A Unified Bayesian Framework for MEG/EEG Source Imaging

David Wipf and Srikantan Nagarajan
Biomagnetic Imaging Lab, University of California, San Francisco
513 Parnassus Avenue, S362
San Francisco, CA 94143 USA
phone: +1.415.476.6888 *fax*: +1.415.502.4302
{dwipf,sri}@mrsc.ucsf.edu

Abstract

The ill-posed nature of the MEG (or related EEG) source localization problem requires the incorporation of prior assumptions when choosing an appropriate solution out of an infinite set of candidates. Bayesian approaches are useful in this capacity because they allow these assumptions to be explicitly quantified using postulated prior distributions. However, the means by which these priors are chosen, as well as the estimation and inference procedures that are subsequently adopted to affect localization, have led to a daunting array of algorithms with seemingly very different properties and assumptions. From the vantage point of a simple Gaussian scale mixture model with flexible covariance components, this paper analyzes and extends several broad categories of Bayesian inference directly applicable to source localization including empirical Bayesian approaches, standard MAP estimation, and multiple variational Bayesian (VB) approximations. Theoretical properties related to convergence, global and local minima, and localization bias are analyzed and fast algorithms are derived that improve upon existing methods. This perspective leads to explicit connections between many established algorithms and suggests natural extensions for handling unknown dipole orientations, extended source configurations, correlated sources, temporal smoothness, and computational expediency. Specific imaging methods elucidated under this paradigm include weighted minimum ℓ_2-norm, FOCUSS, MCE, VESTAL, sLORETA, ReML and covariance component estimation, beamforming, variational Bayes, the Laplace approximation, and automatic relevance determination (ARD). Perhaps surprisingly, all of these methods can be formulated as particular cases of covariance component estimation using different concave regularization terms and optimization rules, making general theoretical analyses and algorithmic extensions/improvements particularly relevant.

I. Introduction

Magnetoencephalography (MEG) and electroencephalography (EEG) use an array of sensors to take electromagnetic field (or voltage potential) measurements from on or near the scalp surface with excellent temporal resolution. In both cases, the observed field is generated by the same synchronous, compact current sources located within the brain. Because the mapping from source activity configuration to sensor measurement is many to one, accurately determining the spatial locations of these unknown sources is extremely difficult. The relevant localization problem can be posed as follows: The measured electromagnetic signal is $B \in \mathbb{R}^{d_b \times n}$, where d_b equals the number of sensors and n is the number of time points at which measurements are made. The unknown sources $S \in \mathbb{R}^{d_s \times n}$ are the (discretized) current values at d_s candidate locations distributed throughout the cortex. These candidate locations are obtained by segmenting a structural MR scan of a human subject and tesselating the gray matter surface with a set of vertices. B and S are related by the likelihood model

$$B = LS + \mathcal{E}, \tag{1}$$

where L is the so-called lead-field matrix, the i-th column of which represents the signal vector that would be observed at the scalp given a unit current source at the i-th vertex with a fixed orientation (flexible orientations can be incorporated by including multiple columns per location, one for each directional component). Multiple methods based on the physical properties of the brain and Maxwell's equations are available for this computation [37]. Finally, $\mathcal{E}$ is a noise-plus-interference term where we assume, for simplicity, that columns are drawn independently from $\mathcal{N}(0, \Sigma_\epsilon)$. However, temporal correlations can easily be incorporated if desired using the method outlined in [13]. Additionally, here we will mostly assume that Σ_ϵ is known; however, robust procedures for its estimation can be found in [25] and can naturally be incorporated into the proposed model.[1]

To obtain reasonable spatial resolution, the number of candidate source locations will necessarily be much larger than the number of sensors ($d_s \gg d_b$). The salient inverse problem then becomes the ill-posed estimation of these activity or source regions, which are reflected by the nonzero rows of the source estimate matrix $\hat{S}$. Because the inverse model is severely underdetermined, all efforts at source reconstruction are heavily dependent on prior assumptions, which in a Bayesian framework are embedded in the distribution $p(S)$.

To appear in *NeuroImage, 2008*. This research was supported by NIH grants R01DC04855 and R01DC006435. Also, special thanks to Hagai Attias, Karl Friston, Jason Palmer, Rey Ramírez, Kensuke Sekihara, and Johanna Zumer for helpful discussions.

[1] Joint estimation of S and Σ_ϵ can also potentially lead to identifiability issues.

If under a given experimental or clinical paradigm this $p(S)$ were somehow known exactly, then the posterior distribution $p(S|B)$ can be computed via Bayes rule:

$$p(S|B) = \frac{p(B|S)p(S)}{p(B)}. \tag{2}$$

This distribution contains all possible information about the unknown S conditioned on the observed data B [4]. Two fundamental problems prevent using $p(S|B)$ for source localization. First, for most priors $p(S)$, the distribution $p(B)$ given by

$$p(B) = \int p(B|S)p(S)dS \tag{3}$$

cannot be computed. Because this quantity, which is sometimes referred to as the model *evidence*, is required to compute posterior moments and is also sometimes used to facilitate model selection [13], [22], this deficiency can be very problematic. Of course if only a point estimate for S is desired, then this normalizing distribution may not be needed. For example, a popular estimator involves finding the value of S that maximizes the posterior distribution, often called the *maximum a posteriori* or MAP estimate, and is invariant to $p(B)$. However MAP estimates may be unrepresentative of posterior mass and are unfortunately intractable to compute for most $p(S)$ given reasonable computational resources. Secondly, we do not actually know the prior $p(S)$ and so some appropriate distribution must be assumed, perhaps based on neurophysiological constraints or computational considerations. In fact, it is this choice, whether implicitly or explicitly, that differentiates a wide variety of localization methods at a very high level.

Such a prior is often considered to be fixed and known, as in the case of minimum ℓ_2-norm approaches [2], minimum current estimation (MCE) [18], [44], FOCUSS [7], [16], sLORETA [30], and minimum variance beamformers [45]. Alternatively, a number of empirical Bayesian approaches have been proposed that attempt a form of model selection by using the data to guide the search for an appropriate $p(S)$. In this scenario, candidate priors are distinguished by a set of flexible hyperparameters $\boldsymbol{\gamma}$ that must be estimated via a variety of data-driven iterative procedures. Examples include hierarchical covariance component models [12], [24], [31], automatic relevance determination (ARD) [22], [26], [33], [34], [41], and several related variational Bayesian methods [13], [14], [27], [29], [36], [38].

While seemingly quite different in many respects, we present a generalized framework that encompasses all of these methods and points to intimate connections between algorithms. The underlying motivation here is to leverage analytical tools and ideas from machine learning, Bayesian inference, and convex analysis that have not as of yet been fully exploited in the context of MEG/EEG source localization. Specifically, we address how a simple Gaussian scale mixture prior [29] with flexible covariance components underlie and generalize all of the above. This process demonstrates a number of surprising similarities or out-right equivalences between what might otherwise appear to be very different methodologies. We also analyze several properties of this framework related to: (i) computational speed and convergence guarantees, (ii) properties of locally- and globally-optimal source reconstructions with respect to emergent cost functions, and (iii) theoretical localization bias recovering basic source configurations. Wherever possible, results are derived in a general setting and therefore can largely propagate down to specific cases such as the methods listed above. Overall, we envision that by providing a unifying theoretical perspective and comprehensive analyses, neuroelectromagnetic imaging practitioners will be better able to assess the relative strengths of many Bayesian strategies with respect to particular applications; it will also help ensure that different methods are used to their full potential and not underutilized. Moreover, this process points to several promising directions for future research.

The remainder of the paper is organized as follows. Section II introduces the basic Gaussian scale mixture model and three primary inference possibilities - empirical Bayes or hyperparameter MAP ($\boldsymbol{\gamma}$-MAP), basic MAP estimation of sources (S-MAP), and variational Bayesian methods (VB). Section III describes $\boldsymbol{\gamma}$-MAP in detail using the notion of automatic relevance determination for determining an optimal weighting of covariance components. Fast algorithms are derived as well as theoretical analysis of convergence, local minima, and localization bias. Section IV re-derives and extends basic S-MAP source estimation procedures in light of the Gaussian scale mixture model and discusses similar properties and connections with $\boldsymbol{\gamma}$-MAP. In particular, S-MAP is shown to be equivalent to empirical Bayes using a somewhat different regularization term for estimating hyperparameters. We then present two types of variational Bayesian algorithms in Section V, the mean-field and Laplace approximations, and discuss close connections with $\boldsymbol{\gamma}$-MAP. Finally, sLORETA and beamforming are analyzed in Section VI while Section VII gives a comprehensive summary of the ideas and contributions of the paper. Portions of this work have appeared previously in conference proceedings [51].

II. Bayesian Modeling Using General Gaussian Scale Mixtures and Arbitrary Covariance Components

In this section, we present a general-purpose Bayesian framework for source localization and discuss a central distinction between fixed-prior MAP estimation schemes and empirical Bayesian approaches that adopt a flexible, parameterized prior. While often derived using different assumptions and methodology, they can be related via a simple hierarchical structure based on general Gaussian scale mixture distributions with arbitrary covariance components. Numerous special cases of this model have been considered previously in the context of MEG source localization and related problems as will discussed in subsequent sections.

A. The Model

To begin we invoke the noise model from (1), which fully defines the assumed likelihood

$$p(B|S) \propto \exp\left(-\frac{1}{2}\|B - LS\|^2_{\Sigma_\epsilon^{-1}}\right), \tag{4}$$

where $\|X\|_{\Sigma_\epsilon^{-1}}$ denotes the weighted matrix norm $\sqrt{\text{trace}[X^T\Sigma_\epsilon^{-1}X]}$. While the unknown noise covariance can also be parameterized and seamlessly estimated from the data via the proposed paradigm, for simplicity we assume that Σ_ϵ is known and fixed. Next we adopt the following source prior for S:

$$p\left(S|\boldsymbol{\gamma}\right) \propto \exp\left(-\frac{1}{2}\text{trace}\left[S^T\Sigma_s^{-1}S\right]\right), \qquad \Sigma_s = \sum_{i=1}^{d_\gamma}\gamma_i C_i. \tag{5}$$

This is equivalent to applying a zero-mean Gaussian distribution with covariance Σ_s to each column of S. Here $\boldsymbol{\gamma} \triangleq [\gamma_1, \ldots, \gamma_{d_\gamma}]^T$ is a vector of d_γ nonnegative hyperparameters that control the relative contribution of each covariance basis matrix C_i. While the hyperparameters are unknown, the set of components

$$\mathcal{C} \triangleq \{C_i : i = 1, \ldots, d_\gamma\} \tag{6}$$

is assumed to be fixed and known. Such a formulation is extremely flexible however, because a rich variety of candidate covariance bases can be proposed as will be discussed in more detail in Section II-C. Moreover, this structure has been advocated by a number of others in the context of neuroelectromagnetic source imaging [12], [24], [31]. Finally, we assume a hyperprior on $\boldsymbol{\gamma}$ of the form

$$p(\boldsymbol{\gamma}) \propto \prod_{i=1}^{d_\gamma}\frac{1}{2}\exp\left[-f_i(\gamma_i)\right], \tag{7}$$

where each $f_i(\cdot)$ is an unspecified function that is assumed to be known. The implicit prior on S, obtained by integrating out (marginalizing) the unknown $\boldsymbol{\gamma}$, is known as a Gaussian scale mixture:

$$p(S) = \int p(S|\boldsymbol{\gamma})p(\boldsymbol{\gamma})d\boldsymbol{\gamma}. \tag{8}$$

B. Estimation and Inference

Estimation and inference using the proposed model can be carried out in multiple ways depending how the unknown quantities S and $\boldsymbol{\gamma}$ are handled. This leads to a natural partitioning of a variety of inverse methods. We briefly summarize three possibilities before discussing the details and close interrelationships in Sections III through V.

1) *Hyperparameter MAP/Empirical Bayes ($\boldsymbol{\gamma}$-MAP):* The first option comes from the recognition that if $\boldsymbol{\gamma}$ were known (and so Σ_s is known as well), then the conditional distribution $p(S|B,\boldsymbol{\gamma}) \propto p(B|S)p(S|\boldsymbol{\gamma})$ is a fully specified Gaussian distribution with mean and covariance given by

$$\begin{aligned} \text{E}_{p(S|B,\boldsymbol{\gamma})}\left[S\right] &= \Sigma_s L^T\left(\Sigma_\epsilon + L\Sigma_s L^T\right)^{-1}B & (9)\\ \text{Cov}_{p(\boldsymbol{s}_j|B,\boldsymbol{\gamma})}\left[\boldsymbol{s}_j\right] &= \Sigma_s - \Sigma_s L^T\left(\Sigma_\epsilon + L\Sigma_s L^T\right)^{-1}L\Sigma_s, \;\; \forall j, & (10)\end{aligned}$$

where $\boldsymbol{s}_j$ denotes the j-th column of S and individual columns are uncorrelated. It is then common to use the simple estimator $\hat{S} = \text{E}_{p(S|B,\boldsymbol{\gamma})}\left[S\right]$ for the unknown sources. However, since $\boldsymbol{\gamma}$ is actually not known, a suitable approximation $\hat{\boldsymbol{\gamma}} \approx \boldsymbol{\gamma}$ must first be found. One principled way to accomplish this is to integrate out the sources S and then solve

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \int p(B|S)p(S|\boldsymbol{\gamma})p(\boldsymbol{\gamma})dS. \tag{11}$$

This treatment is sometimes referred to as empirical Bayes [4], because the $\boldsymbol{\gamma}$-dependent prior on S, $p(S|\boldsymbol{\gamma})$, is empirically learned from the data, often using expectation-maximization (EM) algorithms which treat S as hidden data [8]. Additionally, the process of marginalization provides a natural regularizing mechanism that can shrink many elements of $\boldsymbol{\gamma}$ to exactly zero, in effect pruning the associated covariance component from the model, with only the relevant components remaining. Consequently, estimation under this model is sometimes called automatic relevance determination (ARD) [23], [26], [41] and will be explored further in Section III. This procedure can also be leveraged to obtain a rigorous lower bound on $\log p(B)$. While knowing $p(S|B)$ is useful for source estimation given a particular model, access to $p(B)$ (or equivalently $\log p(B)$) can assist model selection [13], [22].

2) *Source-Space MAP (S-MAP):* The second option follows directly from (8). Specifically, if we integrate out the unknown $\boldsymbol{\gamma}$, we can treat $p(S)$ as the effective prior and attempt to compute a MAP estimate of S via

$$\hat{S} = \arg\max_S \int p\left(B|S\right) p\left(S|\boldsymbol{\gamma}\right) p(\boldsymbol{\gamma}) d\boldsymbol{\gamma} = \arg\max_S p(B|S)p(S). \tag{12}$$

This option will be addressed in detail in Section IV and follows from ideas in [7], [29]. While it may not be immediately transparent, solving S-MAP also leads to a shrinking and pruning of superfluous covariance components. In short, this occurs because the hierarchical model upon which (12) is based leads to a convenient, iterative EM algorithm-based implementation, which treats the *hyperparameters* $\boldsymbol{\gamma}$ as hidden data and computes their expectation for the E-step. Over the course of learning, this expectation collapes to zero for many of the irrelevant hyperparameters, removing them from the model in much the same way as $\boldsymbol{\gamma}$-MAP. In fact, this affiliation will be made much more explicit in Section IV-B, where we convert S-MAP problems involving the proposed hierarchical model to a dual formulation in $\boldsymbol{\gamma}$-space.

3) *Variational Bayesian (VB) Approximations* A third possibility involves finding formal approximations to $p(S|B)$ as well as the marginal $p(B)$ using an intermediary approximation for $p(\boldsymbol{\gamma}|B)$. Because of the intractable integrations involved in obtaining either distribution, practical implementation requires additional assumptions leading to different types of approximation strategies as outlined in Section V. First, the so-called mean-field approximation [1], [3] makes the simplifying assumption that the joint distribution over unknowns S and $\boldsymbol{\gamma}$ factorizes, meaning $p(S, \boldsymbol{\gamma}|B) \approx \hat{p}(S|B)\hat{p}(\boldsymbol{\gamma}|B)$, where $\hat{p}(S|B)$ and $\hat{p}(\boldsymbol{\gamma}|B)$ are chosen to minimize the Kullback-Leibler divergence between the factorized and full posterior. This is accomplished via an iterative process akin to EM, effectively using two E-steps (one for S and one for $\boldsymbol{\gamma}$). It also produces a rigorous lower bound on $\log p(B)$ similar to $\boldsymbol{\gamma}$-MAP. A second possibility proposed in [14], [13] applies a second-order Laplace approximation to the posterior on the hyperparameters (after marginalizing over the sources S), which is then iteratively matched to the true posterior; the result can then be used to approximate $p(S|B)$ and $\log p(B)$. We emphasize two crucial points:

- Perhaps surprisingly, both of the VB methods described above lead to posterior approximations $\hat{p}(S|B) = p(S|B, \boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}})$, where $\hat{\boldsymbol{\gamma}}$ is equivalently computed via $\boldsymbol{\gamma}$-MAP.[2] Consequently, VB enjoys the same component pruning (or ARD) and much of the $\boldsymbol{\gamma}$-MAP analysis of Section III can be applied directly to VB.
- The only meaningful difference between VB and $\boldsymbol{\gamma}$-MAP, at least in the context of the proposed generative model, involves approximations to the model evidence $\log p(B)$, with different VB and $\boldsymbol{\gamma}$-MAP methods giving different estimates. These ideas will be discussed further in Section III-C and Section V.

While details and assumptions may differ $\boldsymbol{\gamma}$-MAP, S-MAP, and VB can all be leveraged to obtain a point estimate of S expressed in the Tikhonov regularized form [35]

$$\hat{S} = \hat{\Sigma}_s L^T \left(\Sigma_\epsilon + L\hat{\Sigma}_s L^T\right)^{-1} B \tag{13}$$

with

$$\hat{\Sigma}_s = \sum_i \hat{\gamma}_i C_i. \tag{14}$$

Importantly, because each method intrinsically sets $\hat{\gamma}_i = 0$ for superfluous covariance components, to an extent that will be made explicit in Sections III through V, we can in principle allow the cardinality of $\mathcal{C}$ to be very large and rely on the data-driven learning processes, either $\boldsymbol{\gamma}$-MAP, S-MAP, or VB, to select the most appropriate subset. Specific choices for $\mathcal{C}$ lead to a wide variety of established algorithms as will be discussed next.

