This article was published in

Nummenmaa, A., Auranen, T., Hämäläinen, M. S., Jääskeläinen, I. P., Sams, M., Vehtari, A., and Lampinen, J. (2007). Automatic relevance determination based hierarchical Bayesian MEG inversion in practice. NeuroImage, 37 (3): 876-889.

ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 37 (2007) 876–889

Automatic relevance determination based hierarchical Bayesian MEG inversion in practice

Aapo Nummenmaa,[a,b,*] Toni Auranen,[a,b] Matti S. Hämäläinen,[c,d] Iiro P. Jääskeläinen,[a,b] Mikko Sams,[a,b,e] Aki Vehtari,[a] and Jouko Lampinen[a]

[a]*Laboratory of Computational Engineering, Helsinki University of Technology, Espoo, Finland*
[b]*Advanced Magnetic Imaging Centre, Helsinki University of Technology, Espoo, Finland*
[c]*Massachusetts General Hospital-Massachusetts Institute of Technology-Harvard Medical School, Athinoula A. Martinos Center for Biomedical Imaging, Charlestown, MA 02139, USA*
[d]*Harvard/MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[e]*Brain Research Unit, Low Temperature Laboratory, Helsinki University of Technology, Espoo, Finland*

Received 29 December 2006; revised 16 March 2007; accepted 10 April 2007
Available online 19 April 2007

In recent simulation studies, a hierarchical Variational Bayesian (VB) method, which can be seen as a generalisation of the traditional minimum-norm estimate (MNE), was introduced for reconstructing distributed MEG sources. Here, we studied how nonlinearities in the estimation process and hyperparameter selection affect the inverse solutions, the feasibility of a full Bayesian treatment of the hyperparameters, and multimodality of the true posterior, in an empirical dataset wherein a male subject was presented with pure tone and checkerboard reversal stimuli, alone and in combination. An MRI-based cortical surface model was employed. Our results show, with a comparison to the basic MNE, that the hierarchical VB approach yields robust and physiologically plausible estimates of distributed sources underlying MEG measurements, in a rather automated fashion.

Keywords: MEG; Inverse problem; Hierarchical modeling; Variational Bayes; Automatic relevance determination

Introduction

Magnetoencephalography (MEG) measures neural activity with temporal resolution of milliseconds, but the *inverse problem* of localising the source currents generating the observed extracranial magnetic fields has no unique solution (for a detailed exposition of MEG, see, *e.g.*, Hämäläinen et al., 1993). Reasonable estimates of the currents can be obtained, however, if suitable constraints on the sources are applied. Estimation methods can be divided into roughly two categories, first of which tries to explain the measurements with a small number of *equivalent current dipoles* whereas the second assumes a distribution of such dipoles throughout the brain and imposes some minimum-norm or maximal smoothness constraints on the current distribution (for a review of most common inverse methods, see, *e.g.*, Baillet et al., 2001).

The assumptions about the distributions of the currents in the *distributed source models* are naturally interpreted in a Bayesian (Bernardo and Smith, 2000) way as *a priori* probabilities implied by the model when no data are yet observed. The prior is accompanied by a *likelihood function* or *observation model* describing how different source configurations give rise to observed fields. The likelihood is operationally constructed by (1) solving the *forward problem*, which consists of assuming a conductor model for the head and numerically solving the Maxwell's equations dictating how currents in a conductor generate electromagnetic fields (see, *e.g.*, Mosher et al., 1999), and (2) specifying a distribution for the measurement noise (*e.g.*, a multivariate Gaussian). After obtaining a set of MEG data, the likelihood and the prior are combined *via* Bayes' rule to obtain the *posterior probability distribution* of the currents given the data, which can be used to make statistical inferences about the parameters of interest, in this case the source currents generated by neural activity. In general, the posterior is proportional to the product of the prior and the likelihood, and the constant of proportionality ensures that posterior probabilities sum up to unity. This important constant is termed the *evidence* or *marginal likelihood* of a model, and it equals the probability of the data when integrated (summed) over all parameter values. It can be used as a criterion for Bayesian model selection by choosing the model which has the largest marginal likelihood.

* Corresponding author. Laboratory of Computational Engineering, Helsinki University of Technology, Espoo, Finland. Fax: +358 9 451 4830.
E-mail address: Aapo.Nummenmaa@hut.fi (A. Nummenmaa).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2007.04.021

Here we examine the hierarchical generalisation of the minimum-norm estimate (MNE) (Hämäläinen and Ilmoniemi, 1984; Dale and Sereno, 1993) introduced by Sato et al. (2004). The hierarchical method assumes *a priori* individual precisions (inverse variances) for the currents, and imposes a further Gamma-distribution hyperprior for the prior precisions. This prior is essentially similar to the Automatic Relevance Determination (ARD) prior used for input selection for neural networks (Neal, 1996). It allows a small number of the currents to take large values and explain a larger proportion of the data, whilst suppressing the others by setting them close to zero. In the original approach of Sato et al. (2004), a Variational Bayesian (VB) approximation was developed for posterior inference (for a review of VB-methods, see, *e.g.*, Ghahramani and Beal, 2001). Usually, computing posterior summary quantities of interest, such as the posterior expectation, requires evaluation of multidimensional integrals which are not analytically tractable; the evidence itself is often such an intractable integral. The most common variational approach assumes some sets of variables to be independent (in this case the currents and their precisions), and maximises iteratively a *free energy* function, or equivalently minimises the Kullback–Leibler divergence (KL-divergence) from the factorised trial distribution to the true posterior. The output of the algorithm is an analytical (tractable) approximate for the true posterior distribution, and a lower bound for the evidence.

In Nummenmaa et al. (in press), we developed an alternative inference scheme for the ARD-prior model based on Markov chain Monte Carlo methods (MCMC) and compared the results to those obtained with the VB-approach. In the MCMC scheme, the posterior is represented by a large set of numerical samples obtained from a Markov chain with the posterior distribution as its stationary distribution (see, *e.g.*, Robert and Casella, 2004). In the previous work, we raised the question related to the hyperprior selection for the precision parameters (Nummenmaa et al., in press). The standard choice of a noninformative hyperprior leads to the marginal likelihood becoming unbounded, and consequently the posterior becoming improper (unnormalisable). We also briefly considered the possibility of estimating these parameters from the data and thus performing a full Bayesian estimation of the model, and demonstrated the multimodality of the true posterior.

The purpose of this article is to elucidate the practical importance of these rather theoretical considerations by using a simple empirical dataset consisting of MEG signals evoked by simple auditory, visual, and audiovisual stimuli. We demonstrate that for the ARD-model, the utility of using the marginal likelihood (or free energy) for model selection is in fact fairly limited and that the hyperparameters must be set by hand to some values which can potentially have a significant effect on the solutions. However, with fixed hyperparameters and fixed reconstruction grid size, the free energy can in principle be used to estimate the posterior mass proportions of different modes, which correspond in this case to "possible solutions to the inverse problem". We also compare the hierarchical method to basic MNE with respect to thresholding of the estimates. As the results are rather similar for both the MCMC and the VB-estimation schemes, but the latter is computationally less intensive than the former, we will adopt the variational framework in this study. To the best of our knowledge this is the first article in which real MEG data is analysed with the hierarchical method.

Materials and methods

The audiovisual dataset

We employed the same audiovisual dataset that has been also analysed by Auranen et al. (in press) using a different inverse method. The data consist of MEG fields evoked by auditory tones and visual checkerboards presented separately (A, V) or simultaneously (AV). The MEG raw data were acquired at 600 Hz sampling frequency with Neuromag Vectorview device, downsampled to 150 Hz, high-pass filtered (cutoff 1 Hz) to remove slow drifts and notch filtered to remove 50 Hz noise. The frequency of the binaural auditory tones was 800 Hz, their duration 80 ms, with 5 ms linear rise/fall. The visual stimuli were square shaped black–white checkerboards located at the centre of the visual field with equal duration to the auditory tones. The task was to passively listen the tones and fixate on the centre of the screen. The inter-stimulus interval was 4 s, and for each stimulus category we averaged ~150 trials (trials with concurrent EOG signal exceeding 150 μV were excluded).

In order to facilitate the comparison of the results with those of Auranen et al. (in press), we first used the multi-pair approximation (Plis et al., 2005; Jun et al., 2005) to obtain the full spatiotemporal noise covariance matrix. Since the present model assumes that the noise covariance does not depend on time, we then estimated the noise covariance matrix as the mean of the noise covariance matrices at different time points. The spatiotemporal noise covariance matrix was estimated from over 1500 data fragments randomly selected from the off-stimulus periods. The averaged MEG evoked fields are illustrated in Fig. 1.

The ARD-prior model

We employed a cortical constraint in constructing the space of possible sources (Dale and Sereno, 1993). White–grey matter boundary surface was segmented from the subject's structural MRI using FreeSurfer software (Dale et al., 1999; Fischl et al., 1999), and the orientations of the current dipoles were assumed to be perpendicular to this surface. As the number of vertices in the FreeSurfer surface is rather high (~150,000), a decimated set of vertices is commonly used in inverse computations. For the model description, let us define the following:

M = Number of MEG sensors
T = Number of time points in the averaged MEG evoked field time series
N = Number of vertices in the decimated cortical surface
$\mathbf{G}$ = $M \times N$-dimensional gain matrix
$\Sigma_{\mathbf{G}}$ = Fixed part of the $M \times M$-dimensional inverse noise covariance matrix
$\mathcal{M}$ = Collective notation for all implicit modeling assumptions and parameters

— — — — — — — — — — — — —

$\boldsymbol{B}(t)$ = $M \times 1$-dimensional vector of averaged MEG evoked fields at time t
$\boldsymbol{B}_{1:T}$ = The set of all $\boldsymbol{B}(t)$'s
β = A common scale parameter in the inverse noise covariance and the current prior

