This article was published in

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

ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 35 (2007) 669–685

Hierarchical Bayesian estimates of distributed MEG sources: Theoretical aspects and comparison of variational and MCMC methods

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

[a]*Laboratory of Computational Engineering, Helsinki University of Technology, PO Box 9203, 02015, HUT, 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*

Received 25 October 2005; revised 10 February 2006; accepted 1 May 2006
Available online 12 February 2007

Magnetoencephalography (MEG) provides millisecond-scale temporal resolution for noninvasive mapping of human brain functions, but the problem of reconstructing the underlying source currents from the extracranial data has no unique solution. Several distributed source estimation methods based on different prior assumptions have been suggested for the resolution of this inverse problem. Recently, a hierarchical Bayesian generalization of the traditional minimum norm estimate (MNE) was proposed, in which the variance of distributed current at each cortical location is considered as a random variable and estimated from the data using the variational Bayesian (VB) framework. Here, we introduce an alternative scheme for performing Bayesian inference in the context of this hierarchical model by using Markov chain Monte Carlo (MCMC) strategies. In principle, the MCMC method is capable of numerically representing the true posterior distribution of the currents whereas the VB approach is inherently approximative. We point out some potential problems related to hyperprior selection in the previous work and study some possible solutions. A hyperprior sensitivity analysis is then performed, and the structure of the posterior distribution as revealed by the MCMC method is investigated. We show that the structure of the true posterior is rather complex with multiple modes corresponding to different possible solutions to the source reconstruction problem. We compare the results from the VB algorithm to those obtained from the MCMC simulation under different hyperparameter settings. The difficulties in using a unimodal variational distribution as a proxy for a truly multimodal distribution are also discussed. Simulated MEG data with realistic sensor and source geometries are used in performing the analyses.

Keywords: MEG source reconstruction; Hierarchical modeling; Variational Bayes; Markov chain Monte Carlo; Sensitivity analysis

Introduction

MEG allows monitoring of brain activity noninvasively with temporal resolution in the millisecond range (for a review, see, e.g., Hämäläinen et al., 1993). The spatial localization of the source currents generating the measured signals necessitates solving the electromagnetic inverse problem, which is known to be ill posed (von Helmholtz, 1853). However, by imposing some additional constraints on the currents, reasonable estimates for the locations of the sources can be obtained. The most common approaches assume either a small number of equivalent current dipoles (ECDs) (Mosher et al., 1992) or a continuous distribution which has some minimum norm (Hämäläinen and Ilmoniemi, 1984; Matsuura and Okabe, 1995) or maximal smoothness (Pascual-Marqui, 2002) properties.

The additional constraints are most naturally interpreted in the framework of Bayesian inference (Bernardo and Smith, 1994) as a priori probabilities reflecting the data analyst's prior beliefs and knowledge on the nature of the possible source configurations. This prior probability distribution is then combined with an observation model (a likelihood function) for obtaining the posterior probability distribution of the currents, to which statistical inferences about the sources are based on. In literature, various kinds of priors based on anatomical, physiological, and temporal information have been suggested (Dale and Sereno, 1993; Baillet and Garnero, 1997; Phillips et al., 2002). In most cases, some point estimate, such as the maximum a posteriori probability (MAP) estimate, has been taken to represent the solution to the inverse problem. For simple enough models, the entire posterior distribution of possible solutions to the inverse problem has been numerically investigated by using Monte Carlo methods (Schmidt et al., 1999; Bertrand et al., 2001a,b; Kincses et al., 2003).

Among the growing body of different approaches, the minimum norm estimate (MNE) has been further developed by utilizing magnetic resonance imaging (MRI)-based anatomical constraints on the locations and orientations of the currents (Dale and Sereno, 1993) and noise sensitivity normalization (Dale et al.,

* Corresponding author. Laboratory of Computational Engineering, Helsinki University of Technology, PO Box 9203, 02015, HUT, 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 © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2006.05.001

2000). Depth weighting has been used to compensate MNE's well-known bias towards superficial solutions (Köhler et al., 1996), whereas functional magnetic resonance imaging (fMRI) spatial information has been incorporated by fMRI-weighted and -guided versions of the MNE (Dale and Sereno, 1993; Liu et al., 1998; Ahlfors and Simpson, 2004). The virtues of MNE are its computational convenience, as an explicit inverse operator is available to compute the MAP estimate, and the generic nature of the prior assumptions from which it can be derived (for several different derivations, see Liu et al., 2002). With a Bayesian interpretation, MNE amounts to assuming that a priori the currents at each location of the discretized brain have a Gaussian distribution with zero mean and a fixed variance across the sources without any correlations among them. However, such prior makes solutions with high amplitudes in few locations and close to zero elsewhere extremely improbable. Hence, MNE produces rather diffuse solutions also for focal sources. The Gaussian prior model corresponding to MNE is a member of a more general family of ℓ^p-norm priors including also the Bayesian analogue of the ℓ^1 minimum-current estimate (MCE) (Uutela et al., 1999). The properties of the inverse estimates under the ℓ^p-norm prior were investigated in more detail by Auranen et al. (2005).

Sato et al. (2004) propose a hierarchical generalization of the Gaussian prior corresponding to the MNE. In the hierarchical approach, individual prior variances are assumed for the currents at each cortical location, and these variances are estimated from the data using automatic relevance determination (ARD) prior (Neal, 1996). The method can naturally incorporate additional prior information from both functional and anatomical MRI. The results show in general decreased localization error and increased spatial resolution in comparison with the traditional MNE.

For the hierarchical case, the estimation problem becomes nonlinear, and an analytic solution is no more available. A variational Bayesian (VB) method is developed in Sato et al. (2004) to obtain an analytical approximate for the true posterior distribution of the model parameters. The VB method assumes that the source currents and their variances are independent, which results via an iterative algorithm in a closed-form factorized distribution used as a proxy for the true posterior. Generally, it is rather difficult to assess how crude such a factorization assumption is for a given model, since correlations and other dependencies among the variables assumed independent are not readily estimated.

To cast some light on these issues, we propose an alternative strategy for performing Bayesian inference with the hierarchical model introduced by Sato et al. (2004). Here, we construct an MCMC scheme for obtaining a numerical representation of the posterior distribution and compare the results with those obtained with the VB approach using simulated MEG data. We also reveal some potential problems in the posterior analysis related to the hyperpriors selected in Sato et al. (2004) and discuss some possible solutions. Consequently, we do not compare the results of the hierarchical model to those of the MNE and its relatives, since this is already done in Sato et al. (2004). Furthermore, we do not give a single numerical quantity indicating which method is "better", as both of the methods have their own virtues and limitations. Our aim is rather to provide a complementary view on various sides of this interesting MEG inverse model.

From the viewpoint of a practical neuroscientist, this paper might be a rather technical one as various modeling aspects are explicitly addressed in detail. These aspects, such as the hyperprior selection, are important for having insight in how the model and the estimation algorithms work. Also, understanding the multimodal nature of the posterior distribution is crucial for proper interpretation of the resulting inverse estimates. We have tried to describe the central results visually, so that an amenable reader can appreciate these even if he or she is not interested in the mathematical subtleties per se.

Methods

Bayesian methods and inference

The starting point of Bayesian data analysis (Gelman et al., 2003) is to consider both the data $\boldsymbol{D}$ and the model parameters $\boldsymbol{\theta}$ as random variables and set up a probability model $P(\boldsymbol{D}, \boldsymbol{\theta}|M)$ based on knowledge about the hypothesized mechanism which generates the data. The joint probability of parameters and data explicitly expresses the fact that it is always conditioned on some set of assumptions M made by the data analyst.

The statistical inference is based on the posterior probability of the model parameters given the data obtained by using Bayes' formula:

$$P(\boldsymbol{\theta}|\boldsymbol{D}, M) = \frac{P(\boldsymbol{D},\boldsymbol{\theta}|M)}{P(\boldsymbol{D}|M)} = \frac{P(\boldsymbol{D}|\boldsymbol{\theta}, M)P_0(\boldsymbol{\theta}|M)}{P(\boldsymbol{D}|M)}. \quad (1)$$

The term $P(\boldsymbol{D}|\boldsymbol{\theta},M)$ is called likelihood as it gives the probability of the data given the parameter values (and modeling assumptions), whereas $P_0(\boldsymbol{\theta}|M)$ is called prior since it reflects the probabilities of different parameter values when no data has arrived. The normalizing constant in the 4 denominator is termed the evidence for the model M and takes care that the posterior probabilities over all parameter values sum to unity:

$$P(\boldsymbol{D}|M) = \int P(\boldsymbol{D}|\boldsymbol{\theta}, M)P_0(\boldsymbol{\theta}|M)\mathrm{d}\boldsymbol{\theta}. \quad (2)$$

Typically the dimension of the parameter space is large, and the posterior distribution is difficult to handle, as one would like to compute some summary quantities such as posterior expectation value and standard deviation of the parameter $\boldsymbol{\theta}$ or compute marginal distributions of some parameters of interest. This requires the evaluation of multidimensional integrals which are not usually analytically tractable. One possible solution to this problem is to construct a Markov chain which has the posterior distribution as its stationary distribution. The chain can then be utilized in generating a large set of numerical samples from the joint posterior distribution, and thus implicitly performing the required integrations (see, e.g., Gilks et al., 1996; Robert and Casella, 2004). With faster computers equipped with large amounts of memory and software development, this approach has gained considerable popularity in the Bayesian statistics community (see, e.g., the BUGS project, http://www.mrcbsu.cam.ac.uk/bugs/). The sampling approach has naturally its own drawbacks, such as difficulties in establishing the convergence of the chain and slow mixing of the sampler caused by correlated variables, which decreases the number of independent samples.

The normalizing constant of the posterior distribution can usually be ignored when performing posterior inferences with fixed data $\boldsymbol{D}$ and modeling assumptions M. Since the process of constructing a statistical model itself is subject to some ad hoc assumptions and uncertainties, we might as well have two or more candidate models M_1, M_2,…, M_m. In order to perform model

averaging or model comparison *via* the Bayes factor, we would have to compute the evidence or marginal likelihood $P(\boldsymbol{D}|M_i)$ for all models (see, Eq. (2); for a recent application of Bayesian model averaging in MEG/EEG imaging, see, Trujillo-Barreto et al. (2004)). The evidence is not usually tractable to compute analytically and difficult to compute numerically (for some possible Monte Carlo strategies, see Gelman and Meng (1998); Neal (2001)).

The variational Bayesian method starts from a different viewpoint (for a review on probabilistic graphical models and VB methods, see, e.g., Ghahramani and Beal, 2001). The evaluation of (the logarithm of) the marginal likelihood is formulated as maximization of a free energy functional over the space of probability distributions. In order to make the maximization procedure feasible, the distribution is usually assumed to factorize over some subsets of variables. As a result of the maximization of the free energy, one obtains a lower bound for the marginal likelihood, and an analytical approximate for the true posterior. The maximization of the free energy is equivalent of minimizing the asymmetric Kullback–Leibler divergence (KL divergence) between the true and the variational posterior. In a sense, the VB algorithm hence finds a tractable distribution which is as close as possible to the original (for discussion on the concept of closeness, see, Appendix A).

Our aim is to compare the results obtained by the two techniques, and we thus employ both MCMC sampling and the VB approach for the posterior inference. The VB method is inherently approximative, whereas the MCMC sampling can in principle give a numerical representation of the true posterior distribution and thereby obtain a more thorough picture of its structure. Because MCMC inverse solutions are based on a finite number of samples, they are in this respect also only approximate numerical estimates. However, no independency assumptions for the currents and their variances have to be made, and our results show that (for reasonable hyperparameter values) the sampler converges quickly and has rather good mixing properties. For the sake of completeness, we review the variational Bayesian estimation process in Appendix A. For the posterior simulation we utilize Gibbs sampling as almost all of the conditional distributions are of standard form. To sample from the conditional distributions of nonstandard form, we use the method of slice sampling (Neal, 2003); details of the sampling scheme are presented in Appendix B.

The forward model

The locations of possible sources were assumed to be cortically constrained (Dale and Sereno, 1993) to the boundary of grey and white matter segmented from a subjects structural MRI using FreeSurfer software (Dale et al., 1999; Fischl et al., 1999). The orientations of the current dipoles at the vertices of the segmented surface were assumed to be perpendicular to the cortical mantle. This resulted in a linear model for the MEG measurements at a single timepoint (Hämäläinen et al., 1993):

$$\boldsymbol{B}(t) = \boldsymbol{G}\boldsymbol{J}(t) + \boldsymbol{N}(t). \tag{3}$$

where $\boldsymbol{G}$ is the gain matrix (each column comprises of the magnetic signals produced by unit dipole placed at a given location), $\boldsymbol{J}(t)$ is the vector of dipole amplitudes, and $\boldsymbol{N}(t)$ is the measurement noise vector. Neuromag Vectorview (Neuromag Ltd., Finland) sensor geometry and a single compartment boundary element model was used in the forward model computations. Assuming that the noise distribution is a multivariate Gaussian and independent of time, the observation model (3) leads to a likelihood function

$$P(\boldsymbol{B}_{1:T}|\boldsymbol{J}_{1:T},\, \beta\Sigma_{\boldsymbol{G}}) = \left(\frac{1}{2\pi}\right)^{MT/2} |\beta\Sigma_{\boldsymbol{G}}|^{T/2} \exp\left(-\frac{1}{2}\sum_{t=1}^{T} (\boldsymbol{B}(t) - \boldsymbol{G}\boldsymbol{J}(t))'(\beta\boldsymbol{R}_{\boldsymbol{G}})(\boldsymbol{B}(t) - \boldsymbol{G}\boldsymbol{J}(t))\right), \tag{4}$$

where $(\beta\Sigma_{\boldsymbol{G}})^{-1}$ is the noise covariance matrix. That is, we assume that the inverse noise covariance is known up to a scale factor β. N is the number of possible source locations, M is the number of MEG channels, and T is the number of timepoints. $\boldsymbol{B}_{1:T} = \{\boldsymbol{B}(t)|t = 1{:}T\}$ and $\boldsymbol{J}_{1:T} = \{\boldsymbol{J}(t)|t = 1{:}T\}$ represent all of the observed MEG data and the modeled currents, respectively.