C. Selections for $\mathcal{C}$

In the simplest case, the single-component assumption $\Sigma_s = \gamma_1 C_1 = \gamma_1 I$, where I is an identity matrix, leads to a weighted minimum-ℓ_2-norm solution [2]. More interesting covariance component terms have been used to effect spatial smoothness, depth bias compensation, and candidate locations of likely activity [24], [31]. With regard to the latter, it has been suggested that prior information about a source location can be codified by including a second term C_2, i.e., $\mathcal{C} = \{C_1, C_2\}$, with all zeros except a patch of 1's along the diagonal signifying a location of probable source activity, perhaps based on fMRI data [31]. For $\boldsymbol{\gamma}$-MAP, S-MAP, and VB, we will obtain a source estimate representing a trade-off (modulated by the relative values of the associated $\hat{\gamma}_1$ and $\hat{\gamma}_2$) between honoring the prior information imposed by C_2 and the smoothness implied by C_1. The limitation of this proposal is that we generally do not know, *a priori*, the regions where activity is occurring with both high spatial and temporal resolution. Therefore, we cannot reliably known how to choose an appropriate location-prior term C_2 in many situations.

[2]The associated hyperprior required for equality here need not be the same however; see Section V.

A potential solution to this dilemma is to try out many different (or even all possible) combinations of location priors. For example, if we assume the underlying source currents are formed from a collection of dipolar point sources located at each vertex of the lead-field grid, then we may choose $\mathcal{C} = \{\boldsymbol{e}_i \boldsymbol{e}_i^T : i = 1, \ldots, d_s\}$, where each $\boldsymbol{e}_i$ is a standard indexing vector of zeros with a '1' for the i-th element (and so $C_i = \boldsymbol{e}_i \boldsymbol{e}_i^T$ encodes a prior preference for a single dipolar source at location i).[3] This specification for the prior involves the counterintuitive addition of an unknown hyperparameter for every candidate source location which, on casual analysis may seem prone to severe overfitting (in contrast to [31], which uses only one or two fixed location priors). As suggested previously however, $\boldsymbol{\gamma}$-MAP, S-MAP, and VB all possess an intrinsic, sparsity-based regularization mechanism. This ameliorates the overfitting problem substantially and effectively reduces the space of possible active source locations by choosing a small relevant subset of active dipolar locations. Such a procedure has been empirically successful in the context of neural networks [26], compressed sensing [19], kernel machines [11], [41], and multiple dipole fitting for MEG [33], a significant benefit to the latter being that the optimal number of dipoles need not be known a priori. For S-MAP algorithms, we will show that this selection of $\mathcal{C}$ leads to a generalized version of FOCUSS and MCE [7], for $\boldsymbol{\gamma}$-MAP and VB, a general form of sparse Bayesian learning (SBL) [46], a variant of automatic relevance determination (ARD) [26].

In contrast, to model sources with some spatial extent, we can choose $C_i = \boldsymbol{\psi}_i \boldsymbol{\psi}_i^T$, where each ψ_i represents, for example, a $d_s \times 1$ geodesic neural basis vector that specifies an *a priori* weight location *and* activity extent [32], [34]. In this scenario, the number of covariance components satisfies $d_\gamma = v d_s$, where v is the number of scales we wish to examine in a multi-resolution decomposition, and can be quite large ($d_\gamma \approx 10^6$). The net result of this formulation is a source prior composed of a mixture of Gaussian kernels of varying scales and locations. The number of mixture components is learned from the data and is naturally forced to be small (sparse) to a degree which will be quantified later. In general, the methodology is quite flexible and other prior specifications can be included as well, such as temporal and spectral constraints [21] or the covariance components from [13].

In conclusion, there are two senses with which to understand the notion of covariance component selection. First, there is the selection of which components to include in the model before any estimation takes place, i.e., the choice of $\mathcal{C}$. Second, there is the selection that occurs *within* $\mathcal{C}$ as a natural byproduct of many hyperparameters being driven to zero during the learning process. Such components are necessarily pruned by the model; those that remain have therefore been 'selected' in some sense. Here we have argued that in many cases the later, data-driven selection can be used to ease the burden of often ad hoc user-specified selections. This notion will be examined in detail in later sections where theoretical and empirical evidence will be presented.

III. Hyperparameter MAP Estimation ($\boldsymbol{\gamma}$-MAP)

$\boldsymbol{\gamma}$-MAP obtains a point estimate for the unknown $\boldsymbol{\gamma}$ by first integrating out the unknown sources S producing the hyperparameter likelihood equation

$$p(B|\boldsymbol{\gamma}) = \int p\left(B|S\right) p\left(S|\boldsymbol{\gamma}\right) dS \propto \exp\left(-\frac{1}{2} B^T \Sigma_b^{-1} B\right), \tag{15}$$

where

$$\Sigma_b = \Sigma_\epsilon + L \Sigma_s L^T. \tag{16}$$

To estimate $\boldsymbol{\gamma}$ we then solve

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma}|B) = \arg\max_{\boldsymbol{\gamma}} p(B|\boldsymbol{\gamma}) p(\boldsymbol{\gamma}), \tag{17}$$

which is equivalent to minimizing the cost function

$$\mathcal{L}(\boldsymbol{\gamma}) \triangleq -2 \log p(B|\boldsymbol{\gamma}) p(\boldsymbol{\gamma}) \equiv \underbrace{\operatorname{trace}\left[C_b \Sigma_b^{-1}\right]}_{\text{data fit}} + \underbrace{\log |\Sigma_b|}_{\substack{\text{volume-based} \\ \text{regularization}}} + \underbrace{\tfrac{1}{n} \textstyle\sum_{i=1}^{d_y} f_i\left(\gamma_i\right),}_{\substack{\text{hyperprior-based} \\ \text{regularization}}} \tag{18}$$

where $C_b \triangleq n^{-1} B B^T$ is the empirical covariance. This cost function is composed of three parts. The first is a data fit term based on the dissimilarity between the empirical covariance C_b and the model covariance Σ_b; in general, this factor encourages $\boldsymbol{\gamma}$ to be large. The second term provides the primary regularizing or sparsifying effect, penalizing a measure of the volume formed by the model covariance Σ_b.[4] Since the volume of any high dimensional space is more effectively reduced by collapsing individual dimensions as close to zero as possible (as opposed to reducing all dimensions isometrically), this penalty term promotes a model covariance that is maximally degenerate (or non-spherical), which pushes elements of $\boldsymbol{\gamma}$ to exactly zero. Much more detail on this phenomena can be found in [46].

[3]Here we assume dipoles with orientations constrained to be orthogonal to the cortical surface; however, the method is easily extended to handle unconstrained dipoles [36], [49].

[4]The determinant of a matrix is equal to the product of its eigenvalues, which is a well-known volumetric measure.

Finally, the third term follows directly from the hyperprior, which we have thus far assumed to be arbitrary. This term can be useful for incorporating specific prior information, perhaps from fMRI data or some other imaging modality. It can also be used to expedite hyperparameter pruning or conversely, to soften the pruning process or prevent pruning altogether. One popular choice is the inverse-Gamma hyperprior given by

$$p_i(\gamma_i^{-1}) = \text{Gam}(a_i)^{-1} b_i^{a_i} \gamma_i^{1-a_i} \exp\left(-\frac{b_i}{\gamma_i}\right) \tag{19}$$

for the i-th hyperparameter, where $\text{Gam}(x) = \int_0^\infty z^{x-1}\exp(-z)dz$ is a standard Gamma function and $a_i, b_i \geq 0$ are shape parameters. In [38], the latter are determined by fMRI data. In the limit as $a_i, b_i \to 0$, then (19) converges to a noninformative Jeffreys prior; when optimization is performed in $\log\gamma_i$ space as is customary in some applications [41], the effective prior is flat and can therefore be ignored. In contrast, for $a_i = b_i \to \infty$, all hyperparameters are constrained to have equal value and essentially no learning (or pruning) occurs. Consequently, the standard weighted minimum ℓ_2-norm solution can be seen as a special case. A detailed treatment of what types of hyperpriors lead to covariance component pruning can be found in [50].

For simplicity of exposition, and because of proven effectiveness in practice, we will often concern ourselves with flat hyperpriors when considering the $\boldsymbol{\gamma}$-MAP option. Further theoretical justification for this selection can be found in Section III-B and [46]. Hence the third term in (18) vanishes and the only regularization will come from the $\log|\cdot|$ term. In this context, the optimization problem with respect to the unknown hyperparameters is sometimes referred to as type-II maximum likelihood [4]. It is also equivalent to the restricted maximum likelihood (ReML) cost function [12]. However, the proposed optimization methods (Section III-A) can be easily modified to handle a wide variety of alternative choices for $f_i(\cdot)$.

Regardless of how $\boldsymbol{\gamma}$ is optimized, once some $\hat{\boldsymbol{\gamma}}$ is obtained, we compute $\hat{\Sigma}_s$ using (14), which fully specifies our assumed empirical prior on S. To the extent that the 'learned' prior $p(S|\hat{\boldsymbol{\gamma}})$ is realistic, the resulting posterior quantifies regions of significant current density and point estimates for the unknown sources can be obtained by evaluating the posterior mean computed using (13).

A. Optimization

The primary objective of this section is to minimize (18) with respect to $\boldsymbol{\gamma}$. For simplicity, we will first present updates with $f_i(\gamma_i) = 0$ (i.e., a flat hyperprior).[5] We then address natural adaptations to more general cases (e.g., not just conjugate priors). Of course one option is to treat the problem as a general nonlinear optimization task and perform gradient descent or some other generic procedure. In contrast, here we will focus on methods specifically tailored for minimizing (18) using principled methodology. We begin with methods based directly on the EM algorithm and then diverge to alternatives that draw on convex analysis to achieve faster convergence. The algorithms presented here can also be applied to VB as discussed in Section V.

A common approach to this problem in the context of EEG/MEG comes from [12], [13] and is implemented via the SPM software platform.[6] Here a restricted maximum likelihood (ReML) method is proposed for optimization which utilizes what amounts to EM-based updates treating S as hidden data [8]. For the E-step, the mean and covariance of S are computed given some estimate of the hyperparameters $\hat{\boldsymbol{\gamma}}$. For the M-step, we then must update $\hat{\boldsymbol{\gamma}}$ using these moments as the true values. Unfortunately, the optimal value of $\hat{\boldsymbol{\gamma}}$ cannot be obtained in closed form for arbitrary $\mathcal{C}$ (although for certain special cases it can), so a second-order Fisher scoring procedure is adopted to approximate the desired solution. While effective for estimating small numbers of hyperparameters [24], [31], this approach requires inverting a $d_\gamma \times d_\gamma$ Fisher information matrix, which is not computationally feasible for large d_γ. Moreover, unlike exact EM implementations, there is no guarantee that (18) will be decreased at each iteration.

Consequently, here we present alternative optimization procedures that expand upon ideas from [22], [38], [41], [51], apply to the arbitrary covariance model discussed above, and naturally guarantee that $\gamma_i \geq 0$ for all i. All of these methods rely on reparameterizing the generative model such that the implicit M-step can be solved in closed form.

First we note that $\mathcal{L}(\boldsymbol{\gamma})$ only depends on the data B through the $d_b \times d_b$ sample correlation matrix C_b. Therefore, to reduce the computational burden, we replace B with a matrix $\mathsf{B} \in \mathbb{R}^{d_b \times \text{rank}(B)}$ such that $\mathsf{B}\mathsf{B}^T = BB^T$. This removes any per-iteration dependency on n, which can potentially be large, without altering that actual cost function. It also implies that, for purposes of computing $\boldsymbol{\gamma}$, the number of columns of S is reduced to match $\text{rank}(B)$.

Next we introduce the decomposition

$$S = \sum_{i=1}^{d_\gamma} A_i \mathsf{S}_i = A\mathsf{S}, \tag{20}$$

where each A_i is selected such that $A_i A_i^T = C_i$ and $A \triangleq [A_1, \cdots, A_{d_\gamma}]$, $\mathsf{S} \triangleq [\mathsf{S}_1^T, \cdots, \mathsf{S}_{d_\gamma}^T]^T$. In Section II-C, both $\boldsymbol{e}_i$ and $\boldsymbol{\psi}_i$ were special cases of A_i. Letting $\mathsf{L} \triangleq [\mathsf{L}_1, \ldots, \mathsf{L}_{d_\gamma}] = L[A_1, \cdots, A_{d_\gamma}]$, this allows us to re-express the original hierarchical

[5] Note also that as n becomes large, the effect of the hyperprior becomes inconsequential anyway.

[6] www.fil.ion.ucl.ac.uk/spm/software/

Bayesian model as

$$\begin{aligned} p(\mathsf{B}|\mathsf{S}) &\propto \exp\left(-\frac{1}{2}\|\mathsf{B}-\mathsf{L}\mathsf{S}\|^2_{\Sigma_\epsilon^{-1}}\right) \\ p(\mathsf{S}_i|\gamma_i) &\propto \exp\left(-\frac{1}{2\gamma_i}\|\mathsf{S}_i\|^2_{\mathcal{F}}\right), \ \ \forall i=1,\ldots,d_\gamma, \end{aligned} \tag{21}$$

where $\|X\|_{\mathcal{F}}$ is the standard Frobenius norm $\sqrt{\text{trace}[X^TX]}$. The hyperprior remains unaltered. It is easily verified by the rules for transformation of random variables that (21) and the original model are consistent. It also follows that

$$\Sigma_b = \Sigma_\epsilon + L\left(\sum_{i=1}^{d_\gamma}\gamma_i C_i\right)L^T = \Sigma_\epsilon + \sum_{i=1}^{d_\gamma}\gamma_i \mathsf{L}_i\mathsf{L}_i^T = \Sigma_\epsilon + \mathsf{L}\Sigma_{\mathsf{s}}\mathsf{L}^T, \tag{22}$$

where Σ_{s} is the diagonal, $\boldsymbol{\gamma}$-dependent prior covariance of the pseudo-sources.[7]

1) EM Algorithm: We will now treat S as the hidden data as opposed to the actual sources S and exploit the diagonal structure of the pseudo-source covariance Σ_{s} in implementing the EM algorithm. For the E-step, given the value of $\boldsymbol{\gamma}$ at the k-th iteration, the pseudo-sources are Gaussian with mean and covariance given by

$$\begin{aligned} \mathrm{E}_{p(\mathsf{S}|\mathsf{B},\boldsymbol{\gamma}^{(k)})}[\mathsf{S}] &= \Sigma_{\mathsf{s}}^{(k)}\mathsf{L}^T\left(\Sigma_\epsilon + \mathsf{L}\Sigma_{\mathsf{s}}^{(k)}\mathsf{L}^T\right)^{-1}\mathsf{B} \\ \mathrm{Cov}_{p(\mathsf{s}_j|\mathsf{B},\boldsymbol{\gamma}^{(k)})}[\mathsf{s}_j] &= \Sigma_{\mathsf{s}}^{(k)} - \Sigma_{\mathsf{s}}^{(k)}\mathsf{L}^T\left(\Sigma_\epsilon + \mathsf{L}\Sigma_{\mathsf{s}}^{(k)}\mathsf{L}^T\right)^{-1}\mathsf{L}\Sigma_{\mathsf{s}}^{(k)}, \ \ \forall j, \end{aligned} \tag{23}$$

where s_j is the j-the column of S (and columns are independent of one another). The M-step then solves

$$\begin{aligned} \boldsymbol{\gamma}^{(k+1)} &\rightarrow \arg\min_{\boldsymbol{\gamma}} \mathrm{E}_{p(\mathsf{S}|\mathsf{B},\boldsymbol{\gamma}^{(k)})}\left[-\log p\left(\mathsf{B},\mathsf{S},\boldsymbol{\gamma}\right)\right], \\ &= \arg\min_{\boldsymbol{\gamma}} \mathrm{E}_{p(\mathsf{S}|\mathsf{B},\boldsymbol{\gamma}^{(k)})}\left[-\log p\left(\mathsf{S}|\boldsymbol{\gamma}\right)\right] \\ &= \arg\min_{\boldsymbol{\gamma}} \sum_i \left(nr_i\log\gamma_i + \frac{[\bar{\mathsf{S}}_i^{(k+1)}]^2}{\gamma_i}\right), \end{aligned} \tag{24}$$

where

$$\bar{\mathsf{S}}_i^{(k+1)} \triangleq \sqrt{\mathrm{E}_{p(\mathsf{S}_i|\mathsf{B},\boldsymbol{\gamma}^{(k)})}[\|\mathsf{S}_i\|^2_{\mathcal{F}}]} = \sqrt{\left\|\mathrm{E}_{p(\mathsf{S}_i|\mathsf{B},\boldsymbol{\gamma}^{(k)})}[\mathsf{S}_i]\right\|^2_{\mathcal{F}} + \text{trace}\left[\mathrm{Cov}_{p(\mathsf{S}_i|\mathsf{B},\boldsymbol{\gamma}^{(k)})}[\mathsf{S}_i]\right]}. \tag{25}$$

The expression (24) can be computed analytically because the diagonal pseudo-source covariance effectively decouples the maximization problem, allowing a closed-form solution for each γ_i individually. This leads to the M-step updates

$$\gamma_i^{(k+1)} \rightarrow \frac{1}{nr_i}\left[\bar{\mathsf{S}}_i^{(k+1)}\right]^2. \tag{26}$$

The combined update for the $(k+1)$-th iteration (obtained by plugging (23) into (26) is then

$$\gamma_i^{(k+1)} \rightarrow \frac{1}{nr_i}\left\|\gamma_i^{(k)}\mathsf{L}_i^T\left(\Sigma_b^{(k)}\right)^{-1}\mathsf{B}\right\|^2_{\mathcal{F}} + \frac{1}{r_i}\text{trace}\left[\gamma_i^{(k)}I - \gamma_i^{(k)}\mathsf{L}_i^T\left(\Sigma_b^{(k)}\right)^{-1}\mathsf{L}_i\gamma_i^{(k)}\right]. \tag{27}$$

Given the appropriate simplifying assumptions on the form of Σ_s and some additional algebraic manipulations, (27) reduces to the algorithm from [38], even though the later was derived under a very different variational Bayes (VB) framework. Section V will examine why this should be the case.

The EM procedure can also be modified to provide an alternative update that handles the case where $f_i(\gamma_i) \neq 0$ in a variety of cases, including when $p(\boldsymbol{\gamma})$ is not a conjugate prior (see Appendix B for an alternative derivation of EM). For example, if any of the following conditions hold, then a simple EM algorithm (or generalized EM algorithm [8]) can be derived:

- $\min_{\gamma_i}\left[\frac{c_1}{\gamma_i} + c_2\log\gamma_i + f_i(\gamma_i)\right]$ has an analytical solution for all i and $c_1, c_2 \geq 0$.
- $f_i(\gamma_i)$ is concave with respect to γ_i.
- $f_i(\gamma_i)$ is concave with respect to γ_i^{-1}.

