we illustrate its application to a real data set that is available at http://www.fil.ucl.ac.uk/spm.

Theory

This section describes the model and inversion scheme. In brief, we use ReML to estimate covariance hyperparameters at both the sensor and source levels. Once these hyperparameters have been optimized, the posterior mean and covariance of the parameters (source activity) are given by simple functions of the data and hyperparameters. Here, ReML can be regarded as operating in an evidence optimization framework, which leads to ARD phenomena and the elimination of redundant sources. We will show exactly how a Laplace approximation to the posterior of the hyperparameters allows one to invert models with multiple sparse priors, quickly and efficiently.

[1] Although we anticipate further developments to cover full hierarchical models for multiple subject analysis and non-stationary spatial priors.

A parametric empirical Bayes model

We start with a hierarchical linear model of EEG or MEG data $Y \in \Re^{n\times s}$ over n channels and s samples.[2]

$$\begin{aligned} Y &= L\theta + X\beta + \varsigma \\ \theta &= \varepsilon \\ \varsigma &\sim N(0, V, \Sigma^{\varsigma}) \\ \varepsilon &\sim N(0, V, \Sigma^{\varepsilon}) \\ \Sigma^{\varsigma}(\lambda) &= \exp(\lambda_1^{\varsigma})Q_1^{\varsigma} \\ \Sigma^{\varepsilon}(\lambda) &= \exp(\lambda_1^{\varepsilon})Q_1^{\varepsilon} + \ldots + \exp(\lambda_m^{\varepsilon})Q_m^{\varepsilon} \end{aligned} \quad (1)$$

where $L=\Re^{n\times d}$ is a known gain or lead-field matrix and $\theta=\Re^{d\times s}$ are the unknown source dynamics at d dipoles. We have also included some confounds, X and their parameters, β as fixed effects in this model. X could be a column of ones, which restricts the estimation of covariance components to data that are mean corrected over channels (i.e., re-referenced to the mean). The terms ς and ε represent random fluctuations in channel and source space respectively. Their temporal correlations are denoted by V, which, for simplicity, we assume are fixed and known. Their spatial covariances are mixtures of covariance components $Q=\{Q^{\varsigma},Q^{\varepsilon}\}$ at each level, controlled by unknown hyperparameters, $\lambda=\{\lambda^{\varsigma},\lambda^{\varepsilon}\}$. The first-level hyperparameters λ^{ς} encode the covariance of measurement or sensor noise; here we will consider only one component, $Q^{\varsigma}=I$, noting that independence over channels does not preclude serial correlations over time. Similarly, λ^{ε} encodes the contribution of multiple empirical covariance priors Q^{ε} on the sources. Note that we are parameterizing the covariance components in both sensor and source space in exactly the same way. The scalar function $\exp(\lambda_i^{\varepsilon})$ returns the covariance scale parameters as a non-negative function of the hyperparameters. Eq. (1) allows us to specify a full generative model whose parameters and hyperparameters we seek to infer. This model comprises a likelihood and priors:

$$\begin{aligned} p(y|\theta,\beta,\lambda) &= N(L\theta + X\beta, V, \Sigma^{\varsigma}(\lambda)) \\ p(\theta) &= N(0, V, \Sigma^{\varepsilon}(\lambda)) \\ p(\lambda) &= N(\eta, \Pi^{-1}) \end{aligned} \quad (2)$$

with flat priors on the parameters of the confounds. Note that this model entails the specification of hyperpriors; $p(\lambda)=N(\eta,\Pi^{-1})$ with mean η and precision Π. It is these (shrinkage) hyperpriors that lead to ARD and automatic model selection (see below).

Empirical priors

Under a given lead-field and the form of spatiotemporal correlations in the noise, the model is defined by the number and composition of the empirical priors on the sources; $Q^{\varepsilon}=\{Q_1^{\varepsilon},\ldots,Q_m^{\varepsilon}\}$. It is these priors and the ensuing model space we want to explore. The number of components could range from one, e.g., $Q^{\varepsilon}=I$ as a classical minimum norm model, to thousands, with one component for each source (cf. Sato et al., 2004; Wipf et al., 2006). The composition of each component encodes the a priori deployment of source activity, for example, $Q^{\varepsilon}=KK^T=G$ (where $K=\Re^{d\times d}$ is a spatial convolution matrix) would correspond to a smooth or coherence prior, a component with off-diagonal terms could model two correlated sources and so on. In fact, we can model a distributed pattern of sources, $q_i=\Re^{d\times 1}$ with a separate covariance component, $Q_i^{\varepsilon}=q_iq_i^T$. In this framework, the conventional minimum norm prior $Q^{\varepsilon}=I$ encodes the highly improbable prior that sources are expressed everywhere, are of equal amplitude and are never correlated. We will show that there are much better a priori models for EEG responses. We will focus on source components with compact support and introduce the idea of modeling correlated sources explicitly, with components that have two or more regions of compact support.

The basic idea behind this approach is that any combination of prior components can be optimized and, critically, different combinations can be compared using their evidence (i.e., using Bayesian model comparison). If each component corresponds to a mixture of patterns; i.e., $Q_i^{\varepsilon}=\sum_{j\in i} q_jq_j^T$, we face the problem of searching a large model space, where each model corresponds to one partition of the patterns. Clearly, the number of partitions is enormous, ranging from one component with m patterns to m components with one pattern. These extremes could correspond to a minimum norm constraint and ARD respectively. In our current software implementation, we provide two approaches to this problem; first a greedy search that successively splits the component with the highest variance (hyperparameter) into two, using the activity (parameter) of the component's constituent patterns. The hyperparameters of the new partition (i.e., model) are optimized and the procedure repeated until the evidence stops increasing (this procedure is described in Friston et al., 2007). Alternatively, one can start with one component per pattern and use ARD to eliminate patterns. This effectively assigns redundant patterns to a null component with negligible variance. In short, the greedy search starts with one component and considers increasing numbers, while the ARD scheme starts with the maximum number, which it tries to reduce; both are guided by the model evidence. In this paper, we focus on the second (ARD) approach. In what follows, we describe how the model parameters and hyperparameters are optimized and how this furnishes the model evidence.

Model inversion

Inversion of this model proceeds in two variational steps that, for linear models of this sort, corresponds to expectation maximization (**EM**; Dempster et al., 1977): an **E**-step optimizes the parameters, holding the hyperparameters constant and the **M**-step optimizes the hyperparameters, holding the parameters constant. The reason that there are two steps is that we assume that the conditional density on the parameters and hyperparameters can be factorized. In statistical physics, this is called a mean-field assumption. In what follows, we will eliminate the parameters by substituting the source level into the sensor level of the model. This means we only have to iterate the **M**-step to optimize the hyperparameters. In this instance, the **E**-step reduces to a single operation, after the **M**-step has converged.

First, the model is reduced by projection using spatial U and temporal S projector matrices (cf. Phillips et al., 2002a,b). The temporal matrix defines a temporal subspace spanned by the signal; in our implementation, we use the principal eigenvectors of

[2] Where we use a matrix-normal notation $\varsigma \sim N(0,V,\Sigma^{\varsigma}) \Leftrightarrow \mathrm{vec}(\varsigma) \sim N(0, V\otimes\Sigma^{\varsigma})$ to denote a multivariate Gaussian density on a matrix and $\otimes$ is the Kronecker tensor product.

the sample covariance matrix over time.[3] The temporal projection can also incorporate any band-pass filtering (we use 1→64 Hz). The spatial projector would normally be the residual forming matrix, $\mathbf{U}=I-XX^{-}$ that restricts the estimation to the null space of confounds,[4] hence ReML. This gives a simplified model in which the confounds disappear because $\mathbf{U}^T X=0$:

$$\begin{aligned}\tilde{Y} &= \tilde{L}\tilde{\theta}+\tilde{\varsigma}\\ \tilde{\theta} &= \tilde{\varepsilon}\\ \\ \tilde{\varsigma} &\sim N(0,\tilde{V},\tilde{\Sigma}^{\varsigma})\\ \tilde{\varepsilon} &\sim N(0,\tilde{V},\Sigma^{\varepsilon})\\ \tilde{\Sigma}^{\varsigma}(\lambda) &= \exp(\lambda^{\varsigma})\mathbf{U}^T Q^{\varsigma}\mathbf{U}\\ \tilde{V} &= S^T V S\end{aligned} \tag{3}$$

where $\tilde{Y}=\mathbf{U}^T YS \quad \tilde{L}=\mathbf{U}^T L \quad \tilde{\theta}=\theta S \quad \tilde{\varsigma}=\mathbf{U}^T\varsigma S \quad \tilde{\varepsilon}=\varepsilon S$.

Notice that the spatial projection in channel or sensor space only affects the first-level terms and does not change the spatial priors on source space, encoded by Σ^{ε}. This reduction projects all the spatiotemporal information in $\tilde{Y}\in\Re^{u\times v}$ into u spatial and v temporal modes.

Second, the source level is projected onto measurement space to give random effects that are a mixture of first-level and second-level components

$$\begin{aligned}\tilde{Y} &= \tilde{L}\tilde{\varepsilon}+\tilde{\varsigma}\\ \\ \tilde{Y} &\sim N(0,\tilde{V},\tilde{\Sigma})\\ \Sigma(\lambda) &= \exp(\lambda_1)Q_1+\ldots+\exp(\lambda_{m+1})Q_{m+1}\\ \\ Q &= \mathbf{U}^T Q^{\varsigma}\mathbf{U},\tilde{L}Q_1^{\varepsilon}\tilde{L}^T,\ldots,\tilde{L}Q_m^{\varepsilon}\tilde{L}^T\\ \lambda &= \lambda^{\varsigma},\lambda_1^{\varepsilon},\ldots,\lambda_m^{\varepsilon}\end{aligned} \tag{4}$$

Eq. (4) is exactly the same as Eq. (3) but we have eliminated the parameters by substituting the second level into the first. This means we can dispense with the **E**-step because only the hyperparameters remain. Furthermore, we have reduced the problem to estimating the covariance components of the (projected) data, which can proceed in channel space as opposed to high-dimensional source space (cf. Gaussian process modeling). In this form, the covariance components from the source level are now simply covariance components $\tilde{L}Q_i^{\varepsilon}\tilde{L}^T$ of the data.

Finally, the moments of the hyperparameters are estimated iteratively in an **M**-step that is formally equivalent to ReML. These are then used to evaluate the conditional density of the parameters (in a single **E**-step) and a bound on the model evidence. ReML or restricted maximum likelihood was introduced by Patterson and Thompson (1971) as a technique for estimating variance components, which accounts for the loss in degrees of freedom that result from estimating fixed effects (Harville, 1977). Here, we simply supplement ReML with Gaussian hyperpriors; $p(\lambda)=N(\eta,\Pi^{-1})$, converting the maximum likelihood estimates into conditional modes (and the ReML objective function into a free energy bound on the log-evidence). Gaussian hyperpriors are equivalent to placing log-normal priors on the scale parameters. Critically, as the conditional mode of the scale parameter $\exp(\mu_i^{\lambda})$ goes to zero, so does its conditional variance and the corresponding component is switched off, enabling ARD.