The minimum norm estimate

Let us assume a Gaussian prior with zero mean and precision (inverse variance) $\alpha = 1/\sigma^2$ for dipole amplitudes at all $\boldsymbol{N}$ locations of the discretized grey–white matter boundary:

$$P_0(\boldsymbol{J}(t)|\boldsymbol{\alpha}) = \left(\frac{\boldsymbol{\alpha}}{2\pi}\right)^{N/2} \exp\left(-\frac{\boldsymbol{\alpha}}{2}\boldsymbol{J}(t)'\boldsymbol{J}(t)\right),$$
$$P_0(\boldsymbol{J}_{1:T}|\boldsymbol{\alpha}) = \prod_{t=1}^{T} P_0(\boldsymbol{J}(t)|\boldsymbol{\alpha}). \tag{5}$$

Combining likelihood (4) and prior (5) to a posterior via Bayes' rule, and maximizing with respect to $\boldsymbol{J}(t)$ results in the minimum norm (MAP) estimate

$$\hat{\boldsymbol{J}}_{\mathrm{MNE}}(t) = \boldsymbol{L}(\boldsymbol{\alpha},\beta)\boldsymbol{B}(t), \tag{6}$$

where the inverse operator $\boldsymbol{L}(\alpha, \beta)$ is

$$\boldsymbol{L}(\boldsymbol{\alpha},\beta) = (\boldsymbol{G}'(\beta\boldsymbol{\Sigma}_{\boldsymbol{G}})\boldsymbol{G} + \boldsymbol{\alpha}\boldsymbol{I})^{-1}\boldsymbol{G}'(\beta\boldsymbol{\Sigma}_{\boldsymbol{G}}) = \frac{1}{\boldsymbol{\alpha}}\boldsymbol{G}'\left(\frac{1}{\boldsymbol{\alpha}}\boldsymbol{G}\boldsymbol{G}' + \frac{1}{\beta}\sum\nolimits_{\boldsymbol{G}}^{-1}\right)^{-1}, \tag{7}$$

assuming that all necessary matrix inverses exist (see, e.g., appendix of Liu et al., 2002). It is immediately seen that this depends only on the ratio α/β, which is in the literature often referred to as the regularization parameter.

Hierarchical approach: the model

In the hierarchical generalization of Sato et al. (2004), it is assumed that each of the currents has a Gaussian prior distribution with an individual precision parameter α_i:

$$P_0(\boldsymbol{J}(t)|\boldsymbol{\alpha},\beta) = \left(\frac{1}{2\pi}\right)^{N/2} |\beta\boldsymbol{A}|^{1/2} \exp\left(-\frac{\beta}{2}\boldsymbol{J}(t)'\boldsymbol{A}\boldsymbol{J}(t)\right), \tag{8}$$

$$P_0(\boldsymbol{J}_{1:T}|\boldsymbol{\alpha},\beta) = \prod_{t=1}^{T} P_0(\boldsymbol{J}(t)|\boldsymbol{\alpha},\beta), \tag{9}$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_N)$ is a vector comprising of the precision parameters, and $\boldsymbol{A} = \mathrm{diag}(\boldsymbol{\alpha})$ is the corresponding diagonal matrix.

Following Sato et al. (2004), we have included the parameter β also to the prior precision matrix to facilitate the VB-estimation. We also assume the noninformative prior on β

$$P_0(\beta) = 1/\beta, \tag{10}$$

and impose the ARD-prior (Neal, 1996) on the precision parameters α_i,

$$P_0(\boldsymbol{\alpha}_i|\boldsymbol{\alpha}_{0i},\gamma_{0i}) = \mathrm{Gamma}(\boldsymbol{\alpha}_i|\boldsymbol{\alpha}_{0i},\gamma_{0i}), \tag{11}$$

$$\mathrm{Gamma}(\boldsymbol{\alpha}_i|\boldsymbol{\alpha}_{0i},\gamma_{0i}) = \frac{1}{\boldsymbol{\alpha}_i}\left(\frac{\boldsymbol{\alpha}_i\gamma_{0i}}{\boldsymbol{\alpha}_{0i}}\right)^{\gamma_{0i}}\Gamma(\gamma_{0i})^{-1}\exp\left(-\frac{\boldsymbol{\alpha}_i\gamma_{0i}}{\boldsymbol{\alpha}_{0i}}\right), \tag{12}$$

$$P_0(\boldsymbol{\alpha}|\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0) = \prod_{i=1}^{N} P(\boldsymbol{\alpha}_i|\boldsymbol{\alpha}_{0i},\gamma_{0i}), \tag{13}$$

where $\Gamma(\cdot)$ is the Euler Gamma function and $\boldsymbol{\alpha}_0 = (\alpha_{01},\ldots,\alpha_{0N})$, $\boldsymbol{\gamma}_0 = (\gamma_{01} = \gamma_{0N})$. In order to complete the model specification, some values for the hyperprior parameters α_{0i}, γ_{0i} must be assumed. In Sato et al. (2004), this is solved by imposing a noninformative prior by setting $\gamma_{0i} = 0$ for all i. This leads to the prior

$$P_0(\boldsymbol{\alpha}_i) = \frac{1}{\boldsymbol{\alpha}_i}, \tag{14}$$

as is immediately seen from the Eq. (12). This is an improper (unnormalizable) probability distribution meaning that its integral over the domain of the random variable is not finite. Improper priors are often used, but in this case, the improper prior also leads to an improper posterior distribution. This is discussed in Gelman (2005) and Gelman et al. (2003, pp. 136, 390), where it is stated that imposing a standard noninformative distribution for the prior deviation parameters σ_i,

$$P_0(\log\sigma_i) \propto 1, \text{ or } P_0(\sigma_i) = \frac{1}{\sigma_i},\ \boldsymbol{\alpha}_i = \frac{1}{\sigma_i^2}, \tag{15}$$

produces an improper posterior. By the rule for transformation of a random variable, it is readily computed that the noninformative prior (14) on the precision parameters α_i corresponds to prior

$$P_0(\sigma_i) = \left|\frac{\mathrm{d}\boldsymbol{\alpha}_i}{\mathrm{d}\sigma_i}\right| P_0(\boldsymbol{\alpha}_i(\sigma_i)) = \left|\frac{-2}{\sigma_i^3}\right|\sigma_i^2 = \frac{2}{\sigma_i} \tag{16}$$

for σ_i, which is essentially the same as (15). Since a detailed proof of the improperness of the posterior distribution under these circumstances is omitted in Gelman et al. (2003), we demonstrate this in Appendix C.

To avoid this unpleasant situation, one possibility is to assume some ad hoc nonzero values for these hyperparameters or impose a further prior for α_{0i}, γ_{0i} and try to estimate their posterior distribution from the data as well. We perform a sensitivity analysis on the model by comparing the estimates obtained with different hyperparameter settings. The possibility of introducing these parameters as truly random variables is also studied. More specifically, we consider the case $\alpha_{0i} = \alpha_0$, γ_{0i} = fixed $\neq 0$ for all i, and assuming a uniform prior for α_0:

$$P_0(\boldsymbol{\alpha}_0) \propto 1. \tag{17}$$

In summary, in the most general case the joint probability of data, parameters and hyperparameters is

$$\begin{aligned} &P(\boldsymbol{B}_{1:T},\boldsymbol{J}_{1:T},\beta,\boldsymbol{\alpha},\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0) \\ &= P(\boldsymbol{B}_{1:T}|\boldsymbol{J}_{1:T},\beta)P_0(\boldsymbol{J}_{1:T}|\beta,\boldsymbol{\alpha})P_0(\boldsymbol{\alpha}|\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0)P_0(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0)P_0(\beta), \end{aligned} \tag{18}$$

and the corresponding posterior distribution is

$$\begin{aligned} &P(\boldsymbol{J}_{1:T},\beta,\boldsymbol{\alpha},\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0|\boldsymbol{B}_{1:T}) \\ &= P(\boldsymbol{B}_{1:T},\boldsymbol{J}_{1:T},\beta,\boldsymbol{\alpha},\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0)/P(\boldsymbol{B}_{1:T}) \end{aligned} \tag{19}$$

$$\propto P(\boldsymbol{B}_{1:T}|\boldsymbol{J}_{1:T},\beta)P_0(\boldsymbol{J}_{1:T}|\beta,\boldsymbol{\alpha})P_0(\boldsymbol{\alpha}|\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0)P_0(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0)P_0(\beta). \tag{20}$$

Often it is convenient to use the negative natural logarithm of the posterior probability in numerical computations, which is termed the *posterior energy*:

$$\begin{aligned} &E(\boldsymbol{J}_{1:T},\beta,\boldsymbol{\alpha},\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0|\boldsymbol{B}_{1:T}) \\ &= -\ln P(\boldsymbol{J}_{1:T},\beta,\boldsymbol{\alpha},\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0|\boldsymbol{B}_{1:T}) \\ &= -\ln P(\boldsymbol{B}_{1:T},\boldsymbol{J}_{1:T},\beta,\boldsymbol{\alpha},\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0) + \ln P(\boldsymbol{B}_{1:T}) \\ &= -\ln P(\boldsymbol{B}_{1:T},\boldsymbol{J}_{1:T},\beta,\boldsymbol{\alpha},\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0) + \text{constant}. \end{aligned} \tag{21}$$

Sato et al. (2004) introduce also a spatial smoothness prior on the current. Because of the increased computational burden, the model with the spatial prior is applied only after estimating the model with no spatial prior as described above and localizing the areas containing large current amplitudes based on this. A finer discretization grid is used in the estimation of the spatial model, and the areas showing no large current amplitudes in the nonspatial model are assumed to contain no sources. Our analysis is restricted to the nonspatial model at this stage, since this is the cornerstone of the hierarchical approach.

The simulated data

We used simulated data to study the structure of the hierarchical model. Two typical cortical patches were generated to the triangulated cortical surface by selecting a center point and adding its nearest and second nearest neighbors consecutively to obtain sources with some spatial extent. Irrespective of the patch area, a total amount of 80 nAm of source current was assigned to each cortical patch. The time courses of the sources were assumed to be identical (Fig. 1).

The number of timepoints was 51, and the simulated measurements were computed using Eq. (3). To avoid the most obvious kind of an inverse crime, a much denser realization of the gain matrix $\boldsymbol{G}$ (~306 × 90,000) was used in the data generation than in the inverse estimation. Inverse crime is a collective term for all those elements which are fixed in the data generation model and later assumed to be exactly known in the inverse model. The most important and common of these is using the same discretization of the model for both generating the simulated data and performing the inverse estimation, which always produces overoptimistic results (see, e.g., Kaipio and Somersalo, 2005).

Finally, Gaussian noise was added to the simulated fields so that the signal-to-noise ratio (SNR) defined by

$$\mathrm{SNR} = \frac{\boldsymbol{B}_s'\boldsymbol{B}_s}{N_s\sigma_s^2}, \tag{22}$$

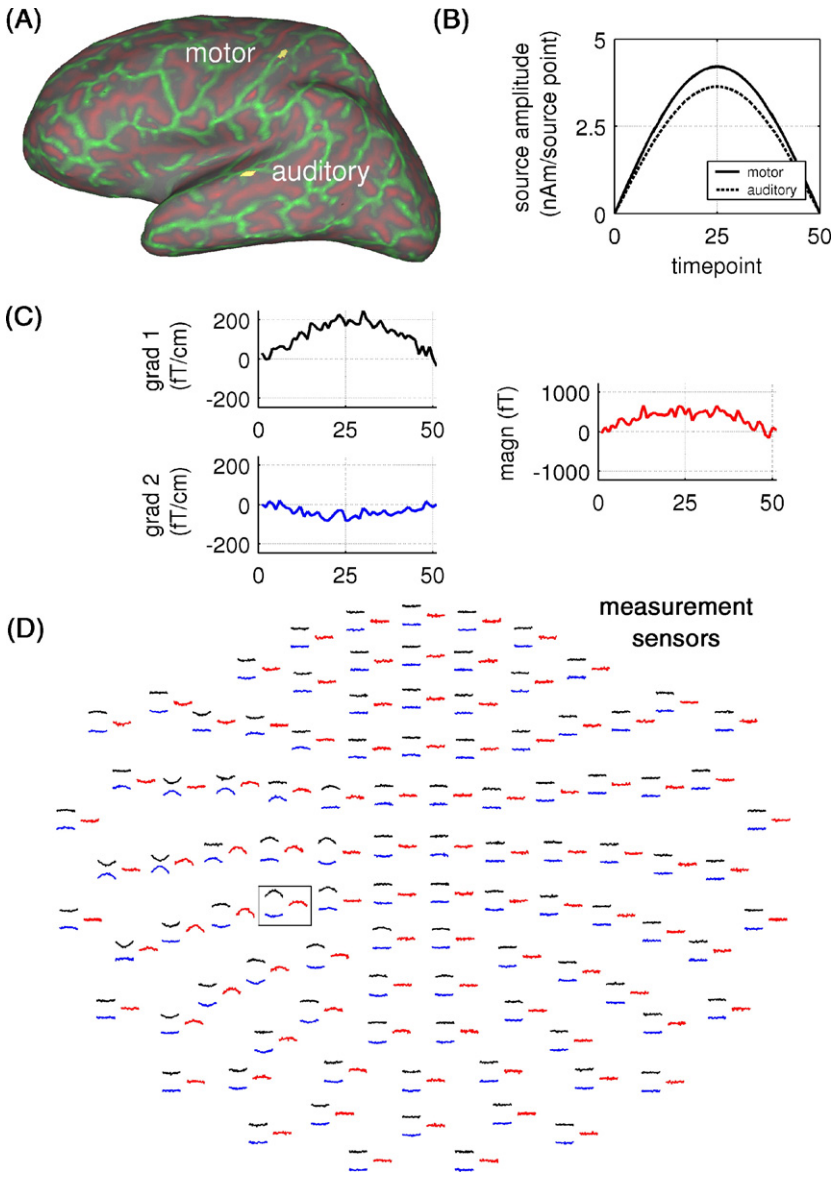

Fig. 1. (A) Locations of the simulated source patches (gyri are depicted as green and sulci as red). (B) The source amplitude time courses at the two locations. (C) and (D) The corresponding simulated gradiometer and magnetometer measurements plotted on the sensor grid with a closer view of the sensor enclosed by the black box.

where N_s is the number of sensors and σ_s^2 is the noise variance, was 5 for magnetometers and 10 for gradiometers at the peak value of the simulated data.