While space precludes a detailed treatment of each of these here, the end result is a straightforward modification to the M-step. For example, assuming the third condition above holds, then we can derive the generalized EM update

$$\gamma_i^{(k+1)} \rightarrow \frac{1}{nr_i}\left\|\gamma_i^{(k)}\mathsf{L}_i^T\left(\Sigma_b^{(k)}\right)^{-1}\mathsf{B}\right\|^2_{\mathcal{F}} + \frac{1}{r_i}\text{trace}\left[\gamma_i^{(k)}I - \gamma_i^{(k)}\mathsf{L}_i^T\left(\Sigma_b^{(k)}\right)^{-1}\mathsf{L}_i\gamma_i^{(k)}\right] + \frac{1}{nr_i}\phi_i'\left[\left(\gamma_i^{(k)}\right)^{-1}\right], \tag{28}$$

[7]Because each S_i (or more accurately, each column of S_i) is effectively described by a diagonal prior covariance parameterized by γ_i, and since each S_i is assumed independent, the total covariance of S will be diagonal. The diagonal section associated with S_i of length r_i is given by $\gamma_i\mathbf{1}_{r_i}$, where $\mathbf{1}_{r_i}$ is an $r_i \times 1$ vector of ones.

where $\phi_i(\gamma_i^{-1}) \triangleq f_i(\gamma_i)$ and $\phi_i'(\gamma_i^{-1})$ is the derivative with respect to γ_i^{-1}. This is gauranteed to reduce (or leave unchanged) the cost function (18) at every iteration.

Noting that off-diagonal elements of the trace term on the righthand side of (27) and (28) need not be computed, the per-iteration cost of EM updates is at most $\mathrm{O}\left(d_b^2 \sum_{i=1}^{d_\gamma} r_i\right) \leq \mathrm{O}\left(d_b^3 d_\gamma\right)$. This expense can be significantly reduced still further in cases where different pseudo lead-field components, e.g., some L_i and L_j, contain one or more columns in common. This situation occurs if we desire to use the geodesic basis functions with flexible orientation constraints, as opposed to the fixed orientations assumed in Section II-C. In general, the linear dependence on d_γ is one of the attractive aspects of this method, effectively allowing for extremely large numbers of hyperparameters and covariance components.

The problem then with (27) is not the per-iteration complexity but the convergence rate, which we have observed to be prohibitively slow in practical situations with high-resolution lead-field matrices and large numbers of hyperparameters (see Appendix A for a toy example and Appendix B for a brief description of why EM can be slow). The only reported localization results using this type of EM algorithm are from [38], where a relatively low resolution lead-field matrix is used in conjunction with a simplifying heuristic that constrains some of the hyperparameter values.

2) MacKay Updates: To avoid constraints based on computational limitations, which can potentially degrade the quality of source estimates, a faster update rule is needed. To this end, we modified the procedure of [22], which involves taking the gradient of $\mathcal{L}(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$, rearranging terms, and forming the fixed-point update

$$\gamma_i^{(k+1)} \rightarrow \frac{1}{n} \left\| \gamma_i^{(k)} \mathsf{L}_i^T \left(\Sigma_b^{(k)}\right)^{-1} \mathsf{B} \right\|_{\mathcal{F}}^2 \left(\mathrm{trace}\left[\gamma_i^{(k)} \mathsf{L}_i^T \left(\Sigma_b^{(k)}\right)^{-1} \mathsf{L}_i \right] \right)^{-1}. \tag{29}$$

The complexity of each iteration is the same as before, only now the convergence rate can be orders of magnitude faster as shown in Appendix A. And the updates can be extended to a variety of hyperpriors $f_i(\gamma_i)$, although we omit details here. As an additional point of interest, for the appropriate selection of $\mathcal{C}$, the hyperparameters obtained after a single iteration of (29) lead to the sLORETA estimate of source activity. This connection will be discussed further in Section VI-B.

Unlike the EM method, one criticism of (29) is that there currently exists no proof that it represents a proper descent function (meaning $\mathcal{L}\left(\boldsymbol{\gamma}^{(k+1)}\right) \leq \mathcal{L}\left(\boldsymbol{\gamma}^{(k)}\right)$ at every iteration), although we have never observed an exception in practice. While we can show that (29) is equivalent to iteratively solving a particular min-max problem in search of a saddle point (see Appendix B), provable descent is still suspect.

3) Convexity-Based Approach: We now derive a related update rule that is both significantly faster than EM *and* is proven to produce $\boldsymbol{\gamma}$ vectors such that $\mathcal{L}\left(\boldsymbol{\gamma}^{(k+1)}\right) \leq \mathcal{L}\left(\boldsymbol{\gamma}^{(k)}\right)$ for every iteration k. Using a dual-form representation of $\mathcal{L}(\boldsymbol{\gamma})$ that leads to a tractable auxiliary cost function, this update is given by

$$\gamma_i^{(k+1)} \rightarrow \frac{\gamma_i^{(k)}}{\sqrt{n}} \left\| \mathsf{L}_i^T \left(\Sigma_b^{(k)}\right)^{-1} \mathsf{B} \right\|_{\mathcal{F}} \left(\mathrm{trace}\left[\mathsf{L}_i^T \left(\Sigma_b^{(k)}\right)^{-1} \mathsf{L}_i \right] \right)^{-1/2}. \tag{30}$$

Details of the derivation are given in Appendix B and empirical validation on a simple toy problem is shown in Appendix A. The auxiliary function upon which (30) is based can also be iteratively minimized by solving a series of convex, weighted second-order cone (SOC) problems [48], although specifics are beyond the scope of this paper. Consequently, $\boldsymbol{\gamma}$-MAP can be implemented using standard convexity-based optimization tools designed for large problems. This perspective also leads to a variety of desirable analytical properties [48].

Finally, the extension to non-zero $f_i(\gamma_i)$ is very straightforward in two situations:

- $\min_{\gamma_i}\left[\frac{c_1}{\gamma_i} + c_2\gamma_i + f_i(\gamma_i)\right]$ has an analytical solution for all i and $c_1, c_2 \geq 0$.
- $f_i(\gamma_i)$ is concave for all i.

The second condition from above leads to the modified update

$$\gamma_i^{(k+1)} \rightarrow \frac{\gamma_i^{(k)}}{\sqrt{n}} \left\| \mathsf{L}_i^T \left(\Sigma_b^{(k)}\right)^{-1} \mathsf{B} \right\|_{\mathcal{F}} \left(\mathrm{trace}\left[\mathsf{L}_i^T \left(\Sigma_b^{(k)}\right)^{-1} \mathsf{L}_i + \frac{1}{n} f_i'(\gamma_i) \right] \right)^{-1/2}, \tag{31}$$

which is guaranteed to reduce (18) at each iteration (this result follows from the derivation in Appendix B).

B. Analysis of $\boldsymbol{\gamma}$-MAP

Previously, we have claimed that the $\boldsymbol{\gamma}$-MAP process naturally forces excessive/irrelevant hyperparameters to converge to zero, thereby reducing model complexity. Note that, somewhat counterintuitively, this occurs even when a flat hyperprior is assumed. While this observation has been verified empirically by ourselves and others in various application settings, there has been relatively little corroborating theoretical evidence, largely because of the difficulty in analyzing the potentially multimodal, non-convex $\boldsymbol{\gamma}$-MAP cost function. As such, we provide the following result:

Theorem 1: Every local minimum of the generalized $\boldsymbol{\gamma}$-MAP cost function (18) is achieved at a solution with at most $\text{rank}(B)d_b \leq d_b^2$ nonzero hyperparameters if $f_i(\gamma_i)$ is concave and non-decreasing for all i (this includes $f_i(\gamma_i) = 0$, or equivalently $p_i(\gamma_i) = 1$).

The proof follows from a straightforward extension of [52, Theorem 2] and the fact that the $\boldsymbol{\gamma}$-MAP cost only depends on the $\text{rank}(B)$ matrix $C_b = n^{-1}BB^T$. Theorem 1 comprises a worst-case bound that is only tight in very nuanced situations; in practice, for any reasonable value of Σ_ϵ, the number of nonzero hyperparameters is typically much smaller than d_b. The bound holds for all Σ_ϵ, including $\Sigma_\epsilon = 0$, indicating that some measure of hyperparameter pruning, and therefore covariance component pruning, is built into the $\boldsymbol{\gamma}$-MAP framework irrespective of the noise-based regularization. Moreover, the number of nonzero hyperparameters decreases monotonically to zero as Σ_ϵ is increased. And so there is always some $\Sigma_\epsilon = \Sigma'_\epsilon$ sufficiently large such that all hyperparameters converge to exactly zero. Therefore, we can be confident that the pruning mechanism of $\boldsymbol{\gamma}$-MAP is not merely an empirical phenomena. Nor is it dependent on a particular sparse hyperprior, the result holds when a flat (uniform) hyperprior is assumed.

The number of observation vectors n also plays an important role in shaping $\boldsymbol{\gamma}$-MAP solutions. Increasing n has two primary benefits: (i) it facilitates convergence to the global minimum (as opposed to getting stuck in a suboptimal extrema) and (ii), it improves the quality of this minimum by mitigating the effects of noise [46]. With perfectly correlated (dipolar) sources, primarily only the later benefit is in effect. For example, with low noise and perfectly correlated sources, the estimation problem reduces to an equivalent problem with $n = 1$, so the local minima profile of the cost function does not improve with increasing n. Of course standard $\boldsymbol{\gamma}$-MAP can still be very effective in this scenario [34], [46]. In contrast, geometric arguments can be made to show that uncorrelated sources with large n offer the best opportunity for local minima avoidance. More details about how $\boldsymbol{\gamma}$-MAP functions when temporally correlated sources are present can be found in [47] as well as Section VI-A where beamforming is discussed. Finally, a third benefit to using $n > 1$ is that it leads to temporal smoothing of estimated time courses (i.e., rows of $\hat{S}$). This occurs because the selected covariance components do not change across time, as would be the case if a separate set of hyperparameters were estimated at each time point. This distinction will be illustrated in subsequent papers.

Further theoretical support for $\boldsymbol{\gamma}$-MAP is possible in the context of localization bias assuming simple source configurations. For example, substantial import has been devoted to quantifying localization bias when estimating a single dipolar source. Recently it has been shown, both empirically [30] and theoretically [40], that the sLORETA algorithm has zero location bias under this condition at high SNR. Because $\boldsymbol{\gamma}$-MAP can be viewed as an iterative enhancement of the sLORETA as mentioned previously (and discussed further in Section VI-B), the question naturally arises whether $\boldsymbol{\gamma}$-MAP retains this desirable property. In fact, it can be shown that this is indeed the case in two general situations. We assume that the lead-field matrix L represents a sufficiently high sampling of the source space such that any active dipole aligns with some lead-field column. Unbiasedness can also be shown in the continuous case for both sLORETA and general $\boldsymbol{\gamma}$-MAP, but the discrete scenario is more straightforward and of course more relevant to any practical task.

Theorem 2: Assume that $\mathcal{C}$ includes (among others) d_s covariance components of the form $C_i = \boldsymbol{e}_i\boldsymbol{e}_i^T$. Then in the absence of noise (high SNR), and assuming $f_i(\gamma_i) = 0$, $\boldsymbol{\gamma}$-MAP has zero localization bias when estimating a single dipolar source, regardless of the value of n.

If we are willing to tolerate some additional assumptions, then this theorem can be significantly expanded. For example, multiple dipolar sources can be localized with zero bias if they are perfectly uncorrelated (orthogonal) across time and assuming some mild technical conditions [46]. This result formalizes the notion, mentioned above, that $\boldsymbol{\gamma}$-MAP performs best with uncorrelated sources. Turning to the more realistic scenario where noise is present gives the following:

Theorem 3: Let $\mathcal{C}$ be constructed as above and assume the noise covariance matrix Σ_ϵ is known up to a scale factor and $f_i(\gamma_i) = 0$. Then given a single dipolar source, in the limit as n becomes large the $\boldsymbol{\gamma}$-MAP cost function is unimodal, and a source estimate with zero localization bias achieves the global minimum.

For most reasonable lead-fields and covariance components, this global minimum will be unique, and so the unbiased solution will be found as in the noiseless case. As for proofs, all the theoretical results pertaining to localization bias in this section follow from local minima properties of ML covariance component estimates. While details have been deferred to [47], the basic idea is that if the outerproduct BB^T can be expressed as some non-negative linear combination of the available covariance components, then the $\boldsymbol{\gamma}$-MAP cost function is unimodal and $\hat{\Sigma}_b = n^{-1}BB^T$ at any minimizing solution. This $\hat{\Sigma}_b$ in turn produces unbiased source estimates in a variety of situations.

While theoretical results of this kind are admittedly limited in applicability, many iterative Bayesian schemes in fact fail to exhibit similar performance. For example, all of the S-MAP algorithms we are aware of, including FOCUSS and MCE methods, provably have a localization bias in the general setting, although in particular cases there may be no bias. When we move to more complex source configurations with possible correlations and noise, theoretical results are not available;

however, empirical tests provide a useful means of comparison (results will be presented in forthcoming papers).

C. An Alternative Perspective on γ-MAP and a Bound on $\log p(B)$

For purposes of model selection, a rigorous bound on $\log p(B)$ can be derived using principles from convex analysis that have been successfully applied in general-purpose probabistic graphical models [20]. The basic idea follows from an intrinsic property of Gaussian scale mixtures outlined in [29] that readily adapts to the present context. As in Section III-A, we choose to work with the pseudo-source parameterization for convenience. To begin, we define the function $g_i(\cdot)$ via

$$g_i(\|\mathsf{S}_i\|_{\mathcal{F}}^2) \triangleq -2\log \int p\left(\mathsf{S}_i|\gamma_i\right) p_i(\gamma_i) d\gamma_i, \tag{32}$$

which is always possible since $p(\mathsf{S}_i|\gamma_i)$ only depends on S_i through the Frobenius norm operator.[8] It can be shown that, for any $p_i(\gamma_i)$, the $g_i(\cdot)$ which results from the above marginalization will be a concave, non-decreasing function of its argument [29]. From convex analysis [6], this property implies that $g_i(\cdot)$ can always be expressed as a minimum over upper-bounding lines or

$$g_i(z) = \min_{\alpha \geq 0} \left[\alpha z - g_i^*(\alpha)\right], \tag{33}$$

where $g_i^*(\cdot)$ is the concave conjugate of $g_i(\cdot)$ defined via the duality relationship

$$g_i^*(\alpha) = \min_{z \geq 0} \left[\alpha z - g_i(z)\right]. \tag{34}$$

For what follows, it is not crucial that these expression be fully digested; however, [20] contains a very accessible treatment of conjugate duality and related concepts that are applicable here. The implication of (33) is that now $p(\mathsf{S}_i)$ can be written in the variational form

$$\begin{aligned} p(\mathsf{S}_i) &= \exp\left[-\frac{1}{2} g_i\left(\|\mathsf{S}_i\|_{\mathcal{F}}^2\right)\right] \\ &= \exp\left[-\frac{1}{2}\min_{\alpha \geq 0}\left(\alpha \|\mathsf{S}_i\|_{\mathcal{F}}^2 - g_i^*(\alpha)\right)\right] \\ &= \max_{\gamma_i \geq 0} \exp\left[-\frac{1}{2\gamma_i}\|\mathsf{S}_i\|_{\mathcal{F}}^2\right] \exp\left[\frac{1}{2} g_i^*\left(\gamma_i^{-1}\right)\right], \end{aligned} \tag{35}$$

where we have used the substitution $\alpha = \gamma_i^{-1}$ to highlight the connection with $\boldsymbol{\gamma}$-MAP. Now suppose that we would like to compute a tractable approximation to $p(B, \mathsf{S})$ that can be normalized to produce some $\hat{p}(\mathsf{S}|B)$ (and therefore a posterior estimate $\hat{p}(S|B)$ of the real sources as well) and provides a bound on $\log p(B)$ that can be used for model selection. Both objectives are easily accomplished using (35). By dropping the maximization, we obtain the rigorous lower bound on $p(B, \mathsf{S})$ given by

$$p(B, \mathsf{S}) = p(B|\mathsf{S})p(\mathsf{S}) \geq p(B, \mathsf{S}; \boldsymbol{\gamma}) \triangleq p(B|\mathsf{S}) \prod_i \exp\left[-\frac{1}{2\gamma_i}\|\mathsf{S}_i\|_{\mathcal{F}}^2\right] \exp\left[\frac{1}{2} g_i^*\left(\gamma_i^{-1}\right)\right]. \tag{36}$$

To find the optimal approximation, we minimize

$$\begin{aligned} \mathcal{L}(\boldsymbol{\gamma}) &\triangleq \int \left|p(B, \mathsf{S}) - p(B, \mathsf{S}; \boldsymbol{\gamma})\right| d\mathsf{S} \equiv -\int p(B, \mathsf{S}; \boldsymbol{\gamma}) d\mathsf{S} \\ &\equiv \operatorname{trace}\left[C_b \Sigma_b^{-1}\right] + \log|\Sigma_b| - \sum_{i=1}^{d_y}\left[r_i \log\gamma_i + \frac{1}{n} g_i^*\left(\gamma_i^{-1}\right)\right] \end{aligned} \tag{37}$$

to obtain some $\hat{\boldsymbol{\gamma}}$. This is of course the same as the $\boldsymbol{\gamma}$-MAP cost function with $f_i(\gamma_i) = -nr_i \log\gamma_i - g_i^*\left(\gamma_i^{-1}\right)$. If these $f_i(\gamma_i)$ so obtained are concave and non-decreasing, then the sparsity bound from Theorem 1 will also hold. Additionally, none of the $\boldsymbol{\gamma}$-MAP optimization methods detailed in Section III-A require the hyperprior to be normalized, so these approaches can be applied equally well here when convenient for optimizing (37) to give some $\hat{\boldsymbol{\gamma}}$. Once this value is determined, the approximate distribution

$$\hat{p}(\mathsf{S}|B) = \frac{p(B, \mathsf{S}; \hat{\boldsymbol{\gamma}})}{\int p(B, \mathsf{S}; \hat{\boldsymbol{\gamma}}) d\mathsf{S}} \tag{38}$$

[8]Thus far we have assumed that γ_i parameterizes a linear weighting on covariance components. However, the function $g_i(\cdot)$ is invariant to this parameterization because of the marginalization. Consequently, for what follows, the particular parameterization (e.g., precision vs. variance or some other non-linear variant) is irrelevant.

is Gaussian with moments given by (23) computed using $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}$. Finally, for purposes of model comparison, the convex variational formulation gives the following rigorous lower bound:

$$\begin{aligned} \log p(B) &\geq \log \int p(B, \mathsf{S}; \hat{\boldsymbol{\gamma}}) d\mathsf{S} \\ &= -\frac{n}{2}\left[\text{trace}\left[C_b \hat{\Sigma}_b^{-1}\right] + \log\left|\hat{\Sigma}_b\right| - \sum_{i=1}^{d_y}\left[r_i \log \hat{\gamma}_i + g_i^*\left(\hat{\gamma}_i^{-1}\right)\right] + \left(d_b - \sum_i r_i\right)\log 2\pi\right]. \end{aligned} \tag{39}$$

Later in Section V we will describe related bounds produced by alternative variational Bayesian methods.