Under a Laplacian fixed-form assumption the conditional density of the hyperparameters is simply $q(\lambda)=N(\mu^{\lambda},\Sigma^{\lambda})$, where μ^{λ} and Σ^{λ} are the conditional mode or expectation and covariance of the hyperparameters respectively. Under this assumption the free energy bound on the log-evidence is

$$\begin{aligned}F = &-\frac{v}{2}tr\left(\Sigma\left(\mu^{\lambda}\right)^{-1}C\right)-\frac{v}{2}\ln|\Sigma\left(\mu^{\lambda}\right)|-\frac{uv}{2}\ln 2\pi\\ &+\frac{1}{2}\ln|\Sigma^{\lambda}\Pi|-\frac{1}{2}\left(\mu^{\lambda}-\eta\right)^T\Pi\left(\mu^{\lambda}-\eta\right)\end{aligned} \tag{5}$$

where $C=\frac{1}{v}\tilde{Y}\tilde{V}^{-1}\tilde{Y}^T$ is the sample covariance matrix of the data over time bins, trials or conditions being analyzed. See Friston et al. (2007) for a detailed discussion of this objective function and its role in expectation maximization and variational Bayes.

*The **M**-step: hyperparameter estimation*

The hyperparameters are optimized by entering the sample covariance matrix **C** into the following **M**-step and iterating, until convergence

$$\begin{aligned}F_{\lambda i} &= -\frac{v}{2}tr\left(\mathbf{P}_i\left(\mathbf{C}-\Sigma\left(\mu^{\lambda}\right)\right)\right)-\Pi_{ii}\left(\mu_i^{\lambda}-\eta_i\right)\\ \\ F_{\lambda\lambda ij} &= -\frac{v}{2}tr\left(\mathbf{P}_i\Sigma\left(\mu^{\lambda}\right)\mathbf{P}_j\Sigma\left(\mu^{\lambda}\right)\right)-\Pi_{ij}\\ \\ \Delta\mu^{\lambda} &= -F_{\lambda\lambda}^{-1}F_{\lambda}\\ \\ \Sigma^{\lambda} &= -F_{\lambda\lambda}^{-1}\end{aligned} \tag{6}$$

This is simply a Fisher-scoring scheme that optimizes the free energy with respect to the hyperparameters. F_{λ} and $F_{\lambda\lambda}$ are the gradient and expected curvature of the free energy.[5] Notice that we only have to update the conditional expectation of the hyperparameters because their conditional covariance is simply the curvature of the free energy; this follows from the Laplace assumption. Here the matrix $\mathbf{P}_i=-\exp(\lambda_i)\Sigma^{-1}Q_i\Sigma^{-1}$ is the derivative of the data precision, $\Sigma(\mu^{\lambda})^{-1}$ with respect to the i-th hyperparameter, evaluated at the conditional expectations (see Eq. (4)). See Friston et al. (2007) for derivations of Eq. (6) and associated variables. When there are large numbers of covariance components, with small rank, it is computationally more efficient to compute these derivatives using a singular value decomposition of the components; see Appendix A for details. This is particularly useful when components have the form; $Q_i^{\varepsilon}=q_i q_i^T$.

This optimization scheme is very simple and reasonably efficient because it uses the expected curvature of the objective function. Wipf et al. (2006) present some alternative schemes, under additional assumptions, but consider only Gaussian process priors that are linear in the hyperparameters. By using the nonlinear hyperparameterization $\Sigma(\lambda)=\exp(\lambda_1)Q_1+\ldots$ in Eq. (4), we can

[3] The principal vectors are defined operationally as those eigenvectors whose normalized eigenvalues are greater than 1/512 (cf. the Kaiser criterion). This generally retains over 99% of the data variance.

[4] Or the principal singular vectors of the adjusted data, $Y-XX^{-}Y$; if one wanted to put an upper bound on the number of spatial modes. X^{-} is the generalized inverse of X.

[5] Here, and in Appendix A, we denote differentiation using a subscript notation, such that $F_{\lambda}\equiv\partial_{\lambda}F\equiv\partial F/\partial\lambda$.

ensure positive semi-definite covariances, while exploiting the ReML scheme for optimization. It is this nonlinear hyperparameterization that underlies ARD behavior (see below).

Technically speaking, optimizing the log-evidence bound with ReML is a much simpler alternative to standard variational schemes, which entail the use of (improper) conjugate hyperpriors. It allows one to use any hyperprior and formulate an efficient ascent, using gradients and curvatures in the normal way. The reason that the **M**-step does not need quantities from the **E**-step is that any dependence on the parameters has been eliminated by collapsing the hierarchical model down to the data level. This is a key advantage of ReML and renders it formally the same as Gaussian process modeling, which uses a density on the space of functions causing the data. In our case, the data are a mixture of covariance components, which are a nonlinear function of the hyperparameters (and only the hyperparameters).

E-step: parameter estimation

After the **M**-step has converged the hyperparameter estimates are used to reconstitute the source covariance in Eq. (1) to compute the conditional density $q(\tilde{\theta})=N(\tilde{\mu}^{\theta},\tilde{\Sigma}^{\theta})$ of the sources using the matrix inversion lemma:

$$\mathbf{M} = \Sigma^{\varepsilon}(\mu^{\lambda})\mathbf{L}^{T}\Sigma(\mu^{\lambda})^{-1}$$

$$\tilde{\mu}^{\theta} = \mathbf{M}\tilde{Y}$$

$$\tilde{\Sigma}^{\theta} = \Sigma^{\varepsilon} - \mathbf{M}\tilde{\mathbf{L}}\Sigma^{\varepsilon} \tag{7}$$

Notice that, at no point in the **M-** or **E**-step, do we have to invert matrices larger than $n\times n$, where n is the number of channels. A key aspect of this scheme is that the maximum *a posteriori* or MAP estimator matrix $\mathbf{M}=\Sigma^{\varepsilon}L^{T}\Sigma^{-1}$ needs only to be computed once for any trial or collection of trials (depending on the data one wants to optimize the estimates over). This estimator can then be applied to any mixture of responses over time bins, trials or conditions to obtain the conditional expectation of that mixture or contrast. For example, the conditional expectation of a contrast of responses in source space that is specified by a time–frequency contrast[6] matrix $W\in\Re^{s\times u}$, over time bins, is

$$\mathbf{E}\{\theta W\} = \tilde{\mathbf{M}}Y\tilde{W}$$

$$\tilde{W} = SS^{T}W$$

$$\tilde{M} = \mathbf{U}\mathbf{M}\mathbf{U}^{T} \tag{8}$$

These contrasts generally test for specific time–frequency components by defining a temporal subspace of interest (e.g., gamma oscillations between 300 and 400 ms after stimulus onset). This contrast has conditional covariance, $\tilde{W}^{T}V\tilde{W}\otimes\mathbf{U}\tilde{\Sigma}^{\theta}\mathbf{U}^{T}$. The contrast matrix can be a simple vector; for example a Gaussian window $W\in\Re^{s\times 1}$ over a short period of peristimulus time or cover specified frequency ranges (with one frequency per column) over extended periods of peristimulus time (Kiebel and Friston, 2004). The conditional expectation of the energy in a contrast is

$$\mathbf{E}(\theta W W^{T}\theta^{T}) = \tilde{M}Y\,\tilde{W}\tilde{W}^{T}Y^{T}\tilde{M}^{T} + \Sigma^{\theta}tr(\tilde{W}^{T}V\tilde{W}) \tag{9}$$

See Friston et al. (2006) for details. Both the conditional estimates of contrasts and their energy can then be used as summaries of condition-specific responses for each subject and entered into statistical models of between subject responses in the usual way.

Conditional contrasts

Although we will focus on reconstructing event-related potentials (ERP) averaged over a single trial-type in this paper, contrasts can cover both peristimulus time and conditions. This extension induces a factorization of the contrast matrix $\tilde{W}\rightarrow c\otimes\tilde{W}$ into a peristimulus time factor, $\tilde{W}$ over time bins and a contrast vector over trial-types or conditions; e.g., $c=[1;-1]$ would test for a larger response in the first condition. In this instance the conditional expectation of the contrast is $\tilde{\mathrm{M}}[Y_1,\ldots,Y_N](c\otimes\tilde{\mathrm{W}})$ for N condition-specific event-related potentials in $Y=[Y_1,\ldots,Y_N]$.

This factorization of the contrast matrix into within and between-trial effects can be particularly useful for estimating energy over individual trials of the same type (cf. induced responses[7]). In this instance, $\tilde{W}\rightarrow I\otimes\tilde{W}$ and the induced response is

$$\begin{aligned}\tilde{M}Y(I\otimes\tilde{W}\tilde{W})^{T}Y^{T}\tilde{M}^{T} + \Sigma^{\theta}tr(I\otimes\tilde{W}^{T}V\tilde{W})\\ = \sum_{i}\tilde{M}Y_i\tilde{W}\tilde{W}^{T}Y_i^{T}\tilde{M}^{T} + N\Sigma^{\theta}tr(\tilde{W}^{T}V\tilde{W})\end{aligned} \tag{10}$$

This is simply the sum of squared conditional estimates from each trial plus a term that depends on the conditional covariance. This term corrects for bias due to the implicit shrinkage priors we have used (see Friston et al., 2006 for more details). In this paper, we will not consider contrasts over trials or conditions further because our focus is on comparing the models that underlie the estimators. Furthermore, most inference in ERP research is at the between-subject level and only requires a summary of each subject's condition-specific response. This summary is usually the conditional estimates considered here.

Model comparison and the log-evidence

To compare different models (defined by different priors) we need their log-evidence or marginal likelihood (see also Serinagaoglu et al., 2005). The log-evidence is bounded by the free energy optimized in the **M**-step, such that when the free energy is maximized, so is the model evidence and we have the approximate equality $\ln p(\tilde{Y}|m)\approx F$. This rests on the Laplace approximation for both the parameters and hyperparameters and is derived from basic principles in Friston et al. (2007).

It is interesting to consider the behavior of the last two (complexity) terms of the free energy in Eq. (5); when a covariance component is not necessary to explain data its hyperparameter approaches its prior expectation; $\mu^{\lambda}\rightarrow\eta$. In this instance, the

[6] A contrast matrix refers to a set of weights that are applied to parameters to form a mixture or compound (if this compound is estimable, it is referred to as a contrast in classical statistics).

[7] In this paper, we make no distinction between the total energy induced by a stimulus or event and the energy remaining after the evoked (i.e., average) response has been removed.

