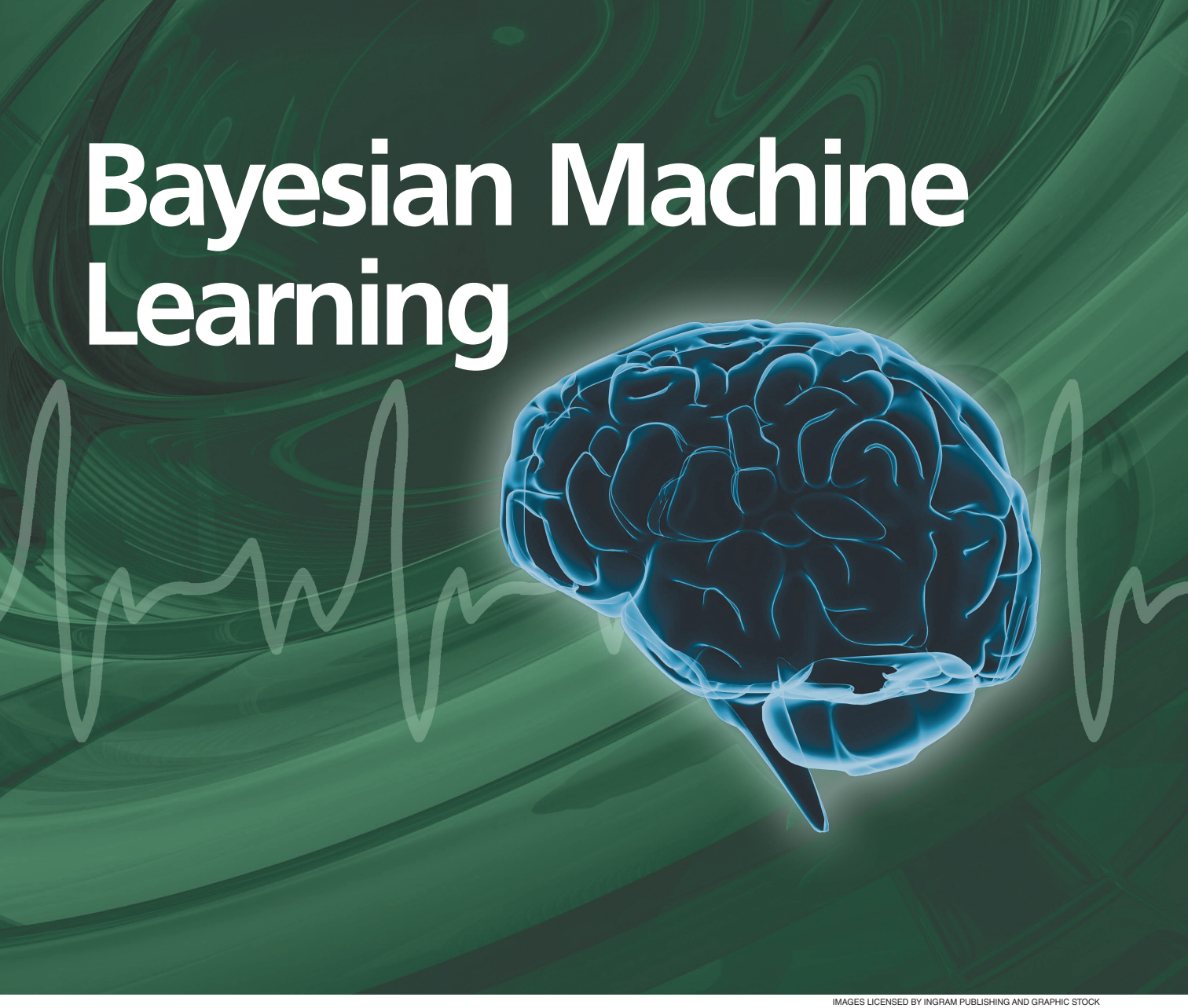
Bayesian Machine Learning



IMAGES LICENSED BY INGRAM PUBLISHING AND GRAPHIC STOCK

Wei Wu, Srikantan Nagarajan, and Zhe Chen

EEG/MEG signal processing measurements

Electroencephalography (EEG) and magnetoencephalography (MEG) are the most common noninvasive brain-imaging techniques for monitoring electrical brain activity and inferring brain function. The central goal of EEG/MEG analysis is to extract informative brain spatio-temporal–spectral patterns or to infer functional connectivity between different brain areas, which is directly useful for neuroscience or clinical investigations. Due to its potentially complex nature [such as nonstationarity, high dimensionality, subject variability, and low signal-to-noise ratio (SNR)], EEG/MEG signal processing poses some great challenges for researchers. These challenges can be addressed in a principled manner via Bayesian machine learning (BML). BML is an emerging field that integrates Bayesian statistics, variational methods, and machine-learning