IV. Source-Space MAP Estimation (S-MAP)

S-MAP methods operate in source space by intergrating out $\boldsymbol{\gamma}$ and then computing a MAP estimate $\hat{S}$. There are (at least) two ways to go about this depending on how the integration is performed with respect to the hierarchical model from Section II. The most direct method involves solving

$$\hat{S} = \arg\max_S \int p\left(B|S\right) p\left(S|\boldsymbol{\gamma}\right) p(\boldsymbol{\gamma}) d\boldsymbol{\gamma}. \tag{40}$$

However, in a general setting this can be a difficult optimization problem and furthermore, the nature of the underlying cost function is not immediately transparent. Consequently, we advocate an indirect alternative utilizing the pseudo-source decomposition given by S described previously, which leads to an efficient EM implementation and a readily interpretable cost function. It also demonstrates that both FOCUSS and MCE can be viewed as EM algorithms that are readily generalized to handle more complex spatio-temporal constraints. Explicitly, we will minimize

$$\begin{aligned} \mathcal{L}(\mathsf{S}) &\triangleq -2\log \int p\left(B|\mathsf{S}\right) p\left(\mathsf{S}|\boldsymbol{\gamma}\right) p(\boldsymbol{\gamma}) d\boldsymbol{\gamma} \\ &= -2\log p\left(B|\mathsf{S}\right) p(\mathsf{S}) \\ &\equiv \|B - \mathsf{L}\mathsf{S}\|^2_{\Sigma_\epsilon^{-1}} + \sum_{i=1}^{d_{\boldsymbol{\gamma}}} g_i\left(\|\mathsf{S}_i\|^2_{\mathcal{F}}\right), \end{aligned} \tag{41}$$

where $g_i(\cdot)$ is defined as in (32).

For many choices of the hyperprior, the associated $g_i(\cdot)$ may not be available in closed form. Moreover, it is often more convenient and transparent to directly assume the form of $g_i(\cdot)$ rather than infer its value from some postulated hyperprior. In fact, based on results in [29], virtually any non-decreasing, concave function $g_i(\cdot)$ of interest can be generated by the proposed hierarchical model. In other words, there will always exist some $p_i(\gamma_i)$, possibly improper, such that the stated Gaussian mixture representation will produce any desired concave $g_i(\cdot)$.[9] A generalized version of MCE and FOCUSS can be produced from the selection

$$g_i(z) = c_i z^{p/2}, \tag{42}$$

which is concave and amenable to a Gaussian scale-mixture representation for any $p \in (0, 2]$ and constant $c_i > 0$. This is key because, even if the corresponding $p_i(\gamma_i)$ that produced a particular $g_i(\cdot)$ is unavailable or unwieldy, effective EM update rules can still be developed based on the associated unknown scale-mixture [29] as will be discussed in Section IV-A. The value of c_i can be chosen to account for dimensionality differences of the pseudo-sources S_i, provide symmetry with $\boldsymbol{\gamma}$-MAP, and to allow update rules and cost functions that reduce to regular FOCUSS and MCE with the appropriate simplifications. While space precludes a detailed treatment, the selection $c_i = \frac{2}{p} r_i^{\frac{2-p}{2}}$ accomplishes these objectives. To retrieve the exact FOCUSS and MCE cost functions, we must further assume $n = 1$, $\Sigma_\epsilon = \lambda I$ (where λ is a non-negative scalar), and $A_i = \boldsymbol{e}_i$ for $i = 1, \ldots, d_s$. Under these assumptions (41) reduces to

$$\mathcal{L}(\boldsymbol{s}) = \|\boldsymbol{b} - L\boldsymbol{s}\|_2^2 + \frac{2\lambda}{p}\sum_{i=1}^{d_s} |s_i|^p, \tag{43}$$

where $\boldsymbol{b}$ is the observation vector (consistent with $n = 1$) and $\boldsymbol{s}$ is the associated source vector. With $p = 1$, (43) is the regularized MCE cost function. In the limit $p \to 0$, the second term converges (up to a constant which does not affect the final solution) to $2\lambda \sum_{i=1}^{d_s} \log|s_i|$ and we retrieve the basic FOCUSS cost function.

Before we proceed, a couple of points are worth noting regarding the general case. First, if A is invertible, then solving (40) and (41) lead to an identical estimate for S. Secondly, while an infinite set of decompositions $C_i = A_i A_i^T$ can be found for each covariance component, the cost function (41) is invariant to the particular one that is chosen, meaning that the estimate $\hat{S}$ will be the same. Consequently, A_i can be computed in the most convenient fashion.

[9]The variational representation from Section III-C allows any concave $g_i(\cdot)$ to be generated and also yields identical EM-like updates (to those discussed in the next section).

A. Optimization

Presumably, there are a variety of ways to optimize (41). One particularly straightforward and convenient method exploits the hierarchical structure inherent in the assumed Bayesian model. This leads to simple and efficient EM-based update rules as follows [7], [11], [29].[10] It also demonstrates that the canonical FOCUSS iterations are equivalent to principled EM updates. Likewise, regularized MCE solutions can also be obtained in the same manner.

Ultimately, we would like to estimate each S_i, which in turn gives us the true sources S. If we knew the values of the hyperparameters $\boldsymbol{\gamma}$ this would be straightforward; however, these are of course unknown. Consequently, in the EM framework $\boldsymbol{\gamma}$ is treated as hidden data whose distribution (or relevant expectation) is computed during the E-step. The M-step then computes the MAP estimate of S assuming $\boldsymbol{\gamma}$ equals the appropriate expectation.

For the $(k+1)$-th E-step, the expected value of each γ_i^{-1} under the distribution $p\left(\boldsymbol{\gamma}|\mathsf{B},\mathsf{S}^{(k)}\right)$ is required (see the M-step below) and can be computed analytically assuming $g_i(\cdot)$ is differentiable, regardless of the underlying form of $p(\boldsymbol{\gamma})$. Using results pertaining to Gaussian scale mixtures [29] and the assumption $g_i(z) \propto z^{p/2}$, it can be shown that the relevant sufficient statistic is

$$\bar{\gamma}_i^{(k+1)} \triangleq \mathrm{E}_{p(\gamma_i|\mathsf{B},\mathsf{S}^{(k)})}\left[\gamma_i^{-1}\right]^{-1} = \left[\frac{1}{r_i n}\|\mathsf{S}_i^{(k)}\|_{\mathcal{F}}^2\right]^{\frac{2-p}{2}}. \tag{44}$$

The associated M-step is then readily computed using

$$\begin{aligned}\mathsf{S}^{(k+1)} &\rightarrow \arg\max_{\mathsf{S}} \mathrm{E}_{p(\boldsymbol{\gamma}|\mathsf{B},\mathsf{S}^{(k)})}\left[-\log p\left(\mathsf{B},\mathsf{S},\boldsymbol{\gamma}\right)\right] \\ &= \arg\max_{\mathsf{S}} p(B|\mathsf{S}) \prod_i p(\mathsf{S}_i|\gamma_i = \bar{\gamma}_i^{(k+1)}),\end{aligned} \tag{45}$$

giving the i-th pseudo-source update rule

$$\mathsf{S}_i^{(k+1)} \rightarrow \bar{\gamma}_i \mathsf{L}_i^T \left(\Sigma_\epsilon + \sum_{i=1}^{d_\gamma} \bar{\gamma}_i \mathsf{L}_i \mathsf{L}_i^T\right)^{-1} \mathsf{B}. \tag{46}$$

$S^{(k+1)} = A\mathsf{S}^{(k+1)}$ is then the $(k+1)$-th estimate of the true sources. Each iteration of this procedure decreases the cost function (41) and convergence to some fixed point is guaranteed. Also, by substituting (46) into (44) and treating $\gamma_i^{(k+1)} \equiv \bar{\gamma}_i^{(k+1)}$ as an estimate of γ_i, the combined iteration expressed in terms of $\boldsymbol{\gamma}$ alone becomes

$$\gamma_i^{(k+1)} \rightarrow \left[\frac{1}{nr_i}\left\|\gamma_i^{(k)} \mathsf{L}_i^T \left(\Sigma_b^{(k)}\right)^{-1} B\right\|_{\mathcal{F}}^2\right]^{\frac{2-p}{2}} = \left[\frac{1}{nr_i}\left\|\gamma_i^{(k)} \mathsf{L}_i^T \left(\Sigma_b^{(k)}\right)^{-1} \mathsf{B}\right\|_{\mathcal{F}}^2\right]^{\frac{2-p}{2}}. \tag{47}$$

We note that B-dependency is limited to BB^T, so we can replace B with B as before. Each iteration then produces a refined estimate of the unknown hyperparameters $\boldsymbol{\gamma}$ just as with the $\boldsymbol{\gamma}$-MAP algorithms from Section III-A, implying that in some sense we are performing covariance component estimation akin to $\boldsymbol{\gamma}$-MAP. The updates, while not equivalent, are very related. For example, when $p \rightarrow 0$, (47) equals the first term of the $\boldsymbol{\gamma}$-MAP update (27) or the numerator in (29). The exact connection between S-MAP and $\boldsymbol{\gamma}$-MAP will be made explicit in Section IV-B. Regardless, from a computational standpoint, the S-MAP updates described here are of the same per-iteration complexity as the $\boldsymbol{\gamma}$-MAP updates from Section III-A.

As a final point of comparison, given the simplifying assumptions $n = 1$ and $A_i = \boldsymbol{e}_i$ as before, (47) reduces to

$$\gamma_i^{(k+1)} \rightarrow \left[\gamma_i^{(k)} \ell_i^T \left(\lambda I + L \mathrm{diag}[\boldsymbol{\gamma}^{(k)}] L^T\right)^{-1} \boldsymbol{b}\right]^{2-p}, \tag{48}$$

where ℓ_i is the i-th column of L. Not surprisingly, we recover the exact FOCUSS updates when $p \rightarrow 0$ and a FOCUSS-like update for MCE when $p = 1$.[11] Note however, that while previous applications of MCE and FOCUSS to neuroelectromagnetic source imaging using $n = 1$ require a separate iterative solution to be computed at each time point in isolation, here the entire S can be computed at once with $n > 1$ for about the same computational cost as a single FOCUSS run. The benefits of the generalized versions of FOCUSS and MCE, beyond mere computational expediency, will be discussed further below.

[10] Traditionally, the EM algorithm has been employed to obtain iterative maximum likelihood (ML) parameter estimates; however, the procedure is easily adapted to compute MAP estimates as well.

[11] Technically, the updates can be combined in hyperparameter space, as we have done above for consistency with $\boldsymbol{\gamma}$-MAP, or directly in source space, as is often displayed in the literature [16]. For the latter, we substitute the $\boldsymbol{\gamma}$ update (44) into (46) resulting in an iteration over the sources that is identical to regularized FOCUSS [16].

B. Analysis of S-MAP

The nature of the EM updates for S-MAP, where an estimate of $\boldsymbol{\gamma}$ is obtained via the E-step, suggest that this approach is indirectly performing some form of covariance component estimation. But if this is actually the case, it remains unclear exactly what cost function these covariance component estimates are minimizing. This is unlike $\boldsymbol{\gamma}$-MAP, where $\boldsymbol{\gamma}$ is explicitly selected to minimize (18). The following theorem details the relationship between the S-MAP solution and a particular covariance component-based cost function.

Theorem 4: Let $\hat{\boldsymbol{\gamma}}$ be a minimum of the cost function

$$\mathcal{L}_p(\boldsymbol{\gamma}) \triangleq \underbrace{\operatorname{trace}\left[C_b \Sigma_b^{-1}\right]}_{\text{data fit}} + \underbrace{\tfrac{2-p}{p} \sum_{i=1}^{d_y} r_i \gamma_i^{\frac{p}{2-p}}}_{\text{regularization}}, \tag{49}$$

where $p \in (0, 2)$ and $\boldsymbol{\gamma} \geq 0$. Then

$$\hat{S} \triangleq \hat{\Sigma}_s L^T \left(\Sigma_\epsilon + L \hat{\Sigma}_s L^T\right)^{-1} B, \quad \hat{\Sigma}_s \triangleq \sum_i \hat{\gamma}_i C_i \tag{50}$$

satisfies, $\hat{S} = \sum_i A_i \mathsf{S}_i^*$, where S^* is a minimum of (41).

The proof is beyond the scope of this paper; however, it follows by extension of convexity results in [50] to the proposed hierarchical model.[12] Moreover, while Theorem 4 only pertains to the selection $g_i(z) \propto z^{p/2}$ for simplicity, the result is readily generalized to any concave, non-decreasing $g_i(\cdot)$, and therefore any S-MAP problem with a source prior that can be expressed as a Gaussian scale mixture.

Several things are worth noting regarding $\mathcal{L}_p(\boldsymbol{\gamma})$. First, the data fit term is equivalent to that employed by the $\boldsymbol{\gamma}$-MAP cost function. Consequently, the fundamental difference between S-MAP and $\boldsymbol{\gamma}$-MAP ultimately lies in the regularization mechanism of the covariance components. Unlike $\boldsymbol{\gamma}$-MAP, the penalty term in (49) is a separable summation that depends on the value of p to affect hyperparameter pruning; there is no volume-based penalty as in (18). For $p \leq 1$, this penalty is concave in $\boldsymbol{\gamma}$ and leads to the following result:

Theorem 5: Assuming $p \leq 1$, every local minimum of $\mathcal{L}_p(\boldsymbol{\gamma})$ is achieved at a solution with at most $\operatorname{rank}(B) d_b \leq d_b^2$ nonzero hyperparameters.

The proof is directly analogous to that of Theorem 1; both demonstrate how concave regularization terms lead to hyperparameter pruning. As a direct consequence of this result, we note that hyperparameter pruning directly leads to pruning of pseudo-sources as previously claimed:

Corollary 1: Assuming $p \leq 1$, every local minimum of the generalized MAP cost function (41) is achieved at a solution $\hat{S} = \sum_i A_i \hat{\mathsf{S}}_i$ such that at most $\operatorname{rank}(B) d_b \leq d_b^2$ pseudo-sources $\hat{\mathsf{S}}_i$ satisfy $\|\hat{\mathsf{S}}_i\|_{\mathcal{F}} > 0$.

This result implies that, rather than promoting sparsity at the level of individual source elements at a given voxel and time (as occurs with standard MCE and FOCUSS when $n = 1$, $A_i = \boldsymbol{e}_i$), here sparsity is encouraged at the level of the pseudo-sources S_i.[13] The functions $g_i(\cdot)$ operate on the Frobenius norm of each S_i and favor solutions with $\|\mathsf{S}_i\|_{\mathcal{F}} = 0$ for many indeces i. As before, the bound from Theorem 5 is extremely weak; in practice many more components will be pruned. However, an equally key feature of this model is that *within* a nonzero S_i, smooth (non-sparse) solutions are favored by virtue of the Frobenius norm operator. This is why, for example, S-MAP (like $\boldsymbol{\gamma}$-MAP) does not produce overly discontinuous source estimates across time as will occur if regular FOCUSS (or MCE) is applied individually to each time point. Because each S_i spans all time points, and smoothness is encouraged within each S_i, the resulting source estimate is a superposition of smooth temporal bases. This is also explains why, when given the appropriate A_i [36], dipole orientations need not be biased to the coordinate axis [18].

While the sparsity bounds for S-MAP do not provide any distinction from $\boldsymbol{\gamma}$-MAP, this does not imply that actual performance in practical situations will be the same or even similar. This is because Theorems 1 and 5, which are worst-case bounds applying to component pruning at any *local* minimum, say nothing about the probability of achieving the *global* minimum or the quality of this global minimum even if it is achieved. In these respects the selection of $g_i(\cdot)$, or in particular the value of p, can have a very significant impact.

[12] Roughly speaking, the EM iterations (44) and (46) can be viewed as coordinate descent over a particular auxiliary cost function dependent on both S and $\boldsymbol{\gamma}$. This auxiliary function can be solved analytically for S giving the above expression, which only depends on $\boldsymbol{\gamma}$. So the EM algorithm for S-MAP can be viewed equivalently as minimizing (49) over $\boldsymbol{\gamma}$ or (41) over S. Additionally, this duality can be extended the other way as well; $\boldsymbol{\gamma}$-MAP can be expressed explicitly in S-space as a proper S-MAP method [48].

[13] While bounds such as Corollary 1 have been derived for the special case $n = 1$, $A_i = \boldsymbol{e}_i$ using concavity results directly in S-space [35], the methodology used previously is insufficient for the general case because the S-MAP cost function is not concave in S-space.

The simple choice $p = 1$ leads to a convex cost function devoid of non-global minima; the update rules from Section IV-A will converge to a global solution (alternatively a second-order-cone (SOC) program can be used to solve the $p = 1$ case). However, the resulting generalized MCE estimate associated with this global minimum can potentially have problems recovering certain standard source configurations. In fact, MCE solutions, commonly referred to as minimum ℓ_1-norm solutions, have been studied exhaustively in the statistics and signal processing communities regarding the ability to recover sparse (i.e., dipolar-like) sources. Specifically, conditions are stipulated whereby minimum ℓ_1-norm solutions are guaranteed to locate some number of nonzero coefficients (or sources) in an underdetermined linear model [9], [15], [17], [42]. While derived primarily for the simplified case where $n = 1$, $A_i = \boldsymbol{e}_i$, they nontheless elucidate concerns in the more general setting. For example, one of the earliest and most straightforward of these results, which applies when $\Sigma_\epsilon = 0$, can be posed as follows:

Assume we have observed a single measurement vector $\boldsymbol{b}$ produced by the generative model $\boldsymbol{b} = L\boldsymbol{s}_0$, where $\boldsymbol{s}_0$ is a dipolar source vector meaning that most elements are equal to exactly zero or equivalently, $\|\boldsymbol{s}_0\|_0$ is small.[14] Moreover, assume that no other solution to $\boldsymbol{b} = L\boldsymbol{s}$ satisfies $\|\boldsymbol{s}\|_0 \leq \|\boldsymbol{s}_0\|_0$, meaning no solution with fewer number of equivalent dipoles could have produced $\boldsymbol{b}$. Then the following result from [9] dictates a rigorous, computable condition whereby the MCE solution is guaranteed to successfully locate all dipoles.

Theorem 6: (*adapted from Donoho and Elad, 2003*) Given an arbitrary lead-field L with columns ℓ_i normalized such that $\ell_i^T \ell_i = 1, \forall i = 1, \ldots, d_s$, and given $G \triangleq L^T L$ and $\kappa \triangleq \max_{i \neq j} |G_{ij}|$, if $\boldsymbol{s}_0$ satisfies

$$\|\boldsymbol{s}_0\|_0 < 1/2(1 + 1/\kappa), \tag{51}$$

then the MCE solution is guaranteed to equal $\boldsymbol{s}_0$.