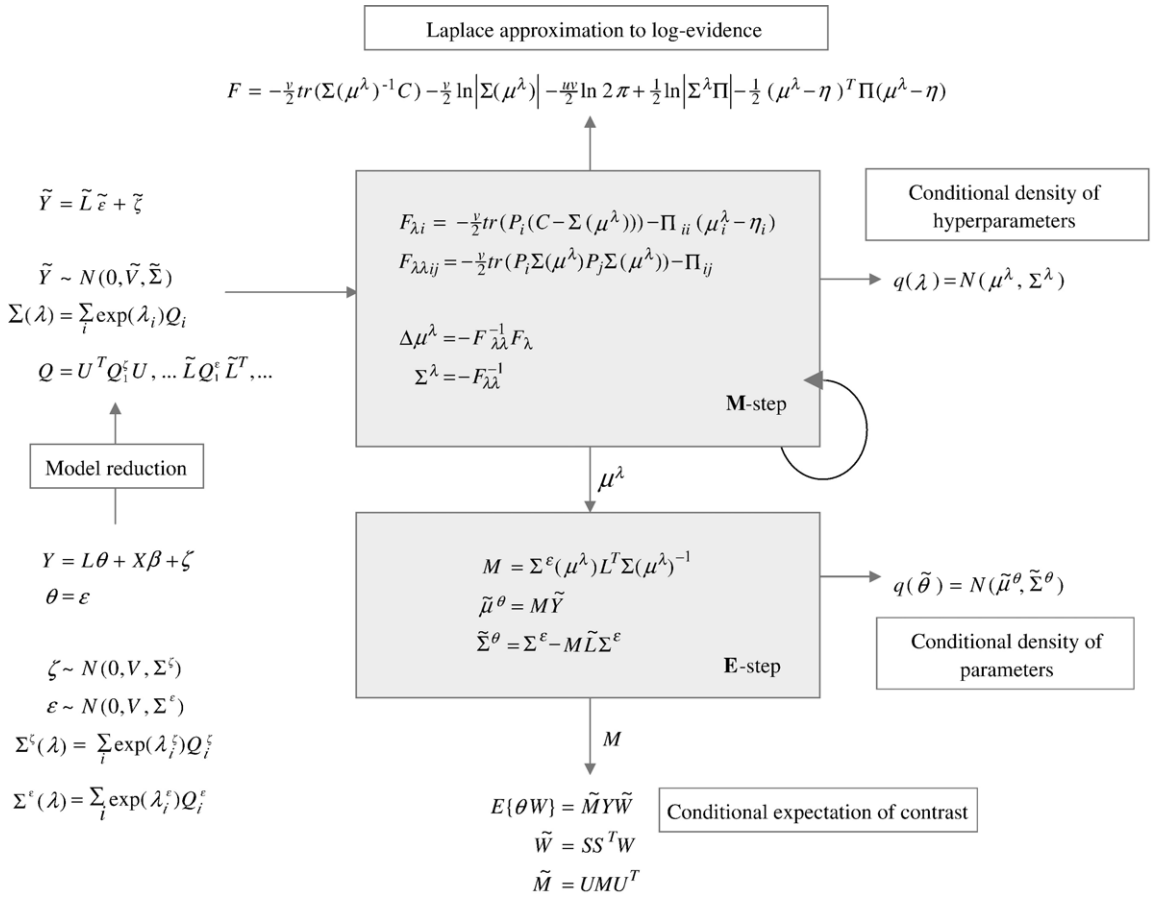

Fig. 1. Schematic showing the architecture of the inversion scheme. This comprises a model reduction, by projection onto a spatiotemporal subspace, followed by variational inversion of the ensuing hierarchical or parametric empirical Bayes model. In this instance, the variational scheme, under a Laplace assumption, reduces to expectation maximization. The products of this scheme, which rests on iteration of the **M**-step, are the conditional density of the model parameters (and hyperparameters) and a variational bound on the model's log-evidence or marginal likelihood. Note the central role of the **M**-step in furnishing the sufficient statistics necessary for evaluating the free energy bound and the conditional expectations of the parameters or sources and of any contrasts.

conditional precision approaches its upper bound, $\Sigma^{\lambda-1} \to \Pi$, which is the prior precision and both the complexity terms approach zero. In other words, provided one uses a non-informative hyperprior with a very low prior expectation, redundant covariance components will be switched off and will not affect the free energy bound. This is because they do not increase accuracy or complexity. In our work,[8] we use $\eta = -32$ and $\Pi = 1/256$.

A relatively flat Gaussian hyperprior effectively places a Jeffreys prior on the scale parameters, $\exp(\mu^\lambda)$ of the covariance components; this hyperprior is always a proper density, which may be useful in a sampling context. However, the Gaussian is not completely flat and enables us to preclude unreasonably large scale parameters; for example, for a scale parameter of one, the hyperparameter must be two prior standard deviations; $32 = 2\sqrt{256}$ from its prior mean. Strictly speaking, this is a weakly informative hyperprior, as characterized by Gelman (2006); i.e., a weakly informative proper distribution "that is set up so that the information it does provide is intentionally weaker than whatever actual prior knowledge is available".

[8] To ensure this prior is only weakly informative, we scale the sample covariance and its components so that their trace is one.

To compare two or more models we simply look at the difference in log-evidence; this is the log of the Bayes factor comparing two models (Kass and Raftery, 1995). By convention, a difference in log-evidence of about three or more is taken as strong evidence in favor of the model with the greater likelihood. We will demonstrate this approach to model comparison in the final section to look at different models and to optimize a model within a given class. This concludes the specification of the forward model and its inversion. See Fig. 1 for a schematic summary and the central role of the **M**-step (augmented ReML). In the next section, we look at some specific examples, with a special focus on model comparison.

Simulations

In this section, we use simulations to evaluate the performance of various models. We will look at model evidence, variance explained and pragmatic measures of spatial and temporal accuracy. We describe how the data were simulated and the models considered and then report comparative analyses. First, we consider the generative models used to simulate data. These models comprise the conventional forward model encoded in the lead-field matrix and the priors on the sources.

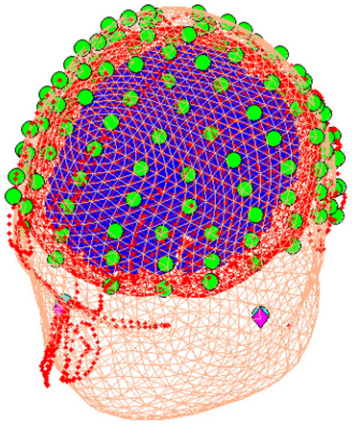

Fig. 2. Meshes and locations used to define the forward model. The three concentric meshes correspond to the scalp, the skull and cortex; the cortex mesh comprises 4004 vertices, which constitute source space. The green dots indicate the position of the 128 channels. Channel locations were registered to the meshes using fiducials in both spaces (cyan cortex and magenta diamonds) and the subject's head shape, as digitized with a Polhemus Isotrak (red dots). This display format is the standard SPM output.

The forward model

The forward models in this paper use a high-density canonical cortical mesh. These meshes were obtained by warping a template mesh to match the structural anatomy of an individual subject as described in Mattout et al. (2007). The template mesh of a neurotypical male was extracted from a structural MRI, using BrainVISA and Anatomist.[9] This furnished a high-density mesh with a uniform and discrete coverage of the gray–white matter interface. After down-sampling to various sizes, this mesh corresponds to the templates currently available in the latest release of the SPM software package (see Software note). Here, we use a mesh down-sampled to 4004 vertices. For any given mesh, each vertex location corresponds to a dipole position, whose orientation is fixed perpendicular to the surface.

The lead-field or gain matrix was computed for the canonical mesh and coregistered channel locations using a three-sphere head model for EEG using routines from BrainStorm (http://neuroimage.usc.edu/brainstorm/). The coregistration and forward model was computed within SPM5 (http://www.fil.ion.ucl.ac.uk/spm). This provided the gain matrix; $\mathbf{L}\otimes\Re^{128\times 4004}$ coupling 4004 cortical sources to 128 EEG channels, where each source has a unique location in the standard anatomical space of Talairach and Tournoux (1988). See Fig. 2 for the spatial configuration of sources and sensors used for the simulations. In fact, this configuration is from the single subject, whose data are analyzed in the next section.

Three spatiotemporal prior models

Fig. 3 shows the three prior models for the deployment of activity over cortical sources considered in this paper. For all three models, temporal priors were fixed by assuming Gaussian autocorrelations; $V(\tau)$:τ=4 among channel noise with a standard deviation of four milliseconds; i.e.,

$$V(\tau) = K(\tau)K(\tau)^T$$

$$K(\tau)_{ij} = \exp(-\frac{1}{2}(i-j)^2\tau^{-2}) \tag{11}$$

This is roughly the autocorrelation of white noise that has been filtered with the band-pass filter we used in pre-processing. The models differed in terms of their empirical spatial priors. These models included a conventional minimum norm model (MNM) where $Q^\varepsilon = I$. As mentioned above, this model asserts that all sources are active, with equal a priori probability and that none are correlated. We then consider a more realistic model (COH) with two components modeling independent and coherent sources respectively; $Q^\varepsilon = \{I,G\}$ (cf. Pascual-Marqui, 2002), where

$$G(\sigma) = \exp(\sigma A) \approx \sum_{i=0}^{8} \frac{\sigma^i}{i!}\mathbf{A}^i \tag{12}$$

is a spatial coherence prior, which is a Green function of an adjacency matrix **A**. This matrix encodes the neighborhood relationships, $A_{ij}\in[0,1]$, between nodes of the cortical mesh defining the solution space; see LeSage and Pace (2000) and Harrison et al. (2007) for more details on this Gaussian process prior. The Taylor approximation above ensures that only eighth-order neighbors (i.e., nodes connected by eight or less edges) have non-zero values. This enforces priors with compact and sparse support on the cortical mesh nodes. The smoothness parameter, σ, can be thought of as an autoregression coefficient and varies between zero and one. In this paper, we used a spatial coherence prior, $G(\sigma)$:σ=0.6, which propagates spatial dependences over three or four mesh vertices that are, on average, about 6 mm apart.

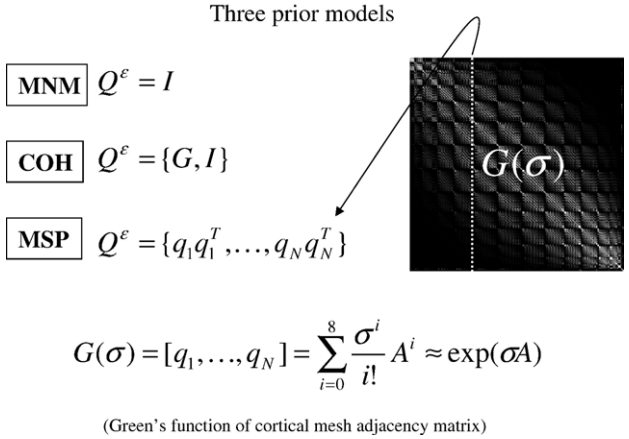

Fig. 3. Schematic illustrating the three models we focus on in this paper. These include an MNM model with a single covariance component encoding identically and independently distributed sources, a model accommodating spatial dependencies through an additional component modeling spatial coherence. This component is a Greens function based on the adjacency matrix of a cortical mesh modeling source space; finally we consider multiple sparse priors by modeling multiple source components as patterns with compact support. In fact, these are sparsely sampled columns of the Green function matrix.

[9] Cointepas et al. (2001) and http://brainvisa.info/doc/brainvisa/en/processes/aboutBrainVISA.html.

Finally, we consider a multiple sparse prior (MSP) model with components $Q^e = \{q_1 q_1^T, \ldots, q_N q_N^T\}$ modeling activity in N patterns, $q_1, \ldots, q_N$. These source components were formed by sampling from evenly spaced columns of the coherence matrix above. This ensured that each source component had compact support and was locally coherent. Note that there are many fewer patterns than dipoles. One can imagine these components as being formed by selecting a dipole and propagating a proportion (i.e., σ) of its activity to connected dipoles and then repeating this eight times. We sampled components from homologous (nearest neighbor) nodes in each hemisphere to give right, q_i^{right} and left, q_i^{left} hemispheric components. We then added the homologues to give a bilateral component; $q_i^{\text{both}} = q_i^{\text{right}} + q_i^{\text{left}}$, modeling correlated sources in each hemisphere. All three components entered the model. Unless otherwise stated, we use 128 components per hemisphere. This number is based on the model optimization procedure described in the next section. Note that these components are not proper (i.e., they have a rank of only one); however, this is not an issue, provided we restrict ourselves to inference using the marginal posteriors at each dipole.

Note that if bilateral sources are truly correlated the unilateral components are redundant and will not be selected; conversely, in the absence of correlations, the bilateral component will be irrelevant. The motivation for using this particular set of components is based on prior knowledge about extrinsic cortico-cortical connections in the brain that mediate long-range synchrony and coherence; these can be loosely classified as intra-hemispheric 'U' fibers and inter-hemispheric trans-callosal connections coupling homologous regions (Salin and Bullier, 1995). These two sources of correlation are reflected in the local coherence modeled in all components and the possibility for inter-hemispheric correlations that are accommodated by the bilateral source components. Clearly, there are many other priors on functional anatomy that we could explore; however, the current MSP model is sufficiently different from the conventional and smoothness-constraint models to make a comparative evaluation interesting.