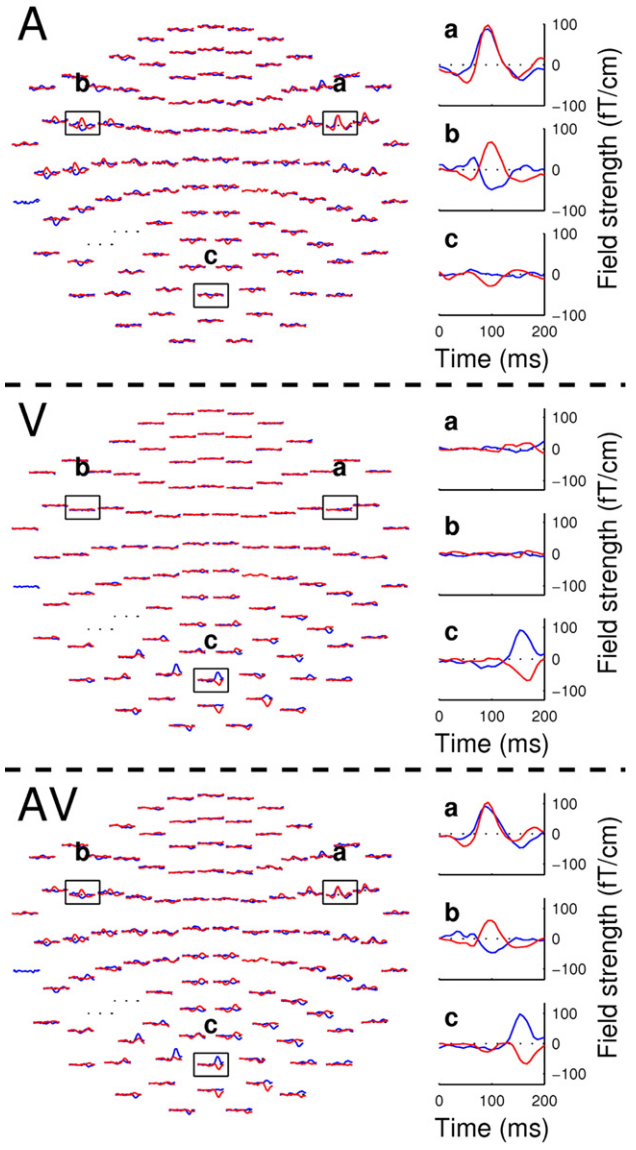

Fig. 1. For each stimulus type A, V, and AV, the timeseries of the two planar gradiometers (red and blue) are depicted on the sensor grid (viewed from the top, nose pointing up). For three sensor locations (a), (b), and (c), a closer view is also provided to facilitate comparisons between the three conditions. A sensor location with a dotted line indicates a noisy channel excluded from the analysis.

$\boldsymbol{J}(t)$	=	$N\times 1$-dimensional vector of distributed currents at time t
$\boldsymbol{J}_{1:T}$	=	The set of all $\boldsymbol{J}(t)$'s
$\boldsymbol{\alpha}$	=	$(\alpha_1,\ldots,\alpha_N)=N\times 1$-dimensional vector of prior precisions
$\boldsymbol{A}$	=	diag($\boldsymbol{\alpha}$)
α_0, γ_0	=	Mean and degrees-of-freedom parameters of the gamma-distribution prior for the α_i's

The variables above the dashed line are assumed to be fixed from the point of view of estimating the hierarchical model. The symbol $\mathcal{M}$ is introduced to remind of several more or less arbitrary modeling assumptions, such as using the cortical constraint (**G**), Gaussian noise ($\Sigma_{\mathbf{G}}$), choosing a specific time window and sampling frequency of the evoked response (T) and using only gradiometer MEG channels (M). In the following model description, all of the variables above the dashed line along with other implicit modeling assumptions are embedded into this important symbol $\mathcal{M}$.

The hierarchical model comprises of the following blocks:

1. *Observation model (likelihood)*. The statistical model gives the probability of obtaining a set of observations due to a particular realisation of noise, assuming that we know the underlying current configuration. The model stems from the linear relationship between the amplitudes of the currents and the measured fields:

$$\boldsymbol{B}(t) = \mathbf{G}\boldsymbol{J}(t) + \boldsymbol{N}(t), \tag{1}$$

where the gain matrix **G** is computed by using one-layer boundary element model (see, *e.g.*, Hämäläinen et al., 1993). The measurement noise $\boldsymbol{N}(t)$ is assumed to be independent of time and to have a Gaussian distribution with zero mean and inverse covariance $\beta\Sigma_{\mathbf{G}}$:

$$\boldsymbol{N}(t) \sim \mathrm{N}(\mathbf{0}, (\beta\Sigma_{\mathbf{G}})^{-1}). \tag{2}$$

The fixed part of the inverse noise covariance matrix, $\Sigma_{\mathbf{G}}$, is estimated from the raw MEG data during the preprocessing stage, whereas β is an unknown scale parameter to be estimated within the VB-algorithm. Since the noise is assumed to be independent of time, the likelihood of parameters at all time points is obtained by multiplying the likelihoods associated with the single time point measurements. We will denote these functions respectively by

$$P(\boldsymbol{B}_{1:T}|\boldsymbol{J}_{1:T}, \beta, \mathcal{M}) \quad \text{and} \quad P(\boldsymbol{B}(t)|\boldsymbol{J}(t), \beta, \mathcal{M}). \tag{3}$$

2. *Prior for* $\boldsymbol{J}(t)$. The hierarchical prior assumes that current amplitude at cortical location i at time t has a Gaussian distribution with zero mean and precision $\beta\alpha_i$:

$$\boldsymbol{J}(t)_i \sim \mathrm{N}(0, (\beta\alpha_i)^{-1}), \quad \text{or in vector form} \quad \boldsymbol{J}(t) \sim \mathrm{N}(\mathbf{0}, (\beta\boldsymbol{A})^{-1}). \tag{4}$$

The parameter β has been incorporated to the prior also in order to facilitate the VB-estimation. The prior precisions are assumed to be time-independent, and hence the prior for $\boldsymbol{J}_{1:T}$ is the product of the priors for $\boldsymbol{J}(t)$ at different time points; these are respectively denoted as

$$P_0(\boldsymbol{J}_{1:T}|\boldsymbol{\alpha}, \beta, \mathcal{M}) \quad \text{and} \quad P_0(\boldsymbol{J}(t)|\boldsymbol{\alpha}, \beta, \mathcal{M}). \tag{5}$$

3. *Prior for* β. The precision scale parameter β is assumed to have the "noninformative" prior

$$P_0(\beta|\mathcal{M}) = 1/\ \beta. \tag{6}$$

The improper prior does not lead to improper posterior for this parameter, an argument which is not proved here but made intuitively plausible since the posterior of β is directly influenced by the observed data (noise).

4. *Prior for* $\boldsymbol{\alpha}$**.** The ARD-prior (Neal, 1996) is imposed on the α_i's; this prior is called a hyperprior as it is a prior for the parameters of the prior:

$$\alpha_i \sim \mathrm{Gamma}(\alpha_i|\alpha_0, \gamma_0), \tag{7}$$

with the Gamma-distribution parameterised as

$$\mathrm{Gamma}(\alpha_i|\alpha_0, \gamma_0) = \frac{1}{\alpha_i}\left(\frac{\alpha_i\gamma_0}{\alpha_0}\right)^{\gamma_0}\Gamma(\gamma_0)^{-1}\exp\left(-\frac{\alpha_i\gamma_0}{\alpha_0}\right), \tag{8}$$

and $\Gamma(\cdot)$ being the Euler Gamma function.

The joint prior of the $\boldsymbol{\alpha}$ is obtained again by multiplying the independent priors for the individual α_i's. We denote this by

$$P_0(\boldsymbol{\alpha}|\alpha_0, \gamma_0, \mathcal{M}). \tag{9}$$

5. *Prior for* α_0, γ_0. The next step would be to continue the hierarchy and specify a prior for the parameters of the hyperprior. At this stage we do not specify the prior, but denote it generically as

$$P_0(\alpha_0, \gamma_0|\mathcal{M}). \tag{10}$$

Collecting the pieces of the model introduced above section, the "probability of all" (with fixed α_0, γ_0) becomes

$$\begin{aligned} &P(\boldsymbol{B}_{1:T}, \boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}|\alpha_0, \gamma_0, \mathcal{M}) \\ &\quad = P(\boldsymbol{B}_{1:T}|\boldsymbol{J}_{1:T}, \beta, \mathcal{M}) P_0(\boldsymbol{J}_{1:T}|\beta, \boldsymbol{\alpha}, \mathcal{M}) P_0(\boldsymbol{\alpha}|\alpha_0, \gamma_0, \mathcal{M}) P_0(\beta|\mathcal{M}). \end{aligned} \tag{11}$$

The joint posterior of the unknown quantities can be formally obtained as

$$\begin{aligned} &P(\boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}|\boldsymbol{B}_{1:T}, \alpha_0, \gamma_0, \mathcal{M}) \\ &\quad = \frac{P(\boldsymbol{B}_{1:T}, \boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}, |\alpha_0, \gamma_0, \mathcal{M})}{P(\boldsymbol{B}_{1:T}|\alpha_0, \gamma_0, \mathcal{M})}, \end{aligned} \tag{12}$$

which is just "probability of all" divided by the marginal probability that the data $\boldsymbol{B}_{1:T}$ comes from this model, given the values of α_0, γ_0 and the set of other assumptions $\mathcal{M}$:

$$\begin{aligned} &P(\boldsymbol{B}_{1:T}|\alpha_0, \gamma_0, \mathcal{M}) \\ &\quad = \int P(\boldsymbol{B}_{1:T}, \boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}|\alpha_0, \gamma_0, \mathcal{M}) \mathrm{d}\boldsymbol{J}_{1:T} \mathrm{d}\beta \mathrm{d}\boldsymbol{\alpha}. \end{aligned} \tag{13}$$

This term is the evidence for model $\mathcal{M}$ and it tells how probably the data come from this model. The integrations involved in computing the evidence are not tractable, and hence a Variational Bayesian method is developed in Sato et al. (2004) to perform approximate posterior inference. In the variational approach, the posterior is assumed to factorise in two parts, $Q_{J,\beta}(\boldsymbol{J}_{1:T}, \beta)$ and $Q_\alpha(\boldsymbol{\alpha})$. Then, the Q-distributions which maximise the *free energy* functional

$$\begin{aligned} \mathcal{F}(Q; \alpha_0, \gamma_0, \mathcal{M}) = &\log P(\boldsymbol{B}_{1:T}|\alpha_0, \gamma_0, \mathcal{M}) - \mathrm{KL}[Q_{J,\beta}(\boldsymbol{J}_{1:T}, \beta) \\ &\times Q_\alpha(\boldsymbol{\alpha}) || P(\boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}, |\boldsymbol{B}_{1:T}, \alpha_0, \gamma_0, \mathcal{M})] \end{aligned} \tag{14}$$

are searched, where KL(||) is the asymmetric KL-divergence from the first argument distribution to the second. In practice, The VB-algorithm operates by iteratively estimating the parameters of the approximate factor distributions until the free energy converges (to a local or global maximum). This procedure is visually described in Fig. 2, explicit equations can be found in Sato et al. (2004) and Nummenmaa et al. (in press).