Results

The following conventions are adopted in the figures throughout the results section: The unit of current amplitude is nAm. The gyri are depicted as white and sulci as grey. Negative amplitude values mean source orientation pointing to the inside of the brain and positive pointing to the outside, with direction perpendicular to the cortical surface. In addition, for some figures the color map range does not contain the whole range of the plotted current values in order to make visual comparison easier. When displaying the VB or MCMC trends of the currents or the prior current precisions, a colorcode is used to indicate the corresponding source point index. In every case, the distributed currents were estimated for all timepoints, even though the results are shown only for some of these.

Behavior of the model with the noninformative hyperprior

Here, we present briefly the effects of using the noninformative hyperprior (see, Eq. (14)). As is demonstrated in Appendix C, in this case, the posterior distribution is in fact improper. For an improper posterior distribution the concepts of Markov chain theory (see, e.g., Gilks et al., 1996, Chapter 4) such as convergence become meaningless as the posterior is no longer a probability distribution. Nevertheless, as the MCMC method operates on unnormalized distributions (bound to be proper) and the VB algorithm also runs without any apparent problems, we applied both approaches to the simulated dataset. For the initialization, we

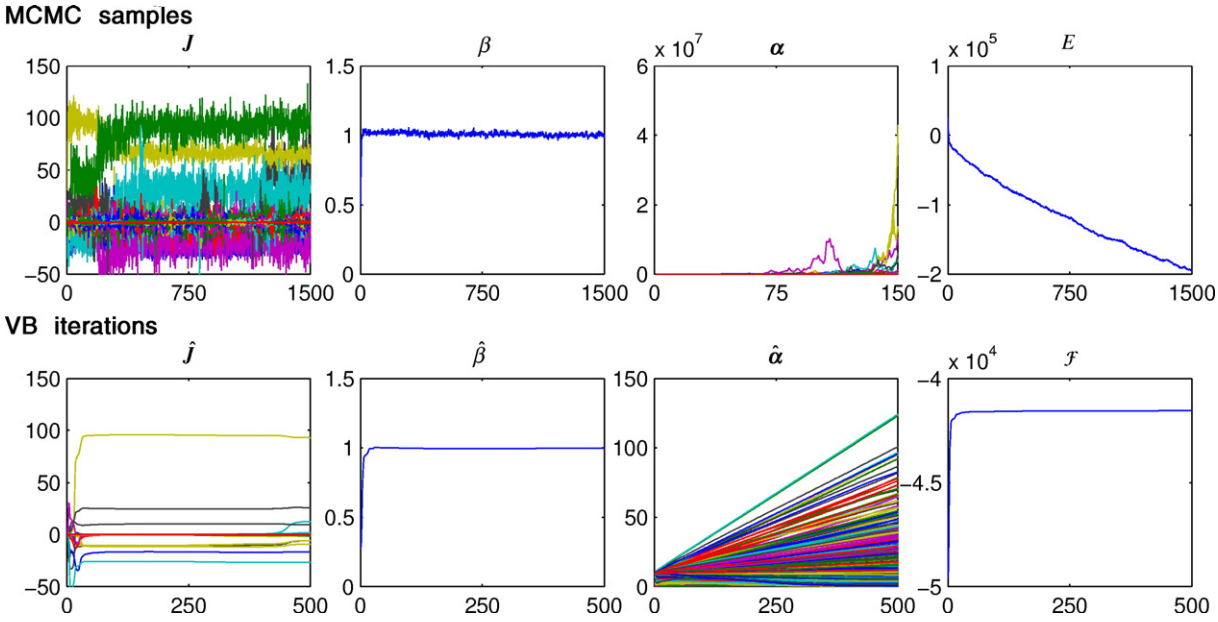

Fig. 2. In the upper row the trends of the MCMC run are shown for $\boldsymbol{J}(t=25)$, β, $\boldsymbol{\alpha}$ and E, respectively. The lower row displays the VB algorithm time courses of the corresponding quantities of the variational posterior $\hat{\boldsymbol{J}}(t=25)$, $\hat{\beta}$, $\hat{\boldsymbol{\alpha}}$, and $\boldsymbol{F}$ (see also Appendix A). Note that only 150 samples of α are plotted for the sake of better visualization of the divergence.

set the prior variances to value $\alpha_i = 10$, for all i, and for the MCMC chain we also set $\beta = 1$. For this case, we use a rather sparse grid (~700 points) for the source reconstruction, in order to ease the computational burden. The VB algorithm was run for 500 iterations and 1500 samples were obtained from the joint posterior distribution. The results are shown in Fig. 2.

At the level of visual inspection, the trends of the currents $\boldsymbol{J}(t=25)$ and β pose no immediate problem for either of the methods. The apparent convergence of the algorithms is rapid and robust enough. On the other hand, the behavior of the second level parameters α_i reveals the cause of the improperness the posterior: there is an infinite amount of probability mass in the tails of the posterior distribution in the direction of the α_i's. Interestingly, we see from the rightmost column of Fig. 2, that the posterior energy does not really converge but the free energy does or is growing at least extremely slow. In fact, the relative change of the free energy at iteration k

$$\Delta\mathcal{F}_k = \left| \frac{\mathcal{F}_k - \mathcal{F}_{k-1}}{\mathcal{F}_{k-1}} \right| \tag{23}$$

is of the order of magnitude 10^{-6} at the end of the run. The reason for this is probably that the variational posterior is always a proper distribution, and that the free energy has a well defined limit as $\gamma_{0i} \to 0$ for all i (see Appendix A). Still, the current parameters also change after 450 iterations of the VB algorithm, although one might have stopped the algorithm after 50 iterations after the free energy had converged. For the remainder of this paper, we use only proper distributions in the computations.[1]

Estimating the hyperparameters α_0 and γ_0 from the data

Next, we consider the possibility of including the hyperparameters α_0 and γ_0 also as true random variables and estimating their posterior distributions. We begin with fixing $\gamma_0 = 2$ and trying to estimate the posterior distribution of α_0. The variational update rule for updating α_0 is presented in Appendix A, whereas the conditional distribution of α_0 given the rest of the parameters is shown in Appendix B. Both of the algorithms were initialized at $\alpha_i = \alpha_0 = 10$ for all i; for the MCMC scheme, we also set initially $\beta = 1$. Some 500 VB iterations were performed, whereas 30,000 samples were obtained from the joint posterior of which only every 20th was saved due to memory limitations. The grid size was the same as previously (~700 points). Qualitatively, the source parameters $\boldsymbol{J}(t)$ and the parameter β behave quite similarly to the previous case with the noninformative hyperprior, so we concentrate on the α_i's and α_0. The trends of these parameters are shown in Fig. 3.

By looking at the trends, it is obvious, that introducing the parameter α_0 as a random variable causes problems to both of the methods. Firstly, the MCMC methods starts to suffer from serious autocorrelations rendering the sampling approach very inefficient. In fact, the integrated autocorrelation time for the parameter α_0 in the thinned chain is about 100, reducing the number of independent samples to 15 from the original 30,000 samples. Also, looking at the posterior energy, it is not at all clear whether the chain has actually converged. Finally, the posterior distribution of α_0 appears to be quite diffuse, indicating that the data do not contain much information about it.

Similar problems arise in the VB framework. The α_0 has not settled into any specific value in the 500 iterations but keeps growing slowly and also drags the α_i's with it. The free energy, on the other hand, seems to have reached a plateau value, implying that it is very flat in the direction of increasing α_0 (and the consequently increasing α_i's). It is evident that the data are not specifically informative on the prior mean of the α_i's, and this causes a serious slowing down of both posterior inference schemes.

In principle, one could treat γ_0 as a random variable and derive the VB update equations and necessary conditional distributions for the MC approach. However, this is likely only to accentuate the

[1] The uniform prior for α_0 can be restricted to some finite interval, say $0 < \alpha_0 \leq 10^6$, to guarantee that all distributions remain proper.

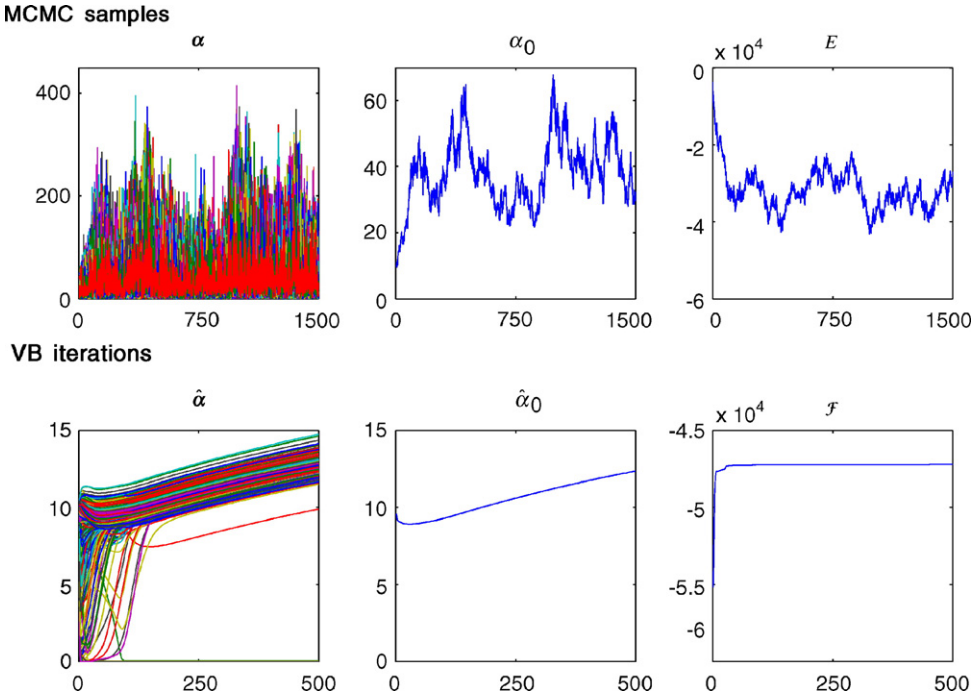

Fig. 3. In the upper row, the MCMC trends are shown for $\boldsymbol{\alpha}$, α_0 and E, respectively, when attempting to estimate α_0 from the data. The lower row illustrates the evolution of the corresponding variational quantities in the VB algorithm ($\hat{\boldsymbol{\alpha}}$, $\hat{\alpha}_0$ and $\mathcal{F}$).

problems described above. Also, since at the limit $\gamma_0 \rightarrow 0$ the marginal likelihood of the data becomes infinite, the most likely scenario is that γ_0 will slowly drift towards that value. It is as well likely that using a more realistic reconstruction grid, say of 3000 source space points, will also make the estimation of α_0 and γ_0 all the more difficult and, above all, slow. In conclusion, it seems that with these methods the estimation of the parameters α_0 and γ_0 is not feasible in practice. Consequently, we assume fixed and nonzero values for these parameters in all subsequent analyses.

Sensitivity of the model on the choice of the γ_0

Next, we study the effect of tuning the hyperparameter γ_0. The prior (12) can be thought of incorporating the information that we have $2\gamma_0$ observations of the prior precisions α_i with average precision α_0. Three values for γ_0 were assumed: 10, 1 and 0.1. We set $\alpha_0 = 10$, which was used also for the initialization, along with $\beta = 1$ for the MCMC algorithm. Grid size of ~1500 was chosen for this analysis. The results were qualitatively similar for both VB and MCMC methods, and we focus on the output of the former approach. The results are shown in Fig. 4.

The inverse solutions in the first column of Fig. 4 given as the variational posterior expectation value of the current show, that the inference procedure is somewhat sensitive on the value chosen for γ_0. This is quite natural recalling the fact that lowering the value of γ_0 takes the true posterior closer to the improper case. The second column of Fig. 4 shows how the variational posterior expected values of α_i's become more dispersed as the prior on α_i's becomes less informative with decreasing γ_0. Also, with smaller γ_0 the parameterwise convergence of the VB algorithm becomes slower, with $\gamma_0 = 0.1$ not even quite getting there. The plateau value which the free energy reaches seems to somewhat increase with decreasing γ_0. This is not a surprise, since the free energy lower bounds the marginal likelihood, which becomes infinite as $\gamma_0 \rightarrow 0$ (still, the free energy itself has a finite limit, see Appendix A). Extrapolating these results a bit, this model appears to be a case in which performing hyperparameter (model) selection based on marginal likelihood or evidence maximization leads to difficulties by suggesting to choose a model with an improper posterior density ($\gamma_0 = 0$). So, by practical considerations again, we set γ_0 to some reasonable value like 10 to speed up the convergence of the algorithms in what follows.

Sensitivity of the model on the choice of the α_0

Here, we consider briefly the influence of different values of α_0 on the inverse estimates obtained from the MCMC and VB algorithms. We set $\gamma_0 = 10$ and use values $\alpha_0 = 10$, 1 and 0.1 for inference. As before, the algorithms were initialized with $\alpha_i = \alpha_0$ for all i, and for the MCMC also $\beta = 1$. The grid size was ~1500 as in the previous analysis. The results are shown in Fig. 5.