While elegant in theory, in practice Theorem 6 is very difficult to apply. For example, the lead-fields L required for MEG/EEG source localization (when suitably normalized as required by the theorem) typically have $\kappa \approx 1$, meaning some columns tend to be highly correlated. This implies that only sparse solutions with at most one nonzero element are guaranteed to be found, and even this will only occur for suitably normalized lead-fields. Therefore with unnormalized lead-fields, localization bias can still occur when estimating single dipolar sources with no noise (unlike $\boldsymbol{\gamma}$-MAP which can be invariant to to column normalization schemes in this respect). Additionally, the above bound is tight, meaning a lead-field matrix can always be constructed such that the MCE solution will sometimes fail to find $\boldsymbol{s}_0$ if $\|\boldsymbol{s}_0\|_0 \geq 1/2(1 + 1/\kappa)$. Further analysis of general equivalence conditions and performance bounds for MCE-like algorithms are derived in [10], [43]. However again, these are all more applicable to applications such as compressed sensing, where the columns of the analogous L are not highly correlated, than to neuroelectromagnetic localization problems. Moreover, the potential difficulty dealing with correlated (or coherent) lead-field columns persists in the more general scenario involving $n > 1$ and $A_i \neq \boldsymbol{e}_i$.

To summarize then, the problem with MCE is not the existence of local minima. Rather, it is that the global minimum may be unrepresentative of the true source distribution even for simple dipolar source configurations. In this situation, and especially when lead-field columns are highly correlated, the MCE solution may fail to find sufficiently sparse source representations consistent with the assumption of a few equivalent current dipoles. We anticipate that this issue could persist with more complex covariance components and source configurations. Regardless, the unimodal, convex nature of the generalized MCE cost function remains a very attractive advantage over $\boldsymbol{\gamma}$-MAP.

In contrast, if $p < 1$ (which implies that $g_i(z^2)$ is strictly concave in z), then more pseudo sources will be pruned at any global solution, which often implies that those which remain may be more suitable than the MCE estimate. Certainly this is true when estimating dipolar sources, but it likely holds in more general situations as well. However, local minima can be an unfortunate menace with $p < 1$. For example, the canonical FOCUSS algorithm, which implies $p \to 0$, has a combinatorial number of local minima satisfying

$$\binom{d_s - 1}{d_b} + 1 \leq \text{ \# of FOCUSS Local Minima} \leq \binom{d_s}{d_b}. \tag{52}$$

Obviously this number will be huge for practical lead-field matrices, which largely explains the sensitivity of FOCUSS to initialization and noise. While the FOCUSS cost function can be shown to have zero localization bias at the global solution, because of the tendency to become stuck at local optima, in practice a bias can be observed when recovering even a single dipolar source. Other selections of p between zero and one can lead to a similar fate.

In the general case, a natural trade-off exists with S-MAP procedures: the greater the sparsity of solutions at the global minimum the less possibility that this minimum is biased, but the higher the chance of suboptimal convergence to a biased local minimum. In this regard the optimal balance could well be application dependent; however, these factors crucially disallow any theoretical bias results analogous to those given for $\boldsymbol{\gamma}$-MAP in Section III-B. So in this respect $\boldsymbol{\gamma}$-MAP maintains a distinct advantage, which is further elaborated in [46] in a somewhat different context. However, S-MAP is nonetheless capable of successfully handling large numbers of diverse covariance components, and therefore simultaneous constraints on the source space. This addresses many of the concerns raised in [24] pertaining to existing MAP methods.

[14] $\|\cdot\|_0$ denotes the ℓ_0 quasi-norm, which is simply a count of the number of nonzero elements in a vector.

Finally, to link (49) to more traditional weighted minimum ℓ_2-norm methods, we note that

$$\lim_{p\to 2} \frac{2-p}{p} r_i \gamma_i^{\frac{p}{2-p}} = \begin{cases} \infty, & \gamma_i > 1 \\ 0 & \gamma_i \leq 1 \end{cases} \tag{53}$$

When this expression is plugged into (49), subsequent minimization forces all $\gamma_i \to 1$, and the resulting estimate $\hat{S}$ is the standard weighted minimum norm estimate.

V. Variational Bayes

Sections III and IV describe two complementary yet distinctive means of utilizing a simple Gaussian scale mixture model to affect source localization. From the perspective of a Bayesian purist however, the pursuit of MAP estimates for unknown quantities of interest, whether parameters S or hyperparameters $\boldsymbol{\gamma}$, can be misleading since these estimates discount uncertainty and may not reflect regions of significant probability mass, unlike (for example) the posterior mean. Variational Bayesian methods, which have successfully been applied to a wide variety of hierarchical Bayesian models in the machine learning literature [1], [3], offer an alternative to $\boldsymbol{\gamma}$-MAP and S-MAP that putatively tackle these concerns. The principle idea here is that all unknown quantities should either be marginalized (intergrated out) when possible or approximated with tractable distributions that reflect underlying uncertainty and have computable posterior moments.[15] Practically, we would like to account for ambiguity regarding $\boldsymbol{\gamma}$ when estimating $p(\mathsf{S}|B)$, and potentially, we would also like a good approximation for $p(B)$, or a bound on the model evidence $\log p(B)$, for application to model selection [13], [22].

In this section we will briefly consider two possible variational approximations germane to the source localization problem: the mean-field approximation (VB-MF), and a fixed-form, Laplace approximation (VB-LA). It turns out that both are related to $\boldsymbol{\gamma}$-MAP but with important distinctions.

A. Mean-Field Approximation (VB-MF)

The basic strategy here is to replace intractable posterior distributions with approximate ones that, while greatly simplified and amenable to simple inference procedures, still retain important characteristics of the full model. In the context of our presumed model structure, both the posterior source distribution $p(\mathsf{S}|B)$, which is maximized with S-MAP, as well as the hyperparameter posterior $p(\boldsymbol{\gamma}|B)$, which is maximized via $\boldsymbol{\gamma}$-MAP, are quite complex and can only be expressed up to some unknown scaling factor (the integration required for normalization is intractable). Likewise, the joint posterior $p(\mathsf{S},\boldsymbol{\gamma}|B)$ over all unknowns is likewise intractable and complex. VB attempts to simplify this situation by finding an approximate joint posterior that factorizes as

$$p(\mathsf{S},\boldsymbol{\gamma}|B) \approx \hat{p}(\mathsf{S},\boldsymbol{\gamma}|B) = \hat{p}(\mathsf{S}|B)\hat{p}(\boldsymbol{\gamma}|B), \tag{54}$$

where $\hat{p}(S|B)$ and $\hat{p}(\boldsymbol{\gamma}|B)$ are amenable to closed-form computation of posterior quantities such as means and variances (unlike the full posteriors upon which our model is built). This is possible because the enforced factorization, often called the *mean-field* approximation reflecting its origins in statistical physics, simplifies things significantly [1], [3]. The cost function optimized to find this approximate distribution is

$$\hat{p}(\mathsf{S}|B), \hat{p}(\boldsymbol{\gamma}|B) = \arg \min_{q(\mathsf{S}),q(\boldsymbol{\gamma})} \mathrm{KL}\left[q(\mathsf{S})q(\boldsymbol{\gamma}) \| p(\mathsf{S},\boldsymbol{\gamma}|B)\right], \tag{55}$$

where $q(\mathsf{S})$ and $q(\boldsymbol{\gamma})$ are arbitrary probability distributions and $\mathrm{KL}[\cdot\|\cdot]$ indicates the Kullback-Leibler divergence measure.

Recall that $\boldsymbol{\gamma}$-MAP iterations effectively compute an approximate distribution for S (E-step) and then a point estimate for $\boldsymbol{\gamma}$ (M-step); S-MAP does the exact opposite. In contrast, here an approximating distribution is required for both parameters S and hyperparameters $\boldsymbol{\gamma}$. While it is often contended that conjugate hyperpriors must be employed such that (55) is solvable [5], in fact this problem can be solved by coordinate descent over $q(\mathsf{S})$ and $q(\boldsymbol{\gamma})$ for virtually any hyperprior [29] in the following sense.

Assume first that $q(\boldsymbol{\gamma})$ is held fixed to some $q^{(k)}(\boldsymbol{\gamma})$ and we are optimizing (55) with respect to $q(\mathsf{S})$. This can be solved via

$$\begin{aligned} q^{(k+1)}(\mathsf{S}) &\rightarrow \arg\min_{q(\mathsf{S})} \mathrm{KL}\left[q(\mathsf{S})q^{(k)}(\boldsymbol{\gamma}) \| p(\mathsf{S},\boldsymbol{\gamma}|B)\right] \\ &= \arg\min_{q(\mathsf{S})} \mathrm{KL}\left[q(\mathsf{S})q^{(k)}(\boldsymbol{\gamma}) \| p(\mathsf{S},\boldsymbol{\gamma},B)\right] \\ &= \arg\min_{q(\mathsf{S})} \left[\int q(\mathsf{S}) \log q(\mathsf{S}) d\mathsf{S} - \int q(\mathsf{S})q^{(k)}(\boldsymbol{\gamma}) \log p(B|\mathsf{S})p(\mathsf{S}|\boldsymbol{\gamma})p(\boldsymbol{\gamma}) d\mathsf{S}\right]. \end{aligned} \tag{56}$$

If we define

$$\bar{\gamma}_i^{(k)} \triangleq \mathrm{E}_{q^{(k)}(\boldsymbol{\gamma})}\left[\gamma_i^{-1}\right]^{-1}, \quad \bar{\boldsymbol{\gamma}}^{(k)} \triangleq [\bar{\gamma}_1^{(k)}, \ldots, \bar{\gamma}_{d_\gamma}^{(k)}]^T, \tag{57}$$

[15] $\boldsymbol{\gamma}$-MAP accomplishes this in some sense given the convex bounding perspective of Section III-C.

then (56) simplifies to

$$\begin{aligned} q^{(k+1)}(\mathsf{S}) &\rightarrow \arg\min_{q(\mathsf{S})} \left[\int q(\mathsf{S}) \log q(\mathsf{S}) d\mathsf{S} - \int q(\mathsf{S}) \log p(B|\mathsf{S}) p(\mathsf{S}|\boldsymbol{\gamma} = \bar{\boldsymbol{\gamma}}^{(k)}) d\mathsf{S} \right] \\ &= \arg\min_{q(\mathsf{S})} \mathrm{KL}\left[q(\mathsf{S}) \| p(B, \mathsf{S}|\boldsymbol{\gamma} = \bar{\boldsymbol{\gamma}}^{(k)}) \right] \\ &= \arg\min_{q(\mathsf{S})} \mathrm{KL}\left[q(\mathsf{S}) \| p(\mathsf{S}|B, \boldsymbol{\gamma} = \bar{\boldsymbol{\gamma}}^{(k)}) \right] \\ &= p\left(\mathsf{S}|B, \boldsymbol{\gamma} = \bar{\boldsymbol{\gamma}}^{(k)} \right). \end{aligned} \tag{58}$$

The final distribution is Gaussian with moments given by (23) using $\boldsymbol{\gamma} = \bar{\boldsymbol{\gamma}}^{(k)}$. Note that we can also use B instead of B for computational convenience. So this update is equivalent to the E-step of $\boldsymbol{\gamma}$-MAP.

Updates for $q^{(k+1)}(\boldsymbol{\gamma})$ proceed in a similar fashion with fixed $q(\mathsf{S}) = q^{(k+1)}(\mathsf{S})$ giving

$$\begin{aligned} q^{(k+1)}(\boldsymbol{\gamma}) &\rightarrow \arg\min_{q(\boldsymbol{\gamma})} \mathrm{KL}\left[q^{(k+1)}(\mathsf{S}) q(\boldsymbol{\gamma}) \| p(\mathsf{S}, \boldsymbol{\gamma}|B) \right] \\ &= \prod_i p\left(\gamma_i | B, \|\mathsf{S}_i\|_{\mathcal{F}} = \bar{\mathsf{S}}_i^{(k+1)} \right), \quad \bar{\mathsf{S}}_i^{(k+1)} \triangleq \sqrt{\mathrm{E}_{q^{(k+1)}(\mathsf{S}_i)}\left[\|\mathsf{S}_i\|_{\mathcal{F}}^2 \right]}, \end{aligned} \tag{59}$$

where the factorization over γ_i is a natural consequence of the mean-field approximation; it does not follow from an additional assumption. Regardless, the posterior distributions over each γ_i are not available in closed form, except when a conjugate prior is used for $p(\boldsymbol{\gamma})$. However, the update for $q^{(k+1)}(\mathsf{S})$ above only requires $\bar{\boldsymbol{\gamma}}^{(k+1)}$ at each iteration, not the entire distribution. The latter can be computed using

$$\bar{\gamma}_i^{(k+1)} \rightarrow \mathrm{E}_{q^{(k+1)}(\gamma_i)}\left[\gamma_i^{-1} \right]^{-1} = \left. \frac{x \bar{p}(x)}{-\bar{p}'(x)} \right|_{x = \bar{\mathsf{S}}_i^{(k+1)}} \tag{60}$$

where $\bar{p}(\|\mathsf{S}_i\|_{\mathcal{F}}) \triangleq p(\mathsf{S}_i)$ and $\bar{p}'(\cdot)$ is the associated first derivative. The derivation follows from [28], [29]. (60) is an alternative form of the M-step used with the EM version of $\boldsymbol{\gamma}$-MAP, so convergence can potentially be slow as was discussed in Section III-A.

Two things are worth pointing out with respect to this result. First, it implies that we can maximize (55) with respect to $\hat{p}(\mathsf{S}|B)$ without ever explicitly computing $\hat{p}(\boldsymbol{\gamma}|B)$ in full. This is accomplished by iteratively updating each $\bar{\mathsf{S}}_i$ and $\bar{\gamma}_i$. Secondly, we have the following:

Theorem 7: Let $\hat{p}(\mathsf{S}|B)$ denote the distribution that solves (55) assuming some arbitrary hyperprior $p_i(\gamma_i) \propto \exp[-\frac{1}{2} f_i(\gamma_i)]$. Then there exists a hyperprior $\tilde{p}_i(\gamma_i) \propto \exp[-\frac{1}{2}\tilde{f}_i(\gamma_i)]$ and corresponding $\boldsymbol{\gamma}$-MAP estimate $\tilde{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} p(B|\boldsymbol{\gamma}) \prod_i \tilde{p}_i(\gamma_i)$ such that

$$\hat{p}(\mathsf{S}|B) = p(\mathsf{S}|B, \boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}). \tag{61}$$

The selection $\tilde{f}_i(\gamma_i) = -n r_i \log \gamma_i - g_i^*\left(\gamma_i^{-1}\right)$ achieves this result, where $g_i^*(\cdot)$ is the concave conjugate of $g_i(\cdot)$ defined via (32).

The proof, which can be derived from results in [29], has been omitted for brevity. Theorem 7 implies that the $\hat{p}(S|B)$ obtained from the mean-field approximation offers nothing new that could not already be obtained via $\boldsymbol{\gamma}$-MAP (although the hyperprior needed for equivalence will generally not be the same as used by VB-MF, i.e., $f_i(\gamma_i) \neq \tilde{f}_i(\gamma_i)$). Consequently, although we are in principle accounting for uncertainty in $\boldsymbol{\gamma}$ when arriving at the posterior on S, we do not actually obtain a new class of approximations by this added sophistication.

The analysis here sheds some light on existing localization algorithms. For example, in [38] a VB-MF algorithm is developed for localizing sources with some spatial extent and potentially some prior knowledge from fMRI. The underlying cost function can be shown to be a special case of (55) assuming that Σ_s is parameterized via

$$\Sigma_s = \sum_{i=1}^{d_s} \gamma_i \boldsymbol{e}_i \boldsymbol{e}_i + \sum_{j=1}^{d_s} \gamma_{(d_s+j)} \boldsymbol{\psi}_j \boldsymbol{\psi}_j^T, \tag{62}$$

and so $d_{\boldsymbol{\gamma}} = 2 d_s$.[16] When fMRI data is available, it is incorporated into a conjugate inverse Gamma hyperprior on $\boldsymbol{\gamma}$, as is also commonly done with $\boldsymbol{\gamma}$-MAP methods [5]. Optimization is then performed using update rules that reduce to special cases of (58) and (60). Unfortunately, because these EM updates are prohibitively slow in practice, some heuristics are proposed to facilitate the estimation process [38]. Of course in light of the fact that minimizing the VB-MF cost can be accomplished via

[16]Recall from Section II-C that $\boldsymbol{\psi}_j$ is a Gaussian geodesic basis function.

$\boldsymbol{\gamma}$-MAP for this model (assuming the appropriate $f(\boldsymbol{\gamma})$), the faster updates described in Section III-A could easily be substituted. With these results in mind then, the general methods of [12], [24], [31] and [33], [34], [41] as well as the variational method of [38] are all identical with respect to their underlying source estimates $\hat{p}(S|B)$; they differ only in which covariance component set $\mathcal{C}$ and possibly hyperpriors are assumed, and in how optimization is performed.

Before proceeding, however, we should point out that there is a drawback to using (60) in place of a full update on $q(\boldsymbol{\gamma})$. For purposes of model comparison, it is sometimes desirable to have an estimate of $\log p(B)$ [13], [22]. One popular estimator which strictly lower bounds $\log p(B)$ and falls out of the VB framework is given by

$$\begin{aligned}\log p(B) \geq F &\triangleq \log p(B) - \mathrm{KL}\left[\hat{p}(S|B)\hat{p}(\boldsymbol{\gamma}|B) \| p(\mathsf{S}, \boldsymbol{\gamma}|B)\right] \\ &= \int \hat{p}(\mathsf{S}|B)\hat{p}(\boldsymbol{\gamma}|B) \log \frac{p(B, \mathsf{S}, \boldsymbol{\gamma})}{\hat{p}(\mathsf{S}|B)\hat{p}(\boldsymbol{\gamma}|B)} d\boldsymbol{\gamma},\end{aligned} \tag{63}$$

where the inequality follows by the non-negativity of the Kullback-Leibler divergence. The quantity F is sometimes referred to the variational free energy. Evaluation of F requires the full distribution $\hat{p}(\boldsymbol{\gamma}|B)$ and therefore necessitates using conjugate priors or further approximations. One possibility, which involves the use of a fixed-form Gaussian assumption, is discussed in the next section and in Appendix D.