Digital Object Identifier 10.1109/MSP.2015.2481559
Date of publication: 29 December 2015

1053-5888/16©2016IEEE

techniques to solve various problems from regression, prediction, outlier detection, feature extraction, and classification. BML has recently gained increasing attention and widespread successes in signal processing and big-data analytics, such as in source reconstruction, compressed sensing, and information fusion. To review recent advances and to foster new research ideas, we provide a tutorial on several important emerging BML research topics in EEG/MEG signal processing and present representative examples in EEG/MEG applications.

INTRODUCTION

EEG and MEG are two dominant technologies for the noninvasive measurement of dynamic whole-brain electrical activity. The central goal of EEG/MEG is to extract a wealth of spatiotemporal–spectral patterns of brain activity and functional connectivity from large and complex time-series data to address important questions in neuroscience, neural engineering, and clinical studies. The analysis of EEG/MEG signals poses enormous challenges in signal processing and statistical methods because of their intrinsic high dimensionality, large or sparse sample size, low SNR, nonstationarity (across time/trials/conditions/subjects/groups), nonlinearity (of the feature space with regard to the signal channel space), and other structures afforded by a variety of signal types [e.g., event-related potentials (ERPs) and oscillations]. This article is partly motivated to address these issues in the emerging field of BML. Specifically, we focus on the probabilistic graphical models (within a generative modeling framework) and Bayesian inference algorithms that are used in EEG/MEG signal processing. We identify three emerging and distinct yet somewhat overlapping research lines as the focus of this tutorial: 1) variational Bayesian (VB) methods, 2) sparse Bayesian learning (SBL) methods, and 3) nonparametric Bayesian (NB) methods.

EEG/MEG ORIGINS AND ATTRIBUTES

Neuronal activity gives rise to extracellular electric and magnetic fields, which are detected in MEG and EEG [1]. Electric current contributions from all active cellular processes within a volume of brain tissue superimpose at a given location in the extracellular medium and generate an electric potential, V_e (a scalar measured in volts), with respect to a reference potential. The difference in V_e between two locations gives rise to an electric field (a vector whose amplitude is measured in volts per unit of distance). EEG measures the summation of the synchronous activity (i.e., population V_e in voltage) of millions of neurons that have similar spatial orientation, which can be detected on the scalp surface using conductive electrodes and sensitive bioelectric amplifiers. Synchronized neuronal currents also induce weak magnetic fields that can be detected outside the head using MEG, which makes use of highly sensitive magnetic-field sensors. Most current MEG systems use superconducting quantum interference device sensors, which have a sensitivity on the order of 5–10 fT/square root of Hz, with many more emerging magnetic-sensing technologies that have comparable sensitivity. MEG and EEG signals represent complementary information about the underlying sources that contribute to electric potentials on the scalp and magnetic fields outside the head. Importantly, both EEG and MEG measurements have an excellent temporal resolution (~1 millisecond) in sampling the rich temporal dynamics of neuronal population activity. The spatial sampling precision of EEG and MEG signals is typically around 1–2 centimeters outside the head, but the ultimate spatial resolution of the methods depends on the algorithms and applications.

MOTIVATION

At a very broad level, the two dominant signal processing applications of MEG and EEG are 1) electromagnetic brain imaging and 2) brain-state classification. Signal processing algorithms that reconstruct and visualize the neuronal activities (and functional connectivity) based on MEG/EEG sensor measurements are referred to as the *electromagnetic brain-imaging algorithms*. Electromagnetic brain imaging has diverse applications in basic and clinical neuroscience, such as neurophysiological oscillations imaging associated with normal brain function and how these processes may be altered in disease or during an intervention. Algorithms that make inferences about the behavioral/mental state of a subject are called *brain-state classification algorithms*. Brain-state classification has widespread neuroscience applications, including methods for diagnosis of abnormal brain activity, as well as the development of brain–computer interfaces (BCIs).