Fig. 5(A) shows the behavior of the algorithms for the case $\gamma_0 = 10$, $\alpha_0 = 1$. The convergence is fast for both of the methods, and the mixing of the Markov chain is also quite rapid allowing robust estimation of posterior expectations. The fast mixing is evident from the absence of slow trends in the time series of the parameters (compare to Fig. 3). The inverse estimates as given by the posterior expectation values seem to be rather similar, at least qualitatively for case $\alpha_0 = 10$ and also quantitatively for $\alpha_0 = 1$, as can be seen from the two uppermost rows of Fig. 5(B). For the value $\alpha_0 = 0.1$ the MCMC estimate deviates quite significantly from the VB estimate, which still resembles the estimates obtained by different settings of α_0. This behavior could be expected, since the variational posterior is an approximation and is likely to vary in accuracy within a model with manually controllable parameters. We observed that free energy is increasing slightly with increasing

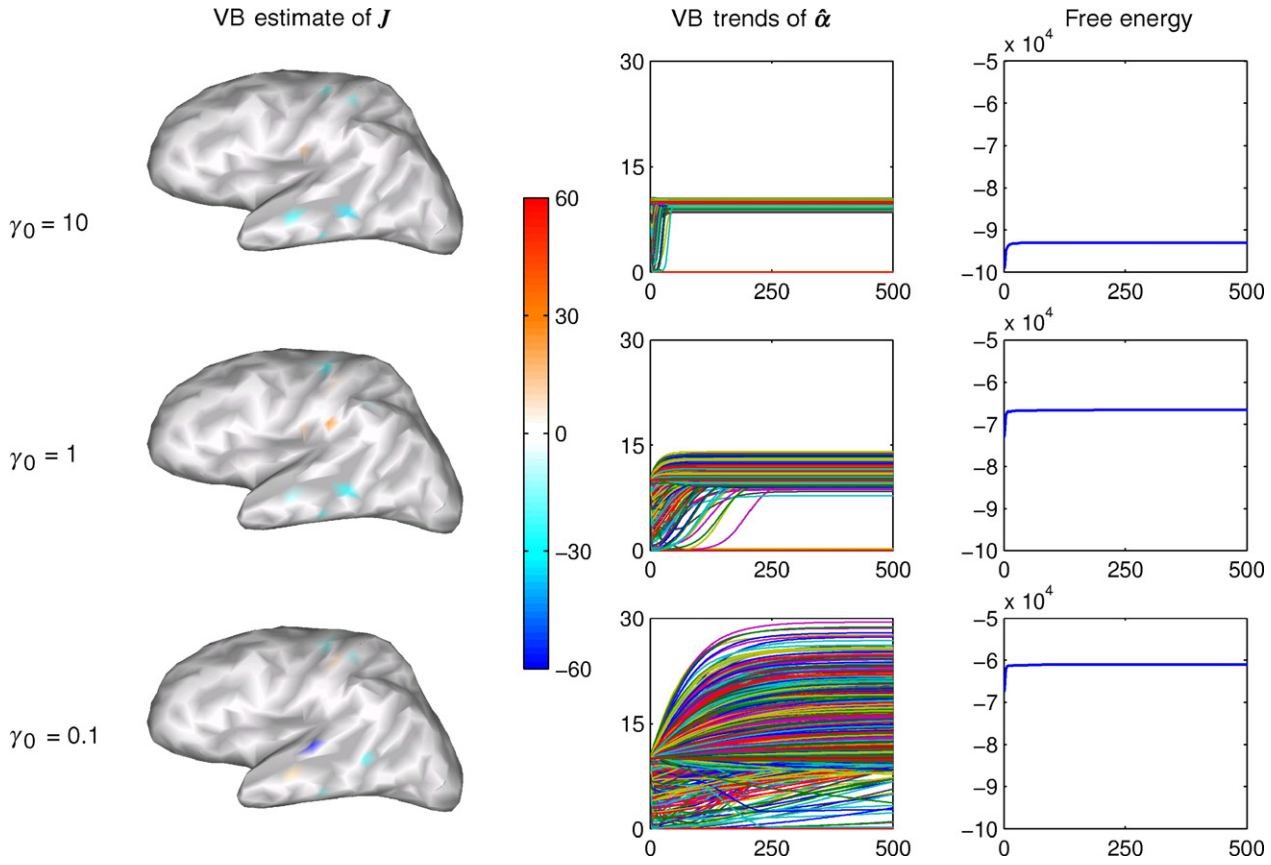

Fig. 4. The leftmost column shows VB-expected values of $\boldsymbol{J}(t = 25)$ plotted on the inflated brain obtained by using different values of γ_0. Similarly, the middle and rightmost columns display the VB trends of α_i's and $\mathcal{F}$, respectively.

α_0, which is in unison with the previous results from trying to estimate α_0 from the data. All in all, it seems that the model is not so sensitive to the choice of α_0 as it is for γ_0; this is especially true for the VB approach.

Structure of the posterior distribution with fixed γ_0 and α_0

Finally, we investigate the structure of the posterior distribution by fixing the hyperparameters γ_0 and α_0 to the value 10 and initializing both of the algorithms by setting $\alpha_i = 1$, $\alpha_i = 10$ or drawing the α_i's uniformly from interval $[1/\alpha_0, \alpha_0]$. Again, the value $\beta = 1$ was also used for the initialization of the MCMC chain. This time, the reconstruction grid was assumed to consist of ~3000 discretization points.

The results for the MCMC-scheme are shown in Fig. 6. Here, we can observe clearly the multimodality of the posterior distribution even with fixed α_0 and γ_0; the Markov chains sample from different regions of the parameter space depending on the starting point. Each of these modes represents a possible solution to the source reconstruction problem. With fixed γ_0 and α_0, we can compare the relative posterior energies of the different solutions displayed in the last column of Fig. 6. It seems that the posterior energies of the different modes are of the same order of magnitude. However, the relative posterior probabilities of the uppermost and lowermost solutions are about $\exp(1.7 \times 10^5 - 1.6 \times 10^5) = \exp(10{,}000)$, which renders the relative probability of the latter solution to practically zero. On the other hand, it is not feasible to directly calculate the amount of probability *mass* contained in the vicinity of the different modes by using samples from the separate MCMC simulation runs. In very high-dimensional cases such as this, the posterior probability ratios of different modes tend to be huge, but the mass proportions may still be comparable. The second column shows the posterior expectation values of the prior current deviation parameters $\sigma_i = \alpha_i^{-1/2}$, which reflect those source points that are estimated to be active at some stage (the prior precisions are assumed to be same for all timepoints). Since the sources do not move spatially, the very close resemblance of the expected values of the current prior deviations and the currents themselves at the timepoint of maximal signal should be expected.

The corresponding quantities for the VB algorithm can be seen in Fig. 7. We observe that the output of the VB algorithm is far less sensitive to the initialization, even though there are slight differences in the obtained variational distributions. This is most likely due to the fact that the factorization assumption acts as an extra regularization term smoothing the solutions. While the variational posterior itself is always unimodal and the process of minimizing the KL divergence takes into account the probability mass proportions of the modes of the true posterior, it is possible for the VB algorithm to get stuck in a narrow local mode. Since the height of the modes of the true posterior were of the same order of magnitude, and the variational posterior finds its way close to the uppermost mode in Fig. 7, one might speculate that this is the most "significant" mode containing most of the posterior probability

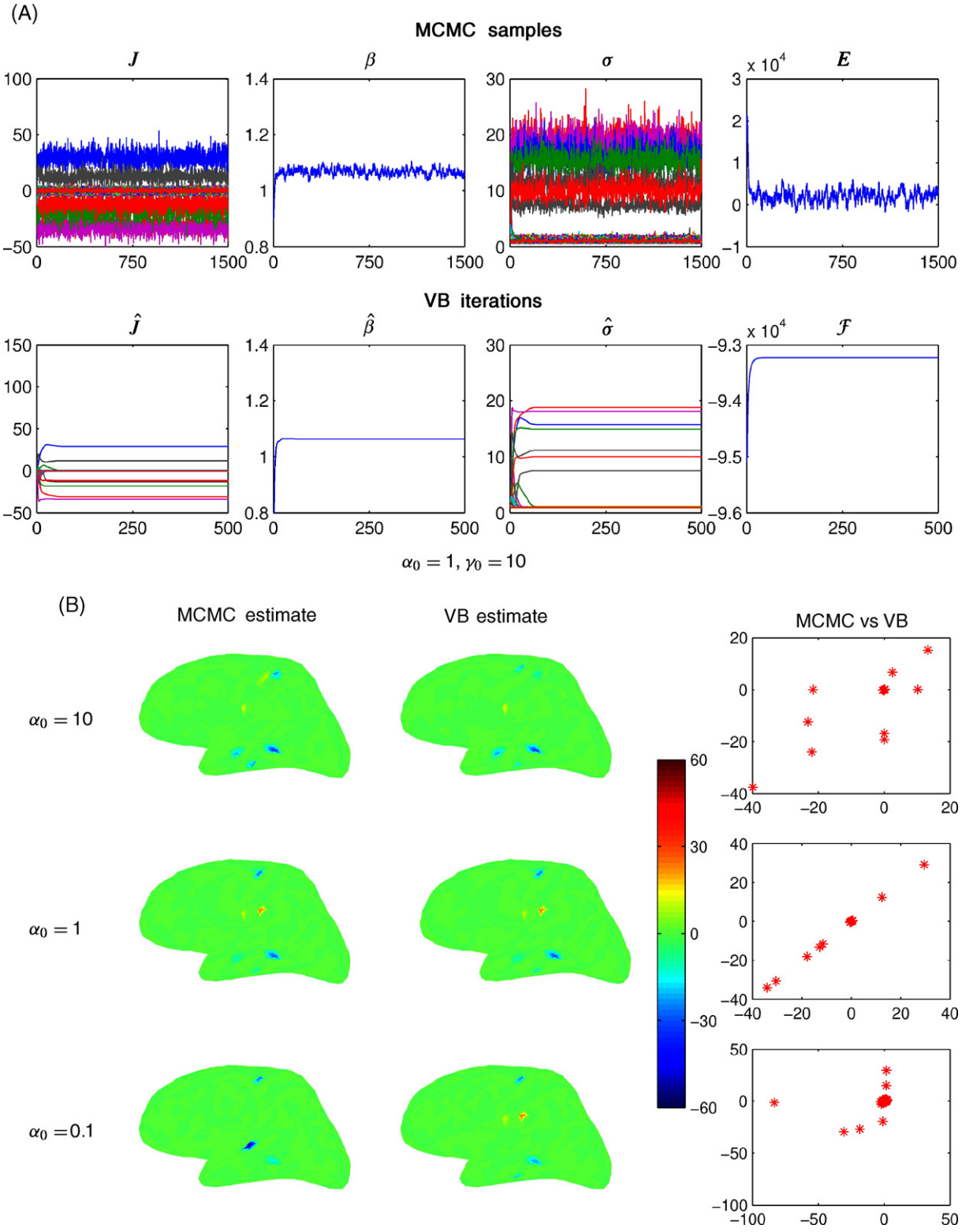

Fig. 5. (A) The MCMC and VB trends for the case $\alpha_0 = 1$, $\gamma_0 = 10$. (B) The leftmost and middle columns show respectively the MCMC- and VB-expected values of $\hat{\boldsymbol{J}}$ ($t = 25$) plotted on the inflated brain obtained by using different values of α_0. Rightmost column shows the VB and MCMC solutions as a scatterplot for the same hyperparameter values. Note the very close resemblance of the estimates for the case $\alpha_0 = 1$, $\gamma_0 = 10$.

mass. But we initialized the algorithm by choosing three arbitrary starting points, and there might be many more yet unseen modes. A more systematical analysis of the number and properties of different modes of the posterior under different source and noise conditions is beyond the scope of the present study. From both practical and theoretical point of view, it can be dangerous to use a

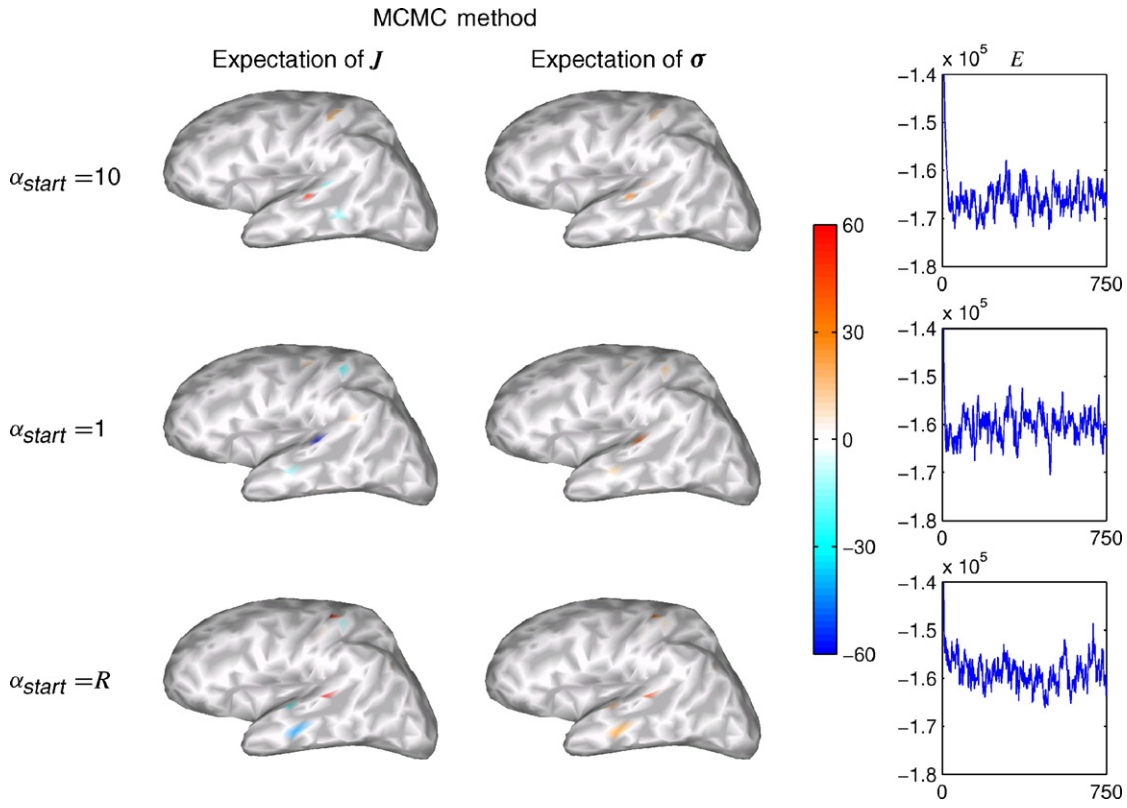

Fig. 6. The leftmost and the middle columns show respectively the MCMC-expected value of the currents $\boldsymbol{J}(t = 25)$ and the prior deviations σ_i plotted on the inflated brain, obtained by using different initializations of the algorithm (R refers to the random initialization, see text). The rightmost column displays the corresponding MCMC trends of the posterior energy E.

unimodal distribution as an approximate for a truly multimodal distribution.