Synthetic data

We took care to simulate realistic data using as many empirical constraints as possible. Our strategy was to use empirical data to define temporal dynamics of evoked responses (and the level of noise) and assign these dynamics to distributed but contiguous

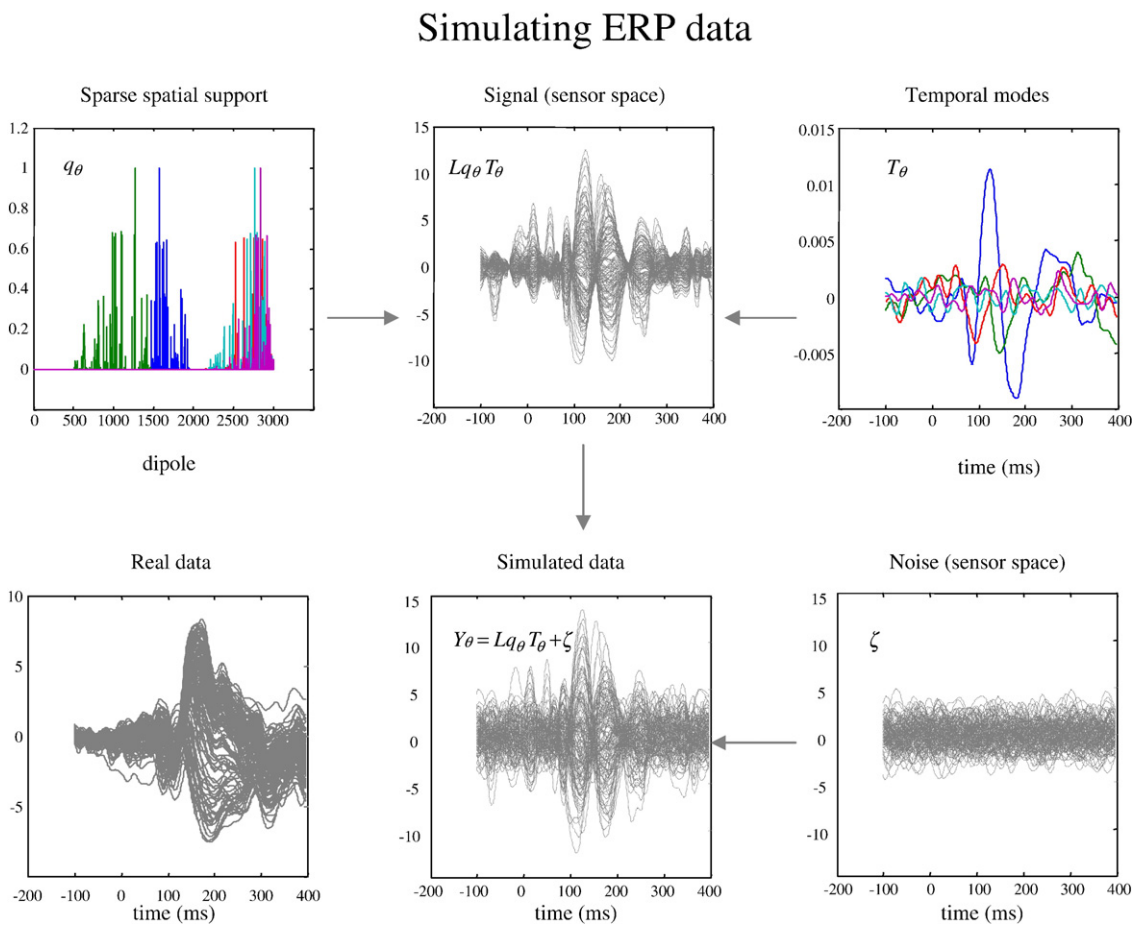

Fig. 4. Schematic illustrating the construction of synthetic data. The upper panels pertain to the signal; the middle panel shows the synthetic signal in channel space, which was obtained by projecting the signal in source space through the lead-field matrix. Source activity is composed of five compact sources distributed over dipoles on the cortical mesh (left panel). The dynamics of these sources conform to principal eigenvariates of real data (right panel) computed following a singular value decomposition of real channel data (the color of the time course encodes the corresponding spatial support in the left panel). Gaussian noise (lower right) is added to the synthetic signal (lower middle) to provide simulated data. For comparison, real data (used in the third section) is also shown (lower left). Data are shown for 100 ms before stimulus onset to 400 ms after.

nodes on a cortical mesh. Using the real EEG data described in the next section, we performed a singular value decomposition in channel space to identify its principal time courses over 821 (~1 ms) time bins, starting 200 ms before the presentation of a stimulus. We retained the first five singular vectors, $T_\theta \in \Re^{5\times 821}$ and deployed these over five distributed sources. These sources, $q_\theta \in \Re^{4004\times 5}$ were columns of a smooth spatial coherence prior; $G(\sigma)$:$\sigma = 0.8$. Critically, these columns were selected at random and were not the same components used by the MSP model, which used different columns of the spatial coherence prior and a different value of spatial coherence, 0.8 vs. 0.6. By construction the time courses of the five sources were orthogonal; however, no special measures were taken to prevent them overlapping in space or time.

The ensuing source activity was projected through the lead-fields to simulate signals in channel space, $Lq_\theta T_\theta$. Serially correlated noise $\varsigma \in \Re^{128\times 821}$ was created by sampling from a Gaussian distribution and smoothing with a Gaussian convolution matrix, $K(\tau)$: $\tau = \sqrt{128}$, which modeled serial correlations with a correlation length of about 10 ms. The noise was scaled to a tenth of the L_1-norm of the simulated signal; this provided a signal to noise ratio of about ten, in terms of relative power or variance. The signal and noise were mixed to provide simulated data. An example from this procedure is shown in Fig. 4 for 512 time bins.

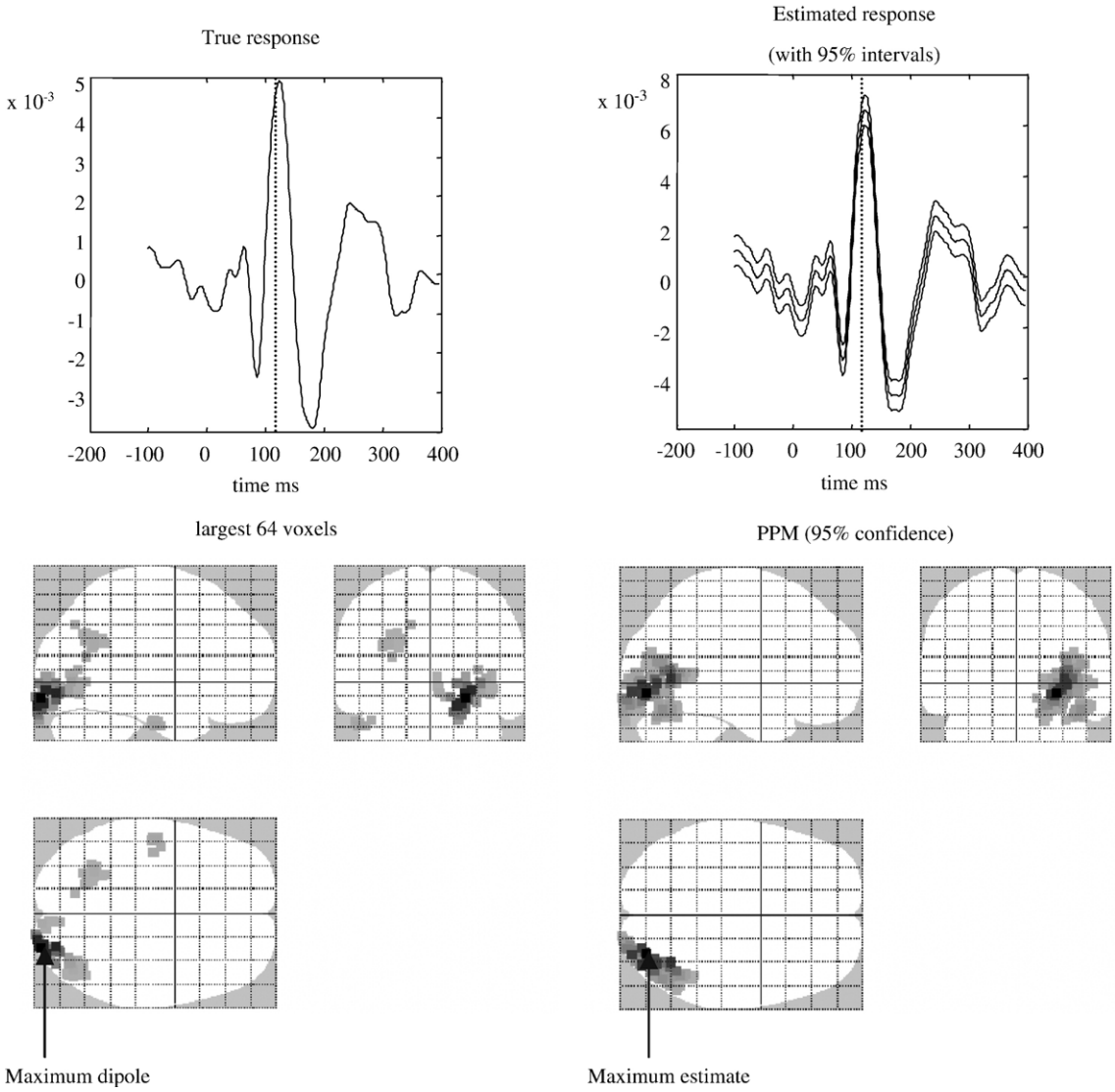

Fig. 5. True (left) and estimated (right) responses for one realization. The upper panels show the time course from the dipole with the maximum activity over all dipoles and time bins. The conditional estimate of this response is shown on the right, along with its 95% confidence intervals. The agreement is self-evident. The lower panels show the spatial deployment of activity for the time bin with the biggest absolute response (inducted by the broken lines in the upper panels). These conditional reconstructions are shown in maximum intensity projection format and highlight the 64 dipoles with the largest activity (true response; left) and those which are estimated with 95% confidence to be greater than zero (estimated response; right); this constitutes a posterior probability map or PPM. The location and time course of the dipole with the true and estimated maximum responses (indicated with arrows) are used to measure spatial and temporal accuracy respectively. Data are shown for 100 ms before stimulus onset to 400 ms after.

Simulations and accuracy measures

We employed four comparative metrics: the log of the marginal likelihood or *model evidence* as approximated with free energy in Eq. (5); the percent variance explained over all channels and time bins by the conditional estimates (cf. coefficient of determination); *spatial accuracy* was assessed with the Euclidean distance in millimeters between the most active [unsigned] dipole over time bins and the dipole with the largest [unsigned] conditional expectation at the same time. *Temporal accuracy* was assessed with the squared correlation (i.e., the coefficient of determination) between the true time course of the most active dipole and the conditional estimate from the dipole used to assess spatial accuracy. It should be noted that this is a rather severe test of localization performance because it requires the inversion to find the correct (i.e., largest) among five distributed but compact sources. Furthermore, the temporal accuracy can be subverted, if the wrong dipole has been identified due to spatial inaccuracies. The Euclidean distance is a rather unsophisticated measure (a Geodesic metric would be more principled in a single-subject setting); however, this assessment of localization error is commonplace and is more relevant to inference at the between-subject level in three-dimensional anatomical space (see Henson et al., 2007 for a fuller discussion).