Classical signal processing methods for EEG/MEG analyses have mostly included digital filtering, spectral analysis, or source separation [2]. To date, EEG/MEG analyses have encountered many challenges, and the past several decades have witnessed significant progress in the two application domains mentioned previously. Despite the proliferation of BML algorithms for EEG/MEG signal processing, a comprehensive tutorial that provides an overview of important concepts and state-of-the-art development is still lacking in this fast-growing field. This article provides a review on some up-to-date BML algorithms for electromagnetic brain imaging and brain-state classification. Specifically, we focus on a few emerging theme topics, which may help researchers to gain a coherent picture of the assorted methods, develop a deep understanding of their mechanisms, appreciate the most relevant research themes, and spark further research in this area. To this end, our tutorial provides a self-contained methodological guide by disseminating representative BML methods for statistical analysis of EEG/MEG signals, with accessible yet rigorous mathematics describing their central ideas and carefully chosen examples demonstrating their efficacy in practical applications.

LATENT VARIABLE GENERATIVE MODELING FRAMEWORK FOR EEG/MEG SIGNAL PROCESSING

The main applications of this tutorial's focus are EEG/MEG-based brain imaging and brain-state classification, which can be cast under a unifying latent variable modeling framework

$$X = f(Z, \boldsymbol{\theta}) + E, \tag{1}$$

where X is the observed variables; Z is the unobserved latent variables; f is a function of Z that can be linear or nonlinear, parameterized by a known or unknown parameter $\boldsymbol{\theta}$; and E is an error term that consists of uncorrelated or correlated noise and/or interference. The analysis goal is to infer the latent

NOTATIONS

The bold font is used to denote vectors or matrices. Data and parameters are displayed by italic and regular font, respectively. Subscript index t denotes the discrete time, whereas superscript indices s, k, and l denote the subject, condition, and trial, respectively. $\mathbb{E}[\cdot]$ and $\mathbb{C}[\cdot]$ denote the expectation and covariance operators, respectively. Some common notations for probability distributions and stochastic processes used in this article are listed in Table 1.

variables and unknown parameters for data interpretation. When the data likelihood $p(X|Z,\boldsymbol{\theta})$ is a Gaussian distribution, the problem becomes regression, which includes the problem of electromagnetic imaging. When the data likelihood consists of category label c, such as $p(X|Z,\boldsymbol{\theta},c)$ (e.g., the brain-state classification problem), the inference problem becomes $p(c|X) = \int p(c,Z|X)\,dZ = \int p(c|X,Z)p(Z|X)\,dZ$, which can be decomposed into two steps: estimation of $p(Z|X)$ and estimation of $p(c|Z)$.

For electromagnetic brain imaging, f can be approximated as a linear function of Z as follows:

$$X = \mathbf{A}Z + E \text{ or } x_t = \mathbf{A}z_t + e_t, \tag{2}$$

where $X = [x_1, \cdots, x_T] \in \mathbb{R}^{N\times T}$ is the EEG/MEG sensor data matrix with N sensors and T time points. The source activity matrix, $Z = [z_1, \cdots, z_T] \in \mathbb{R}^{M\times T}$ is associated with M latent brain sources. Unless stated otherwise, throughout this tutorial, we assume that the EEG/MEG data have been preprocessed with a proper clean-up procedure to have nonbrain biological artifacts removed. In the spatiotemporal decomposition problem, $\mathbf{A} \in \mathbb{R}^{N\times M}$ is interpreted as a mixing matrix. In the inverse problem, $\mathbf{A}$ is known as the *lead-field matrix*, which can be obtained by solving the forward problem (from Maxwell's equations) based on the structural information of the subject's head, as well as electric and geometric properties of the electric sources and the volume conductor. Each unknown source Z often represents the magnitude of a neural current dipole, projecting from an rth (discretized) voxel or candidate location distributed throughout the brain. These candidate locations can be obtained by segmenting a structural magnetic resonance (MR) scan of a human subject and tesselating the brain volume with a set of vertices. Since the number of brain sources largely outnumber the sensors, i.e., $M \gg N$, reconstructing brain sources from EEG/MEG data is a highly ill-posed problem with an infinite number of solutions. Further anatomical or functional constraints should, therefore, be incorporated to restrict the solution space. Anatomically, in reasonable settings, the dipole sources are restricted to be situated on the cerebral cortex and their orientations perpendicular to the cortical surface. Functionally, spatial smoothness and sparsity are the most widely used constraints.