Fig. 8 displays the time courses of the VB and MCMC solutions corresponding to the case where α_i's were initialized to value $\alpha_0 = 10$. Both algorithms produce very similar results, where mostly the amplitude varies in the course of time as the simulated sources do not move spatially. The temporal profiles of the largest currents in the estimates follow closely those of the original sources but are somewhat less smooth; this is due to noise and the fact that consecutive data points are modeled as independent.

Discussion

We have studied the recently proposed hierarchical Bayesian approach for solving the MEG inverse problem (Sato et al., 2004). We introduced an alternative posterior inference strategy by using Markov chain Monte Carlo methods, and compared the results to the variational Bayesian approach. We pointed out some potential problems related to imposing a noninformative hyperprior for the current precisions and studied the possibility of circumventing this choice by a fully Bayesian treatment which turned out to be computationally too inefficient for practical purposes. The sensitivity of the estimates to different hyperparameter settings was studied, as well as the differences between the "true" posterior (as represented by the numerical samples) and the variational posterior which makes a factorization assumption about the currents and their variances.

Our results clearly show that the choice of the parameters for the hyperprior is a nontrivial issue. In comparison with the traditional MNE, this choice is not perhaps as crucial, since the parameters which are manually fixed are the parameters of the prior of the prior precisions, not the prior precisions themselves. The solutions produced by the hierarchical model are close in nature to those obtained by multidipole models. This is most likely linked to multimodality of the posterior distribution as several source combinations can produce similar measurements. The hierarchical prior is very flexible, and in fact, integration over the prior variances results in an effective Student t-distribution prior for the currents (see, e.g., Gelman et al., 2003, pp. 303–305), which is heavy tailed and hence favors focal solutions. There appears to be a natural trade-off between choosing a method providing smoother but unique solution, and the hierarchical approach with better spatial resolution and a multitude of candidate solutions. The inclusion of the spatial prior to the hierarchical methods may possibly remedy the situation but not completely, as Sato et al. (2004) use the model without the spatial prior to find the peak values of the current in a coarser grid, around which the finer analysis with the spatial prior is to be carried out. Therefore, the uncertainty about the current distribution is not perhaps completely represented in the final variational posterior. Furthermore, the true posterior might still contain many modes even when equipped with the spatial smoothness prior.

Qualitatively, and in some circumstances also quantitatively, the VB and MCMC methods produce very similar results. The true posterior as revealed by the MCMC scheme seems to contain many modes, whereas the VB estimates show much more resemblance. This could be due to the fact that one of the modes of the posterior distribution contains most of the posterior probability mass, since

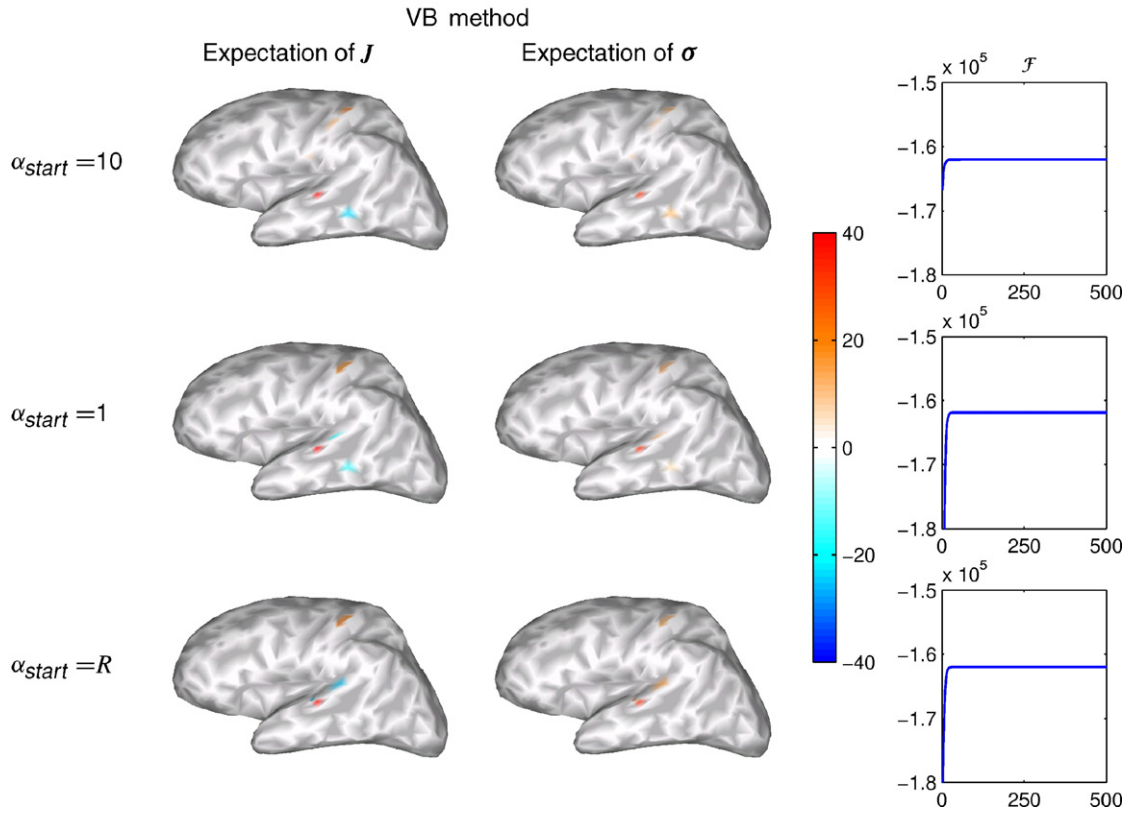

Fig. 7. The leftmost and middle columns display respectively the VB-expected value of the currents $\boldsymbol{J}(t = 25)$ and the prior deviations σ_I plotted on the inflated brain, obtained by using different initializations of the algorithm (R refers to the random initialization, see text). The rightmost figure shows the evolution of the free energy $\boldsymbol{F}$ in the corresponding cases.

the VB algorithm has a tendency to gravitate towards such modes. On the other hand, the Gibbs sampler is not likely to move between the different modes (the currents are updated conditional on the other parameter values, see, Appendix B), and so it is not possible to say directly whether one of the modes is more prominent than the others. Moreover, only a few different initializations were used, and it could be that the rather high-dimensional parameter space contains a vast number of regions with significant proportions of posterior probability mass. At this stage all this is just speculation, and the issue needs to be looked at in more detail and also by using real data. In any case, for reasonable values of α_0 and γ_0, the variational posterior models the marginal distributions of the parameters locally very well, perhaps even surprisingly well taken into account the complexity of the underlying model. As noted before, it still might be dangerous to use the unimodal variational posterior as a proxy for a truly multimodal posterior distribution. In summary, the possible presence of several modes with equal amounts of posterior probability mass may give rise to over-interpretation of the results and poses a challenge for both of the estimation methods.

As a curiosity, we note that the free energy values that the VB algorithm reaches seem to decrease with increasing grid size. Since the free energy provides a lower bound to the (logarithm of the) evidence of the data, which should be maximized at the model selection process, we should then use as sparse grid as possible. By visual inspection of the solutions, we see that quite obviously the quality of the solutions is the best for the most dense grid. This apparent conflict is due to the underdetermined nature of the inverse problem (i.e., the parameters outnumber the data), which results in that the data is always very accurately explained. The automatic Occam's razor arising from the marginalization over all parameters in the evaluation of the evidence suggests then using the most parsimonious model which can explain the data.

In practice, the most dense grid of 3000 points used also in the original approach seems to be suitable in many aspects: increasing the grid size causes slowing down of the algorithms and violates the assumption of a priori independent currents, whereas using a more sparse grid com-promises the accuracy of the source localization and the meaningfulness of the cortical orientation constraint. When using a sparse grid, it is of course straightforward to drop the orientation constraint completely or introduce a loose orientation constraint (Lin et al., 2006). With empirical data, the estimation on the sparse grid is most likely used only as a prelocalization method for the final analysis on the dense grid, and the exact locations of the current peaks on the coarse grid are perhaps not so decisive.

Possible topics for future work include combining the VB and the MCMC approach in order to utilize the rapid deterministic convergence of the VB algorithm and the possibility of performing inferences based on the unfactored posterior offered by the computationally more intensive MCMC method. The performance of the hierarchical model has to be also evaluated with real data. Since the model tries to explain the data with few localized activations, it could be rather sensitive to artefacts and the correctness of the noise model. The authors' concluding view is that the hierarchical Bayesian approach offers an interesting and

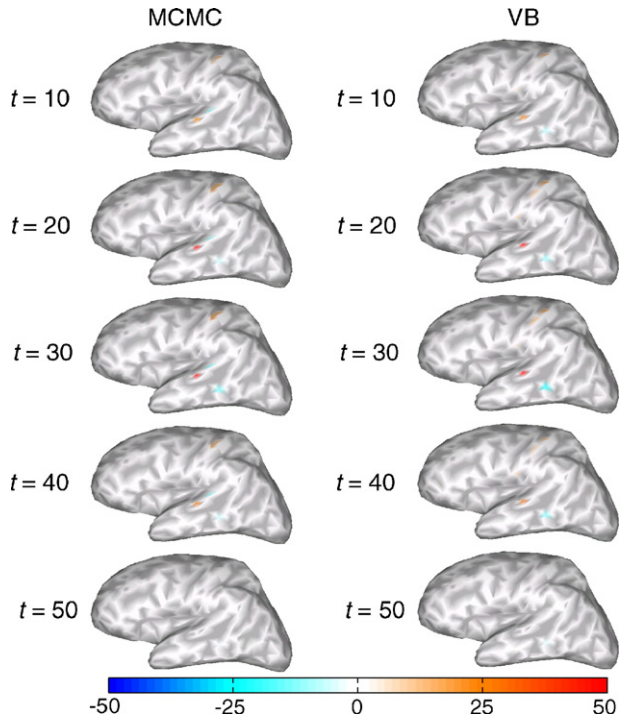

Fig. 8. The left column shows the MCMC inverse solution given as the posterior expected value of the currents when starting the algorithm from $\alpha_{i,\mathrm{start}} = 10$, displayed at timepoints 10, 20, 30, 40 and 50. Similarly, the left column displays the VB inverse solution given as the variational posterior expected value of the currents when starting the algorithm from $\alpha_{i,\mathrm{start}} = 10$, displayed at timepoints 10, 20, 30, 40 and 50.

novel methodology for modeling the distributed currents underlying magnetic measurements of neural activity.

Acknowledgments

This research was supported in part by Academy of Finland (projects: 200521, 206368, Centre of Excellence: 202871), Instrumentarium Science Foundation, Finnish Cultural Foundation, Tekniikan edistämissäätiö (Technological Foundation), Jenny and Antti Wihuri Foundation, National Institutes of Health, and The MIND Institute. The original MEG data to which the simulation model is based on were collected at Brain Research Unit, Low Temperature Laboratory, Helsinki University of Technology.

Appendix A. The VB algorithm

For the sake of self-sufficiency of the presentation, we begin by reviewing the basis of the VB procedure, following Sato et al. (2004). Then we recapitulate the variational update equations in subsections A.1. and A.2. and present an explicit formula for the free energy in subsection A.3. The variational update rule for α_0 is derived in subsection A.4.

Let us define the free energy functional of a trial distribution $Q(\boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha})$ (we have assumed $\boldsymbol{\alpha}_0$ and $\boldsymbol{\gamma}_0$ fixed)

$$\mathcal{F}(Q) = \int \mathrm{d}\boldsymbol{J}_{1:T}\,\mathrm{d}\beta\;\mathrm{d}\boldsymbol{\alpha} Q(\boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}) \log\left(\frac{P(\boldsymbol{B}_{1:T}, \boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha})}{Q(\boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha})}\right) \tag{24}$$

By using Eq. (19), it can be seen immediately that

$$\mathcal{F}(Q) = \log(P(\boldsymbol{B}_{1:T})) - \mathrm{KL}[Q(\boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}) || P(\boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}|\boldsymbol{B}_{1:T})], \tag{25}$$

where the asymmetric Kullback–Leibler divergence (distance) from Q to P is defined as

$$\mathrm{KL}[Q\boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}) || P(\boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}|\boldsymbol{B}_{1:T})] = \int \mathrm{d}\boldsymbol{J}_{1:T}\mathrm{d}\beta\;\mathrm{d}\boldsymbol{\alpha} Q(\boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}) \log\left(\frac{Q(\boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha})}{P(\boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}|\boldsymbol{B}_{1:T})}\right) \tag{26}$$

Hence, maximizing the free energy with respect to Q corresponds to minimizing the KL divergence from the trial distribution Q to the posterior distribution P. Since the KL divergence is always nonnegative, and because the KL divergence from the posterior distribution to itself is zero, we conclude that the maximum of the free energy over all trial distributions equals to the logarithm of the marginal likelihood $P(\boldsymbol{B}_{1:T})$. However, the VB approach usually proceeds by constraining the space of possible trial distributions in order to make the maximization procedure tractable; in that case one obtains only a lower bound on the log-marginal likelihood. Even though the variational approach is very intuitive, there are some theoretical issues which we discuss briefly in the following.