The VB-method yields an analytical approximate for the posterior distribution (12), and a lower bound for the (logarithm of the) evidence (13).

The ARD-prior has the effect that it enables some of the sources to obtain a small prior precision (large variance), and hence large current amplitudes, whilst suppressing the others. In this manner, the data are explained mostly by few relevant sources, and the resulting hierarchical estimates are more focal than the rather diffuse traditional MNE-estimates. However, the hierarchical framework includes the MNE-model, which is obtained by the limit $\gamma_0 \to \infty$, when all of the prior precisions are constrained to be essentially equal (see also, Fig. 3(B)).

Some commentary on the modeling assumptions: (1) only the distribution of noise is assumed to be time-independent. The inverse method could be rather straightforwardly mapped to

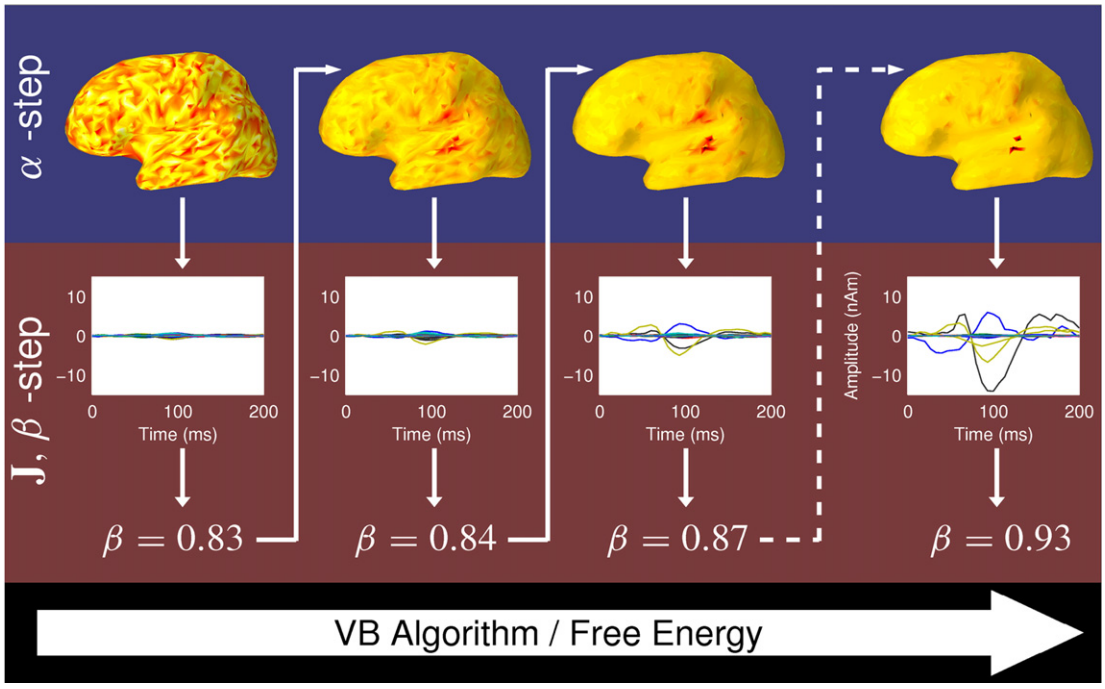

Fig. 2. After an initial guess for the parameters $\boldsymbol{\alpha}$, the algorithm proceeds by estimating the currents by an MNE with the prior precisions $\boldsymbol{\alpha}$ and then computing the (prior/noise) scale parameter β. Given the currents $\boldsymbol{J}(t)$ and the scale parameter β, the prior precisions $\boldsymbol{\alpha}$ are re-estimated and so on, until convergence. The free energy increases with every step by construction. The dashed arrow indicates several VB-steps being performed.

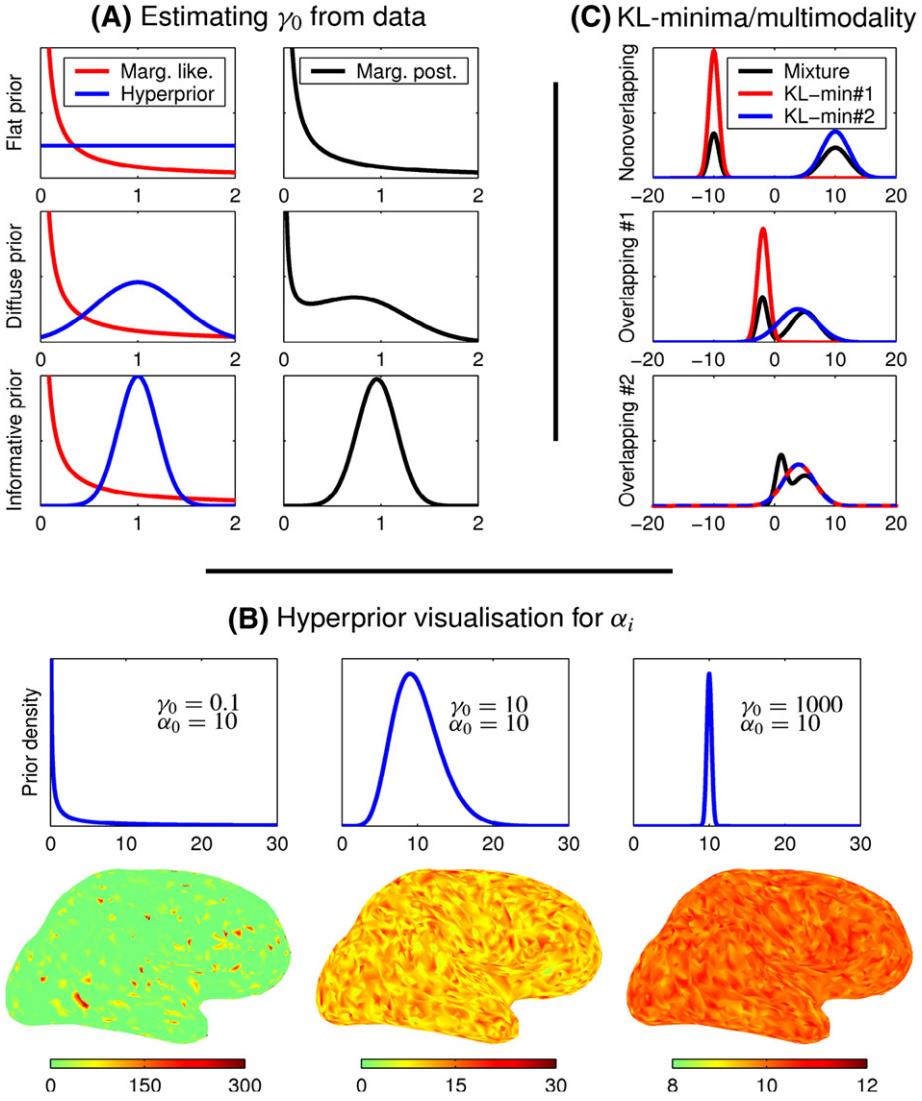

Fig. 3. (A) A schematic illustration of the marginal posterior for γ_0 with three different priors for it. As the marginal likelihood (evidence) has a singularity at γ_0=0, the prior must be rather sharp to render the posterior regular, but then the prior and posterior are essentially equal. (B) Samples from the Gamma-distribution hyperprior plotted on the cortical surface, showing that it controls how similar the currents are assumed to be throughout the brain. The distributions have been scaled for better visualisation of the shape. Note also the different colour scales in the different cortical plots. (C) Results of KL-divergence minimisation when the target is multimodal, and the approximate unimodal, which leads to two local KL-minima. When the modes of the target begin to overlap, the KL-minima also overlap, leading to errors in mass proportion estimation. When the target modes overlap significantly, there is essentially only one KL-minimum.

frequency domain, where the background activity would then form a part of the "signal", even though here we consider only (phase-locked) evoked responses. In the frequency domain, analogously to the statistical independence of consecutive time points, we would assume (as a first approximation) the uncertainty in the estimated spectrum to be independent of frequency, if locations of several frequency components should be simultaneously estimated. (2) The currents are assumed to be independent only *a priori*. This does not mean that the currents could not be correlated *a posteriori*, that is, the estimated posterior source covariance can (and will) in general be nondiagonal. (3) Even though the assumptions of *a priori* independent (implying also uncorrelated) sources, stationary noise distribution, and a data-driven characterisation of the source covariance resemble seemingly those of beamformer techniques, the hierarchical approach is a distributed source estimation method. That is, all currents (and other parameters) are estimated simultaneously, rather than resorting to some spatial filter methodology and projecting the data to each source point separately. (4) A spatial prior could be implemented (Sato et al., 2004), but it causes drastic computational costs, which were relieved in the original approach by first estimating the model with nonspatial prior, finding the current peaks and restricting the source space for the spatial model to the vicinity of these. For simulated data this process can be well justified, but with empirical data the usefulness of such approach is not so clear (see the following section). (5) The parameter β is included to the prior of the currents also because it enables estimation of the *joint* variational posterior of $\boldsymbol{J}_{1:T}$ and β. In practice, if we would estimate the inverse noise covariance $\Sigma_{\mathbf{G}}$ wrong by a factor of 1/2, the

parameter β would take value ~2 to compensate, and the prior precisions $\boldsymbol{\alpha}$ would then in turn adapt to this.

Nonstatistical thresholding

Why then go beyond the basic MNE to more complex models and estimation methods, if it brings also some challenges in interpretation of the results and increases the computational load? We will point out one virtue of the hierarchical approach, related to the nonstatistical thresholding, before embarking a more detailed analysis of the ARD-based inverse estimates.