Fig. 5 shows an example of one simulation and the conditional expectations following its inversion. The upper panels show the time courses for the dipoles used to assess spatial and temporal accuracy. The lower panels show the responses over dipoles using a maximum intensity projection. This uses the same format commonly employed by SPM for other imaging modalities and provides orthogonal, glass brain views of responses or regional effects in the standard anatomical space of Talairach and Tournoux (1988). This is possible because each vertex in the canonical mesh has a direct mapping to standard space. In the figure we have called this a PPM or posterior probability map. This is because we can compute the posterior probability that dipole activity is greater than zero and determine the lower bound on the probability, over dipoles or voxels, at the time-point displayed.

Simulation results

We generated 256 synthetic data sets and inverted the three models above for each realization using $V(\tau)$:τ=4. Fig. 6 (upper panel) shows the distribution of the signal to noise ratio over the simulations; these levels are fairly typical of ERP data that we acquire, with an SNR of about ten (min, 8.2; max, 19.2; mean, 12.1). As might be expected, under these levels of noise the percent of variance explained is about 90%, as shown in the lower panel. Note that, without constraints, one could easily account for all the variance because we are dealing with an over-determined problem. The reason that some variance is unexplained is that the empirical priors are enforcing constraints on the solution. If these empirical priors have been optimized properly, one would like to see about 92.37%=12.1/(1+12.1) of the variance explained because this is the proportion that is true signal. Happily, the MSP model explained almost exactly this proportion (92.30%). The simpler COH and MNM models explained substantially less (50.46% and 52.15% respectively) and showed a greater variability over realizations; see Fig. 6 (lower panel). This is because they are poorer models of the data.

This was confirmed by examining the evidence of the three models over simulations. Fig. 7 (upper panel) shows that the

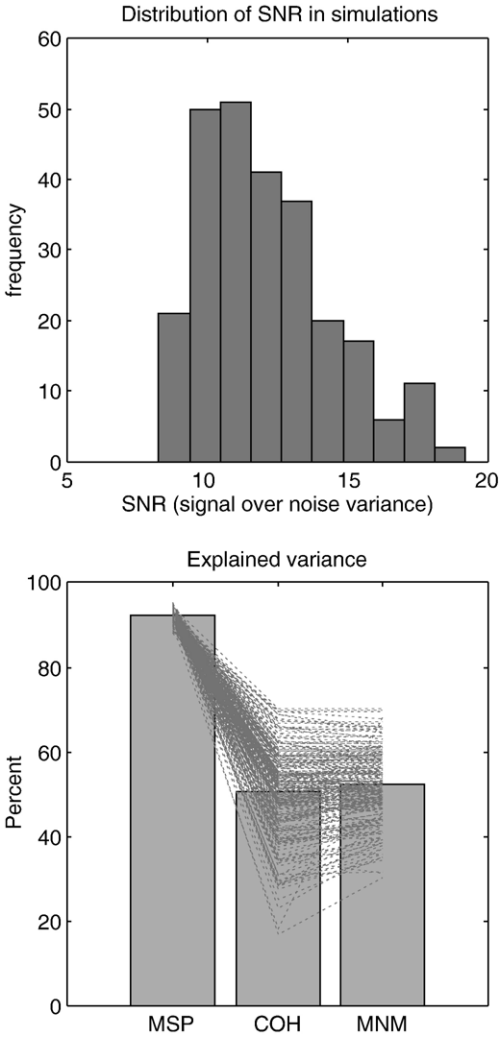

Fig. 6. Top: frequency distribution of signal to noise over the 256 simulations. Lower: the percent variance explained by the conditional estimates for the three models: the bars encode the mean over simulations and the broken lines show the variance explained by the models for each realization.

likelihood of the MSP model is vastly greater than the other two models, although there is strong evidence for COH over MNM. This formal model assessment reflects both accuracy and complexity. However, if we focus on simple measures of accuracy, the MSP model still supervenes. This has to be the case because the MSP is more complex and must pay the price for its extra parameters with an improved fit. The lower panels of Fig. 7 show the average temporal and spatial accuracy measures and the dispersion over realizations for the three models. The same data are plotted in terms of cumulative frequency in Fig. 8 to disclose the quantitative differences more clearly. In terms of spatial accuracy, the main difference appears in the more pronounced mislocalizations; quantitatively, 81.6% of all MSP models located the maximum source within 40 mm of the true maximum. In contrast, only 69.9% of the COH models and 64.4% of the MNM models were able to attain this spatial accuracy. This represents a two-fold increase in false localizations for the simpler models over MSP, which is remarkable. In relation to some simulations, these localization

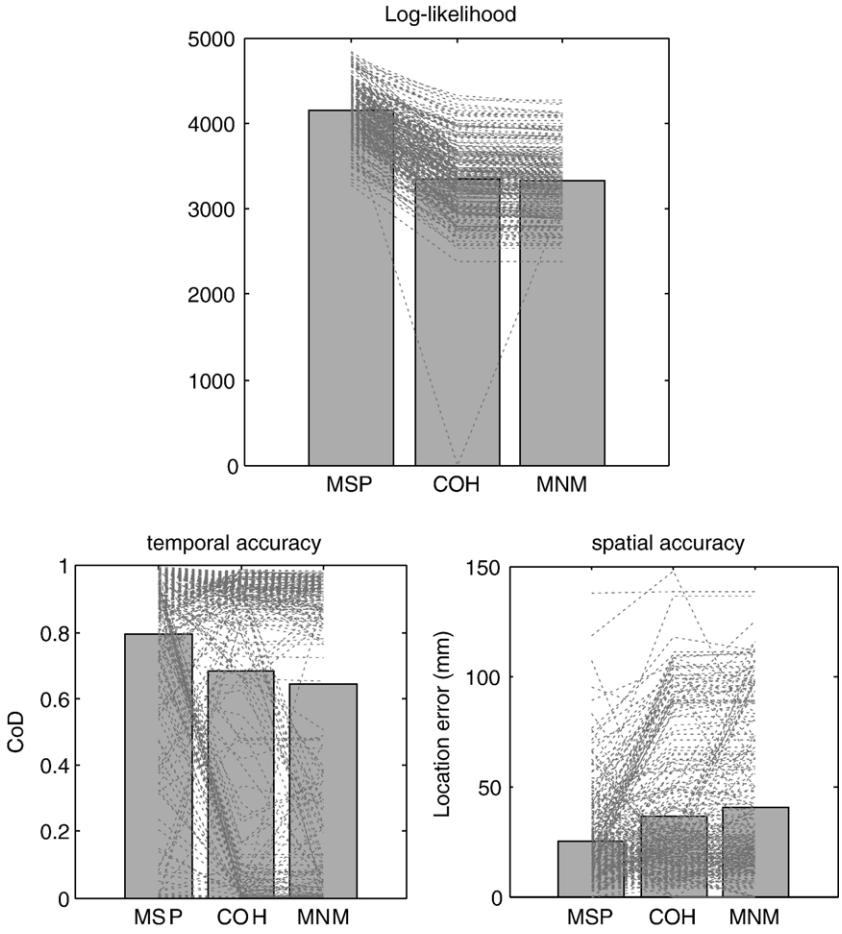

Fig. 7. Top: model comparison in terms of log-evidence or marginal likelihood for the three models under each of the 256 realizations of simulated data. The bars encode the mean log-evidence and the broken lines link the model evidence for each realization (these also indicate any multimodal distribution about the means). Lower panels: accuracy measures for the three models using the same format. Temporal accuracy (left) is measured in terms of the squared correlation or coefficient of determination for the true time course at the true maximum and the time course estimated at the estimated maximum (see Fig. 5). Spatial accuracy (right) is expressed as the Euclidean distance between the true and estimated maximum of source activity in millimeters.

errors may appear to be large; however, recall that we are using multiple distributed sources, where each source is itself dispersed over the cortical sheet. Furthermore, a proportion of realizations had fairly low signal to noise levels. Adding dipoles and noise to simulations characteristically compromises localization performance (see Mosher et al., 1993; Russell et al., 1998). The key

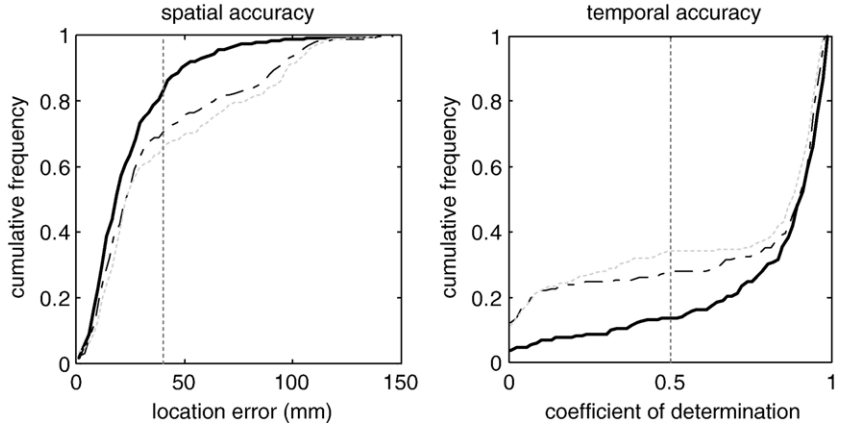

Fig. 8. Representation of spatial and temporal accuracy for the MSP (solid line), COH (dashed line) and MNM (dotted line) models over 256 simulations. Left: cumulative frequency of realizations where the localization error fell below an upper bound (*x*-axis). Right: cumulative frequency over realizations, where the coefficient of determination fell below an upper bound (*x*-axis). The dotted lines represent threshold on accuracy of 32 mm (spatial) and a coefficient of determination of one half (temporal).

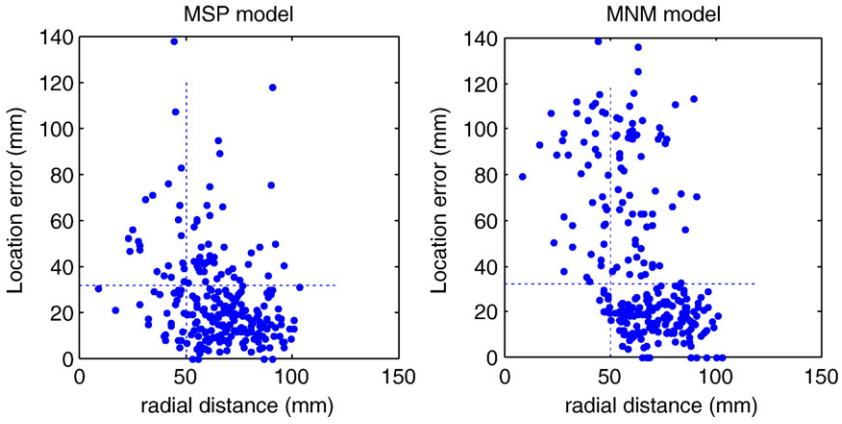

Fig. 9. Detailed analysis of localization error in terms of the depth of the true location of the maximum dipole. This depth is expressed as the distance in millimeters from the origin of standard space. The results on the left (for the MSP model) show that there is a slight depth effect in the sense that the localization errors for deep and superficial sources are roughly the same. In contrast, the MNM model failed to locate deep sources (right panel). The dotted line corresponds to a partitioning of deep and superficial sources at 50 mm and a localization error of 32 mm.

metric here is not absolute error but the improvement afforded by a more accommodating model.