As noted before, the KL divergence is asymmetric, that is in general $\mathrm{KL}[Q \| P] \neq \mathrm{KL}[P \| Q]$. If we adopt a logarithmic score function (a "generalized least squares" cost), the quantity $\mathrm{KL}[Q \| P]$ can be interpreted as expected loss of utility in using the probability density P when the actual probability density is Q (Bernardo and Smith, 1994, pp. 154–155). Thinking in these terms, it would be more consistent to minimize $\mathrm{KL}[P \| Q]$, that is the loss of expected utility in using the approximate posterior Q instead of the true posterior P (rather than vice versa). Of course, the main reason why $\mathrm{KL}[P \| Q]$ is not actually used is that expectation values with respect to P are not tractable to compute, which is the motive for developing an approximative inference scheme in the first place. Even if the KL divergence was computable both ways, it is not clear which one gives more useful approximations for a specific problem. Assume for instance, that we were to approximate a multimodal distribution with a unimodal one. Minimizing $\mathrm{KL}[Q \| P]$ with respect to Q would then result in the variational distribution Q being close to P near one of its modes while neglecting the others. Reversing Q and P and minimizing $\mathrm{KL}[P \| Q]$ with respect to Q would yield a very wide approximative distribution Q, which tries to cover all of the modes of P. In the former case, the true uncertainty is underestimated by neglecting most of the modes, whereas in the second case the approximative posterior Q gives rise to large probabilities also in between the modes of P where the actual posterior probability is close to zero. These different projections of the true posterior to the subspace of factorizable distributions can be also interpreted naturally in the more general framework of information geometry (see, e.g., Tanaka, 2001). In this case, the standard VB approach is followed with factorization

$$Q(\boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}) = Q_{J,\beta}(\boldsymbol{J}_{1:T}, \beta) Q_{\alpha}(\boldsymbol{\alpha}). \tag{27}$$

Plugging this into the free energy (24) one arrives after slight rearrangement at

$$\mathcal{F}(Q) = \langle\langle \log(P(\boldsymbol{B}_{1:T}, \boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}))\rangle_{J,\beta}\rangle_{\alpha} - \langle \log Q_{J,\beta}(\boldsymbol{J}_{1:T}, \beta)\rangle_{J,\beta} - \langle \log Q_{\alpha}(\boldsymbol{\alpha})\rangle_{\alpha} \quad (28)$$

$$= \langle \log P(\boldsymbol{B}_{1:T}|\boldsymbol{J}_{1:T}, \beta)\rangle_{J,\beta} - \mathrm{KL}[Q_{J,\beta}(\boldsymbol{J}_{1:T}, \beta)Q_{\alpha}(\boldsymbol{\alpha})||P_0(\boldsymbol{J}_{1:T}|\beta, \boldsymbol{\alpha})P_0(\boldsymbol{\alpha})], \quad (29)$$

where $<\cdot>_{J,\beta}$,$<\cdot>$ denote expectation value with respect to $Q_{J,\beta}$ and Q_{α}, respectively. The factorization is utilized in maximizing the free energy by alternately maximizing it with respect to $Q_{J,\beta}$ and Q_{α} while keeping the other distribution fixed. Computing the variational derivatives of the free energy, from Eq. (28) for instance, one finds that the extremizing distributions are

$$Q_{J,\beta}(\boldsymbol{J}_{1:T}, \beta) \propto \exp[\langle \log P(\boldsymbol{B}_{1:T}, \boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha})\rangle_{\alpha}], \quad (30)$$

$$Q_{\alpha}(\boldsymbol{\alpha}) \propto \exp[\langle \log P(\boldsymbol{B}_{1:T}, \boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha})\rangle_{J,\beta}]. \quad (31)$$

These equations are coupled, and cannot be solved directly. The VB algorithm operates by starting with an initial guess for $Q_{\alpha}(\boldsymbol{\alpha})$, then computes $Q_{J,\beta}(\boldsymbol{J}_{1:T}, \beta)$ using Eq. (30), obtains a new $Q_{\alpha}(\boldsymbol{\alpha})$ using Eq. (31) and so on, until a fixed point is found. This is by construction a local method, and is guaranteed to arrive only at a local maximum of the free energy.

From Eqs. (30) and (31) one can derive the VB update equations by straightforward computations. We use subscripts "old" and "new" to differentiate between parameters of the variational distributions at successive iterations of the VB algorithm (one iteration consisting of both $\boldsymbol{J}$, β-step and α-step). Furthermore, we use hat to differentiate the parameters of the variational distributions from the corresponding random variables (the notation is admittedly rather heavy but perhaps less ambiguous).

A.1. The $\boldsymbol{J}$, β-step

Without further approximation, the variational posterior of $\boldsymbol{J}_{1:T}$, β can be decomposed as

$$Q_{J,\beta}(\boldsymbol{J}_{1:T}, \beta) = Q_{J|\beta}(\boldsymbol{J}_{1:T}\ |\beta)Q_{\beta}(\beta), \quad (32)$$

$$Q_{\beta}(\beta) = \int \mathrm{d}\boldsymbol{J}_{1:T} Q_{J,\beta}(\boldsymbol{J}_{1:T}, \beta). \quad (33)$$

The conditional posterior of the currents $Q_{J|\beta}(\boldsymbol{J}_{1:T}|\beta)$ is a product of T Gaussians, with each having the same inverse covariance matrix, but different mean:

$$Q_{J|\beta}(\boldsymbol{J}_{1:T}|\beta) = \prod_{T=1}^{T} \boldsymbol{N}\left(\boldsymbol{J}(t)|\hat{\boldsymbol{J}}_{\mathrm{new}}(t), (\beta\hat{\boldsymbol{\Sigma}}_{\boldsymbol{J}\mathrm{new}})^{-1}\right). \quad (34)$$

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{J}\mathrm{new}} = \boldsymbol{G}'\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{G} + \hat{\boldsymbol{A}}_{\mathrm{old}}, \hat{\boldsymbol{A}}_{\mathrm{old}} = \mathrm{diag}(\hat{\alpha}_{l_{\mathrm{old}}}, \ldots, \hat{\alpha}_{N_{\mathrm{old}}}), \quad (35)$$

$$\hat{\boldsymbol{J}}_{\mathrm{new}}(t) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{J}_{\mathrm{new}}}^{-1}\boldsymbol{G}'\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{B}(t). \quad (36)$$

The inversion of the $N \times \boldsymbol{N}$ matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{J}\mathrm{new}}$ can be avoided by using the matrix identity (sometimes referred to as the matrix inversion lemma),

$$(\boldsymbol{G}'\boldsymbol{\Sigma}_{\mathbf{G}}\boldsymbol{G} + \boldsymbol{A})^{-1}\boldsymbol{G}'\boldsymbol{\Sigma}_{\mathbf{G}} = \boldsymbol{A}^{-1}\boldsymbol{G}'(\boldsymbol{G}\boldsymbol{A}^{-1}\boldsymbol{G}' + \boldsymbol{\Sigma}_{\mathbf{G}}^{-1})^{-1}, \quad (37)$$

which can be proved by direct matrix manipulations (see, e.g., the Appendix of Liu et al. (2002)). This leads to update[2]

$$\hat{\boldsymbol{J}}_{\mathrm{new}}(t) = \hat{\boldsymbol{A}}_{\mathrm{old}}^{-1}\boldsymbol{G}'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{B}_{\mathrm{new}}}\boldsymbol{B}(t), \quad (38)$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{B}_{\mathrm{new}}}^{-1} = \boldsymbol{G}\hat{\boldsymbol{A}}_{\mathrm{old}}^{-1}\boldsymbol{G}' + \boldsymbol{\Sigma}_{\mathbf{G}}^{-1}, \quad (39)$$

which requires only inversion of $M \times M$ matrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{B}_{\mathrm{new}}}^{-1}$.

The marginal variational posterior $Q_{\beta}(\beta)$ is obtained by integrating over $\boldsymbol{J}_{1:T}$ in the joint variational posterior $Q_{J,\beta}(\boldsymbol{J}_{1:T}, \beta)$ and performing some matrix algebra. This results in a Gamma distribution with parameters

$$\hat{\gamma}_{\beta_{\mathrm{new}}} = MT/2, \quad (40)$$

$$\hat{\gamma}_{\beta_{\mathrm{new}}}/\hat{\beta}_{\mathrm{new}} = \frac{1}{2}\sum_{t=1}^{T}\Big((\boldsymbol{B}(t) - \boldsymbol{G}\hat{\boldsymbol{J}}_{\mathrm{new}}(t))'\boldsymbol{\Sigma}_{\boldsymbol{G}}(\boldsymbol{B}(t) - \boldsymbol{G}\hat{\boldsymbol{J}}_{\mathrm{new}}(t)) + \hat{\boldsymbol{J}}_{\mathrm{new}}(t)'\hat{\boldsymbol{A}}_{\mathrm{old}}\hat{\boldsymbol{J}}_{\mathrm{new}}(t)\Big). \quad (41)$$

We point out that in Sato et al. (2004), where the equations were originally presented, the first equation was $\hat{\gamma}_{\beta_{\mathrm{new}}} = NT/2$, due to a typographical error. Note also that the degrees-of-freedom parameter does not change in the course of the VB algorithm.

A.2. The α-step

The variational posterior for $\boldsymbol{\alpha}$ is a product of N Gamma-distributions

$$Q_{\alpha}(\boldsymbol{\alpha}) = \prod_{i=1}^{N} \mathrm{Gamma}(\alpha_i|\hat{\alpha}_{i_{\mathrm{new}}}, \hat{\gamma}\alpha_{i_{\mathrm{new}}}), \quad (42)$$

with parameters

$$\hat{\gamma}_{\alpha_{i_{\mathrm{new}}}} = \gamma_{0i} + T/2, \quad (43)$$

$$\hat{\gamma}_{\alpha_{i_{\mathrm{new}}}}/\hat{\alpha}_{i_{\mathrm{new}}} = \frac{\gamma_{0i}}{\alpha_{0i}} + \frac{\hat{\beta}_{\mathrm{new}}}{2}\sum_{t=1}^{T}\hat{\boldsymbol{J}}_{\mathrm{new}}(t)_i^2 + \frac{T}{2}\left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{J}_{\mathrm{new}}}^{-1}\right)_{ii}, \quad (44)$$

which are the update equations of the α-step.

The explicit computation of the $\hat{\boldsymbol{\Sigma}}_{J\mathrm{new}}^{-1}$ can be circumvented by using Eqs. (35), (37) and (39)resulting in the following parameter updates:

$$\hat{\gamma}_{\alpha_{i_{\mathrm{new}}}} = \gamma_{0i} + T/2, \quad (45)$$

$$\hat{\gamma}_{\alpha_{i_{\mathrm{new}}}}/\hat{\alpha}i_{\mathrm{new}} = \frac{\gamma_{0i}}{\alpha_{0i}} + \frac{\hat{\beta}_{\mathrm{new}}}{2}\sum_{t=1}^{T}\hat{\boldsymbol{J}}_{\mathrm{new}}(t)_i^2 + \frac{T}{2}\left(\hat{\boldsymbol{A}}_{\mathrm{old}}^{-1}\left(\boldsymbol{I} - \hat{\boldsymbol{A}}_{\mathrm{old}}^{-1}\boldsymbol{G}'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{B}_{\mathrm{new}}}\boldsymbol{G}\right)\right)_{ii} \quad (46)$$

[2] Here our notation differs slightly from that of Sato et al. (2004) as we use $\boldsymbol{\Sigma}_{\mathbf{x}}$ to denote an *inverse* covariance matrix exclusively.

Again, the degrees of freedom parameter are in fact constant. At the end of the α-step, one sets $\alpha_{i_{\text{new}}} \rightarrow \alpha_{i_{\text{old}}}$ for all i, and continues with the next $\boldsymbol{J}$, β-step.