[TABLE 1] ABBREVIATED NOTATIONS FOR PROBABILITY DISTRIBUTIONS AND STOCHASTIC PROCESSES.

PROBABILITY DISTRIBUTION	NOTATION
$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$	GAUSSIAN (MEAN $\boldsymbol{\mu}$, COVARIANCE $\boldsymbol{\Sigma}$)
$\mathcal{N}^{+}(\boldsymbol{\mu},\boldsymbol{\Sigma})$	HALF-GAUSSIAN (SAME PARAMETERIZATION AS GAUSSIAN)
$\mathcal{G}(a,b)$	GAMMA (SHAPE PARAMETER $a>0$, RATE PARAMETER $b>0$)
$\mathcal{IG}(a,b)$	INVERSE GAMMA (SHAPE PARAMETER $a>0$, SCALE PARAMETER $b>0$)
$\mathcal{W}(\nu,\mathbf{W})$	WISHART (DEGREE OF FREEDOM $\nu>0$, SCALE MATRIX $\mathbf{W}$)
$Be(p)$	BERNOULLI DISTRIBUTION ($0<p<1$)
$\mathcal{B}(a,b)$	BETA DISTRIBUTION (SHAPE PARAMETERS $a>0, b>0$)
GP($\mathbf{0},\mathbf{C}$)	GAUSSIAN PROCESS (GP) (ZERO MEAN FUNCTION, COVARIANCE FUNCTION $\mathbf{C}$)
DP(α, G_0)	DIRICHLET PROCESS (DP) (CONCENTRATION PARAMETER $\alpha>0$, BASE MEASURE G_0)

An extension of the static linear model (2) is a Markovian state-space model, also known as the *dynamic factor analysis model* [3]

$$z_t = \mathbf{F}z_{t-1} + v_t \tag{3}$$
$$x_t = \mathbf{A}z_t + e_t, \tag{4}$$

where $\mathbf{F}$ is a time-invariant state-transition matrix for the latent state z_t, $\mathrm{A} \in \mathbb{R}^{N\times M}$ can be either a factor loading matrix (for modeling low-dimensional sources, where $M \ll N$) or a lead-field matrix (for modeling high-dimensional sources, where $M \gg N$), and $v_t \sim \mathcal{N}(\mathbf{0},\mathrm{I})$ and $e_t \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_e)$ denote zero-mean Gaussian dynamic and measurement noise, respectively. Simple linear algebra will yield $x_t \sim \mathcal{N}(\mathbf{0}, \mathrm{A}\Sigma_z\mathrm{A}^{\mathrm{T}} + \Sigma_e)$, where Σ_z denotes the marginal covariance of z_t.

For the most part, we assume that e_t is a noise-plus-interference term and, for simplicity, that e_ts are drawn independently from $\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_e)$. However, temporal correlations can easily be incorporated using a simple transformation outlined in [4] or using the spatiotemporal framework introduced in [5]. Initially, we assume that $\boldsymbol{\Sigma}_e$ is known; but, in later sections, we also derive how Σ_e is estimated from data.

For brain-state classification, a labeled state or class variable, c, is additionally known given training data. The objective is to determine $p(c|X)$ of the test (unlabeled) data. Three common problems within brain-state classification are disease diagnosis, behavioral state classification, and BCIs. In disease diagnosis, the class or state variable corresponds to the disease diagnosis group, and the observed sensor data are used to make inferences about whether the given EEG/MEG observations carry a signature about the disease group. In behavioral state classification, the problem is to infer the evolving behavioral state from a subject's EEG/MEG data. The most common example is to infer the sleep stage (e.g., awake, slow-wave sleep, and rapid-eye-movement sleep). A second example is to determine the time period when abnormal epileptiform activity are present in EEG/MEG data in a patient with epilepsy. Finally, in BCIs, we learn the brain state associated with the intended state c of the user, and subsequently infer the intended brain state c from new data X.