Similarly, in terms of temporal accuracy, 13.6% of all MSP models failed to reach a coefficient of determination of one half (i.e., half the estimated variance over time could be regarded as veridical; equivalently, the correlation was less than $0.7 = \sqrt{1/2}$). This contrasts with 27.3% and 33.9% of COH and MNM models respectively that failed to reach this temporal accuracy. Again this is a two-fold increase in reconstruction failure. Interestingly, the COH model performed much better than the classical MNM model on temporal accuracy, despite the fact that they differ only in their spatial priors.

To emphasize the superiority of the MSP solutions, we examined them for a common failing of simple models, namely a bias toward superficial sources (Fuchs et al., 1999). Fig. 9 shows the localization error as a function of the distance between the true maximum and the origin of standard anatomical space (i.e., close to the center of the spherical head model). The difference between the MSP solution (left panel) and the simpler MNM solution is obvious; 'deep' dipoles that are close to the origin are more than likely to be misplaced by the MNM model. For example, 81.6% of deep sources (less than 50 mm from the origin) were misplaced by at least 32 mm; this compares to only 28.9% for superficial sources (more than 50 mm from the origin). This bias is reduced substantially in the MSP solutions, with 44.8% and 20.2% of solutions misplaced by 32 mm or more, for deep and superficial sources respectively.

On all the metrics considered, the MSP model outperformed optimized conventional models. This is not too surprising because the data were generated in a way that the MSP model could reproduce. This does not mean that the comparative analysis is specious; it reflects that fact that the MSP model is the most general among the three models; if we had generated data using the COH priors, the MSP would select its components to emulate the performance of the COH model. Note that simulating data with uncorrelated sources of equal amplitude, at every dipole, would not generate very plausible data. In other words, COH priors are not, in all senses, plausible priors but they are certainly much better than MNM priors. We now turn to explorations of model space using real data.

Analyses of real data

In this section, we use real data to provide some provisional validation of the MSP model through model comparison. We start by optimizing the number of MSPs, in the context of real data, and then compare the three models from the previous section, using an optimum MSP model. We conclude with anecdotal evaluations in relation to ECD and fMRI analyses, of the same experimental effects, which try to establish some construct and face validity.

The EEG data

The EEG data were acquired from a subject who participated in a multimodal study on face perception (for detailed description of the paradigm see Henson et al., 2003 and http://www.fil.ion.ucl.ac.uk/spm, where these data can be downloaded). The subject made symmetry judgments on faces and scrambled faces. Faces were presented for 600 ms, every 3600 ms while data were acquired on a 128-channel ActiveTwo system, sampled at 2048 Hz, plus electrodes on left earlobe, right earlobe, to measure eye movement. After artifact rejection,[10] the epochs (80 face trials, collapsing across familiar and unfamiliar faces) were baseline-corrected from −100 ms to 0 ms, averaged and down-sampled to 1024 Hz. The subject's T1-weighted MRI was obtained at a resolution of 1 mm^3 and was used to map a canonical template mesh from standard anatomical space into the individual's space. The subject's head shape was digitized with a Polhemus Isotrak, which was used to coregister the channel locations and cortical mesh using a rigid-body (six-parameter) affine transformation (see Fig. 2).

Optimizing the number of sparse priors

In the previous section, we used 128 patterns or source components per hemisphere (i.e., 384 prior components). This number was chosen on the basis of a model comparison using the real data considered next: Because our Bayesian inversion furnishes the model evidence and the model can be defined in terms of the

[10] Artifacts were defined as epochs in which a time bin exceeded an absolute threshold of 120 μV.

number or form of its components, we can use the free energy bound on the log-evidence to search model space and optimize any model within a specified family. In this case, we can explore the number of MSPs and optimize the model complexity using any empirical data. Fig. 10 shows the results of this analysis for the current data, plotting the log-evidence as a function of the number of components for each hemisphere. It can be seen that the model evidence exhibits a sigmoid behavior jumping sharply after eight and then increasing more slowly (note the logarithmic scaling over the number of components). This suggests that models with fewer components are compromised, in terms of being able to explain observed data, by insufficient degrees of spatial freedom. Increasing the number of MSPs further led to an actual reduction in the free energy. In principle, this should not happen but, in practice, the EM scheme can converge on local minima, or more precisely, a more complex model with additional hyperparameters may increase the chance of local minimum problems.

The lower panels of Fig. 10 show the [unsigned] source activity as a maximum intensity projection. These images show the 512 dipoles with the greatest activity at 170 ms and can be regarded as X-rays of activity. In all cases there are bilateral extrastriate and medial and lateral temporal responses with a more unilateral (right-sided) frontal component. As the model evidence increases the medial sources migrate laterally and posteriorly into the ventral fusiform gyrus. Qualitatively, there is a marked change in source reconstruction (lower panels) as number of components goes from 8 to 32 but a smaller difference from 32 to 128. It takes about 10 s to fit an MSP model with 128 components per hemisphere and we consider this number a useful compromise between the quality of the model and computational cost. In the remainder of this paper, all the MSP modes employ 128 source components per hemisphere (i.e., $384 = 3 \times 128$ components in all).

We performed similar analysis for the coherence parameter of the Green function. These analyses showed much less dependence on coherence, provided that it was in the range; $0.2 \leq s \leq 0.8$. The maximum was usually around 0.6, but changed with the number of MSPs and clearly the number of mesh vertices (which determines the vertex spacing) (results not shown).

Model comparison

Having optimized the **MSP** model, we then compared the three models of the previous section using the ERPs evoked by faces. The results of this analysis are shown in Fig. 11, using the same format as

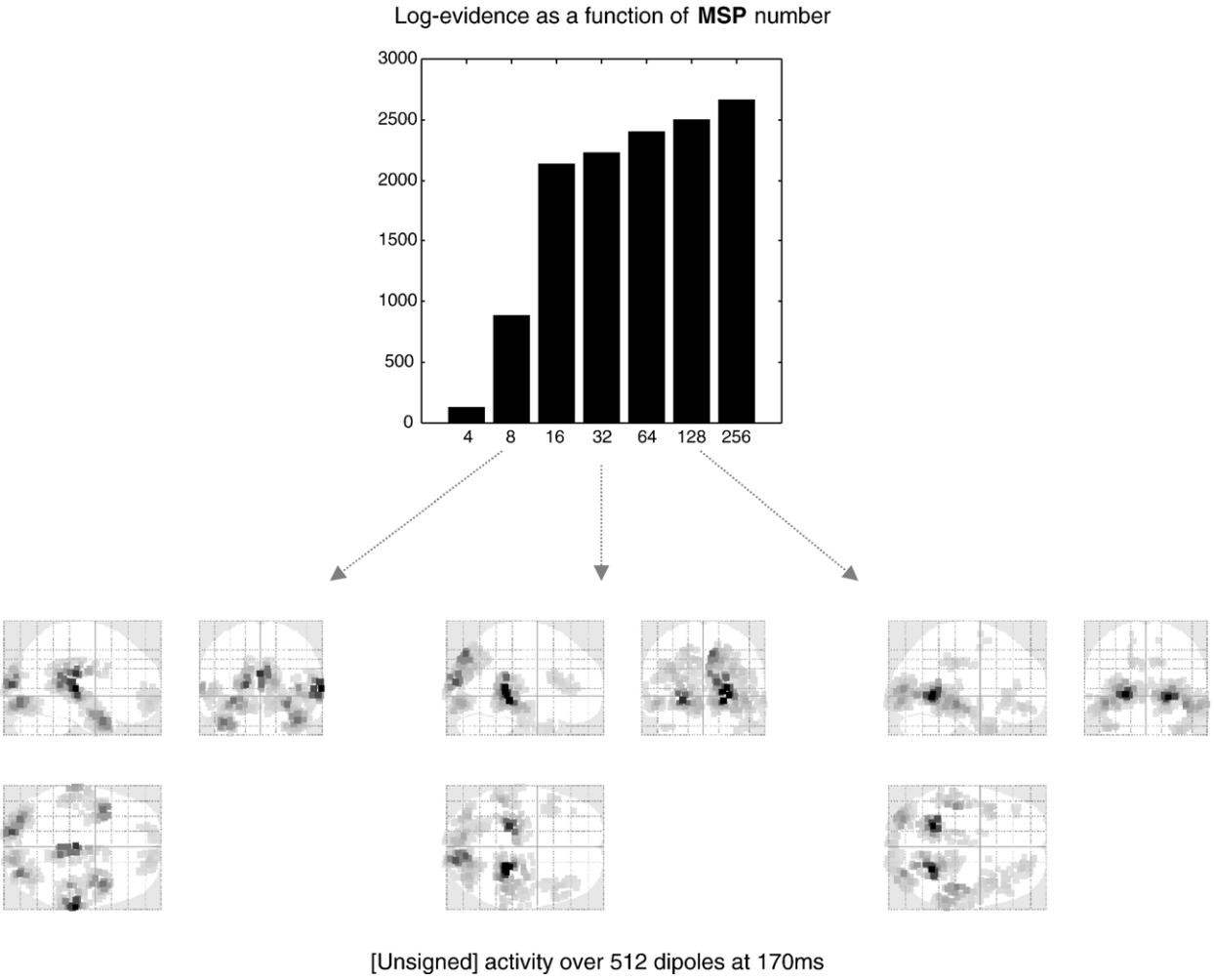

Fig. 10. Exploring model space in terms of the number of multiple sparse priors. Upper panel: the free energy bound or log-evidence for models with increasing number of MSPs (x-axis). These are expressed as the number of source components per hemisphere. Lower panels: maximum intensity projections of the spatial activity at 170 ms for exemplar models (with 8, 32 and 128 MSPs per hemisphere).