A.3. the free energy function

With given variational distributions $Q_{J,\beta}(\boldsymbol{J}_{1:T}, \beta)$ and $Q_\alpha(\boldsymbol{\alpha})$ one can evaluate the free energy by Eq. (28) for instance. The integrals needed in the evaluation of the free energy are not always tractable, even if one is able to derive the VB update rules, but in this case the free energy can be computed in a closed form. It depends on the parameters of the variational distributions $Q_{J,\beta}$ and Q_α (i.e., $\hat{\boldsymbol{J}}(t)$, $\hat{\Sigma}_{\boldsymbol{J}}$, $\hat{\beta}$, $\hat{\gamma}_\beta$ and $\hat{\alpha}_i$, $\hat{\gamma}_{\alpha i}$), as well as the hyperparameters $\boldsymbol{\alpha}_0$, $\boldsymbol{\gamma}_0$ and the fixed part of the inverse noise covariance $\Sigma_{\boldsymbol{G}}$. The following expectation values are needed in the evaluation of the free energy (see, Eq. (18)); the necessary formulae for computing Gaussian integrals of quadratic forms can be found in, e.g., Harville (1999, Chapter 15) whereas the expectation of a logarithm of a Gamma-distributed random variable can be found in e.g., Johnson et al. (1994, Chapter 17). The digamma function is defined by $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log\Gamma(x)$

1.

$$\begin{aligned}\langle\langle \log P(\boldsymbol{B}_{1:T}|\boldsymbol{J}_{1:T},\beta)\rangle_{J,\beta}\rangle_{\boldsymbol{\alpha}} = &\frac{MT}{2}\log\left(\frac{1}{2\pi}\right) + \frac{T}{2}\log|\Sigma_{\boldsymbol{G}}| \\ &+ \frac{MT}{2}\left(\psi(\hat{\gamma}_\beta) - \log\left(\frac{\hat{\gamma}_\beta}{\hat{\beta}}\right)\right) \\ &- \frac{\hat{\beta}}{2}\sum_{t=1}^{T}\left(\left(\boldsymbol{B}(t) - \boldsymbol{G}\hat{\boldsymbol{J}}(t)\right)' \Sigma_{\boldsymbol{G}}(\boldsymbol{B}(t)\right. \\ &\left. - \hat{\boldsymbol{J}}(t)\right)\Big) - \frac{T}{2}\mathrm{Tr}\left[\boldsymbol{G}' \Sigma_{\boldsymbol{G}} \boldsymbol{G} \hat{\Sigma}_{\boldsymbol{J}}^{-1}\right]. \quad (47)\end{aligned}$$

2.

$$\begin{aligned}\langle\langle \log P_0(\boldsymbol{J}_{1:T}|\beta, \boldsymbol{\alpha})\rangle_{J,\beta}\rangle_{\boldsymbol{\alpha}} = &\frac{NT}{2}\log\left(\frac{1}{2\pi}\right) \\ &+ \frac{NT}{2}\left(\psi(\hat{\gamma}_\beta) - \log\left(\frac{\hat{\gamma}_\beta}{\hat{\beta}}\right)\right) \\ &+ \frac{T}{2}\sum_{i=1}^{N}\left(\psi(\hat{\gamma}\boldsymbol{\alpha}_i) - \log\left(\frac{\hat{\gamma}\alpha_i}{\hat{\alpha}i}\right)\right) \\ &- \frac{\hat{\beta}}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}\hat{\boldsymbol{J}}(t)_i^2\,\hat{\alpha}_i \\ &- \frac{T}{2}\sum_{i=1}^{N}\hat{\alpha}_i\left(\hat{\Sigma}_{\boldsymbol{J}}^{-1}\right)_{ii}. \quad (48)\end{aligned}$$

3.

$$\begin{aligned}\langle\langle \log P_0(\boldsymbol{\alpha}|\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0)\rangle_{J,\beta}\rangle_{\alpha} = &\sum_{i=1}^{N}\Bigg\{-\psi(\hat{\gamma}\alpha_i) + \log\left(\frac{\hat{\gamma}\alpha_i}{\hat{\alpha}_i}\right) \\ &+\gamma_{0_i}\left[\log\left(\frac{\gamma_{0_i}}{\alpha_{0_i}}\right) + \left(\psi(\hat{\gamma}\alpha_i)\right.\right. \\ &\left.\left.-\log\left(\frac{\hat{\gamma}\alpha_i}{\hat{\alpha}_i}\right)\right)\right]\Bigg\} + \sum_{i=1}^{N}\left(-\log\Gamma(\gamma_{0_i}) - \frac{\gamma_{0_i}}{\boldsymbol{\alpha}_{0_i}}\hat{\alpha}_i\right).\end{aligned}$$

4.

$$\langle\langle \log P_0(\beta)\rangle_{J,\beta}\rangle_{\boldsymbol{\alpha}} = -\left(\psi(\hat{\gamma}_\beta) - \log\left(\frac{\hat{\gamma}_\beta}{\hat{\beta}}\right)\right) \quad (50)$$

5.

$$\begin{aligned}-\langle\log\, Q_{J,\beta}(\boldsymbol{J}_{1:T},\beta)\rangle_{J,\beta} = &\frac{NT}{2}\log\,(2\pi) + \frac{T}{2}\log|\hat{\Sigma}_{\boldsymbol{J}}^{-1}| \\ &- \frac{NT}{2}\left(\psi(\hat{\gamma}_\beta) - \log\left(\frac{\hat{\gamma}_\beta}{\hat{\beta}}\right)\right) \\ &+ \log\,\Gamma(\hat{\gamma}_\beta) - \log\left(\frac{\hat{\gamma}_\beta}{\hat{\beta}}\right) \\ &+ (1-\hat{\gamma}_\beta)\psi(\gamma_\beta) + \hat{\gamma}_\beta. \quad (51)\end{aligned}$$

6.

$$\begin{aligned}&-\langle\log\, Q_\alpha(\boldsymbol{\alpha})\rangle_\alpha \\ &= \sum_{i=1}^{N}\left(\log\,\Gamma(\hat{\gamma}_{\alpha_i}) - \log\left(\frac{\hat{\gamma}_{\alpha_i}}{\hat{\alpha}i}\right) + (1-\hat{\gamma}_{\alpha_i})\psi(\hat{\gamma}_{\alpha_i}) + \hat{\gamma}_{\alpha_i}\right). \quad (52)\end{aligned}$$

$$\mathcal{F}(\Sigma_{\boldsymbol{G}},\hat{\boldsymbol{J}}(t),\hat{\Sigma}_{\boldsymbol{J}},\hat{\beta},\hat{\gamma}_\beta,\hat{\gamma}_{\alpha_i},\hat{\alpha}_i,\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0) = 1. + 2. + 3. + 4. + 5. + 6. \quad (53)$$

It is easy to show by direct computation, that the VB update equations are reproduced by computing the derivatives of $\mathcal{F}$ with respect to $\hat{\boldsymbol{J}}(t)$, $\hat{\Sigma}_{\boldsymbol{J}}$, $\hat{\beta}$ and $\hat{\alpha}_i$ and setting them to zero, that is maximizing $\mathcal{F}$. Once again, it is possible to avoid the computation of $\hat{\Sigma}_{\boldsymbol{J}}^{-1}$ by utilizing Eqs. (35), (37), (39) and performing some elementary matrix manipulations.

A.4. The α_0 update rule

If we assume that $\alpha_{0i} = \alpha_0$ for all i and treat α_0 as a random variable, the update equation for α_0 can be derived by maximizing $\mathcal{F}$ with respect to it, or equivalently solving the equation

$$\frac{\partial\mathcal{F}}{\partial\boldsymbol{\alpha}_0} = 0. \quad (54)$$

This leads to the following, very natural update rule for α_0:

$$\hat{\boldsymbol{\alpha}}_{0\text{new}} = \frac{1}{N}\sum_{i\,=\,1}^{N}\hat{\boldsymbol{\alpha}}_{i_{\text{new}}}. \quad (55)$$

Assuming a uniform prior for α_0 (say, on interval $(0, 10^6]$) adds a constant term to the free energy, but since it does not affect the update equations, it is omitted.

Appendix B. The MCMC scenario

The conditional distributions used in the MCMC simulation are given below. These are readily derived from the full probability of the data, parameters and hyperparameters:

$$\begin{aligned}P(\boldsymbol{B}_{1:T},\,\boldsymbol{J}_{1:T},\,\beta,\,\boldsymbol{\alpha},\,\boldsymbol{\alpha}_0,\,\boldsymbol{\gamma}_0) = &\left(1/2\pi\right)^{MT/2}\beta^{MT/2}|\Sigma_{\boldsymbol{G}}|^{T/2}\exp \\ &\times\left(-\frac{\beta}{2}\sum_{t=1}^{T}\left(\boldsymbol{B}(t) - \boldsymbol{G}\boldsymbol{J}(t)\right)'\Sigma_{\boldsymbol{G}}\left(\boldsymbol{B}(t) - \boldsymbol{G}\boldsymbol{J}(t)\right)\right) \\ &\times\left(1/2\pi\right)^{NT/2}\beta^{NT/2}|\boldsymbol{A}|^{T/2}\exp\left(-\frac{\beta}{2}\sum_{t=1}^{T}\boldsymbol{J}(t)'\boldsymbol{A}\boldsymbol{J}(t)\right) \\ &\times\prod_{i=1}^{N}\mathrm{Gamma}(\alpha_i|\alpha_{0i},\boldsymbol{\gamma}_{0i})P_0(\boldsymbol{\alpha}_0,\boldsymbol{\gamma}_0)(1/\beta). \quad (56)\end{aligned}$$

In the sampling scheme, we assume that for all i, $\alpha_{0i} = \alpha_0$, $\gamma_{0i} = \gamma_0$, and that these are either fixed, or γ_0 is fixed and α_0 has a uniform prior. Samples from the joint posterior are obtained by sampling in turn from each of the conditional distributions, given the previous values of the other parameters.

1. The conditional posterior distribution of the source currents $\boldsymbol{J}(t)$ given all other parameters and hyperparameters is a multivariate Gaussian with mean $\boldsymbol{\mu_J}(t)$ and covariance $(\beta\boldsymbol{\Sigma_J})^{-1}$, where

$$\boldsymbol{\mu_J}(t) = \boldsymbol{\Sigma_J}^{-1}\,\boldsymbol{G}'\,\boldsymbol{\Sigma_G}\boldsymbol{B}(t), \tag{57}$$

$$\boldsymbol{\Sigma_J} = \boldsymbol{G}'\,\boldsymbol{\Sigma_G}\boldsymbol{G} + \boldsymbol{A}. \tag{58}$$

By using Eq. (37) and similar matrix identities (see, e.g., Kaipio and Somersalo, 2005, p. 78), the direct inversion of $\boldsymbol{\Sigma_J}$ can be avoided even though the covariance matrix $\frac{1}{\beta}\boldsymbol{\Sigma_J}^{-1}$ is explicitly needed in sampling from the conditional (Gaussian) distribution of $\boldsymbol{J}(t)$.

2. For β, the conditional posterior is of Gamma-form, with parameters $\overline{\beta}$ and $\overline{\gamma}_\beta$:

$$\overline{\gamma}_\beta = (M+N)T/2. \tag{59}$$

$$\overline{\gamma}_\beta/\overline{\beta} = \frac{1}{2}\sum_{t=1}^{T}\left((\boldsymbol{B}(t) - \boldsymbol{G}\boldsymbol{J}(t))'\,\boldsymbol{\Sigma_G}(\boldsymbol{B}(t) - \boldsymbol{G}\boldsymbol{J}(t)) + \boldsymbol{J}(t)'\boldsymbol{A}\boldsymbol{J}(t)\right). \tag{60}$$

3. The α_i's have also conditional distributions of Gamma-form, with parameters $\overline{\alpha}_i$, $\overline{\gamma}_\alpha$:

$$\overline{\gamma}_{\boldsymbol{\alpha}} = \gamma_0 + T/2, \tag{61}$$

$$\overline{\gamma}_{\boldsymbol{\alpha}}/\overline{\alpha}_i = \frac{\gamma_0}{\alpha_0} + \frac{\beta}{2}\sum_{t=1}^{T}\boldsymbol{J}\left(t\right)_i^2. \tag{62}$$

4. Finally, the parameter α_0, if assumed a random variable, has conditional posterior

$$P(\alpha_0|\boldsymbol{B}_{1:T}, \boldsymbol{J}_{1:T}, \beta, \alpha, \gamma_0) \propto \prod_{i=1}^{N}\left(\frac{\alpha_i\gamma_0}{\alpha_0}\right)^{\gamma_0}\Gamma\left(\gamma_0\right)^{-1}\exp\left(-\frac{\alpha_i\gamma_0}{\alpha_0}\right). \tag{63}$$

Gibbs sampling can be utilized in simulating from all of the above distributions except for the last one; methods for drawing numerical samples from standard distributions can be found for instance in Appendix A of Gelman et al. (2003). In the case of the conditional distribution of α_0 slice sampling (Neal, 2003) is used instead.

When using Gibbs sampling, the parameters are usually updated in blocks as above, variables in one block being conditioned on the others. Consequently, the Markov chain can not move in all directions of the parameter space at a given step (only to those directions which are being currently updated). This results in that the Gibbs sampler may move between different modes of the posterior with extremely small probability (if at all, see Gilks et al., 1996, p. 53, for a simple example of such behavior). The sampler will thereby get stuck in some of the posterior modes depending on the starting point and only a local picture of the posterior distribution is obtained. In more mathematical terms, the chain is reducible at least for practical computation times, and it can not be said to have converged to the desired distribution globally. Sampling from a multimodal distribution is a very difficult problem, especially if the dimensionality of the parameter space is large (for some possible strategies, see, Liu, 2001, Chapters 10–11).

Appendix C. Improperness of the posterior with the noninformative hyperprior

Here, we demonstrate the improperness of the posterior distribution, in a very hand-waving way, when the noninformative prior is assumed for the precision parameters α_i.

By definition

$$\begin{aligned} P(\boldsymbol{B}_{1:T}) &= \int \mathrm{d}\boldsymbol{J}_{1:T}\,\mathrm{d}\beta\,\mathrm{d}\boldsymbol{\alpha}\,P(\boldsymbol{B}_{1:T}, \boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}) \\ &= \int \mathrm{d}\boldsymbol{J}_{1:T}\,\mathrm{d}\beta\,\mathrm{d}\boldsymbol{\alpha}\,P(\boldsymbol{B}_{1:T}|\boldsymbol{J}_{1:T}, \beta)P_0(\boldsymbol{J}_{1:T}|\beta, \boldsymbol{\alpha})P_0(\boldsymbol{\alpha})P_0(\beta), \end{aligned} \tag{64}$$

where now

$$P_0(\boldsymbol{\alpha}) = \prod_{i=1}^{N}\left(\frac{1}{\alpha_i}\right) = |\boldsymbol{A}|^{-1}, \quad P_0(\beta) = \frac{1}{\beta}, \tag{65}$$

since $\boldsymbol{A} = \mathrm{diag}(\boldsymbol{\alpha})$.