For all of these problems, within a Bayesian generative modeling framework, all prior assumptions are embedded in the distribution $p(Z)$. If $p(Z)$ is fully or partially known under a specific experimental or clinical paradigm, then the posterior distribution $p(Z|X)$ can be computed via Bayes' rule

$$p(Z|X) = \frac{p(X|Z)p(Z)}{p(X)}. \tag{5}$$

The posterior distribution contains all possible information about the unknown Z conditioned on the observed data X. Two fundamental problems exist for computing $p(Z|X)$. First, for most prior $p(Z)$, it is impossible to analytically compute the distribution $p(X)$ by a direct integration

$$p(X) = \int p(X|Z)p(Z)\,\mathrm{d}Z. \tag{6}$$

The quantity $p(X)$, sometimes referred to as the *model evidence*, can be used to facilitate model selection. When a point estimate for Z is desired, $p(X)$ may not be needed. For example, the maximum a posteriori (MAP) estimate of $Z_{\text{MAP}} = \arg\max_z p(Z|X)$ is invariant to $p(X)$. However, Z_{MAP} may be unrepresentative of posterior mass, and it is often intractable to compute for most $p(Z)$. In addition, the prior $p(Z)$ is often assumed based on neurophysiological constraints or computational considerations, which implicitly or explicitly differentiate a wide variety of estimation methods at a very high level.

For example, we can adopt the following source prior for Z:

$$p(Z|\gamma) \propto \exp\left(-\frac{1}{2}\text{trace}\left[Z^{\mathrm{T}}\Sigma_z^{-1}Z\right]\right), \Sigma_z = \sum_{i=1}^{d_\gamma}\gamma_i \mathbf{C}_i. \tag{7}$$

This is equivalent to applying independently, at each time point, a zero-mean Gaussian distribution with covariance Σ_z to each column of Z. Here, $\gamma \triangleq \text{diag}(\gamma_1, \ldots, \gamma_{d_\gamma})$ is a diagonal matrix consisting of d_γ nonnegative hyperparameters that control the relative contribution of each covariance basis matrix $\mathbf{C}_i$. While the hyperparameters are unknown, the set of components $\mathbf{C} \triangleq \{\mathbf{C}_i : i = 1, \ldots, d_\gamma\}$ is assumed to be fixed and known. Such a prior formulation is extremely flexible in that a rich variety of candidate covariance bases can be proposed. Moreover, this structure has been advocated in the context of neuroelectromagnetic source imaging [6], [7]. We can also assume a hyperprior on γ of the form

$$p(\gamma) = \prod_{i=1}^{d_\gamma}\frac{1}{2}\exp\left[-f_i(\gamma_i)\right], \tag{8}$$

where each $f_i(\cdot)$ is a prespecified function. Within such a hierarchical latent modeling framework, the implicit prior on Z, obtained by marginalizing the unknown γ, is known as a *Gaussian scale mixture*

$$p(Z) = \int p(Z|\gamma)p(\gamma)\,\mathrm{d}\gamma. \tag{9}$$

In the condition where the matrix $\mathbf{A}$ is unknown, one also needs to specify a prior distribution $p(\mathbf{A})$. This is used in spatiotemporal decomposition problems for both brain imaging and brain-state classification. In EEG/MEG spatiotemporal decomposition, data are decomposed into the sum of products of spatial and temporal patterns from different components, with each component representing a distinct neurophysiological process. In this case, $\mathbf{A} \in \mathbb{R}^{N\times M}$ is the spatial pattern matrix, and $Z \in \mathbb{R}^{M\times T}$ is the temporal pattern matrix. Spatiotemporal decomposition consists in finding the unknown $\mathbf{A}$ and Z from the signals, typically through some statistical constraints, e.g., statistical independence, sparsity, or nonnegativity of the components.

In the case where $f(\cdot)$ is a nonlinear function of Z in (1), the Gaussianity is lost. In addition, the mapping f may not be invertible and the inverse mapping $\phi: X \to Z$ is also nonlinear and nonunique. In general, Z can be interpreted as the nonlinear feature extracted from X, which can be further used for brain-state classification. We revisit this topic in the section "Applications."

BAYESIAN INFERENCE: A BRIEF OVERVIEW

Estimation and inference can be carried out in multiple ways, depending how the unknown quantities $\mathbf{A}$, Z, and γ are handled. This leads to a natural partitioning of a variety of inverse methods. We briefly summarize three analytical approaches and sampling approaches. For simplicity, we assume $\mathbf{A}$ is known in the following overview.

EMPIRICAL BAYESIAN APPROACHES

If γ and Σ_z are known in advance, then the conditional distribution $p(Z|X,\gamma) \propto p(X|Z)p(Z|\gamma)$ is a fully specified Gaussian distribution with mean and covariance given by

$$\begin{aligned}\mathbb{E}_{p(z_t|X,\gamma)}[z_t] &= \gamma\mathbf{A}^{\mathrm{T}}(\Sigma_e + \mathbf{A}\Sigma_z\mathbf{A}^{\mathrm{T}})^{-1}x_t\\ \mathbb{C}_{p(z_t|X,\gamma)}[z_t] &= \Sigma_z - \Sigma_z\mathbf{A}^{\mathrm{T}}(\Sigma_e + \mathbf{A}\Sigma_z\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A}\Sigma_z,\end{aligned} \tag{10}$$

where z_t denotes the tth column of Z and is uncorrelated with time.