In nonstatistical thresholding some values are set to zero before rendering the results on a (say) segmented cortical surface. The attribute "nonstatistical" is included to differentiate this from (statistical) thresholding of fMRI activity maps, for instance. Thresholding is sometimes simply motivated by practical considerations as the cortical curvature information would be impossible to display simultaneously with current value at all vertices. More often, the small current values are omitted for the sake of better visualisation of the "real activations". If the thresholding is meant to demolish only "insignificant current ripples", it would be rather natural to assume that the displayed "real activity" explains also a significant proportion of the observed data. Taking the basic MNE for example, the matter is not so clear. In the MNE-model, assuming the prior variances to be equal and fixed in all source locations results in the corresponding current values being drastically shrunken towards each other. Hence, all source locations tend to explain roughly equal proportions of the data. On the other hand, taking the hierarchical approach and letting few prior deviations to take large values, we increase the amount of data explained by these source locations, whilst setting the others close to zero yielding in a sense more "robust" estimates. Because small currents can (and usually will) give rise to large fields, when they suitably sum up, this effect pertains also with the hierarchical approach, but to a considerably smaller degree. We demonstrate this in the Results section, where we forward-computed the MNEs and the hierarchical estimates with different thresholds. The Root-Mean-Square-Error (RMSE) is used to quantify the data fit and is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{M}\big(\boldsymbol{B}(t)-\boldsymbol{B}_f(t)\big)'\big(\boldsymbol{B}(t)-\boldsymbol{B}_f(t)\big)}, \tag{15}$$

where $\boldsymbol{B}_f(t)$'s are the forward-computed (predicted) fields, and the mean is hence taken over all time points and sensors.

Even though we consider here only the basic MNE, the thresholding problem touches all distributed inverse methods in which the prior variance is assumed to be rather constant across the cortex, and also their somehow "standardised" versions such as dSPM (Dale et al., 2000) and sLORETA (Pascual-Marqui, 2002). Due to space limitations, a more complete analysis of the thresholding problem must be left for further studies. In this paper, the threshold is set somewhat arbitrarily either to display only X most relevant source locations or to include sources which are above some percentage of the largest source amplitude.

Difficulties with the marginal likelihood and model selection

Unfortunately, as is discussed in Nummenmaa et al. (in press), for this model the conditional posterior distribution becomes improper (and is independent of the value of α_0) with the choice $\gamma_0=0$:

$$P(\boldsymbol{B}_{1:T}|\alpha_0 = \text{not defined},\quad \gamma_0 = 0, \mathcal{M}) = \infty. \tag{16}$$

This is due to the fact that the case $\gamma_0=0$ corresponds to the "noninformative" hyperprior

$$P_0(\alpha_i) = \frac{1}{\alpha_i}, \tag{17}$$

which is an improper distribution, meaning that its integral over the domain of the random variable is not finite. Improper priors are often used, but in this case it leads also to improper *posterior* and hence to Eq. (16) (see Nummenmaa et al., in press; Gelman, 2006; Gelman et al., 2003, pp. 136, 390). Thus, the *type II maximum likelihood* (ML-II) (Berger, 1985) procedure cannot be applied to estimate the value of γ_0 as the evidence $P(\boldsymbol{B}_{1:T}|\alpha_0, \gamma_0, \mathcal{M})$ is apparently maximised by setting $\gamma_0=0$, leading to the improper case. Whenever using improper priors, there are always potential problems in using the evidence (Bayes factor) for model selection, even if all posteriors would be proper (for a related discussion, see, Bernardo and Smith, 2000, pp. 421–424).

In principle, we might pursue also the full Bayesian treatment by imposing a further prior $P_0(\alpha_0, \gamma_0|\mathcal{M})$. Then, the marginal posterior of α_0, γ_0 is proportional to the evidence (marginal likelihood) and the hyperprior (see also Eq. (13)):

$$P(\alpha_0, \gamma_0|\boldsymbol{B}_{1:T}, \mathcal{M}) = \frac{P(\boldsymbol{B}_{1:T}|\alpha_0, \gamma_0, \mathcal{M})P_0(\alpha_0, \gamma_0|\mathcal{M})}{P(\boldsymbol{B}_{1:T}|\mathcal{M})} \tag{18}$$

$$\boldsymbol{\alpha} P(\boldsymbol{B}_{1:T}|\alpha_0, \gamma_0, \mathcal{M})P_0(\alpha_0, \gamma_0|\mathcal{M}). \tag{19}$$

Thus, if we chose the prior to be rather flat with respect to γ_0, the posterior of these parameters would still become unbounded at $\gamma_0=0$. It follows then that we should make such a rather informative prior for γ_0, which goes sufficiently fast to zero as γ_0 goes to zero to render the posterior bounded. This is illustrated in Fig. 3(A).

We could then superficially take into account the uncertainty about these parameters by MCMC-sampling (Nummenmaa et al., in press), but in the case of a relatively flat prior, the sampler would just bang at the smallest admitted value of γ_0. With fixed γ_0, it is possible to estimate α_0 from the data, but as the solutions are more sensitive to γ_0, it is probably not so beneficial taking into account the increased computational burden (Nummenmaa et al., in press).

The ARD-model studied in this paper stems also from a well-known scale mixture representation of the Student t-distribution (Gelman et al., 2003; Geweke, 1993). Namely, with fixed γ_0 and α_0, marginalising the α_i's from the *prior* yields an independent Student t-distribution prior for the distributed current amplitude at each source location with zero mean (exists when $\gamma_0>1/2$), degrees of freedom $2\gamma_0$, and variance $\frac{\gamma_0}{\alpha_0(\gamma_0-1)}$ (exists when $\gamma_0>1$), with the t-distribution parameterised as in Gelman et al. (2003). The conditions for the existence of the prior mean and variance derive directly from the definition of the Student t-distribution, and provide an alternative view on how the prior becomes less restricting and well-behaving as $\gamma_0\to 0$.

In conclusion, we have to assign some values for α_0 and γ_0 by hand in practice. What this means will be explained in the following section. Before moving to the quest of suitable values for

the hyperparameters, we will elaborate the issue of model selection based on the free energy a bit further.

It is often stated in the VB-literature that whatever model selection and hyperparameter optimisation can be done by simply maximising the free energy which lower-bounds the log-marginal likelihood. For simple enough applications and models this probably is the case. For the model under study, the nature of the inverse problem brings in more challenges. Incidentally, whilst the evidence becomes infinite when $\gamma_0 \rightarrow 0$, the free energy remains finite (see Appendix of Nummenmaa et al., in press):

$$\lim_{\gamma_0 \rightarrow 0} \mathcal{F}(Q; \alpha_0, \gamma_0, \mathcal{M}) < \infty . \quad (20)$$

This is due to the asymmetry of the KL-divergence, and the fact that the variational posterior is always a proper distribution. By running the VB-algorithm with several different values of γ_0, it can be seen that the free energy (quite naturally) increases with decreasing γ_0, suggesting consistently to choose the noninformative case $\gamma_0 = 0$, α_0 = undefined. It is now important to emphasise that by looking only at the variational posterior (always proper) and the free energy (always finite) one would conclude that $\gamma_0 = 0$, α_0 = undefined is the correct "Bayesian choice" for the hyperparameter values. Of course, when the evidence is infinite, all lower bounds are equally good or bad as the whole construction becomes ill defined.

As another example, let us consider the problem of deciding how sparse or dense reconstruction grid to use. As can be verified from Nummenmaa et al. (in press), Appendix A, the free energy is an explicit function of the number of vertices in the source space (=dimension of the current amplitude vector), which we have called N. We could embark the free energy maximisation and choose the value of N which gives the maximum:

$$\hat{N} = \arg\max_{N \in \mathbf{N}} \mathcal{F}(Q; \alpha_0, \gamma_0, N, \mathcal{M}^*), \quad (21)$$

where $\mathbf{N}$ is the set of natural numbers and $\mathcal{M}^*$ emphasises that we use now a different set of fixed modeling assumptions. As will be seen in the Results section, the maximisation suggests to use a very sparse grid, leading to visually intolerable inverse estimates. The solution to this "paradox" contains at least three parts.

Firstly, computation of the model evidence embodies the principle of *Occam's Razor*. That is, due to the integrations over model parameters, models with more parameters tend to have smaller evidence. Adding more reconstruction points means adding more current amplitude parameters, and if the data fits are roughly equal in all cases, one should then use as small N as possible.

Secondly, because of the inverse problem, the observed MEG data do not really supply much information about how dense reconstruction grid to use. Very good data fits can be achieved with a very modest (~1000) number of source points. Only by using an extremely small value of N (~100), when the number of source points is smaller than the number of independent MEG measurements, there could be a situation in which the MEG data would not be properly explained due to the limited number of source points.

Thirdly, making N smaller affects the validity of the set of our fixed assumptions $\mathcal{M}^*$. For example, for small enough N, the cortical orientation constraint is surely not valid anymore. That implies that our forward model is not adequate, which is not taken into account in any way when computing the free energy, even though the additional discretisation error could be compensated by modifying the likelihood (see Kaipio and Somersalo, 2005, pp. 181–183).

Once again, we could impose a somewhat informative prior on N, which would reflect the fact that we must have a rather dense grid in order to use the rigid cortical constraint. In addition, we should include the effects of increased discretisation errors in the forward model with more sparse grids. As the process of storing thousands of forward matrices and respective inverse estimates is probably too heavy for practical studies, one is most likely to express the aforementioned prior information by arbitrarily choosing a reconstruction grid size which is known to produce sensible results, without any reference to marginal likelihoods or free energies.

Hyperprior elicitation

The parameters of the prior for the α_i's define how large and similar these prior precisions are assumed to be throughout the brain; this naturally induces a respective constraint on the current amplitudes themselves. As mentioned earlier, the ARD mechanism comes through the hyperprior as it provides the sources which are most "relevant" for explaining the data with small prior precisions (large standard deviations), whilst suppressing the irrelevant by setting their prior precisions to rather large values.