First we perform the integral over $\boldsymbol{J}_{1:T}$:

$$P(\boldsymbol{B}_{1:T}, \beta, \boldsymbol{\alpha}) = \int \mathrm{d}\boldsymbol{J}_{1:T}P(\boldsymbol{B}_{1:T}, \boldsymbol{J}_{1:T}, \beta, \boldsymbol{\alpha}) \tag{66}$$

$$= P_0(\boldsymbol{\alpha})P_0(\beta)\int \mathrm{d}\boldsymbol{J}_{1:T}P(\boldsymbol{B}_{1:T}|\boldsymbol{J}_{1:T}, \beta)P_0(\boldsymbol{J}_{1:T}|\beta, \boldsymbol{\alpha}) \tag{67}$$

$$= P_0(\boldsymbol{\alpha})P_0(\beta) \tag{68}$$

$$\times \int \mathrm{d}\boldsymbol{J}_{1:T}(1/2\pi)^{MT/2}\beta^{MT/2}|\boldsymbol{\Sigma_G}|^{T/2} \times \exp\left(-\frac{\beta}{2}\sum_{t=1}^{T}(\boldsymbol{B}(t) - \boldsymbol{G}\boldsymbol{J}(t))'\,\boldsymbol{\Sigma_G}(\boldsymbol{B}(t) - \boldsymbol{G}\boldsymbol{J}(t))\right) \tag{69}$$

$$\times (1/2\pi)^{NT/2}\beta^{NT/2}|\boldsymbol{A}|^{T/2}\exp\left(-\frac{\beta}{2}\sum_{t=1}^{T}\boldsymbol{J}(t)'\boldsymbol{A}\boldsymbol{J}(t)\right). \tag{70}$$

By combining the terms in the exponentials, and completing the square with respect to $\boldsymbol{J}(t)$ we get

$$\begin{aligned} P(\boldsymbol{B}_{1:T}, \beta, \boldsymbol{\alpha}) &\propto P_0(\boldsymbol{\alpha})P_0(\beta)\beta^{(M+N)T/2}|\boldsymbol{A}|^{T/2} \\ &\times \exp\left(-\frac{\beta}{2}\sum_{t=1}^{T}\left[\boldsymbol{B}(t)'\,\boldsymbol{\Sigma}_{\mathbf{G}}\boldsymbol{B}(t) - \boldsymbol{\mu_J}(t)'\,\boldsymbol{\Sigma_J}\boldsymbol{\mu}\boldsymbol{J}(t)\right]\right) \\ &\times \int \mathrm{d}\boldsymbol{J}_{1:T}\exp\left(-\frac{\beta}{2}\sum_{t=1}^{T}\left((\boldsymbol{J}(t) - \boldsymbol{\mu}\boldsymbol{J}(t))'\,\boldsymbol{\Sigma_J}(\boldsymbol{J}(t) - \boldsymbol{\mu}\boldsymbol{J}(t))\right)\right). \end{aligned} \tag{71}$$

where we have left out numerical constants and (again)

$$\boldsymbol{\Sigma}_{\mathbf{J}} = \boldsymbol{G}'\,\boldsymbol{\Sigma}_{\mathbf{G}}\boldsymbol{G} + \boldsymbol{A}. \tag{72}$$

$$\boldsymbol{\mu_J}(t) = \boldsymbol{\Sigma_J}^{-1}\boldsymbol{G}'\,\boldsymbol{\Sigma_G}\boldsymbol{B}(t). \tag{73}$$

The integral over $\boldsymbol{J}_{1:T}$ gives just the product of inverse normalizing factors of T Gaussian distributions with inverse covariance $\beta\boldsymbol{\Sigma}_{\boldsymbol{J}}$, and we obtain

$$P(\boldsymbol{B}_{1:T},\,\beta,\,\boldsymbol{\alpha}) \propto P_0(\boldsymbol{\alpha})P_0(\beta)\beta^{(M+N)T/2}|\boldsymbol{A}|^{T/2} \tag{74}$$

$$\times \exp\left(-\frac{\beta}{2}\sum_{t=1}^{T}\left[\boldsymbol{B}(t)'\,\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{B}(t) - \boldsymbol{\mu}_{\boldsymbol{J}}(t)'\,\boldsymbol{\Sigma}\boldsymbol{J}\,\boldsymbol{\mu}_{\boldsymbol{J}}(t)\right]\right)|\beta\boldsymbol{\Sigma}_{\boldsymbol{J}}|^{-T/2}, \tag{75}$$

which gives by substituting the prior of β and slight rearrangement

$$P(\boldsymbol{B}_{1:T},\,\beta,\,\boldsymbol{\alpha}) \propto P_0(\boldsymbol{\alpha})|\boldsymbol{A}|^{T/2}|\boldsymbol{\Sigma}_{\boldsymbol{J}}|^{-T/2}\beta^{-1}\beta^{MT/2} \tag{76}$$

$$\times \exp\left(-\frac{\beta}{2}\sum_{t=1}^{T}\left[\boldsymbol{B}(t)'\,\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{B}(t) - \boldsymbol{\mu}_{\boldsymbol{J}}(t)'\,\boldsymbol{\Sigma}_{\boldsymbol{J}}\boldsymbol{\mu}_{\boldsymbol{J}}(t)\right]\right). \tag{77}$$

From this expression, we see that with fixed $\boldsymbol{B}_{1:T}$ and $\boldsymbol{\alpha}$ the part depending on β is proportional to a Gamma distribution with parameters

$$\tilde{\gamma}_{\beta} = MT/2, \tag{78}$$

$$\tilde{\gamma}_{\beta}/\tilde{\beta} = \frac{1}{2}\sum_{t=1}^{T}\left[\boldsymbol{B}(t)'\,\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{B}(t) - \boldsymbol{\mu}_{\boldsymbol{J}}(t)'\,\boldsymbol{\Sigma}_{\boldsymbol{J}}\boldsymbol{\mu}_{\boldsymbol{J}}(t)\right]. \tag{79}$$

Therefore, the integral over β gives just the inverse normalizing factor of this Gamma distribution. Being interested in the behavior of this function as a function of the α_i's, we use Eq. (72) and express the $\boldsymbol{\alpha}$- dependence of $\boldsymbol{\mu}_{\boldsymbol{J}}(t)$ explicitly. Performing the integral over β, this yields

$$P(\boldsymbol{B}_{1:T},\boldsymbol{\alpha}) \propto P_0(\boldsymbol{\alpha})|\boldsymbol{A}|^{T/2}|\boldsymbol{\Sigma}_{\boldsymbol{J}}|^{-T/2} \times \left(\frac{1}{2}\sum_{t=1}^{T}\left[\boldsymbol{B}(t)'\,\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{B}(t) - \boldsymbol{B}(t)'\,\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{G}\boldsymbol{\Sigma}_{\boldsymbol{J}}^{-1}\boldsymbol{G}'\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{B}(t)\right]\right)^{-MT/2}. \tag{80}$$

Let us look at the form of this function at the limit $\boldsymbol{A} \to \infty$ (that is $\boldsymbol{\alpha}_i \to \infty$, for all i):

$$|\boldsymbol{A}|^{T/2}|\boldsymbol{\Sigma}_{\boldsymbol{J}}|^{-T/2} = |\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{J}}^{-1}|^{T/2} = |\boldsymbol{A}(\boldsymbol{G}'\,\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{G} + \boldsymbol{A})^{-1}|^{T/2} = |(\boldsymbol{G}'\,\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{G}\boldsymbol{A}^{-1} + \boldsymbol{I})^{-1}|^{T/2} \to 1,\ \text{as}\ \boldsymbol{A} \to \infty. \tag{81}$$

On the other hand

$$\boldsymbol{\Sigma}_{\boldsymbol{J}}^{-1} = (\boldsymbol{G}'\,\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{G} + \boldsymbol{A})^{-1} = \boldsymbol{A}^{-1}(\boldsymbol{G}'\,\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{G}\boldsymbol{A}^{-1} + \boldsymbol{I})^{-1} \to \boldsymbol{0},\ \text{as}\ \boldsymbol{A} \to \infty, \tag{82}$$

which implies

$$\left(\frac{1}{2}\sum_{t=1}^{T}\left[\boldsymbol{B}(t)'\,\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{B}(t) - \boldsymbol{B}(t)'\,\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{G}\boldsymbol{\Sigma}_{\boldsymbol{J}}^{-1}\boldsymbol{G}'\,\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{B}(t)\right]\right)^{-MT/2} \to \left(\frac{1}{2}\sum_{t=1}^{T}\left[\boldsymbol{B}(t)'\,\boldsymbol{\Sigma}_{\boldsymbol{G}}\boldsymbol{B}(t)\right]\right)^{-MT/2}, \text{as}\ \boldsymbol{A} \to \infty. \tag{83}$$

All in all, we have

$$P(\boldsymbol{B}_{1:T},\boldsymbol{\alpha}) = C \cdot G(\boldsymbol{\alpha},\boldsymbol{B}_{1:T})P_0(\boldsymbol{\alpha}), \tag{84}$$

where C is a numerical nonzero constant, and

$$0 < \lim_{\boldsymbol{\alpha}\to\infty} G(\boldsymbol{\alpha},\,\boldsymbol{B}_{1:T}) < \infty, \text{ for all } \boldsymbol{B}_{1:T}. \tag{85}$$

Thus, if we choose $P_0(\boldsymbol{\alpha})$ as in Eq. (65) we see that the integral of $P(\boldsymbol{B}_{1:T},\,\boldsymbol{\alpha})$ over $\boldsymbol{\alpha}$ diverges logarithmically rendering $P(\boldsymbol{B}_{1:T})$ infinite.

References

Ahlfors, S.P., Simpson, G.V., 2004. Geometrical interpretation of fMRI-guided MEG/EEG inverse estimates. NeuroImage 22, 323–332.

Auranen, T., Nummenmaa, A., Hämäläinen, M.S., Jääskeläinen, I.P., Lampinen, J., Vehtari, A., Sams, M., 2005. Bayesian analysis of the neuromagnetic inverse problem with ℓ^{p} norm priors. NeuroImage 26 (3), 870–884.

Baillet, S., Garnero, L., 1997. A Bayesian approach to introducing anatomo-functional priors in the EEG/MEG inverse problem. IEEE Trans. Biom. Eng. 44 (5).

Bernardo, J.M., Smith, A.F.M., 1994. Bayesian Theory. Wiley, Chichester.

Bertrand, C., Hamada, Y., Kado, H., 2001a. MRI prior computation and parallel tempering algorithm: a probabilistic resolution of the MEG/EEG inverse problem. Brain Topogr. 14 (1).

Bertrand, C., Ohmi, M., Suzuki, R., Kado, H., 2001b. A probabilistic solution to the MEG inverse problem via MCMC methods: The reversible jump and parallel tempering algorithms. IEEE Trans. Biomed. Eng. 48 (5).

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., 2005. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1 (2), 1–19.

Gelman, A., Meng, X.-L., 1998. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Stat. Sci. 13 (2), 163–185.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian data analysis, Chapman and Hall/CRC2nd ed.

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.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Markov Chain Monte Carlo in Practice. Chapman and Hall.

Harville, D.A., 1999. Matrix Algebra From a Statistician's Perspective. Springer-Verlag.

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).

Johnson, N.L., Kotz, S., Balakrishnan, N., 1994. Continuous univariate distributions, volume 1, Wiley Series in Probability and Mathematical Statistics2nd ed. John Wiley and Sons, Inc.

Kaipio, J.P., Somersalo, E., 2005. Statistical and Computational Inverse Problems. Springer.

Köhler, T., Wagner, M., Fuchs, M., Wischmann, H.-A., Drenckhahn, R.,

Theissen, A., 1996. Depth nornalization in MEG/EEG current density imaging. Conference Proceedings of the 18th Annual International Conference of the Engineering in Medicine and Biology Society of the IEEE.

Kincses, W.E., Braun, C., Kaiser, S., Grodd, W., Ackermann, H., Mathiak, K., 2003. Reconstruction of extended cortical sources for EEG and MEG based on a Monte-Carlo-Markov-chain estimator. Hum. Brain Mapp. 18, 100–110.

Lin, F.-H., Belliveau, J.W., Dale, A.M., Hämäläinen, M.S., 2006. Distributed current estimates using cortical orientation constraints. Hum. Brain Mapp. 27, 1–13.

Liu, J.S., 2001. Monte Carlo Strategies in Scientific Computing. Springer.

Liu, A.K., Belliveau, J.W., Dale, A.M., 1998. Spatiotemporal imaging of human brain activity using functional MRI constrained magnetoencephalography data: Monte Carlo simulations. Proc. Natl. Acad. Sci. 95 (15), 8945–8950.

Liu, A.K., Dale, A.M., Belliveau, J.W., 2002. Monte Carlo simulation studies of EEG and MEG localization accuracy. Hum. Brain Mapp. 16, 47–62.

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

Mosher, J.C., Lewis, P.S., Leahy, R.M., 1992. Multiple dipole modeling and localization from spatio-temporal MEG data. IEEE Trans. Biomed. Eng. 39 (6), 541–557.

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

Neal, R.M., 2001. Annealed importance sampling. Stat. Comput. 11, 125–139.

Neal, R.M., 2003. Slice sampling. Ann. Stat. 31 (3), 705–767.

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

Phillips, C., Rugg, M.D., Friston, K.J., 2002. Anatomically informed basis functions for EEG source localization: Combining functional and anatomical constraints. NeuroImage 16, 678–695.

Robert, C.P., Casella, G., 2004. Monte Carlo Statistical Methods, 2nd ed. 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.

Schmidt, D.M., George, J.S., Wood, C.C., 1999. Bayesian inference applied to the electro-magnetic inverse problem. Hum. Brain Mapp. 7, 195–212.

Tanaka, T., 2001. Information geometry of mean-field approximation. In: Opper, M., Saad, D. (Eds.), Advanced Mean Field Methods—Theory and Practice. MIT Press.

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

Uutela, K., Hämäläinen, M., Somersalo, E., 1999. Visualization of magnetoencephalographic data using minimum current estimates. NeuroImage 10, 173–180.

von Helmholtz, H., 1853. Ueber einige Gesetze der Vertheilung elektrischer Ströme in körperlichen Leitern, mit Anwendung auf die thierisch-elektrischen Versuche. Annalen der Physik und Chemie, 89 211–233, 353–377.