A common estimator for the latent sources is to use the posterior mean $\hat{Z} = \mathbb{E}_{p(Z|X,\gamma)}[Z]$. However, since γ is unknown, an approximate solution $\hat{\gamma}$ must first be found. One principled way to accomplish this is to integrate out the sources Z and then solve

$$\hat{\gamma} = \arg\max_{\gamma}\int p(X|Z)p(Z|\gamma)p(\gamma)\,\mathrm{d}Z. \tag{11}$$

This treatment is sometimes referred to as *empirical Bayes* (*EB*) because the prior $p(Z|\gamma)$ is empirically learned from the data, often using expectation-maximization (EM) algorithms. Additionally, marginalization provides a natural regularization that can shrink many elements of γ to zero, in effect pruning irrelevant covariance components from the model. Estimation under this model is sometimes called *automatic relevance determination* (*ARD*). The ARD procedure can also be leveraged to obtain a rigorous lower bound on model evidence $\log p(X)$. While knowing $p(Z|X)$ is useful for source estimation given a particular model, access to $p(X)$ [or, equivalently, $\log p(X)$] can assist model selection.

PENALIZED LIKELIHOOD APPROACHES

The second option is to integrate out the unknown γ, treat $p(Z)$ as the effective prior, and compute the MAP estimate Z_{MAP}:

$$Z_{\text{MAP}} = \arg\max_{Z} \int p(Z|X)p(Z|\gamma)p(\gamma)\,\mathrm{d}\gamma$$
$$= \arg\max_{Z} p(X|Z)p(Z). \quad (12)$$

Solving (12) also leads to a shrinking and pruning of superfluous covariance components. An EM algorithm can be derived for such a hierarchical model. Over the course of learning, this expectation collapses to zero for the irrelevant hyperparameters in γ.

VB APPROACHES

A third possibility involves finding approximate solutions to the posterior $p(Z|X)$ and the marginal $p(X.)$ The idea is that all unknown quantities should either be marginalized when possible or approximated with tractable distributions, and the distributions reflect underlying uncertainty and have computable posterior moments. Practically, we would like to account for the ambiguity regarding γ when estimating $p(Z|X)$, and, potentially, we would prefer a good approximation for $p(X)$ or a bound on the model evidence $\log p(X)$. Because of the intractable integration involved in obtaining either distribution, practical implementation relies on additional assumptions to yield different types of approximation strategies.

VB methods have been successfully applied to a wide variety of hierarchical Bayesian (HB) models in the machine-learning literature, which offer an alternative to EB and penalized likelihood methods [8]. Two of the most popular forms of variational approximations are the mean field approximation (VB-MF) and Laplace approximation (VB-LA). The mean field approximation assumes that the joint distribution over unknowns Z and γ is factorial: $p(Z,\gamma|X) \approx q(Z|X)q(\gamma|X)$, where $q(Z|X)$ and $q(\gamma|X)$ are chosen to minimize the Kullback–Leibler (KL) divergence between the factorized and full posterior. This is accomplished via an iterative process akin to EM, effectively using two E-steps (one for Z and one for γ). It also produces a rigorous lower bound on $\log p(X)$ similar to EB approaches. The second-order Laplace approximation assumes that the posterior on the hyperparameters (after marginalizing over Z is Gaussian, which is then iteratively matched to the true posterior; the result can then be used to approximate $p(Z|X)$ and $\log p(X)$.

Although seemingly different, there are some deep dualities between these three types of analytical approaches [9], [10]. Specifically, it can be shown that the EB approaches are equivalent to penalized likelihood methods when the prior is chosen to be nonfactorial, lead-field, and noise dependent; VB methods are equivalent to EB approaches given an appropriate hyperprior [9], [10]. Theoretical properties concerning the convergence, global and local minima, and localization bias of each of these methods have been analyzed and fast algorithms have been derived that improve upon existing methods [9], [10]. 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. Table 2 lists a high-level comparison of several Bayesian inference paradigms.