The parameter α_0 is the prior mean value of the α_i's, whereas the degrees of freedom parameter γ_0 describes how diffusely they are distributed around their mean α_0. The γ_0 quantifies in a sense how informative prior we are imposing on the α_i's. By setting γ_0 to a very large value (~1000), we effectively constrain all the α_i's to value α_0, which results in the MNE-solution for the $\boldsymbol{J}(t)$ (recall that the MNE-model assumes the prior current variance to be exactly the same throughout the brain). Choosing a very small value for γ_0, on the other hand, lets the α_i vary a lot, corresponding to being uninformative about their distribution. Of course, the completely uninformative and improper case $\gamma_0 = 0$ does not practically constrain the prior precisions at all. Gamma-distribution hyperpriors resulting from different choices of the hyperparameters along with α_i's sampled from these are illustrated in Fig. 3(B).

How should one choose the value for these parameters? As discussed in Nummenmaa et al. (in press), the estimates are not very sensitive for the value of α_0 as long as it is sufficiently large to keep the overall values of the currents small enough. As a smaller value of γ_0 corresponds to a more uninformative prior and larger evidence, we might consider setting γ_0 to a very small (but nonzero) value. This could be done, but our previous experience with the model shows that the convergence of the estimation algorithms (both MCMC and VB) slows down as the posterior of the α_i's becomes more diffuse. This is due to the inverse problem which causes all distributed source models to have a tendency to lean on the prior. For the same reasons, even though not explicitly demonstrated in this article, the multimodality of the posterior is also likely to increase when using a more diffuse hyperprior. In conclusion we suggest to set α_0 to some reasonably large value, such as $\alpha_0 = 10$ and γ_0 to a rather small value for which the estimation algorithm still shows robust convergence.

Multimodality of the posterior

As demonstrated by Nummenmaa et al. (in press), the true posterior distribution is multimodal, each of the modes corresponding to a more or less likely solution to the MEG source reconstruction problem. This is manifested in the VB (or MCMC)

algorithm getting trapped in different regions of the parameter space depending on the starting point. This is of practical importance because the variational posterior is always unimodal, and hence does not represent the whole uncertainty about the currents, which can lead to overinterpretation of the results.

While for this model the free energy is not so useful for hyperprior optimistion, with fixed hyperprior values it can be used to estimate how much posterior probability mass is contained in the vicinity of different posterior modes through the following formula (see Appendix A for a more detailed explanation):

$$w_k \approx \frac{\exp(\mathcal{F}(Q_k))}{\sum_{k=1}^{K} \exp(\mathcal{F}(Q_k))}. \quad (22)$$

In the above equation w_k is the probability mass proportion of the k:th of the K modes, and $\mathcal{F}(Q_k)$ is the free energy value obtained for the corresponding variational posterior Q_k. For this approximate formula to be valid, the posterior modes must be nonoverlapping, the variational posterior must resemble the true posterior locally accurately enough, and we must be able to find all modes containing a significant proportion of the posterior mass.

Because the present model is of rather high complexity, let us look at this issue through a toy example of minimising the KL-divergence from a single Gaussian to a (obviously multimodal) mixture of two Gaussians (recall that this is equivalent to maximising the free energy). As all distributions are normalised, the free energy equals just the negative of the KL-divergence. We may numerically minimise the KL-divergence by standard optimisation techniques, and starting from different parameter values we end up with a different "variational" unimodal Gaussian approximate for the true distribution. We used three degrees of overlap between the two components of the mixture. The true mass proportions were 0.35 and 0.65 in all cases. Resulting distributions and the numerical estimates for the mass proportions are shown in Fig. 3(C) and Table 1.

For this simple example, it is easy to see whether the aforementioned conditions hold or not. In real world applications, the situation is far from trivial. Also, the differences in the mass proportions tend to be huge for high dimensional distributions, as we will see in the Results section. The hyperprior selection also has an effect on the free energy, which tends to exacerbate the problem for this particular model. All in all, it is important to bear in mind the issues raised by multimodality, especially when one gets a particularly pleasing inverse estimate by the $\alpha_i = \alpha_0$ initialisation of the VB-algorithm.

Table 1
Probability mass proportions in the toy example

Case	Nonoverlapping	Overlapping #1	Overlapping #2
KL-min#1	0.3500	0.3334	0.5000
KL-min#2	0.6500	0.6666	0.5000

For the nonoverlapping case, numerically exact probability mass proportions are recovered. When the mixture components begin to overlap, the "variational" distributions overlap as well, and the estimated mass proportions are not correct anymore. If the mixture components overlap significantly, the minima are degenerate, and the posterior mass is equally split between the two.

Results

Thresholding: hierarchical estimate vs. basic MNE

Here we demonstrate the thresholding problem with the A data. We set $\gamma_0 = 10$ to obtain a genuine hierarchical estimate, and $\gamma_0 = 1000$ to yield an effective MNE. The parameter α_0 was set to 10 for both of the cases. We computed the "relevance", or the VB-estimated prior standard deviation of each source, and computed the data fit RMSE (see Eq. (15)) of the thresholded solutions with 1 −8000 "most relevant sources" included to the estimate. The results are shown in Fig. 4.

From the RMSE plot we see that, with the hierarchical method, the error decreases rapidly for the three most relevant sources, which are allowed to take large values by the ARD mechanism. After that the data fits become steadily better (RMSE decreases) by including smaller sources. For the MNE, each source point contributes in roughly equal proportion to the data fit, leading to a smooth, more linear trend in the RMSE curve. In fact, it takes 268 most relevant MNE sources to get an equal RMSE value to that obtained by the 3 most relevant sources in the hierarchical solution. Note also the very different scales of the hierarchical and the MNE solutions; in the latter the relevances differ only in the third decimal place. Furthermore, we see that, for the hierarchical method, the RMSE curve always lies below that of MNE—the difference will only become narrower when most of the sources are included to the "thresholded" estimates (which part of the RMSE graph is not plotted here).

Model selection effects

We studied the effects of grid size and hyperprior selection by assuming three conditions: (1) a very sparse grid with noninformative hyperprior, (2) a realistic grid with a somewhat informative hyperprior, and (3) a realistic grid with an extremely restrictive hyperprior. In all cases the hyperparameter $\alpha_0 = 10$, and the VB-algorithms were initialised by setting $\alpha_i = \alpha_0$ for all i. The analysis was carried out for the A data only. The results are shown in Fig. 5 and Table 2.

Whereas the data fits are rather similar for all cases (both visually and RMSE values, using the *nonthresholded* estimates), the model with sparse grid (smallest N) yields a free energy value which is an order of magnitude larger than for those with the dense, more realistic grid. Taking exponentials would then show overwhelming evidence in favour of this particular solution, which is visually not too convincing. This effect is mostly due to the automatic *Occam's Razor*, as is explained in the Materials and methods section, but the analysis does not take into account the fact that we should have a prior for N as well, excluding unrealistically sparse grids. Furthermore, with the realistic grid cases, smaller γ_0 gives a larger free energy, yielding also visually the most plausible, robust estimate of active locations without requiring careful thresholding. The case $\gamma_0 = 1000$, $N = 8000$ is in fact essentially an MNE, as mentioned before, showing again the characteristic, diffuse activation pattern associated with it. From the histogram of VB-estimated expected prior precisions, we see that it resembles the hyperprior itself in large proportions (see Fig. 3(B)), illustrating the relevancy of the hyperprior selection rather clearly.

To reveal the effects of the hyperparameter selection on the solution in more detail, we assumed five distinct values for the

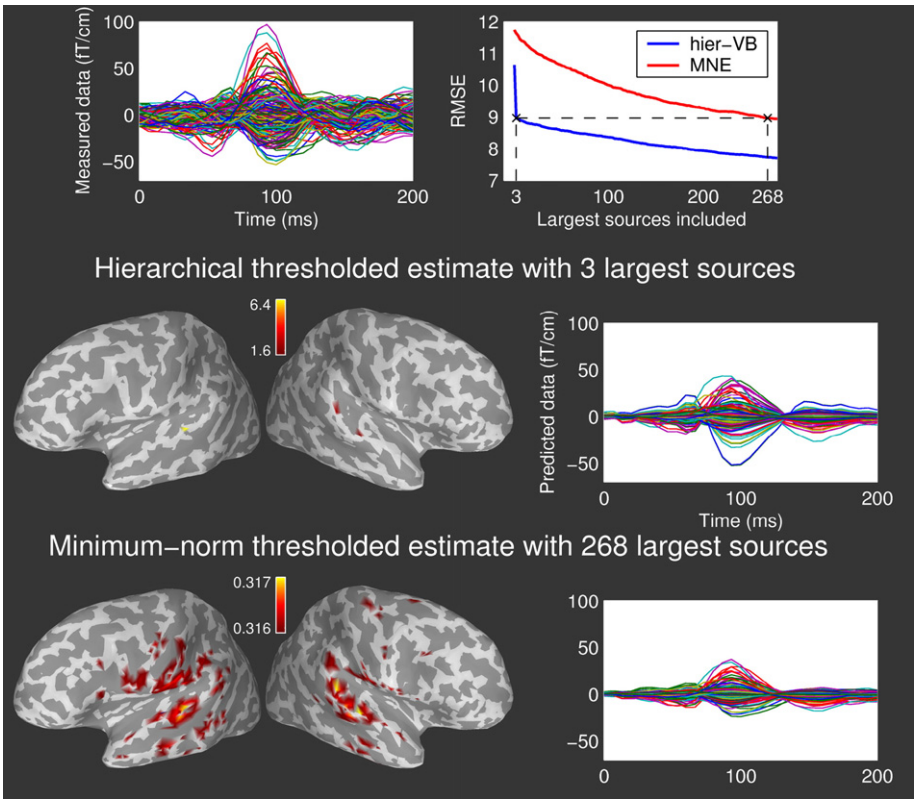

Fig. 4. The left upper plot shows the timeseries of all gradiometer measurements. The right upper plot shows how the RMSE behaves for the MNE and the hierarchical method (see text). The left black cross shows the value of the RMSE obtained by using three largest sources of the hierarchical method, and the right black cross the number of the MNE-sources needed to obtain the same value of RMSE. The middle and lowest row show the corresponding thresholded estimates and the forward-computed (predicted) measurements.

parameter γ_0: 0.1, 1, 5, 10, and 100. The grid size was 8000, and VB-algorithm was initialised by using the value of α_0, which was set to 10. For this part the threshold was set to include the sources for which the estimated prior standard deviation exceeds 5% of the maximal value. The results are shown in Fig. 6.