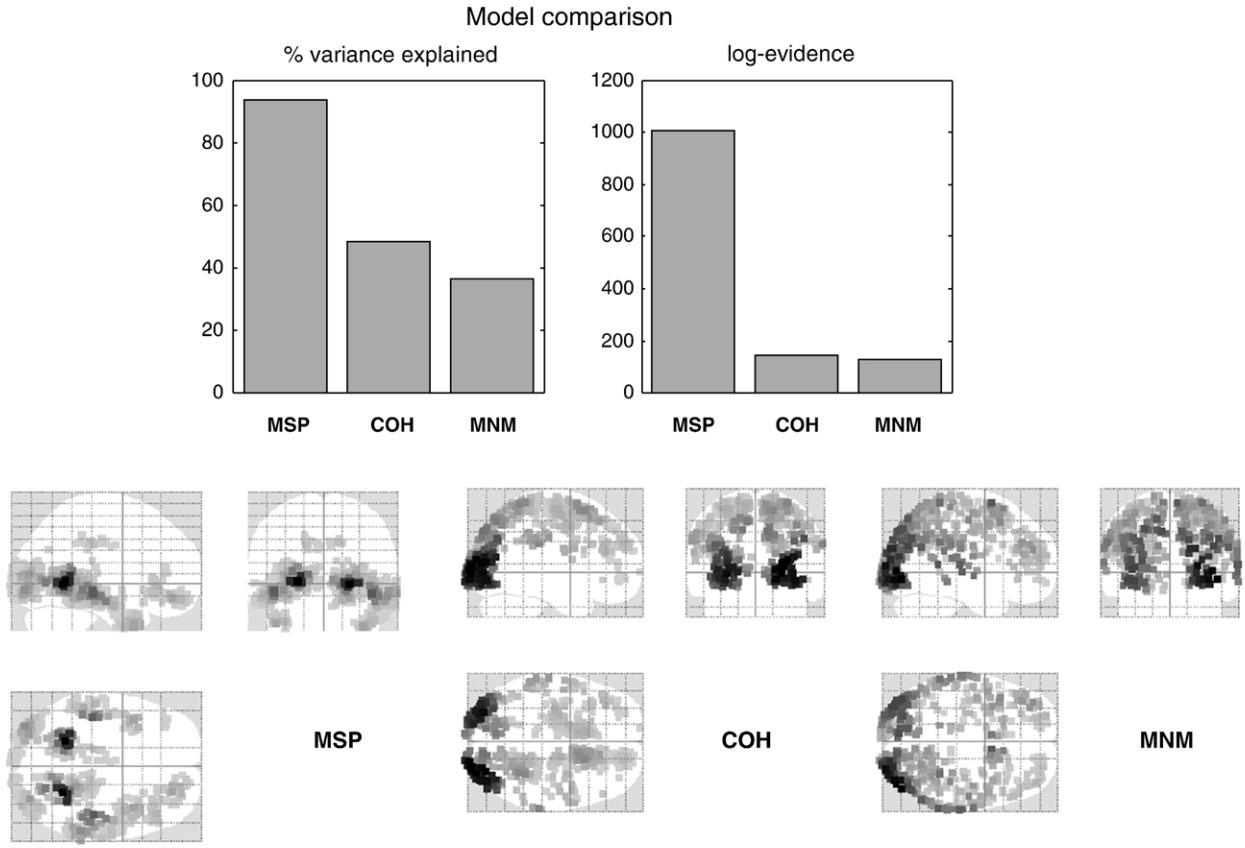

Fig. 11. Model comparison between MSP, COH and MNM models using the empirical ERPs evoked by faces. This uses the same format as the previous figure and shows that the MSP model supervenes. The spatial structure in the maximum intensity projections (lower panels) reflects the complexity of the models employed.

Fig. 10. It is evident that, both in terms of model evidence and accuracy (explained variance) the MSP model is substantially better than the other models. It is also interesting to note that the spatially coherent model is better than the classical minimum norm, although this difference is much smaller. The maximum intensity projections or glass brains show nicely the effect of increasing the constraints on the model as one goes from multiple sparse priors to a single minimum norm prior; as expected the profile of activity becomes simpler and less informative. Specifically, the reconstructed profiles become more superficial and dispersed.

Face validity

The results of the MSP inversion are largely consistent with what we would expect from an ERP elicited by face stimuli. This is illustrated in Fig. 12. In the top middle panel, we show the conditional estimate of the time course over peristimulus time for the maximally responsive dipole in the left fusiform region (at −23, −54, 0 mm). There is a pronounced N170, with reasonably tight conditional confidence intervals. The bilateral maxima at this time bin, in the glass brain, agree very well with the locations of two equivalent current dipoles fitted to the same data (left panels); however, the conventional ECD solution was obtained by averaging the ERP over 150 to 200 ms, during which time the MSP reconstruction changes. There is an equally good correspondence with the profile of activation measured in a group of eighteen subjects using fMRI (see Henson et al., 2007). In both the single-subject MSP reconstruction and the fMRI results, there are three bilateral pairs of ventral occipito-temporal responses in conjunction with a ventral prefrontal source, with right hemisphere emphasis.

Finally, we show in Fig. 13 the correspondence between observed response and conditional expectation in channel space. Although we can only show a single time slice here (the response at 170 ms); the equivalent movies of the evoked dynamics in channel space and their prediction show a pleasing similarity. More importantly, the reconstructed time courses in channel space illustrate a ubiquitous feature of empirical shrinkage priors of the sort we have used here, namely the shrinkage of conditional estimates to their prior expectation of zero. The shrinkage accounts for the fact that only about 90% of the observed variance is explained by optimum models. The remaining variance is, one hopes, largely noise.

Conclusion

This paper has described a new application of hierarchical or empirical Bayes to the distributed source reconstruction problem in EEG and MEG. The key contribution is the automatic selection of multiple cortical sources with compact spatial support that are specified in terms of empirical priors. This obviates the need to use priors with a specific form (e.g., smoothness or minimum norm) or with spatial structure (e.g., priors based on

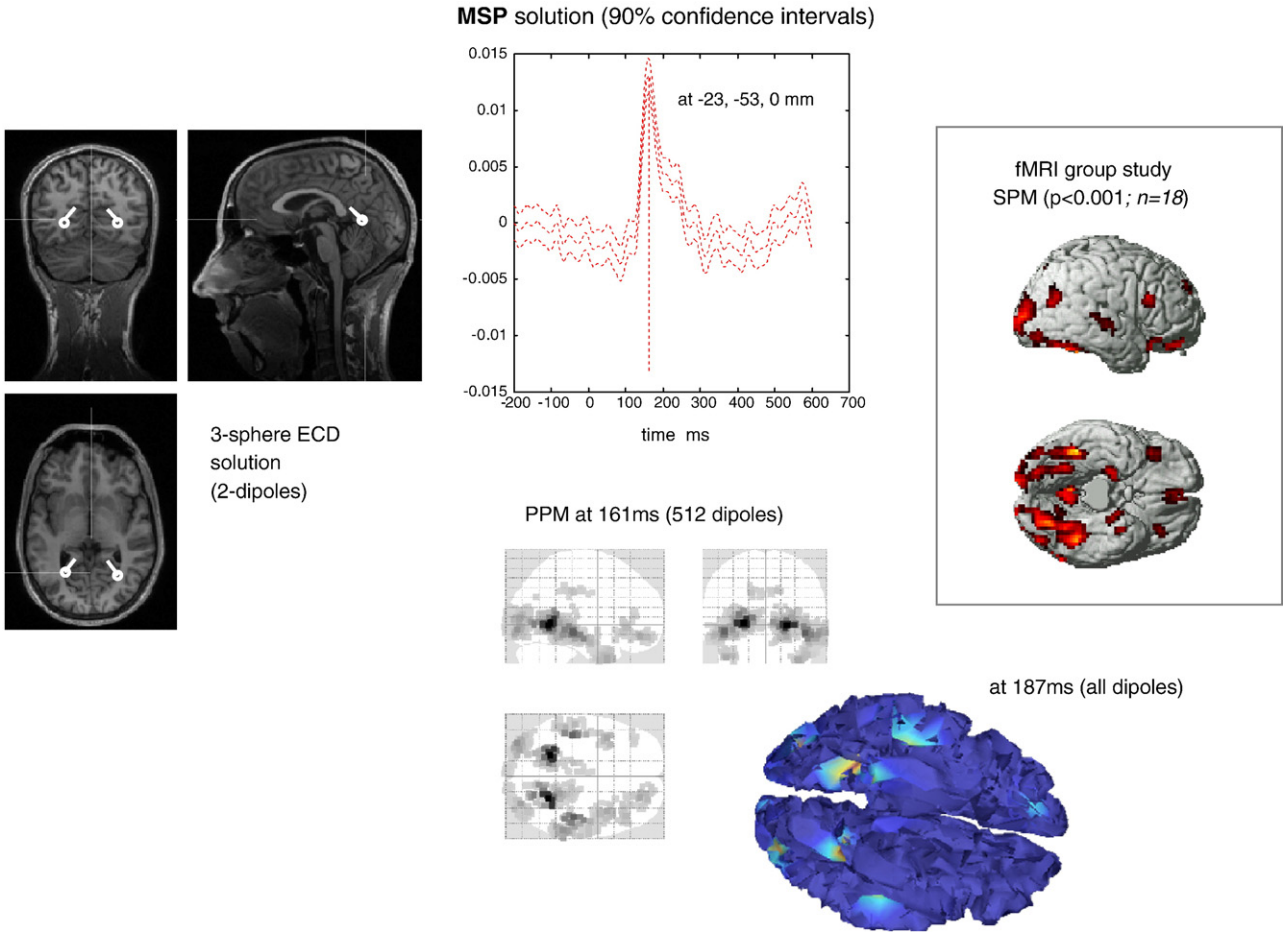

Fig. 12. Comparative results for the N170 as reconstructed using the MSP model (middle panels) and a two-ECD solution (obtained with SPM5 using exactly same data) based on the same three-sphere head model (left panels). The cortical renderings (right panels) show voxels that survived an uncorrected threshold of $p<0.001$, when testing for face-selective responses in a group of eighteen normal subjects using fMRI. The lower inset is of the conditional [unsigned] activity at 187 ms, rendered on the canonical cortical mesh used to model source space (the mesh has been made slightly transparent so that deep sources can be seen easily). Equivalent results are shown in maximum intensity projection format in three-dimensional space (lower panel). This spatial profile is expressed at the same time of the maximum [unsigned] response over all voxels and time bins, at 161 ms. These display formats are the standard SPM output.

depth constraints or functional magnetic resonance imaging results). Furthermore, the inversion scheme allows for a sparse solution, of the sort enforced by equivalent current dipole models, for distributed sources. This means that the approach automatically selects either a sparse or a distributed model, depending on the data.

There are a number of aspects of the scheme we have not demonstrated in this paper, specifically the derivation of conditional contrasts over time–frequency windows and their corresponding conditional energy (cf. induced responses). In a practical setting, we envisage that the source reconstruction described above would be used to summarize the responses of each subject to each experimental trial or condition. In our [academic freeware] implementation, these contrasts are projected from a two-dimensional cortical manifold to a full three-dimensional anatomical space. This enables the conditional estimate to be smoothed and entered into standard SPM analyses for inference at the between-subject level. We deliberately smooth in three-space to ensure that variations in gyral anatomy from subject to subject (which are not confined to a two-dimensional manifold) are accommodated by smoothing, in accord with the matched filter theorem (see Henson et al., 2007 for a fuller discussion).

Further constraints on the solutions under MSP can be enforced by specifying volumes of interest, within which to reconstruct current source density. This effectively forces solutions to occupy volumes of interest and can be a useful device when the regions involved are known in advance. These and other model selection issues will be the subject of future papers that use the techniques described in this paper.

Software note

The inversion scheme and models considered in this paper are implemented in the SPM academic software, which is available freely from http://www.fil.ion.ucl.ac.uk/spm. The MSP and other models are an integral part of the source reconstruction stream, which allows one to create conditional contrasts and their energy for any number of trials or types. The display format used by SPM adopts the same format used in Figs. 2, 12 and 13.

Acknowledgments

The Wellcome Trust, the Medical Research Council and British Council funded this work. Jérémie Mattout is funded by the

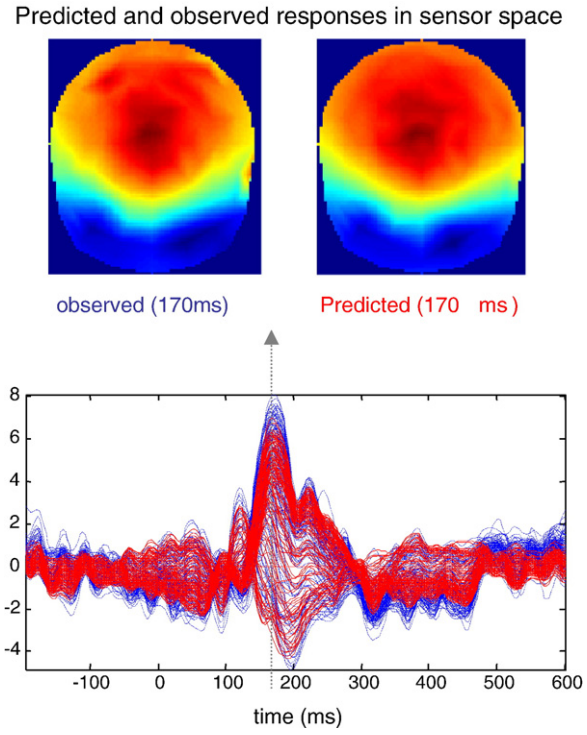

Fig. 13. Observed and predicted responses in channel space. Upper panels: interpolated pseudo-map of activity at 170 ms for true (left) and predicted (right) responses. Lower panel: response in channel space (red, predicted; blue, observed) showing how the predictions are shrunk towards their prior expectation of zero. The dotted line indicates the time bin depicted in the upper panels. This display format is the standard SPM output. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Foundation pour la Recherché Médicale (FRM). Christophe Phillips is funded by the Fonds de la Recherche Scientifique (FNRS).