B. Laplace (Fixed-Form) Approximation (VB-LA)

The Laplace approximation has been advocated in [14] for addressing a variety of general problems in neuroimaging. Application to MEG/EEG source localization as outlined in [13] is tailored to finding a tractable posterior distribution on the hyperparameters and then using this $\hat{p}(\boldsymbol{\gamma}|B)$ to find approximations to $p(S|B)$ and $\log p(B)$ (the pseudo-sources S are not used in the formulation in [13]). To facilitate this process, the hyperparameters are re-parameterized via the transformation $\lambda_i \triangleq \log \gamma_i$, $\forall i$, which now alows them to have positive or negative values. The Laplace approximation then involves the assumption that the posterior distribution of these new hyperparameters $\boldsymbol{\lambda} \triangleq [\lambda_1, \ldots, \lambda_{d_\gamma}]^T$ satisfies

$$p(\boldsymbol{\lambda}|B) \approx \mathcal{N}\left(\boldsymbol{\lambda}|\boldsymbol{\mu}_\lambda, \Sigma_\lambda\right). \tag{64}$$

There are a variety of ways $\boldsymbol{\mu}_\lambda$ and Σ_λ can be chosen to form the best approximation. Although not explicitly presented this way, the method from [13] is equivalent to the following procedure. First, a MAP estimate of $\boldsymbol{\lambda}$ is computed by maximizing $\log p(B, \boldsymbol{\lambda})$ using a seconder-order Fisher scoring procedure, although as before this method is not guaranteed to increase the cost function and requires a very expensive $O(d_\gamma^3)$ matrix inverse. However, because $\log p(B, \boldsymbol{\lambda} = \log \boldsymbol{\gamma}) = \log p(B|\boldsymbol{\gamma}) p(\boldsymbol{\lambda} = \log \boldsymbol{\gamma})$ is equivalent to the $\boldsymbol{\gamma}$-MAP cost function (up to a scaling factor and constant terms), the optimization methods from Section III-A could be applied as well. As stated previously, these methods do not require that the hyperprior be properly normalized, so the fact that $p(\boldsymbol{\lambda} = \log \boldsymbol{\gamma})$ is not a proper density with respect to $\boldsymbol{\gamma}$ is not an issue.

Once the mode $\hat{\boldsymbol{\lambda}}$ is obtained, we set $\boldsymbol{\mu}_\lambda = \hat{\boldsymbol{\lambda}}$. The second-order statistics of the true and approximate distributions are then matched at this mode using $\Sigma_\lambda = -[\ell''(\boldsymbol{\mu}_\lambda)]^{-1}$, where $\ell(\boldsymbol{\mu}_\lambda) \triangleq \log p(B, \boldsymbol{\lambda} = \boldsymbol{\mu}_\lambda)$ and $\ell''(\boldsymbol{\mu}_\lambda)$ is the corresponding Hessian matrix evaluated at $\boldsymbol{\mu}_\lambda$.

We still need approximate distributions for $p(S|B)$ and $p(B)$. For $p(S|B)$, it is suggested in [13] to use $p(S|B, \boldsymbol{\lambda} = \boldsymbol{\mu}_\lambda)$, which is equivalent to a $\boldsymbol{\gamma}$-MAP estimator with $\boldsymbol{\gamma} = \exp(\boldsymbol{\mu}_\lambda)$; however, uncertainty regarding $\boldsymbol{\lambda}$ as reflected by Σ_λ does not effect $p(S|B, \boldsymbol{\lambda} = \boldsymbol{\mu}_\lambda)$. So in this respect, like the previous variational method, there is no advantage over $\boldsymbol{\gamma}$-MAP. Regarding $p(B)$, an approximation to $\log p(B)$ is derived that requires an additional assumption.

Because of the non-negativity of the KL divergence, it always holds that

$$\log p(B) \geq F \triangleq \log p(B) - \mathrm{KL}\left[\mathcal{N}\left(\boldsymbol{\lambda}|\boldsymbol{\mu}_\lambda, \Sigma_\lambda\right) \| p(\boldsymbol{\lambda}|B)\right] = \int \mathcal{N}\left(\boldsymbol{\lambda}|\boldsymbol{\mu}_\lambda, \Sigma_\lambda\right) \log \frac{p(B, \boldsymbol{\lambda})}{\mathcal{N}\left(\boldsymbol{\lambda}|\boldsymbol{\mu}_\lambda, \Sigma_\lambda\right)} d\boldsymbol{\lambda}. \tag{65}$$

Unfortunately however, this bound cannot be computed in closed form, so the computable approximation

$$\hat{F} = \ell(\boldsymbol{\mu}_\lambda) + \frac{1}{2}\left(d_\gamma \log 2\pi + \log|\Sigma_\lambda|\right) \tag{66}$$

is used as a surrogate. The derivation can be found in Appendix C. While the original F does constitute a rigorous bound on $\log p(B)$, $\hat{F}$ does not. This is easily seen by noting that

$$\hat{F} = \log p(B) + \log p(\boldsymbol{\lambda} = \boldsymbol{\mu}_\lambda|B) + \frac{1}{2}\left(d_\gamma \log 2\pi + \log|\Sigma_\lambda|\right). \tag{67}$$

The extra terms can be positive or negative and, because $\log p(\boldsymbol{\lambda} = \boldsymbol{\mu}_\lambda|B)$ is unknown and intractable, there is no way of knowing how $\hat{F}$ relates to $\log p(B)$. Therefore, since $\hat{F}$ can be greater than $\log p(B)$, it could potentially distort efforts to use this approximation for model selecton (e.g., it could artificially inflate the model evidence), although empirical results from [13] indicate that it can be informative in practice nonetheless. A second potential issue with $\hat{F}$ is the computational complexity required to compute $-\log|\ell''(\boldsymbol{\mu}_\lambda)|$; this is a $O(d_\gamma^3)$ calculation like the Fisher scoring procedure. Although this

only need be computed once after $\boldsymbol{\mu}_\lambda$ is obtained (if we use the methods from Section III-A), this can still be an extremely expensive and possibly numerically instable task if d_γ is large.

The Laplace approximation could certainly be implemented with a variety of hyperpriors. In [13], a Gaussian distribution (with user-specified mean and covariance) is assumed on $\boldsymbol{\lambda}$. This is equivalent to assuming a log-normal prior on $\boldsymbol{\gamma}$. With this assumption, (66) is equivalent to Eq. (5) in [13].

As a final consideration, the limitations of this method can potentially be ameliorated by combining the Laplace and mean-field approximations. This possibility is explored in depth in Appendix D, where an alternative learning procedure, with several attractive properties, is developed.

C. Summary

The approximate distribution $\hat{p}(S|B)$ obtained from both variational methods is equivalent to a $\boldsymbol{\gamma}$-MAP estimate given the appropriate equalizing hyperprior. Consequently, the optimization procedures for this purpose described in Section III-A can be directly applied if desired. Likewise, the sparsifying properties of the underlying $\boldsymbol{\gamma}$-MAP cost functions carry over as well (see Section III-B). Additionally, when the mean-field approximation is used, the particular parameterization of the unknown hyperparameters (e.g., $\boldsymbol{\gamma}$ vs. $\boldsymbol{\lambda}$ has no affect on $\hat{p}(S|B)$. This is because the only quantity required for this approximation is the expected value of each γ_i^{-1}; this quantity is invariant to the parameterization. This is quite unlike the Laplace approximation, which crucially depends on the non-linear re-parameterization of $\boldsymbol{\gamma}$.

More substantial differences exist regarding the estimated bound on $\log p(B)$. With VB-MF, we get a principled bound that can only be computed in closed form when conjugate priors on $\boldsymbol{\gamma}$ are present, although Appendix D presents one way to circumvent this difficulty. Because the VB-MF posterior on the hyperparameters naturally factorizes (regardless of the parameterization) via (59), this bound is independent of any correlations between the hyperparameters and is easy to compute. In contrast, VB-LA gives an approximation to $\log p(B)$ that can potentially be utilized with a variety of hyperpriors; however, a rigorous bound is not produced and so formal model comparison can be biased. The approximation also depends on correlations between hyperparameters and can be expensive to compute if d_γ is extremely large.

As discussed in Section III-C, direct $\boldsymbol{\gamma}$-MAP can also be used to obtain a flexible bound on $\log p(B)$ that can be used in a variety of situations and does not depend on conjugacy. In general, it has not yet been established which approximation technique leads to the tightest lower bound on $\log p(B)$ or the most effective metric in practice.

VI. Connections with Other Statistical Tomographic Methods

Thus far, we have attempted to relate and extend three large classes of Bayesian inverse methods, all of which turn out to be performing covariance component estimation/pruning using different sparsity promoting regularization procedures. This section provides further connections to beamforming and sLORETA.

A. Minimum Variance Adaptive Beamforming

Beamformers are spatial filters that pass source signals in particular focused locations while suppressing interference from elsewhere. The widely-used minimum variance adaptive beamformer (MVAB) creates such filters using a sample covariance estimate; however, the quality of this estimate deteriorates when the sources are correlated or the number of samples n is small [39], [53].

But the $\boldsymbol{\gamma}$-MAP strategy can also be used to enhance beamforming in a way that is particularly robust to source correlations and limited data [47]. Specifically, the estimated $\boldsymbol{\gamma}$-MAP data covariance matrix

$$\hat{\Sigma}_b = \Sigma_\varepsilon + \sum_i \hat{\gamma}_i L C_i L^T \tag{68}$$

can be used to replace the problematic sample covariance $C_b = n^{-1} B B^T$. In [47], we prove that this substitution has the natural ability to remove the undesirable effects of correlations or limited data. When n becomes large and assuming uncorrelated sources, this method reduces to the exact MVAB. Simulations using MEG and direction-of-arrival data [47] support these conclusions. Additionally, the method can potentially enhance a variety of traditional signal processing methods that rely on robust sample covariance estimates.

In comparing this idea with the VB approach for estimating source correlations in [36], it is important to make a distinction between estimating the correlation between sources and estimating the locations of correlated sources. In [47] we attempt only the latter using the natural decorrelating mechanism of $\boldsymbol{\gamma}$-MAP, although actual source correlations can still be estimated empirically using the $\hat{S}$ so obtained (which is localized but not actually decorrelated). The simplicity of this approach means that efficient update rules and analyses of global and local minima are possible as discussed in Section III-B and [47]. In contrast, the added complexity involved in explicitly learning source correlations using the method in [36] leads to expensive learning rules (quadratic in d_s) and some ambiguity regarding convergence properties and the nature of minimizing solutions.

B. sLORETA

Standardized low resolution brain electromagnetic tomography (sLORETA) begins with a weighted minimum ℓ_2-norm (WMN) estimate using $\Sigma_s = I$ and $\Sigma_\epsilon = \lambda I$ (where λ is a non-negative scalar) and then standardizes the solution using the covariance of the resulting WMN estimator [30]. This process is designed to compensate for deep current sources, which otherwise may be mapped to superficial cortical locations. The sLORETA procedure is easy to compute analytically and has been show to exhibit zero localization bias if only a single dipolar source is present [40].

We now describe how sLORETA provides an interesting link between the $\boldsymbol{\gamma}$-MAP updates and the iterative FOCUSS source localization algorithm. The connection is most transparent when we assume fixed dipole orientations in each voxel and substitute $\Sigma_\epsilon = \lambda I$ and the prior covariance $\Sigma_s = \sum_{i=1}^{d_s} \gamma_i \boldsymbol{e}_i \boldsymbol{e}_i^T = \mathrm{diag}[\boldsymbol{\gamma}]$ into (29), giving the modified update

$$\begin{aligned} \gamma_i^{(k+1)} &\rightarrow \left[\gamma_i^{(k)} \ell_i^T \left(\lambda I + L \Gamma^{(k)} L^T\right)^{-1} \boldsymbol{b}\right]^2 \left(R_{ii}^{(k)}\right)^{-1}, \\ R^{(k)} &\triangleq \Gamma^{(k)} L^T \left(\lambda I + L \Gamma^{(k)} L^T\right)^{-1} L, \end{aligned} \tag{69}$$

where $\Gamma \triangleq \mathrm{diag}[\boldsymbol{\gamma}]$, ℓ_i is the i-th column of L, $\boldsymbol{b}$ is a single measurement vector, and $R^{(k)}$ is the effective resolution matrix given the hyperparameters at the current iteration. The j-th column of R (called a point-spread function) equals the source estimate obtained using $\hat{\boldsymbol{s}} = \hat{\Sigma}_s L^T \hat{\Sigma}_b^{-1} \boldsymbol{b}$ when the true source is a unit dipole at location j [40].

Continuing, if we assume that initialization of $\boldsymbol{\gamma}$-MAP occurs with $\boldsymbol{\gamma}^{(0)} = \mathbf{1}$ (as is customary), then the hyperparameters produced after a *single* iteration of $\boldsymbol{\gamma}$-MAP are equivalent to computing the sLORETA estimate for standardized current density power [30], [33], [51] (this assumes fixed orientation constraints). In this context, the inclusion of R as a normalization factor helps to compensate for depth bias, which is the propensity for deep current sources within the brain to be underrepresented at the scalp surface [30], [33]. So $\boldsymbol{\gamma}$-MAP can be interpreted as a recursive refinement of what amounts to the non-adaptive, linear sLORETA estimate. In fact, the basic MacKay $\boldsymbol{\gamma}$-MAP update equations have been derived independently from this perspective [33]. Additionally, during this iterative process it typically retains zero location bias in the sense described in Section III-B.

As a further avenue for comparison, if we assume that $R = I$ for all iterations, then the update (69) is equivalent to the original FOCUSS iterations regularized using λ. Therefore, $\boldsymbol{\gamma}$-MAP can be viewed in some sense as taking the recursive FOCUSS update rules and including the sLORETA normalization that, among other things, allows for some level of depth bias compensation.

VII. Discussion

The efficacy of modern Bayesian techniques for quantifying uncertainty and explicitly accounting for prior assumptions make them attractive candidates for source localization. However, it is not always transparent how these methods relate, nor how they can be extended to handle more challenging problems, nor which ones should be expected to perform best in various situations relevant to MEG/EEG source imaging. Starting from a hierarchical Bayesian model constructed using Gaussian scale mixtures with flexible covariance components, we analyze and, where possible, extend three broad classes of Bayesian inference methods: $\boldsymbol{\gamma}$-MAP, which involves integrating out the unknown sources and optimizing the hyperparameters, S-MAP, which integrates out the hyperparameters and directly optimizes over the sources, and variational approximation methods, which attempt to account for uncertainty in all unknowns. Together, these three encompass a surprisingly wide range of existing source reconstruction approaches, which makes general theoretical analyses and algorithmic extensions/improvements pertaining to them particularly relevant. Specific highlights of this endeavor are as follows:

- $\boldsymbol{\gamma}$-MAP, S-MAP, and VB can be viewed as procedures for learning a source covariance model using a set of predetermined symmetric, positive semi-definite covariance components. The number of components in this set, each of which acts as a constraint on the source space, can be extremely large, potentially much larger than the number of sensors. However, a natural pruning mechanism effectively discards components that are unsupported by the data. This occurs because of an intrinsic sparsity preference in the Gaussian scale mixture model, which is manifested in an explicit sparsity-inducing regularization term that is independent of any hyperprior $p(\boldsymbol{\gamma})$. Consequently, it is not crucial that the user/analyst manually determine an optimal set of components *a priori*; many components can be included initially allowing the learning process to remove superfluous ones.
- The wide variety of Bayesian source localization methods that fall under this framework can be differentiated by the following factors:
  1) Selection of covariance component regularization term per the discussion in Sections III-B and IV-B. Note that this will implicitly determine whether we are performing S-MAP, $\boldsymbol{\gamma}$-MAP, or VB.
  2) Choice of initial covariance component set $\mathcal{C}$,
  3) Optimization method/ Update rules, and
  4) Approximation to $\log p(B)$; this determines whether we are ultimately performing $\boldsymbol{\gamma}$-MAP or VB.

- Covariance component possibilities include geodesic neural basis functions for estimating distributed sources [32], [34], spatial smoothing factors [24], indicator matrices to couple dipole components or learn flexible orientations [36], fMRI-based factors [31], and temporal and spectral constraints [21].
- With large numbers of covariance components, S-MAP, $\boldsymbol{\gamma}$-MAP, and VB provably remove or prune a certain number of components which are not necessary for representing the observed data.
- In principle, the noise-plus-interference covariance Σ_ϵ can be jointly estimated as well, competing with all the other components to model the data (see [11] for a special case of the latter in the context of kernel regression). However, identifiability issues can be a concern here [46] and so we consider it wiser to estimate Σ_ϵ via other means (e.g., using VB factor analysis applied to prestimulus data [54]).
- The latent structure inherent to the Gaussian scale-mixture model leads to an efficient, principled family of update rules for $\boldsymbol{\gamma}$-MAP, S-MAP, and VB. This facilitates the estimation of complex covariance structures modulated by very large numbers of hyperparameters (e.g., $100,000+$) with relatively little difficulty.
- Previous focal source imaging techniques such as FOCUSS and MCE display undesirable discontinuities across time as well significant biases in estimating dipole orientations. Consequently, various heuristics have been proposed to address these deficiencies [18], [44]. However, the general spatio-temporal framework of $\boldsymbol{\gamma}$-MAP and S-MAP handles both of these concerns in a robust, principled fashion by the nature of their underlying cost function.
- The standard weighted minimum ℓ_2-norm method can be viewed as a limiting case of both S-MAP and $\boldsymbol{\gamma}$-MAP.
- $\boldsymbol{\gamma}$-MAP provides a useful covariance estimate that can be used to improve the performance of the minimum variance adaptive beamformer when sources are correlated or n is small.
- sLORETA is equivalent to performing a single iteration of a particular $\boldsymbol{\gamma}$-MAP optimization procedure. Consequently, the latter can be viewed as an iterative refinement of sLORETA. This is exactly analogous to the view of FOCUSS as an iterative refinement of a weighted minimum ℓ_2-norm estimate.
- $\boldsymbol{\gamma}$-MAP and VB have theoretically zero localization bias estimating perfectly uncorrelated dipoles given the appropriate hyperprior and initial set of covariance components.
- The role of the hyperprior $p(\boldsymbol{\gamma})$ is heavily dependent on whether we are performing $\boldsymbol{\gamma}$-MAP, S-MAP, or VB. In the S-MAP framework, the hyperprior functions through its role in creating the concave regularization function $g(\cdot)$. In practice, it is much more transparent to formulate a model directly based on a desired $g(\cdot)$ as opposed to working with some supposedly plausible hyperprior $p(\boldsymbol{\gamma})$ and then inferring the what the associated $g(\cdot)$ would be. In contrast, with $\boldsymbol{\gamma}$-MAP and VB the opposite is true. Choosing a model based on the desirability of some $g(\cdot)$ can lead to a model with an underlying $p(\boldsymbol{\gamma})$ that performs poorly. These issues are discussed in detail (albeit in a somewhat different context) in [50].
- Both VB and $\boldsymbol{\gamma}$-MAP give rigorous bounds on the model evidence $\log p(B)$.