The results show that there is a clear "regularising" effect arising from the hyperprior selection. Lower values of γ_0 do not constrain the prior precisions significantly, and the solutions display more variability in the estimated relevances. These solutions also have the largest free energies and smallest RMSE values, as previously explained, and shown in the right lower corner of Fig. 6. The value $\gamma_0=100$ produces already a rather MNE-like solution (note again the different scale in the MNE-plot of VB-estimated prior deviation vector). For $\gamma_0=10$ the visual sources end up in being estimated large and the auditory small, leading to the latter remaining subthreshold. The value of γ_0 which produces the most plausible solution falls between the extremal ones. As the RMSE error decreases and the free energy increases with decreasing γ_0, one should perhaps look for a predictive, cross-validation type of criteria for selecting optimal γ_0 for a more quantitative analysis.

Nonlinearity of the VB-algorithm

To study the nonlinearity of the estimation method in more detail, we computed the inverse estimates for all stimulus types A, V, and AV, with the same hyperprior $\gamma_0=10$, $\alpha_0=10$. Since the stimuli are such that drastic audiovisual interaction effects are not to be expected, the MEG evoked fields add roughly linearly: $A+V\approx AV$ (see Fig. 1). In contrast to the MNE, the hierarchical inverse estimation is a nonlinear process, and this equality is not necessarily preserved amongst the corresponding source estimates. For all cases the VB-algorithm was initialised by setting $\alpha_i=\alpha_0$ for all i.

From Fig. 7 we see that the selected hyperprior is suitable for recovering plausible source locations and amplitude timecourses for the A and V stimuli. For the audiovisual case AV, the hyperprior is too restrictive as the auditory sources are estimated to have very small prior standard deviations and the threshold must be rather delicately set to recover these sources. The auditory part of the solution resembles more of a minimum-norm estimate as the hyperprior does not allow sufficiently many sources to acquire large prior standard deviations. This behaviour illustrates the nonlinearity of the hierarchical estimation procedure, in that for the solutions $A+V\neq AV$. In order to remedy the situation for the AV case, one should then relax the hyperprior and perform the VB-estimation once again.

Multimodality effects

To demonstrate the multimodality effects, we continue with the AV case and loosen the hyperprior by setting $\gamma_0=5$, $\alpha_0=10$. We performed 40 VB-runs with random starting points; the α_i's were drawn from their prior Gamma(10, 5) (see also, Fig. 3). We note that in the case of two VB-runs ending up in the same mode, their mass proportions should be summed up when comparing *different* modes (see Fig. 3(C) and Table 1), but in this particular case one

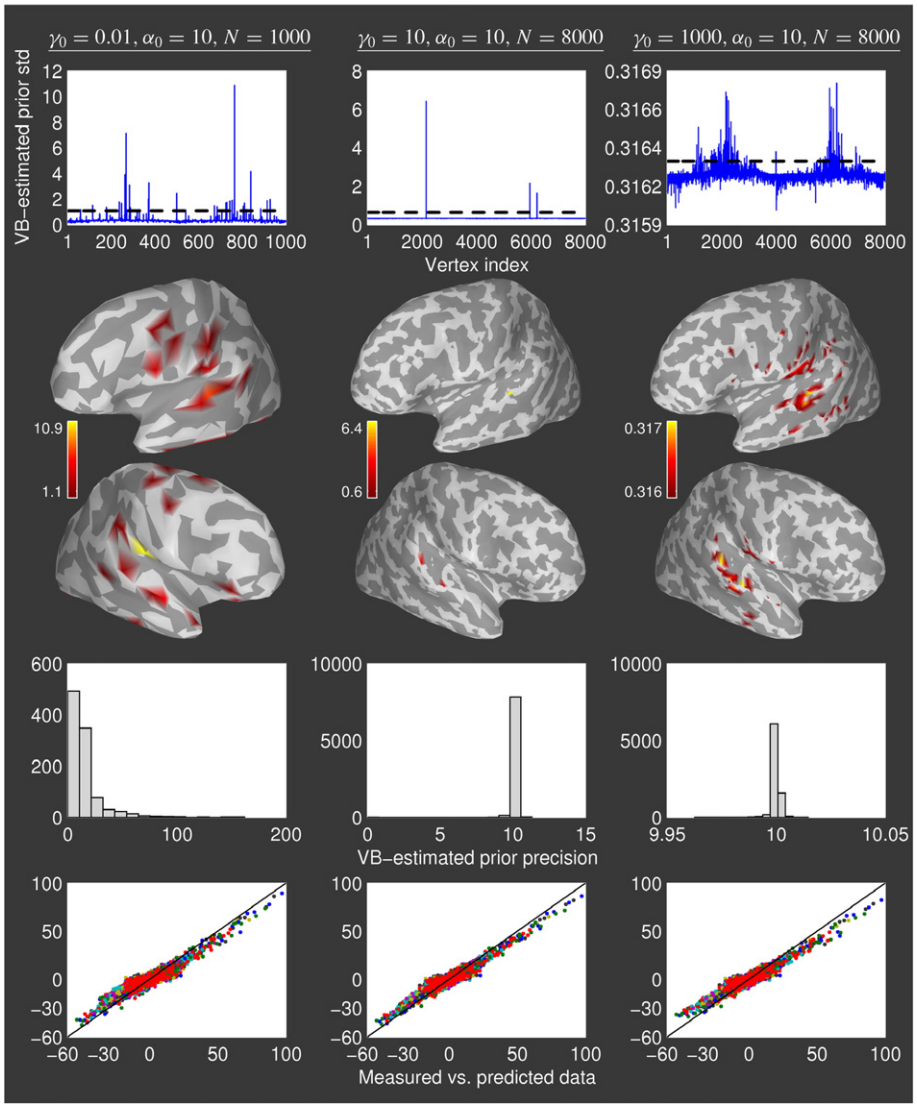

Fig. 5. The upper row shows the raw vectors of the VB-estimated expected prior standard deviations, a large standard deviation indicating high "relevance" of the corresponding cortical location. The black horizontal dashed line indicates the threshold used in the cortical plots shown in the second and third rows. The fourth row shows the corresponding VB-estimated prior precision vectors as histograms, demonstrating the resemblance to the shape of the hyperprior. The bottom row displays the predicted measurements calculated using the VB-estimated expected currents versus the actual MEG measurements as a scatterplot. The black line shows the theoretical case of a "perfect fit".

mode practically contains all the posterior probability mass, so this is neglected in the analysis. The results are shown in Fig. 8.

Again, it appears that most of the solutions yield roughly similar data fits, but one of the solutions has by far the largest posterior mass proportion associated with it (largest free energy). When comparing visually the "best data fit" and "maximal free energy" solutions, it is not actually easy to say which one is more likely to be "the true solution", based on our prior knowledge of activations elicited by the type of stimuli that were used. The reason for this is that the free energy (and the evidence) depends on γ_0; the solution which has the highest free energy fits best to the hyperprior with $\gamma_0 = 5$. This, on the other hand, is selected completely *ad hoc*, by first trying $\gamma_0 = 10$ and finding it too restrictive (see previous section). What perhaps most faithfully represents the "solution" to the inverse problem is the cluster of all modes into which the VB-algorithm converges. This is shown in the second row of Fig. 8. From the clustered modes we see, interestingly, that some solutions also include a more posterior auditory source for the left hemisphere, which is absent in both "minimal RMSE" and "maximal $\mathcal{F}$" solutions. Of course we can compute also a "Bayesian model average" of the modes in the spirit of Trujillo-Barreto et al. (2004) by weighting different modes

Table 2
The free energy and RMSE values for the three different model selection conditions

Case	$\gamma_0 = 0.01$, $N = 1000$	$\gamma_0 = 10$, $N = 8000$	$\gamma_0 = 1000$, $N = 8000$
$\mathcal{F}$	-2.3123×10^4	-2.0946×10^5	7.1469×10^5
RMSE	4.6291	3.9666	4.2734

The RMSE values are rather similar, but for the sparsest grid the free energy value is an order of magnitude larger than those with the realistic grid.