Appendix A

This appendix describes an efficient way to compute the gradients and curvatures of the free energy bound on the log-evidence that is optimized in the **M**-step of the variational scheme described in the main text (see Fig. 1). For clarity, we denote differentiation using a subscript notation, such that $F_\lambda \equiv \partial_\lambda F \equiv \partial F/\partial\lambda$, where λ is a variable (as opposed to an index). An efficient formulation of the **M**-step rests on the equalities

$$F_{\lambda_i} = L_{\lambda_i} - \Pi_{ii}(\mu_i^\lambda - \eta_i)$$

$$F_{\lambda_i\lambda_j} = L_{\lambda_i\lambda_j} - \Pi_{ij}$$

$$L_\lambda = L_\gamma \gamma_\lambda$$

$$L_{\lambda\lambda} = \gamma_\lambda^T L_{\gamma\gamma} \gamma_\lambda \tag{A1}$$

where L_γ and $L_{\gamma\gamma}$ are the gradients and curvatures of $L = \ln p(\tilde{Y}|\gamma, m)$ with respect to $\gamma = [\gamma_1^{(1)}, \gamma_1^{(2)}, \ldots \gamma_2^{(1)}, \gamma_2^{(3)}, \ldots]$, which are the conditional eigenvalues of each component

$$\exp(\lambda_i) Q_i = \sum_j \gamma_i^{(j)} q_i^{(j)} q_i^{(j)T} \Rightarrow$$

$$\gamma_i^{(j)} = \exp(\lambda_i) q_i^{(j)T} Q_i q_i^{(j)}$$

$$q_i^{(j)T} q_i^{(k)} = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases} \tag{A2}$$

Here, $\gamma_i^{(j)}$ and $q_i^{(j)}$ are the j-th eigenvalue and eigenvector of $\exp(\lambda_i) Q_i$. The advantage of this formulation is that the gradients and curvatures are very simple to evaluate[11]

$$L_\gamma = \frac{v}{2} \mathrm{diag}\left(q^T \Sigma^{-1} (C - \Sigma) \Sigma^{-1} q\right)$$

$$L_{\gamma\gamma} = -\frac{v}{2} \left(q^T \Sigma^{-1} q\right) \times \left(q^T \Sigma^{-1} q\right)^T \tag{A3}$$

where $q = [q_1^{(1)}, q_1^{(2)}, \ldots q_2^{(1)}, q_2^{(3)}, \ldots]$ are also the eigenvectors of Q_i. The form of Eq. (A.3) capitalizes on the fact that we are differentiating with respect to the scale parameters of single eigenvectors, where, $tr(A q_i^{(j)} q_i^{(j)T}) = q_i^{(j)T} A q_i^{(j)}$. This means that one can avoid invoking the trace operator for each element of the gradient vector and curvature matrix. The derivatives of the eigenvalues, with respect to the hyperparameters, γ_λ exist only where they pertain to the same component.

$$\frac{\partial \gamma_i^{(j)}}{\partial \lambda_k} = \begin{cases} \exp(\lambda_k) q_i^{(j)T} Q_i q_i^{(j)} & k = i \\ 0 & k \neq i \end{cases} \tag{A4}$$

In practice, the explicit eigen-solution is seldom necessary because components are generally specified as either $Q = I \Rightarrow q = I$ or have the form, $Q_i = q_i q_i^T \Rightarrow q_i^{(1)} = q_i$, where there is only one eigenvalue and $q_i^T q_i = 1$.

References

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.

Cointepas, Y., Mangin, J.-F., Garnero, L., Poline, J.-P., Benali, H., 2001. BrainVISA: Software Platform for Visualization and Analysis of Multi-modality Brain Data. Proc. 7th HBM, Brighton, UK, p. S98.

Daunizeau, J., Friston, K.J., 2007. A mesostate-space model for EEG and MEG. NeuroImage. 2007 Jul 24; [Epub ahead of print].

Daunizeau, J., Grova, C., Marrelec, G., Mattout, J., Jbabdi, S., Pelegrini-Issac, M., Lina, J.M., Benali, H., 2007. Symmetrical event-related EEG/fMRI information fusion in a variational Bayesian framework. NeuroImage 36 (1), 69–87.

Dempster, A.P., Laird, N.M., Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc., Ser. B 39, 1–38.

Efron, B., Morris, C., 1973. Stein's estimation rule and its competitors—an empirical Bayes approach. J. Am. Stat. Assoc. 68, 117–130.

Friston, K.J., Henson, R., Phillips, C., Mattout, J., 2006. Bayesian estimation of evoked and induced responses. Hum. Brain Mapp. 27 (9), 722–735.

Friston, K.J., Mattout, J., Trujillo-Barreto, N., Ashburner, J., Penny, W., 2007. Variational free energy and the Laplace approximation. NeuroImage 34 (1), 220–234.

Friston, K.J., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, G., Penny, W., Ashburner, J., 2007. Bayesian decoding of brain images. NeuroImage Aug 24; [Epub ahead of print].

Fuchs, M., Wagner, M., Kohler, T., Wischmann, HA., 1999. Linear and nonlinear current density reconstructions. J. Clin. Neurophysiol. 16 (3), 267–295.

[11] × is the Hadamard or element-by-element product.

Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Anal. 1 (3), 515–533.

Harrison, L.M., Penny, W., Ashburner, J., Trujillo-Barreto, N., Friston, K.J., 2007. Diffusion-based spatial priors for imaging. NeuroImage (Aug 8; electronic publication ahead of print).

Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Stat. Assoc. 72, 320–338.

Henson, R.N., Goshen-Gottstein, Y., Ganel, T., Otten, L.J., Quayle, A., Rugg, M.D., 2003. Electrophysiological and haemodynamic correlates of face perception, recognition and priming. Cereb. Cortex 13 (7), 793–805.

Henson, R.N., Mattout, J., Singh, K.D., Barnes, G.R., Hillebrand, A., Friston, K.J., 2007. Population-level inferences for distributed MEG source localization under multiple constraints: application to face-evoked fields. NeuroImage 38 (3), 422–438.

Jun, S.C., George, J.S., Plis, S.M., Ranken, D.M., Schmidt, D.M., Wood, CC., 2006. Improving source detection and separation in a spatiotemporal Bayesian inference dipole analysis. Phys. Med. Biol. 51 (10), 2395–2414.

Kass, R.E., Steffey, D., 1989. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). J. Am. Stat. Assoc. 407, 717–726.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795.

Kiebel, S.J., Friston, KJ., 2004. Statistical parametric mapping for event-related potentials (II): a hierarchical temporal model. NeuroImage 22 (2), 503–520.

Kim, H.-C., Ghahramani, Z., 2006. Bayesian Gaussian process classification with the EM-EP algorithm. IEEE Trans. Pattern Anal. Mach. Intell. 28 (12), 1948–1959.

LeSage, J.P., Pace, R.K., 2000. Using Matrix Exponentials to Explore Spatial Structure in Regression Relationships. Working Paper, Department of Economics, University of Toledo.

MacKay, D.J.C., 1992. Bayesian interpolation. Neural Comput. 4 (3), 415–447.

Mattout, J., Phillips, C., Penny, W.D., Rugg, M.D., Friston, KJ., 2006. MEG source localization under multiple constraints: an extended Bayesian framework. NeuroImage 30, 753–767.

Mattout, J., Henson, R.N., Friston, K.J., 2007. Canonical Source Reconstruction for MEG. Computational Intelligence and Neuroscience Article ID 67613.

Mosher, J.C., Spencer, M.E., Leahy, R.M., Lewis, PS., 1993. Error bounds for EEG and MEG dipole source localization. Electroencephalogr. Clin. Neurophysiol. 86 (5), 303–321.

Neal, R.M., 1996. Bayesian Learning for Neural Networks. Springer-Verlag, New York.

Neal, R.M., 1998. Assessing relevance determination methods using DELVE. Neural Networks and Machine Learning. Springer, pp. 97–129.

Nagarajan, S.S., Portniaguine, O., Hwang, D., Johnson, C., Sekihara, K., 2006. Controlled support MEG imaging. NeuroImage 15 (33(3)), 878–885.

Nummenmaa, A., Auranen, T., Hamalainen, M.S., Jaaskelainen, I.P., Lampinen, J., Sams, M., Vehtari, A., 2007. Hierarchical Bayesian estimates of distributed MEG sources: theoretical aspects and comparison of variational and MCMC methods. NeuroImage 35 (2), 669–685.

Pascual-Marqui, R.D., 2002. Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. Methods Find. Exp. Clin. Pharmacol. 24 (Suppl. D), 5–12.

Patterson, H.D., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58, 545–554.

Phillips, C., Rugg, M.D., Friston, KJ., 2002a. Anatomically informed basis functions for EEG source localization: combining functional and anatomical constraints. NeuroImage 16 (3 Pt 1), 678–695.

Phillips, C., Rugg, M., Friston, K.J., 2002b. Systematic regularisation of linear inverse solutions of the EEG source localisation problem. NeuroImage 17, 287–301.

Phillips, C., Mattout, J., Rugg, M.D., Maquet, P., Friston, K.J., 2005. An empirical Bayesian solution to the source reconstruction problem in EEG. NeuroImage 24, 997–1011.

Rasmussen, C.E., 1996. Evaluation of Gaussian Processes and Other Methods for Non-Linear Regression. PhD thesis, Dept. of Computer Science, Univ. of Toronto, 1996. Available from http://www.cs.utoronto.ca~carl/.

Ripley, B.D., 1994. Flexible non-linear approaches to classification. In: Cherkassy, V., Friedman, J.H., Wechsler, H. (Eds.), From Statistics to Neural Networks. Springer, pp. 105–126.

Russell, G.S., Srinivasan, R., Tucker, DM., 1998. Bayesian estimates of error bounds for EEG source imaging. IEEE Trans. Med. Imag. 17 (6), 1084–1089.

Salin, P.-A., Bullier, J., 1995. Corticocortical connections in the visual system: structure and function. Psychol. Bull. 75, 107–154.

Sato, M.A., 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.

Serinagaoglu, Y., Brooks, D.H., MacLeod, RS., 2005. Bayesian solutions and performance analysis in bioelectric inverse problems. IEEE Trans. Biomed. Eng. 52 (6), 1009–1020.

Talairach, J., Tournoux, P., 1988. Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System—An Approach to Cerebral Imaging. Thieme Medical Publishers, New York, NY.

Tipping, M.E., 2001. Sparse Bayesian learning and the Relevance Vector Machine. J. Mach. Learn. Res. 1, 211–244.

Trujillo-Barreto, N., Aubert-Vazquez, E., Valdes-Sosa, P., 2004. Bayesian model averaging. NeuroImage 21, 1300–1319.

Wipf, D.P., Ramırez, R.R., Palmer, J.A., Makeig, S., Rao, B.D., 2006. Automatic Relevance Determination for Source Localization with MEG and EEG Data, Technical Report, University of California, San Diego.