In summary, we hope that these ideas help to bring an insightful perspective to Bayesian source imaging methods, reduce confusion about how different techniques relate, expand the range of feasible applications, and ensure that $\boldsymbol{\gamma}$-MAP, S-MAP, and VB are not underutilized. We also observe a number of surprising similarities or out-right equivalences between what might otherwise appear to be very different methodologies. Additionally, there are numerous promising directions for future research, including time-frequency extensions, alternative covariance component parameterizations, and robust interference suppression.

Appendix A: Empirical Comparison of Convergence Rates for $\boldsymbol{\gamma}$-MAP Algorithms

Several different iterative schemes are discussed in Section III-A for implementing $\boldsymbol{\gamma}$-MAP. While the per-iteration cost is the same for all methods, the convergence rates can be very different. This appendix presents an example of the relative convergence rates on a simple toy problem. A simulated 2D dipolar source was generated and projected to the sensors using the experimental paradigm described in [54]. The signal was corrupted by 10dB additive Gaussian sensor noise. For each algorithm, a flat hyperprior was assumed and $\mathcal{C} = \{\boldsymbol{e}_i \boldsymbol{e}_i^T : i = 1, \ldots, d_\gamma\}$; a single covariance component was used for each voxel and orientation giving $d_\gamma = 3450$ total components. Figure 1 displays the reduction in the $\boldsymbol{\gamma}$-MAP cost function (18) as a function of the iteration number. Consistent with previous observations, the EM updates are considerably slower in reducing the cost. In contrast, the principled updates formed using the convexity-based auxiliary function are relatively similar to the MacKay fixed-point iterations.

Appendix B: Derivation of Alternative $\boldsymbol{\gamma}$-MAP Update Rules

In this section, we re-express the $\boldsymbol{\gamma}$-MAP cost function $\mathcal{L}(\boldsymbol{\gamma})$ in a more convenient form leading to the update rule (30) and a proof that $\mathcal{L}\left(\boldsymbol{\gamma}^{(k+1)}\right) \leq \mathcal{L}\left(\boldsymbol{\gamma}^{(k)}\right)$ at each iteration. In fact, a wide variety of alternative, convergent update rules can be developed by decoupling $\mathcal{L}(\boldsymbol{\gamma})$ using auxiliary functions and an additional set of parameters that can be easily optimized, along with $\boldsymbol{\gamma}$, using coordinate descent.

To begin, the data fit term can be expressed as

$$\operatorname{trace}\left[C_b \Sigma_b^{-1}\right] = \min_X \left\| \frac{\mathsf{B}}{\sqrt{n}} - \sum_{i=1}^{d_\gamma} \mathsf{L}_i X_i \right\|_{\Sigma_\epsilon^{-1}}^2 + \sum_{i=1}^{d_\gamma} \gamma_i^{-1} \|X_i\|_{\mathcal{F}}^2, \tag{70}$$

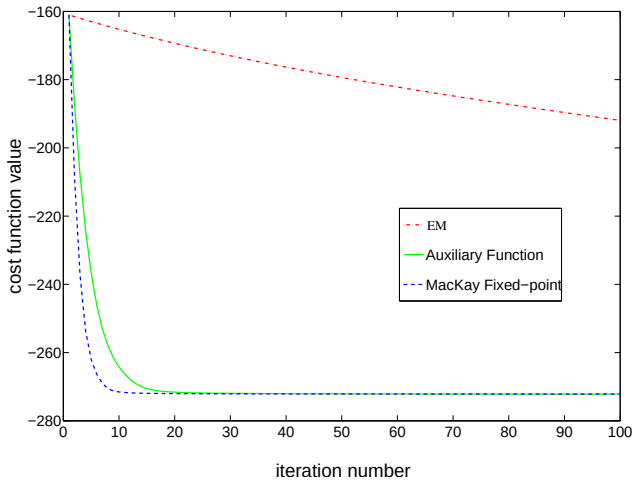

Fig. 1. Convergence comparisons for three methods of performing $\boldsymbol{\gamma}$-MAP.

where $X = \left[X_1^T, \ldots, X_{d_\gamma}^T\right]^T$. Likewise, because the log-determinant term of $\mathcal{L}(\boldsymbol{\gamma})$ is concave in $\boldsymbol{\gamma}$ (see [46]), it can be expressed as an minimum over upper-bounding hyperplanes via

$$\log|\Sigma_b| = \min_{\boldsymbol{z}} \left[\boldsymbol{z}^T\boldsymbol{\gamma} - h^*(\boldsymbol{z})\right], \tag{71}$$

where $h^*(\boldsymbol{z})$ is the concave conjugate of $\log|\Sigma_b|$. For our purposes below, we will never actually have to compute $h^*(\boldsymbol{z})$.

Dropping the minimizations and combining terms from (70) and (71) leads to the modified cost function (or upper bound)

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}, X, \boldsymbol{z}) &\triangleq \left\|\frac{\mathsf{B}}{\sqrt{n}} - \sum_{i=1}^{d_\gamma}\mathsf{L}_i X_i\right\|_{\Sigma_\epsilon^{-1}}^2 + \sum_{i=1}^{d_\gamma}\gamma_i^{-1}\|X_i\|_{\mathcal{F}}^2 + \boldsymbol{z}^T\boldsymbol{\gamma} - h^*(\boldsymbol{z}) \\ &= \left\|\frac{\mathsf{B}}{\sqrt{n}} - \sum_{i=1}^{d_\gamma}\mathsf{L}_i X_i\right\|_{\Sigma_\epsilon^{-1}}^2 + \sum_{i=1}^{d_\gamma}\left[\gamma_i^{-1}\|X_i\|_{\mathcal{F}}^2 + z_i\gamma_i\right] - h^*(\boldsymbol{z}),\end{aligned} \tag{72}$$

where by construction

$$\mathcal{L}(\boldsymbol{\gamma}) = \min_X \min_{\boldsymbol{z}} \mathcal{L}(\boldsymbol{\gamma}, X, \boldsymbol{z}). \tag{73}$$

It is straightforward to show that if $\{\hat{\boldsymbol{\gamma}}, \hat{X}, \hat{\boldsymbol{z}}\}$ is a local minimum to $\mathcal{L}(\boldsymbol{\gamma}, X, \boldsymbol{z})$, then $\hat{\boldsymbol{\gamma}}$ is a local minimum to $\mathcal{L}(\boldsymbol{\gamma})$. Likewise, if $\{\hat{\boldsymbol{\gamma}}, \hat{X}, \hat{\boldsymbol{z}}\}$ is a global minimum of $\mathcal{L}(\boldsymbol{\gamma}, X, \boldsymbol{z})$, then $\hat{\boldsymbol{\gamma}}$ globally minimizes $\mathcal{L}(\boldsymbol{\gamma})$.

Since direct optimization of $\mathcal{L}(\boldsymbol{\gamma})$ may be difficult, we can instead iteratively optimize $\mathcal{L}(\boldsymbol{\gamma}, X, \boldsymbol{z})$ via coordinate descent over $\boldsymbol{\gamma}$, X, and $\boldsymbol{z}$. In each case, when two are held fixed, the third can be globally minimized in closed form. (In the case of $\boldsymbol{\gamma}$ this occurs because each γ_i can be optimized independently given fixed values for X and $\boldsymbol{z}$.) This ensures that each cycle will reduce $\mathcal{L}(\boldsymbol{\gamma}, X, \boldsymbol{z})$, but more importantly, will reduce $\mathcal{L}(\boldsymbol{\gamma})$ (or leave it unchanged if a fixed-point or limit cycle is reached). The associated update rules from this process are as follows.

With $\boldsymbol{z}$ and X fixed, the minimizing $\boldsymbol{\gamma}$ is obtained by solving

$$\nabla_{\boldsymbol{\gamma}}\mathcal{L}(\boldsymbol{\gamma}, X, \boldsymbol{z}) = 0. \tag{74}$$

This leads to the update

$$\gamma_i^{\text{new}} \rightarrow \frac{\|X_i\|_{\mathcal{F}}}{\sqrt{z_i}}. \tag{75}$$

The optimal X (with $\boldsymbol{\gamma}$ and $\boldsymbol{z}$ fixed) is just the standard weighted minimum-norm solution given by

$$X_i^{\text{new}} \rightarrow \gamma_i \mathsf{L}_i^T \Sigma_b^{-1} \frac{\mathsf{B}}{\sqrt{n}} \tag{76}$$

for each i. Finally, the minimizing $\boldsymbol{z}$ equals the slope at the current $\boldsymbol{\gamma}$ of $\log|\Sigma_b|$. As such, we have

$$z_i^{\text{new}} \rightarrow \nabla_{\gamma_i}\log|\Sigma_b| = \text{trace}\left[\mathsf{L}_i^T\Sigma_b^{-1}\mathsf{L}_i\right]. \tag{77}$$

By merging these three rules into a single $\boldsymbol{\gamma}$ update, we arrive at the exact $\boldsymbol{\gamma}$-MAP iteration given by (30).

Using a slightly different set of auxiliary functions other updates, such as the standard EM rule, can be easily derived [46]. For example, using a standard determinant identity, we get

$$\log|\Sigma_b| \equiv \log\left|\mathsf{L}^T\Sigma_\epsilon^{-1}\mathsf{L} + \Sigma_{\mathsf{s}}^{-1}\right| + \sum_i r_i \log\gamma_i \tag{78}$$

up to a constant. The first term is jointly concave in each γ_i^{-1} and so we can write

$$\log|\Sigma_b| \equiv \min_{\boldsymbol{z}}\left[\boldsymbol{z}^T\boldsymbol{\gamma}^{-1} - h^*(\boldsymbol{z})\right] + \sum_i r_i\log\gamma_i \;\leq\; \boldsymbol{z}^T\boldsymbol{\gamma}^{-1} - h^*(\boldsymbol{z}) + \sum_i r_i\log\gamma_i, \tag{79}$$

where $h^*(\boldsymbol{z})$ is the concave conjugate with respect to $\boldsymbol{\gamma}^{-1}$ (elementwise inverse). Substituting this expression into (72) and noting that the optimal $\boldsymbol{z}$, for fixed X and $\boldsymbol{\gamma}$, is $z_i^{\text{new}} \rightarrow \text{trace}\left[\gamma_i I - \gamma_i\mathsf{L}_i^T\left(\Sigma_b\right)^{-1}\mathsf{L}_i\gamma_i\right]$, leads to the exact EM update rules from Section IV-A. This formulation naturally allows us to handle flexible hyperprior kernels $f_i(\gamma_i)$, e.g., $f_i(\gamma)$ concave with respect to γ_i^{-1}. However, this particular bound leads to slower convergence because it is *much* looser around zero than the bound described previously and therefore fails to fully penalize redundant or superfluous components. This prevents the associated hyperparameters from going to zero very quickly, drastically slowing the convergence rate (details will be presented in a subsequent publication).

Also, this process can be used to show that the fixed-point update (29) is iteratively solving a particular min-max problem in search of a saddle point [46]. To accomplish this we define

$$\boldsymbol{\lambda} \triangleq \log\boldsymbol{\gamma}, \tag{80}$$

where the logarithm is understood to apply elementwise. We then note that

$$\log|\Sigma_b| = \log\left|\Sigma_\epsilon + \sum_i \exp(\lambda_i)\mathsf{L}_i\mathsf{L}_i^T\right|. \tag{81}$$

By computing the Hessian with respect to $\boldsymbol{\lambda}$ and rearranging terms, this expression can be shown to be *convex* in $\boldsymbol{\lambda}$. It therefore admits the representation

$$\log\left|\Sigma_\epsilon + \sum_i \exp(\lambda_i)\mathsf{L}_i\mathsf{L}_i^T\right| = \max_{\boldsymbol{z}}\left[\boldsymbol{z}^T\boldsymbol{\lambda} - h^*(\boldsymbol{z})\right] = \max_{\boldsymbol{z}}\left[\boldsymbol{z}^T\log\boldsymbol{\gamma} - h^*(\boldsymbol{z})\right], \tag{82}$$

where now $h^*(\boldsymbol{z})$ is the *convex* conjugate. Substituting this expression into (72) leads to an iterative min-max procedure, whereby a lower bound is repeatedly minimized, followed by subsequent tightening (maximization) of the bound. Unfortunately though, proving convergence in this context is more difficult.

Appendix C: Derivation of Approximate Variational Free Energy from [13]

The variational free energy from Section V-B can be manipulated via

$$\begin{aligned} F &= \int \mathcal{N}\left(\boldsymbol{\lambda}|\boldsymbol{\mu}_\lambda,\Sigma_\lambda\right)\log\frac{p(\boldsymbol{\lambda},B)}{\mathcal{N}\left(\boldsymbol{\lambda}|\boldsymbol{\mu}_\lambda,\Sigma_\lambda\right)}d\boldsymbol{\lambda} \\ &= \int \mathcal{N}\left(\boldsymbol{\lambda}|\boldsymbol{\mu}_\lambda,\Sigma_\lambda\right)\log p(\boldsymbol{\lambda},B)d\boldsymbol{\lambda} - H[\mathcal{N}\left(\boldsymbol{\lambda}|\boldsymbol{\mu}_\lambda,\Sigma_\lambda\right)] \\ &= \int \mathcal{N}\left(\boldsymbol{\lambda}|\boldsymbol{\mu}_\lambda,\Sigma_\lambda\right)\log p(\boldsymbol{\lambda},B)d\boldsymbol{\lambda} + \frac{1}{2}\left(d_\gamma\log 2\pi e + \log|\Sigma_\lambda|\right), \end{aligned} \tag{83}$$

where $H[\cdot]$ is the differential entropy. The remaining integral is intractable however, so a second-order Taylor series approximation of $\ell(\boldsymbol{\lambda}) \triangleq \log p(\boldsymbol{\lambda},B)$ about $\boldsymbol{\mu}_\lambda$ is adopted. Note that this second Laplacian assumption is separate and different from the original assumption that $p(\boldsymbol{\lambda}|B)$ is Gaussian. (Unlike the original, there is no re-normalization after the second-order approximation, which ultimately leads to the extra terms in (67).) This gives

$$\begin{aligned} \ell(\boldsymbol{\lambda}) &= \ell(\boldsymbol{\mu}_\lambda) + \ell'(\boldsymbol{\mu}_\lambda)(\boldsymbol{\lambda}-\boldsymbol{\mu}_\lambda) + \frac{1}{2}(\boldsymbol{\lambda}-\boldsymbol{\mu}_\lambda)^T\ell''(\boldsymbol{\mu}_\lambda)(\boldsymbol{\lambda}-\boldsymbol{\mu}_\lambda) + \ldots \\ &\approx \ell(\boldsymbol{\mu}_\lambda) + \ell'(\boldsymbol{\mu}_\lambda)(\boldsymbol{\lambda}-\boldsymbol{\mu}_\lambda) + \frac{1}{2}(\boldsymbol{\lambda}-\boldsymbol{\mu}_\lambda)^T\ell''(\boldsymbol{\mu}_\lambda)(\boldsymbol{\lambda}-\boldsymbol{\mu}_\lambda), \end{aligned} \tag{84}$$

where $\ell'(\boldsymbol{\mu}_\lambda) \triangleq \left.\frac{\partial \ell(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}}\right|_{\boldsymbol{\lambda}=\boldsymbol{\mu}_\lambda}$ and $\ell''(\boldsymbol{\mu}_\lambda) \triangleq \left.\frac{\partial^2 \ell(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}^2}\right|_{\boldsymbol{\lambda}=\boldsymbol{\mu}_\lambda}$. Plugging this result into (83) gives

$$\begin{aligned}\int \mathcal{N}(\boldsymbol{\lambda}|\boldsymbol{\mu}_\lambda, \Sigma_\lambda) \log p(\boldsymbol{\lambda}, B) d\boldsymbol{\lambda} &\approx \mathrm{E}_{\mathcal{N}(\boldsymbol{\lambda}|\boldsymbol{\mu}_\lambda,\Sigma_\lambda)}\left[\ell(\boldsymbol{\mu}_\lambda) + \ell'(\boldsymbol{\mu}_\lambda)(\boldsymbol{\lambda}-\boldsymbol{\mu}_\lambda) + \frac{1}{2}(\boldsymbol{\lambda}-\boldsymbol{\mu}_\lambda)^T \ell''(\boldsymbol{\mu}_\lambda)(\boldsymbol{\lambda}-\boldsymbol{\mu}_\lambda)\right] \\ &= \ell(\boldsymbol{\mu}_\lambda) + \mathrm{E}_{\mathcal{N}(\boldsymbol{\lambda}|\boldsymbol{\mu}_\lambda,\Sigma_\lambda)}\left[\frac{1}{2}\mathrm{trace}[(\boldsymbol{\lambda}-\boldsymbol{\mu}_\lambda)(\boldsymbol{\lambda}-\boldsymbol{\mu}_\lambda)^T \ell''(\boldsymbol{\mu}_\lambda)]\right] \\ &= \ell(\boldsymbol{\mu}_\lambda) + \frac{1}{2}\mathrm{trace}[\Sigma_\lambda \ell''(\boldsymbol{\mu}_\lambda)]. \end{aligned} \tag{85}$$

Combining (85) with (83), we arrive at

$$F \approx \hat{F} \triangleq \ell(\boldsymbol{\mu}_\lambda) + \frac{1}{2}\mathrm{trace}[\Sigma_\lambda \ell''(\boldsymbol{\mu}_\lambda)] + \frac{1}{2}\left(d_\gamma \log 2\pi e + \log|\Sigma_\lambda|\right). \tag{86}$$

If we assume $\Sigma_\lambda = -\left[\ell''(\boldsymbol{\mu}_\lambda)\right]^{-1}$, then (86) reduces to

$$\begin{aligned}\hat{F} &= \ell(\boldsymbol{\mu}_\lambda) + \frac{1}{2}\mathrm{trace}[I] + \frac{1}{2}\left(d_\gamma \log 2\pi e + \log|\Sigma_\lambda|\right) \\ &= \ell(\boldsymbol{\mu}_\lambda) + \frac{1}{2}\left(d_\gamma \log 2\pi + \log|\Sigma_\lambda|\right), \end{aligned} \tag{87}$$

which is the expression given in [13] (assuming the appropriate hyperprior on $p(\boldsymbol{\lambda})$.

It is worth pointing out that, for fixed $\boldsymbol{\mu}_\lambda$, the selection $\Sigma_\lambda = -\left[\ell''(\boldsymbol{\mu}_\lambda)\right]^{-1}$ minimizes (86) with respect to Σ_λ. However, selecting $\boldsymbol{\mu}_\lambda$ to be the maximum of $\ell(\lambda)$ and then choosing Σ_λ as above is not guaranteed to *jointly* maximize $\hat{F}$ over $\boldsymbol{\mu}_\lambda$ and Σ_λ. This is because of the $\boldsymbol{\mu}_\lambda$-dependency of $\ell''(\boldsymbol{\mu}_\lambda)$. So the Laplace procedure discussed in Section V-B is not explicitly maximizing $\hat{F}$. Of course maximizing $\hat{F}$ is not necessarily the ideal objective anyway, since $\hat{F}$ is not a strict lower bound on $\log p(B)$.