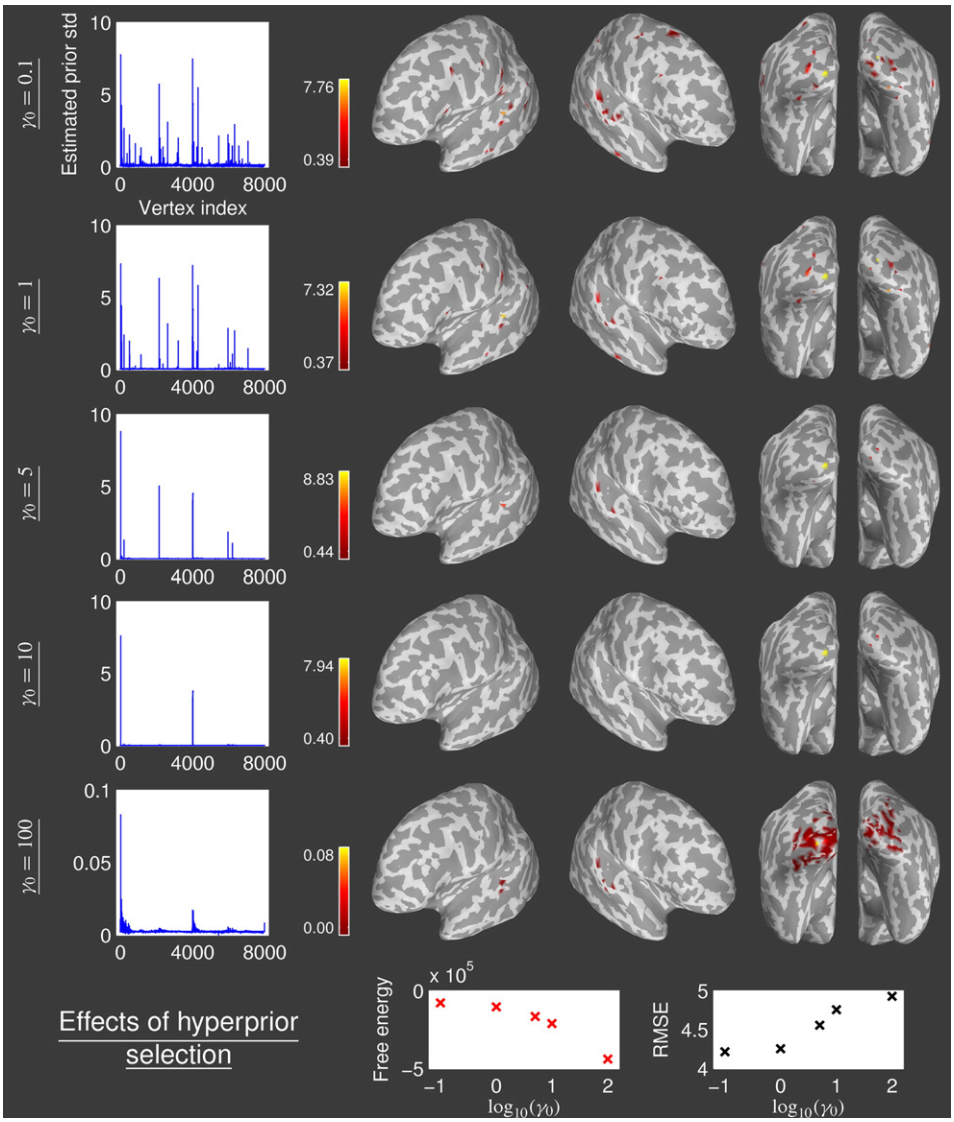

Fig. 6. The five upmost rows display the VB-estimates of the prior standard deviations both as a raw vector and a cortical plot, obtained by using different values of γ_0. The colour bar shows the range of plotted values (see text). The lowest row shows the free energies and RMSE values corresponding to different (logarithmic) values of γ_0.

by their evidence/free energy, but because one mode in this case contains nearly all of the posterior mass, it would be the only one contributing something to the average.

Summary and discussion

We have recapitulated a few theoretical modeling issues regarding the hierarchical approach introduced in Sato et al. (2004) and demonstrated how these issues influence practical analysis of empirical MEG data. First, we studied the problem of thresholding the estimates and showed that the hierarchical method "predicts" the data with a couple of sources equally well as the classical MNE with a couple of hundred sources. Second, we demonstrated that the free energy is somewhat sensitive to the selection of the reconstruction grid and the hyperparameter γ_0, and that it is not thus feasible to set up these aspects of the model in a fully Bayesian way. Third, we varied the parameter γ_0 and demonstrated the regularising effect of this hyperprior shape parameter. Fourth, we studied the nonlinearity of the estimation process by comparing source estimates for the MEG responses to unisensory auditory, unisensory visual, and audiovisual stimuli, by using the same hyperprior. Fifth, we studied the multimodality of the posterior and pointed out the possible influence of the *ad hoc* hyperprior selection on the posterior mass proportions as estimated from the free energy values.

The first two issues can be dealt with rather lightly. One could try several grid sizes and values of γ_0 and choose those which produce best results. In a study with many subjects, the same grid sizes and hyperparameter values should probably be used for all subjects to diminish the bias caused by tweaking the hyperparameter values until a hypothesis-supporting inverse estimate falls out of the analysis for each individual. Of course, different conditions may require different hyperprior settings. In practice one might set γ_0 to as low a value as possible, with the algorithm still showing robust convergence. Even though not explicitly demonstrated in this paper, the multimodality is likely to increase

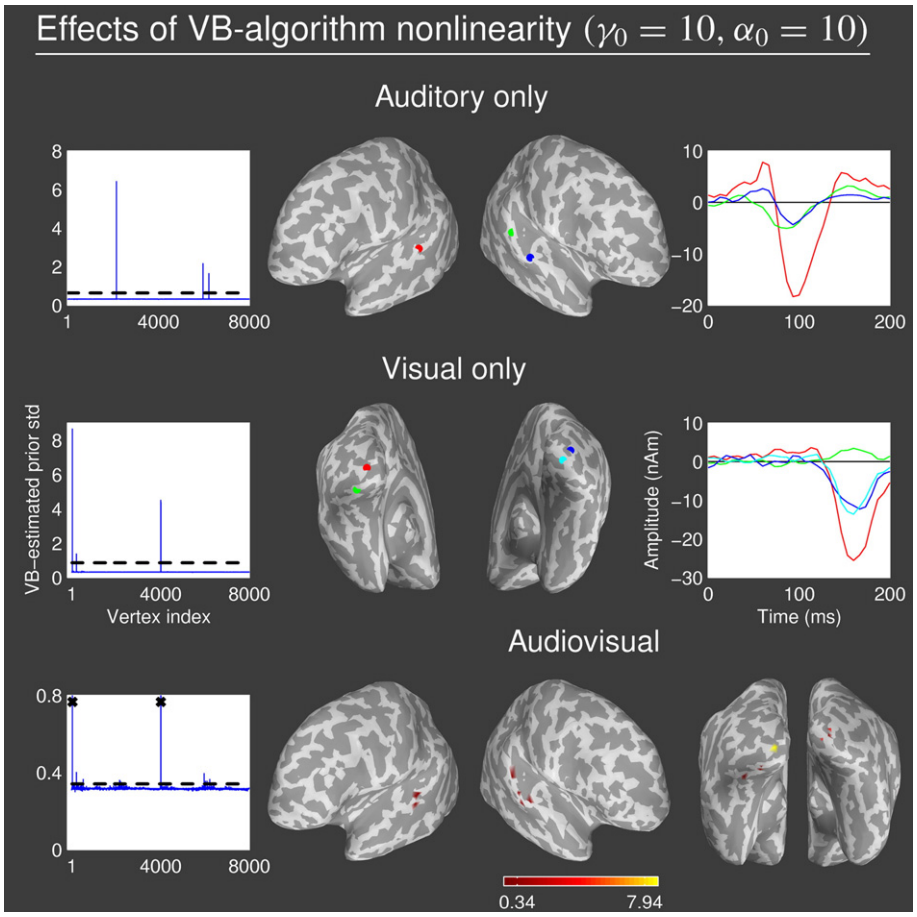

Fig. 7. The first row displays from left to right the raw VB-estimated expected prior standard deviations (with the applied threshold), the locations of the active sites on the cortical surface, and the timecourses of the sources for the A data case. The colours of the amplitude timecourses correspond to those of the cortical locations. The middle row shows the same information for the V case. The bottom row displays the VB-estimated expected prior standard deviations as a raw vector with the used threshold for the AV condition. The black crosses indicate that the *y*-axis has been truncated to make the small auditory sources visible. In the left is the same information plotted on the brain, showing again the smallness of the auditory sources, due to the restrictive hyperprior, in comparison with the largest visual source.

when the hyperparameter γ_0 is moved to lower, more uninformative direction. This is intuitively plausible as the posterior distribution of the prior precisions, which dictates how similar the currents are throughout the brain, becomes increasingly diffuse giving space for more different solution configurations.

The question of γ_0's exact role as a regularisation parameter is nontrivial, however. Changing γ_0 is not directly related for instance to the L-curve method used with linear inverse regularisation (Hansen, 1992). The L-curve method, incidentally, corresponds to letting $\gamma_0 \rightarrow \infty$, so that the algorithm operates in the linear inverse mode, and changing α_0 (which constraints the overall magnitude of the currents) to obtain a value which compromises over the ℓ^2 norm of the solution versus the data fit residual. The parameter γ_0 controls how much the current precisions (and consequently currents themselves) can vary around γ_0, that is the shape of the hyperprior. With simulated data, one might rather safely set γ_0 to a small value because there is a true solution, consistent with the forward model, arising from the few simulated sources. The real data, on the other hand, can contain sources of variability not best explained by few local sources, such as background activity and effects of signal preprocessing (filtering), necessitating a more informative hyperprior to obtain robust results. As a smaller value of γ_0 corresponds to both better data fit and larger free energy (evidence), one should perhaps look for a cross-validation type predictive measures if a more quantitative method for selecting this parameter is desired.

The issue of multimodality calls also some attention. Since the central aim of Bayesian analysis is to represent uncertainty about the quantities under investigation, we might argue that choosing one of the solutions, even the one with overwhelming evidence/free energy, is overfitting (this is so because the free energies depend on our very uncertain choice of γ_0). Instead, one should perhaps seek for several modes at first and try to cluster these in order to display all the potentially activated sites. A group analysis of some sort might then reveal some of the solutions to be more abundant in the population, based on which one might then leave out the rest of the candidate solutions from the final analysis. Also, to help dealing with multiple solutions, combining fMRI data with MEG stands out as an obvious candidate. One might think of at least three immediate ways to utilise spatial fMRI information: fMRI weighting (Sato et al., 2004), initialising the VB-algorithm according to the fMRI data, or choosing the solution which best matches the fMRI activation pattern; this is a topic currently under investigation.