[TABLE 2] COMPARISON OF BAYESIAN INFERENCE APPROACHES.

BAYESIAN INFERENCE APPROACH	LATENT VARIABLE ESTIMATE	HIERARCHICAL PRIOR	SPARSITY, MODEL SELECTION	SPEED, CONVERGENCE
EB	FULL POSTERIOR	NO	YES	FAST
PENALIZED LIKELIHOOD	POINT ESTIMATE	YES	YES	FAST
VB	VARIATIONAL POSTERIOR	YES	YES	FAST
MCMC	FULL POSTERIOR	YES	YES	SLOW

SAMPLING APPROACHES

In contrast to VB approaches (deterministic version of approximate Bayesian inference), sampling approaches can be viewed as a stochastic version of approximate Bayesian inference. The basic idea is simple: instead of computing intractable integration in the posterior computation, draw random samples from the distribution and use plug-in samples for calculation. Nevertheless, drawing Monte Carlo samples from an exact target distribution P is still nontrivial.

When direct sampling from a target distribution is difficult, one can instead draw random samples from a proposal distribution Q. For instance, the importance sampling method first draws samples $\xi_i \sim Q$, and then weighs each sample with a nonnegative importance weight function $w_i(\xi_i) = P(\xi_i)/Q(\xi_i)$ [11]. Alternatively, the Markov chain Monte Carlo (MCMC) method draws samples sequentially from a Markov chain; when the Markov chain converges to the equilibrium point, the samples are viewed approximately from the target distribution. The most common MCMC sampling methods include the Metropolis–Hastings (MH) algorithm [12] and Gibbs sampling [13]. The MH algorithm is the simplest yet the most generic MCMC method to generate samples using a random walk and then to accept them with a certain acceptance probability. For example, given a random-walk proposal distribution $g(z \to z')$ (which defines a conditional probability of moving state z to z'), the MH acceptance probability $\mathcal{A}(z \to z')$ is computed as

$$\mathcal{A}(z \to z') = \min\left(1, \frac{p(z')g(z' \to z)}{p(z)g(z \to z')}\right),$$

which gives a simple MCMC implementation and satisfies the "detailed balance" condition. Gibbs sampling is another popular MCMC method that requires no parameter tuning. Given a high-dimensional joint distribution $p(z) = p(z_1, \ldots, z_M)$, the Gibbs sampler draws samples from the individual conditional distribution $p(z_i|z_{-i})$ in turn while holding others fixed (where z_{-i} denotes the $M-1$ variables in z except for z_i).

Since the random-walk model is highly inefficient in a high-dimensional space, other sampling methods attempt to exploit side information of the likelihood, such as the Hamiltonian sampling method and gradient-based Langevin MCMC [14]. The topics of interest in this article include latent variable modeling, efficient inference, model selection, sparsity (compressed sensing), group analysis, nonstationarity, and NB modeling. We will illustrate ideas using some examples and also provide

reference pointers to state-of-the-art research. Many topics are inherently interrelated (see Figure 1):

- Latent variable modeling and hierarchy are the central themes of probabilistic or Bayesian modeling.
- Hierarchy is a natural way to characterize random effects, nonstationarity, and group variability.
- Depending on the hyperprior/prior assumption or optimization criterion (e.g., posterior, marginal likelihood, and variational free energy), one can derive various exact or approximate Bayesian inference algorithms.
- Sparsity is a prior assumption imposed on the model, which can also be used for model selection.

VB METHODS

This section presents several efficient VB methods developed for EEG/MEG signal processing. Let Θ denote collectively all the unknown parameters to be inferred in (1). We focus on the VB-MF approximation, which seeks a factorable distribution q to approximate the true posterior p by minimizing the KL divergence

$$\mathrm{KL}(q\|p) = -\int q(\Theta)\log\left[\frac{p(\Theta\,|\,X)}{q(\Theta)}\right]\mathrm{d}\Theta, \tag{13}$$

where $q(\Theta) = \prod_{l=1}^{L} q(\Theta_l)$, and $\Theta_l (l = 1, \cdots, L)$ denote the disjoint groups of variables in Θ. Direct minimization of (13) is difficult, but it is easy to show that the log marginal likelihood of the data can be written as