Appendix D: Combining the Laplace (Fixed-Form) and Mean-Field Approximations

The mean-field approximation assumes that $p(\mathsf{S}, \boldsymbol{\gamma}|B) \approx \hat{p}(\mathsf{S}|B)\hat{p}(\boldsymbol{\gamma}|B)$. While the optimal $\hat{p}(\mathsf{S}|B)$ can be computed in closed-form using the method outlined in Section V-A, $\hat{p}(\boldsymbol{\gamma}|B)$, and therefore the variational bound on $\log p(B)$ cannot in general (unless a conjugate hyperprior is used). In contast, the fixed-form assumption from Section V-B works with a variety of hyperpriors but does not lead to a rigorous bound on $\log p(B)$ and can be expensive to compute.

One potential remedy is to combine the two approaches. This involves reparameterizing via $\lambda_i \triangleq \log \gamma_i$ as before and then assuming the approximate posterior distribution of these new hyperparameters $\boldsymbol{\lambda} \triangleq [\lambda_1, \ldots, \lambda_{d_\gamma}]^T$ is

$$q(\boldsymbol{\lambda}) \approx \mathcal{N}(\boldsymbol{\lambda}|\boldsymbol{\mu}_\lambda, \Sigma_\lambda). \tag{88}$$

Note that the optimal distribution for $q(\mathsf{S})$ is Gaussian with no fixed-form assumptions (we will again work with the pseudo-source parameterization for convenience). We will also assume that the (non-conjugate) prior on $\boldsymbol{\lambda}$ is iid Gaussian, with

$$p(\lambda_i) = \mathcal{N}\left(\lambda_i|\mu_0, \sigma_0^2\right), \quad \forall i. \tag{89}$$

This assumption provides consistency with [13] but is by no means required by what follows. Given these provisions, the goal is to minimize $\mathrm{KL}[q(\mathsf{S})q(\boldsymbol{\lambda})\|p(\mathsf{S}, \boldsymbol{\lambda}|B)]$ which is equivalent to maximizing

$$F = \int q(\mathsf{S})q(\boldsymbol{\lambda}) \log \frac{p(B|\mathsf{S})p(\mathsf{S}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{q(\mathsf{S})q(\boldsymbol{\lambda})} d\mathsf{S} d\boldsymbol{\lambda}. \tag{90}$$

Optimization over $q(\mathsf{S})$ given a fixed $q(\boldsymbol{\lambda}) = q^{(k)}(\boldsymbol{\lambda}) = \mathcal{N}\left(\boldsymbol{\lambda}|\boldsymbol{\mu}_\lambda^{(k)}, \Sigma_\lambda^{(k)}\right)$ is simple. All we need to know about $q^{(k)}(\boldsymbol{\lambda})$ is

$$\bar{\gamma}_i^{(k)} = \mathrm{E}_{q^{(k)}(\boldsymbol{\lambda})}[\gamma_i^{-1}]^{-1} = \mathrm{E}_{q^{(k)}(\boldsymbol{\lambda})}[\exp(-\lambda_i)]^{-1} = \exp\left[\mu_{\lambda,i}^{(k)} - \frac{\Sigma_{\lambda,ii}^{(k)}}{2}\right], \quad \forall i. \tag{91}$$

The optimal $q^{(k+1)}(\mathsf{S})$ is then Gaussian, with moments given by (23) with $\gamma_i = \bar{\gamma}_i^{(k)}$ from above.

Optimization over $q(\boldsymbol{\lambda})$ given fixed $q(\mathsf{S}) = q^{(k+1)}(\mathsf{S})$ is a bit more complicated. We have

$$\begin{aligned} q^{(k+1)}(\boldsymbol{\lambda}) &\rightarrow \arg\max_{q(\boldsymbol{\lambda})} \int q^{(k+1)}(\mathsf{S}) q(\boldsymbol{\lambda}) \log \frac{p(B, \mathsf{S}, \boldsymbol{\lambda})}{q^{(k+1)}(\mathsf{S}) q(\boldsymbol{\lambda})} d\boldsymbol{\lambda} d\mathsf{S} \\ &= \arg\max_{q(\boldsymbol{\lambda})} \int q(\boldsymbol{\lambda}) \log \frac{p(\mathsf{S}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{q(\boldsymbol{\lambda})} d\boldsymbol{\lambda} d\mathsf{S}, \\ &= \arg\max_{q(\boldsymbol{\lambda})} \left[\int q(\boldsymbol{\lambda}) \log \prod_i \bar{p}\left(\|\mathsf{S}_i\|_{\mathcal{F}} = \bar{\mathsf{S}}_i^{(k+1)}|\lambda_i\right) p(\lambda_i) d\boldsymbol{\lambda} - \int q(\boldsymbol{\lambda}) \log q(\boldsymbol{\lambda}) d\boldsymbol{\lambda}\right], \end{aligned} \tag{92}$$

where $\bar{p}\left(\|\mathsf{S}_i\|_{\mathcal{F}} = \bar{\mathsf{S}}_i^{(k+1)}|\lambda_i\right)$ is defined as in Section V-A. Because of the natural factorization in (92), the optimal $q(\boldsymbol{\lambda})$ will factor as well and we can solve for each $q(\lambda_i)$ separately. Thus, (92) reduces to

$$\begin{aligned} q^{(k+1)}(\lambda_i) &\rightarrow \arg\max_{q(\lambda_i)} \left[\int q(\lambda_i) \log \bar{p}\left(\|\mathsf{S}_i\|_{\mathcal{F}} = \bar{\mathsf{S}}_i^{(k+1)}|\lambda_i\right) p(\lambda_i) d\lambda_i - \int q(\lambda_i) \log q(\lambda_i) d\lambda_i \right] \\ &= \arg\min_{q(\lambda_i)} \left[\int q(\lambda_i) \left[n r_i \lambda_i + \left(\bar{\mathsf{S}}_i^{(k+1)}\right)^2 e^{-\lambda_i} + \frac{1}{\sigma_0^2}\left(\lambda_i^2 - 2\mu_0\lambda_i\right) \right] d\lambda_i \;\; - \;\; 2\mathrm{H}[q(\lambda_i)] \right] \qquad (93) \\ &= \arg\min_{q(\lambda_i)} \left[n r_i \mathrm{E}_{q(\lambda_i)}[\lambda_i] + \left(\bar{\mathsf{S}}_i^{(k+1)}\right)^2 \mathrm{E}_{q(\lambda_i)}[e^{-\lambda_i}] + \frac{1}{\sigma_0^2}\left(\mathrm{E}_{q(\lambda_i)}[\lambda_i^2] - 2\mu_0 \mathrm{E}_{q(\lambda_i)}[\lambda_i]\right) - 2\mathrm{H}[q(\lambda_i)] \right]. \end{aligned}$$

Thus far we have placed no restrictions on $q(\boldsymbol{\lambda})$; we now apply the fixed-form, Gaussian approximation to (93). Using

$$\begin{aligned} \mathrm{E}_{q(\lambda_i)}[e^{-\lambda_i}] &= \exp\left[-\mu_{\lambda,i} + \frac{\Sigma_{\lambda,ii}}{2} \right] \\ \mathrm{E}_{q(\lambda_i)}[\lambda_i^2] &= \mu_{\lambda,i}^2 + \Sigma_{\lambda,ii} \\ H[q(\lambda_i)] &= \frac{1}{2}[\log 2\pi e + \log \Sigma_{\lambda,ii}], \end{aligned} \qquad (94)$$

the relevant optimization problem becomes

$$\mu_{\lambda,i}^{(k+1)}, \Sigma_{\lambda,ii}^{(k+1)} \rightarrow \arg\min_{\mu_{\lambda,i}, \Sigma_{\lambda,ii}} \left[n r_i \mu_{\lambda,i} + \left(\bar{\mathsf{S}}_i^{(k+1)}\right)^2 \exp\left[-\mu_{\lambda,i} + \frac{\Sigma_{\lambda,ii}}{2} \right] + \frac{\mu_{\lambda,i}^2 + \Sigma_{\lambda,ii} - 2\mu_0\mu_{\lambda,i}}{\sigma_0^2} - \log \Sigma_{\lambda,ii} \right]. \qquad (95)$$

This optimization problem is jointly convex in $\mu_{\lambda,i}$ and $\Sigma_{\lambda,ii}$ and so can be minimized (or partially minimized for generalized EM) via a variety of simple procedures. This must of course be repeated for each i.

This formulation is attractive for several reasons. First, at each iteration, the approximation $q(\mathsf{S})$ is dependent on the posterior variance Σ_λ; the more uncertain we are about λ_i, the more likely it is that we force the estimated covariance component to zero through (91). Secondly, we get strict, lower bound on $\log p(B)$ that can be derived in closed-form and does not require the expensive computation of a full posterior covariance on $\boldsymbol{\lambda}$. Finally, a conjugate prior is not required to proceed.

If we are averse to the computation of (95) at every iteration, we could instead follow the Laplace approximation from Section V-B to get $\mathcal{N}(\boldsymbol{\lambda}|\boldsymbol{\mu}_\lambda, \Sigma_\lambda)$ and then compute a single update of (91) to find $q(\mathsf{S})$ and subsequently F. This would still give a proper bound on $\log p(B)$ and a Σ_λ-dependent estimate of $p(S|B)$, but would only require the diagonal elements of Σ_λ.

REFERENCES

[1] H. Attias, "A variational Bayesian framework for graphical models," *Advances in Neural Information Processing Systems 12*, MIT Press 2000.

[2] S. Baillet, J. Mosher, and R. Leahy, "Electromagnetic brain mapping," *IEEE Signal Processing Magazine*, pp. 14–30, Nov. 2001.

[3] M. Beal and Z. Ghahrarmani, "The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures," *Bayesian Statistics 7*, Oxford University Press 2002.

[4] J. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. New York: Springer-Verlag, 1985.

[5] C. Bishop and M. Tipping, "Variational relevance vector machines," *Proc. 16th Conf. on Uncertainty in Artificial Intelligence*, pp. 46–53, 2000.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[7] S. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Processing*, vol. 53, no. 7, pp. 2477–2488, April 2005.

[8] A. Dempster, D. Rubin, and R. Tsutakawa, "Estimation in covariance components models," *J. American Statistical Association*, vol. 76, pp. 341–353, June 1981.

[9] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proc. National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, March 2003.

[10] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Information Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

[11] M. Figueiredo, "Adaptive sparseness using Jeffreys prior," *Advances in Neural Information Processing Systems 14*, pp. 697–704, MIT Press 2002.

[12] K. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, and J. Ashburner, "Classical and Bayesian inference in neuroimaging: Theory," *NeuroImage*, vol. 16, pp. 465–483, 2002.

[13] K. Friston, L. Harrison, J. Daunizeau, S. Kiebel, C. Phillips, N. Trujillo-Barreto, R. Henson, G. Flandin, and J. Mattout, "Multiple Sparse Priors for the MEG/EEG Inverse Problem," NeuroImage, vol. 39, pp. 1104-1120, 2008.

[14] K. Friston, J. Mattout, N. Trujillo-Barreto, J. Ashburner, and W. Penny, "Variationl free energy and the Laplace approximation," *NeuroImage*, vol. 34, pp. 220–234, 2006.

[15] J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Transactions on Information Theory*, vol. 50, no. 6, pp. 1341–1344, June 2004.

[16] I. Gorodnitsky, J. George, and B. Rao, "Neuromagnetic source imaging with FOCUSS: A recursive weighted minimum norm algorithm," *Journal Electroencephalography and Clinical Neurophysiology*, vol. 95, no. 4, pp. 231–251, Oct. 1995.

[17] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Information Theory*, vol. 49, pp. 3320–3325, Dec. 2003.

[18] M. Huang, A. Dale, T. Song, E. Halgren, D. Harrington, I. Podgorny, J. Canive, S. Lewis, and R. Lee, "Vector-based spatial-temporal minimum ℓ_1-norm solution for MEG," *NeuroImage*, vol. 31, no. 3, pp. 1025–37, 2006.

[19] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Processing*, (to appear).

[20] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.

[21] T. Limpiti, B. V. Veen, and R. Wakai, "Cortical patch basis model for spatially extended neural activity," *IEEE Trans. Biomedical Engineering*, vol. 53, no. 9, pp. 1740–1754, Sept. 2006.

[22] D. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, 1992.

[23] D. MacKay, "Bayesian non-linear modeling for the energy prediction competition," *ASHRAE Transactions*, vol. 100, no. 2, pp. 1053–1062, 1994.

[24] J. Mattout, C. Phillips, W. Penny, M. Rugg, and K. Friston, "MEG source localization under multiple constraints: An extended Bayesian framework," *NeuroImage*, vol. 30, pp. 753–767, 2006.

[25] S. Nagarajan, H. Attias, K. Hild, and K. Sekihara, "A graphical model for estimating stimulus-evoked brain responses in noisy MEG data with large background brain activity," *NeuroImage*, vol. 30, no. 2, pp. 400–416, 2006.

[26] R. Neal, *Bayesian Learning for Neural Networks*. New York: Springer-Verlag, 1996.

[27] A. Nummenmaa, T. Auranen, M. Hämäläinen, I. Jääskeläinen, J. Lampinen, M. Sams, and A. Vehtari, "Hierarchical Bayesian estimates of distributed MEG sources: Theoretical aspects and comparison of variational and MCMC methods," *Neuroimage*, vol. 35, no. 2, pp. 669-685, 2007

[28] J. Palmer and K. Kreutz-Delgado, "Modeling and estimation of dependent subspaces with non-radially symmetric and skewed densities," in preparation.

[29] J. Palmer, D. Wipf, K. Kreutz-Delgado, and B. Rao, "Variational EM algorithms for non-Gaussian latent variable models," *Advances in Neural Information Processing Systems 18*, pp. 1059–1066, MIT Press 2006.

[30] R. Pascual-Marqui, "Standardized low resolution brain electromagnetic tomography (sLORETA): Technical details," *Methods and Findings in Experimental and Clinical Pharmacology*, vol. 24, Suppl D, pp. 5–12, 2002.

[31] C. Phillips, J. Mattout, M. Rugg, P. Maquet, and K. Friston, "An empirical Bayesian solution to the source reconstruction problem in EEG," *NeuroImage*, vol. 24, pp. 997–1011, January 2005.

[32] C. Phillips, M. Rugg, and K. Friston, "Anatomically informed basis functions for EEG source localization: combining functional and anatomical constraints," *NeuroImage*, vol. 13, no. 3, pt. 1, pp. 678–95, July 2002.

[33] R. Ramírez, "Neuromagnetic source imaging of spontaneous and evoked human brain dynamics," PhD Thesis, New York University, May 2005.

[34] R. Ramírez and S. Makeig, "Neuroelectromagnetic source imaging using multiscale geodesic neural bases and sparse Bayesian learning," *Human Brain Mapping, 12th Annual Meeting*, Florence, Italy, June 2006.

[35] B. Rao, K. Engan, S. F. Cotter, J. Palmer, and K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Trans. Signal Processing*, vol. 51, no. 3, pp. 760–770, March 2003.

[36] M. Sahani and S. Nagarajan, "Reconstructing MEG sources with unknown correlations," *Advances in Neural Information Processing Systems 16*, MIT Press 2004.

[37] J. Sarvas, "Basic methematical and electromagnetic concepts of the biomagnetic inverse problem," *Phys. Med. Biol.*, vol. 32, pp. 11–22, 1987.

[38] M. Sato, T. Yoshioka, S. Kajihara, K. Toyama, N. Goda, K. Doya, and M. Kawato, "Hierarchical Bayesian estimation for MEG inverse problem," *NeuroImage*, vol. 23, pp. 806–826, 2004.

[39] K. Sekihara, S. Nagarajan, D. Poeppel, and A. Marantz, "Performance of an MEG adaptive-beamformer technique in the presence of correlated neural activities: Effects on signal intensity and time-course estimates," *IEEE Trans. Biomedical Engineering*, vol. 40, no. 10, pp. 1534–1546, 2002.

[40] K. Sekihara, M. Sahani, and S. Nagarajan, "Localization bias and spatial resolution of adaptive and non-adaptive spatial filters for MEG source reconstruction," *NeuroImage*, vol. 25, pp. 1056–1067, 2005.

[41] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[42] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Information Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.

[43] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals," *IEEE Trans. Information Theory*, vol. 52, no. 3, pp. 1030–1051, March 2006.

[44] K. Uutela, M. Hämäläinen, and E. Somersalo, "Visualization of magnetoencephalographic data using minimum current estimates," *NeuroImage*, vol. 10, pp. 173–180, 1999.

[45] B. V. Veen, W. van Drongelen, M. Yuchtman, and A. Suzuki, "Localization of brain electrical activity via linearly constrained minimum variance spatial filtering," *IEEE Trans. Biomedical Engineering*, vol. 44, no. 9, pp. 867–880, 1997.

[46] D. Wipf, "Bayesian methods for finding sparse representations," PhD Thesis, University of California, San Diego, 2006.

[47] D. Wipf and S. Nagarajan, "Beamforming using the relevance vector machine," *International Conference on Machine Learning*, June 2007.

[48] D. Wipf and S. Nagarajan, "A new view of automatic relevance determination," *Advancies in Neural Information Processing Systems 20* (to appear), 2008.

[49] D. Wipf, J. Owen, H. Attias, K. Sekihara, and S. Nagarajan, "Estimating the Location and Orientation of Complex, Correlated Neural Activity using MEG," submitted.

[50] D. Wipf, J. Palmer, B. Rao, and K. Kreutz-Delgado, "Performance analysis of latent variable models with sparse priors," *IEEE International Conf. Acoustics, Speech, and Signal Processing*, April 2007.

[51] D. Wipf, R. Ramírez, J. Palmer, S. Makeig, and B. Rao, "Analysis of empirical Bayesian methods for neuroelectromagnetic source localization," *Advances in Neural Information Processing Systems 19*, MIT Press 2007.

[52] D. Wipf and B. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Processing*, vol. 52, no. 8, pp. 2153–2164, August 2004.

[53] M. Zoltowski, "On the performance analysis of the MVDR beamformer in the presence of correlated interference," *IEEE Trans. Signal Processing*, vol. 36, pp. 945–947, 1988.

[54] J. Zumer, H. Attias, K. Sekihara, and S. Nagarajan, "A probabilistic algorithm for interference suppression and source reconstruction from MEG/EEG data," *Advancies in Neural Information Processing Systems 19*, MIT Press 2007.