Only one parameter γ_0 must be manually set (apart from the "reconstruction grid/forward model" selection); if it is possible to

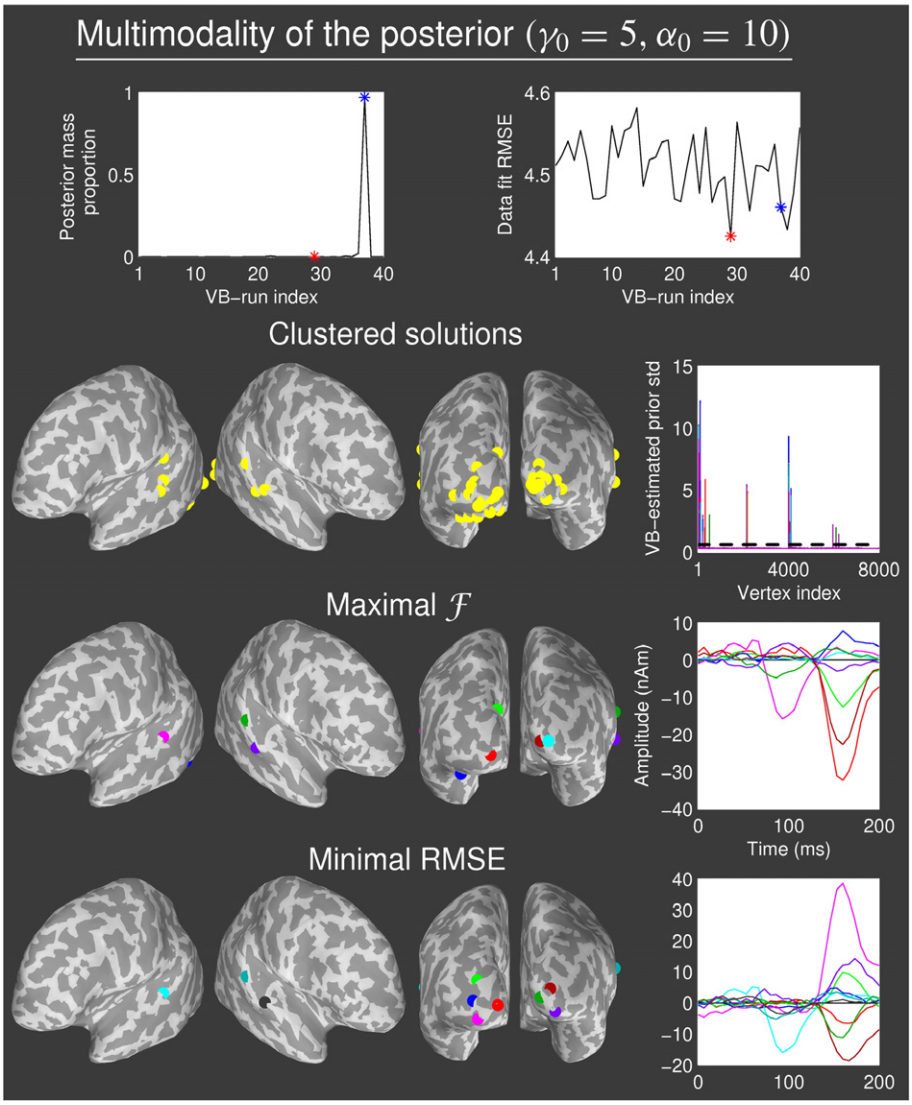

Fig. 8. The left upper corner subfigure shows the posterior mass proportions estimated from the free energy for the 40 randomly initialised VB-runs. The right upper corner displays the corresponding data fit errors. Red asterisk denotes the VB-run index with minimal data fit RMSE, blue asterisk the one with maximal free energy. The second row shows all the sites on the cortical surface which exceed the threshold, when applied to the multitude of candidate solutions, which are shown as raw vectors in the rightmost subfigure. The third row shows the active cortical locations and their timecourses with corresponding colours, for the solution with maximal free energy. The fourth row displays the same information for the solution with minimal data fit RMSE.

add some computational cost, the parameter α_0 could be estimated from the data. It means that the hierarchical method is not much more difficult to apply than the MNE—the basic MNE is recovered from the hierarchical model in the limit $\gamma_0 \to \infty$. After setting α_0 and γ_0, the whole MEG evoked field timeseries can be plugged into the model, and the estimates will come out without any manual user intervention. In conclusion, we have shown that with proper understanding of the virtues and limitations of the hierarchical approach, it offers effective and robust estimates of empirical MEG data in a rather automated fashion.

Acknowledgments

This research was supported in part by Academy of Finland (projects: 200521, 206368, Centre of Excellence: 202871, 213470), The Center for Functional Neuroimaging Technologies (NIH grant P41 RR14075), NIH grants 5R01HD040712-04 and 5R01NS044319-04, as well as Department of Energy under Award Number DE-FG02-99ER62764 to The MIND Institute.

Appendix A. Free energy and probability mass proportions

Here we briefly explain how the free energies of unimodal variational posteriors relate to probability mass proportions of a multimodal target distribution. Let us suppose that we have a normalised distribution $P(\boldsymbol{x})$, such that it is a mixture of K nonoverlapping normalised distributions $P_k(\boldsymbol{x})$, $k=1,\ldots,K$:

$$P(\boldsymbol{x}) = \sum_{k=1}^{K} w_k P_k(\boldsymbol{x}), \tag{23}$$

where the w_k's are the probability mass proportions of the mixture components satisfying

$$\sum_{k=1}^{K} w_k = 1. \quad (24)$$

The property that the distributions do not overlap is defined in this informal treatment as existence of K disjoint sets $\mathcal{I}_k$, $k=1,\ldots,K$ such that their union spans the whole domain of the random variable $\boldsymbol{x}$, and

$$P(\boldsymbol{x}) \approx w_k P_k(\boldsymbol{x}), \quad \text{when } \boldsymbol{x} \in \mathcal{I}_k. \quad (25)$$

Now suppose that we have a unimodal variational distribution $Q_k(\boldsymbol{x})$, which for practical purposes vanishes outside I_k. Then because of Eq. (25), the free energy is

$$\mathcal{F}(Q_k) = \log(Z_P) - KL[Q_k \| P] \approx \log(Z_P) - KL[Q_k \| w_k P_k] \quad (26)$$

$$= \log(Z_P) + \log(w_k) - KL[Q_k \| P_k], \quad (27)$$

where we have formally included the normalising constant of P, $Z_P=1$ to keep the notation similar to the case where the normalising constant is not known.

If the variational posterior Q_k is of sufficiently similar functional form to P_k, the KL-divergence of Q_k and P_k will get close to zero during the optimisation of the free energy, in which circumstances we get

$$\mathcal{F}(Q_k) \approx \log(Z_P) + \log(w_k) \quad \text{or} \quad (28)$$

$$w_k \approx \frac{\exp(\mathcal{F}(Q_k))}{Z_P} \alpha \exp(\mathcal{F}(Q_k)) \quad \text{or} \quad (29)$$

$$w_k \approx \frac{\exp(\mathcal{F}(Q_k))}{\sum_{k=1}^{K} \exp(\mathcal{F}(Q_k))}. \quad (30)$$

To sum up, the probability mass proportions can be computed directly from the free energy values with reasonable precision assuming three conditions:

1. Modes of the target distribution are not significantly overlapping.
2. Variational distribution resembles the target distribution locally accurately enough.
3. We can find all the modes of the target distribution containing a significant proportion of the total probability mass.

In the mixture of Gaussians toy example in the Materials and methods section, all the above conditions can be satisfied with large precision, in which case the mass proportions are numerically exactly given by the above relationship. In more complex applications, it is generally very hard to show that the three conditions hold, but if one is to trust the variational method in the first place they are in a sense already assumed to be valid.

References

Auranen, T., Nummenmaa, A., Hämäläinen, M.S., Jääskeläinen, I.P., Lampinen, J., Vehtari, A., Sams. M., in press. Bayesian inverse analysis of neuromagnetic data using cortically constrained multiple dipoles. Hum. Brain Mapp.

Baillet, S., Mosher, J.C., Leahy, R.M., 2001. Electromagnetic brain mapping. IEEE Signal Process. Mag. 14–30.

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis, second edition. Springer-Verlag New York, Inc.

Bernardo, J.M., Smith, A.F.M., 2000. Bayesian Theory. John Wiley & Sons Ltd.

Dale, A.M., Sereno, M.I., 1993. Improved localization of cortical activity by combining EEG and MEG with MRI cortical surface reconstruction: a linear approach. J. Cogn. Neurosci. 5 (2), 162–176.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis I: segmentation and surface reconstruction. NeuroImage 9, 179–194.

Dale, A.M., Liu, A.K., Fischl, B.R., Buckner, R.L., Belliveau, J.W., Lewine, J.D., Halgren, E., 2000. Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. Neuron 26, 55–67.

Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical surface-based analysis II: inflation, flattening, and a surface-based coordinate system. NeuroImage 9, 195–207.

Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Anal. 1 (3), 515–534.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian Data Analysis, second edition. Chapman and Hall/CRC.

Geweke, J., 1993. Bayesian treatment of the independent Student-t linear model. J. Appl. Econ. 8, S19–S40 (Supplement).

Ghahramani, Z., Beal, M., 2001. Graphical models and variational methods. In: Opper, M., Saad, D. (Eds.), Advanced Mean Field Methods—Theory and Practice. MIT Press.

Hansen, P.C., 1992. Analysis of discrete ill-posed problems by means of the L-curve. SIAM Rev. 34 (4), 561–580.

Hämäläinen, M.S., Ilmoniemi, R.J., 1984. Interpreting measured magnetic fields of the brain: estimates of current distributions. Technical Report TKK-F-A559, Helsinki University of Technology, Department of Technical Physics.

Hämäläinen, M.S., Hari, R., Ilmoniemi, R.J., Knuutila, J., Lounasmaa, O.V., 1993. Magnetoencephalography—Theory, instrumentation, and applications to noninvasive studies of the working human brain. Rev. Modern Phys. 65 (2).

Jun, S.C., George, J.S., Paré-Blagoev, J., Plis, S.M., Ranken, D.M., Schmidt, D.M., Wood, C.C., 2005. Spatiotemporal Bayesian inference dipole analysis for MEG neuroimaging data. NeuroImage 28 (1), 84–98.

Kaipio, J.P., Somersalo, E., 2005. Statistical and Computational Inverse Problems, volume 160 of Applied Mathematical Sciences. Springer.

Mosher, J.C., Leahy, R.M., Lewis, P.S., 1999. EEG and MEG: forward solutions for inverse methods. IEEE Trans. Biomed. Eng. 46 (3).

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

Nummenmaa, A., Auranen, T., Hämäläinen, M.S., Jääskeläinen, I.P., Lampinen, J., Sams, M., Vehtari, A., in press. Hierarchical Bayesian estimates of distributed MEG sources: theoretical aspects and comparison of variational and MCMC methods. NeuroImage, http://www.lce.hut.fi/nummenma/papers/Nummenmaa_2006_NI.pdf.

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

Plis, S.M., George, J.S., Jun, S.C., Paré-Blagoev, J., Ranken, D.M., Schmidt, D.M., Wood, C.C., 2005. Spatiotemporal noise covariance model for MEG/EEG data source analysis. Technical Report LAUR-043643, Los Alamos National Laboratory.

Robert, C.P., Casella, G., 2004. Monte Carlo Statistical Methods, second edition. Springer.

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, 806–826.

Trujillo-Barreto, N.J., Aubert-Vázquez, E., Valdés-Sosa, P.A., 2004. Bayesian model averaging in EEG/MEG imaging. NeuroImage 21, 1300–1319.