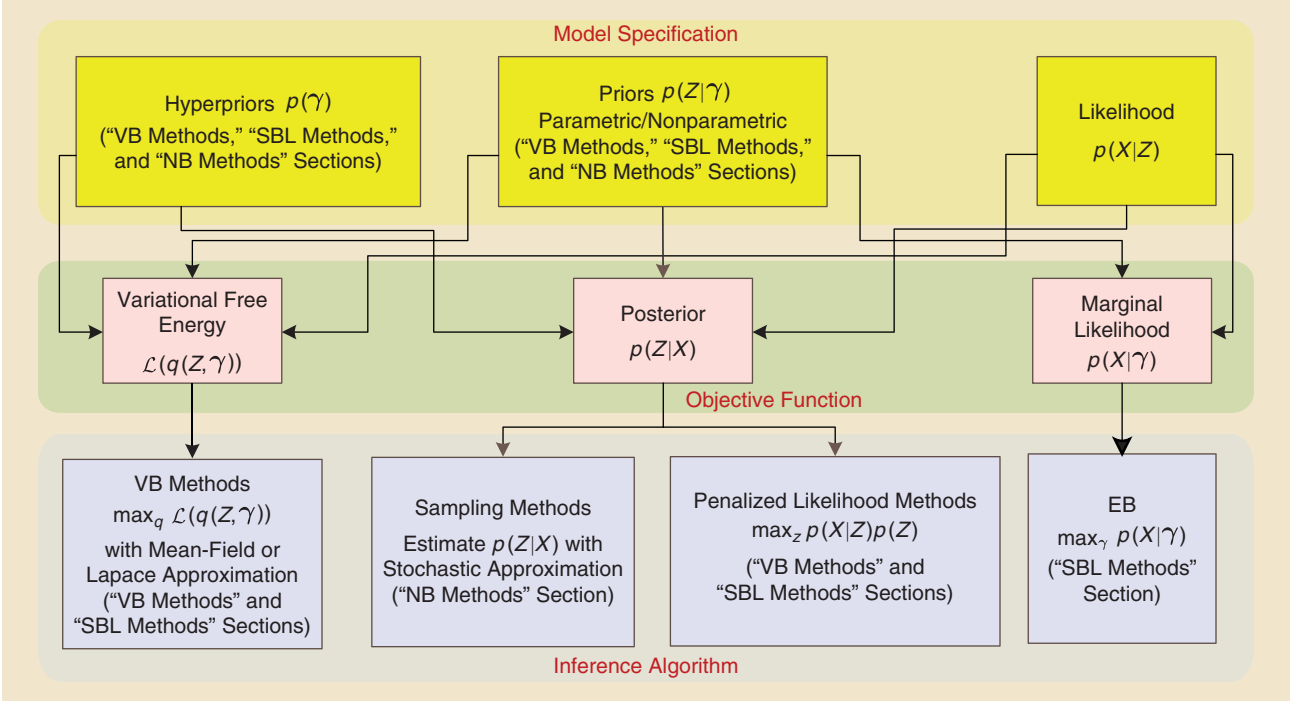
$$\log p(X) = \mathcal{L}(q) + \mathrm{KL}(q\|p), \tag{14}$$

with

$$\mathcal{L}(q) = \int q(\Theta)\log\left[\frac{p(X,\Theta)}{q(\Theta)}\right]\mathrm{d}\Theta. \tag{15}$$

Since $\mathrm{KL}(q\|p) \geq 0$, $\mathcal{L}(q)$ is a lower bound of the log marginal likelihood, which is termed the *variational free energy*. In light of (14), minimizing the KL divergence is equivalent to maximizing $\mathcal{L}(q)$, which is more computationally tractable using numerical optimization algorithms. The most widely used algorithm is coordinate ascent, which alternately updates the approximate distribution of each parameter $\{q(\Theta_l)\}$ until convergence. In brief, this coordinate ascent can be thought of as the solution for one approximate posterior marginal distribution that is expressed in terms of the others. By stepping through the different subsets of unknown parameters, we can iteratively update the approximate marginals.

In the presence of latent variable Z, the variational posterior is often assumed to have a factorial form: $q(Z,\Theta) = q(Z)q(\Theta)$. Similarly, maximizing the variational free energy with respect to two functions $q(Z)$ and $q(\Theta)$ alternatively gives rise to the so-called VB-EM (expectation-maximization) algorithm [15]. In the VB-E (expectation) step, set $[\partial\mathcal{L}/\partial q(Z)] = 0$ and update the variational posterior $q(Z)$; in the VB-M (maximization) step, set $[\partial\mathcal{L}/\partial q(\Theta)] = 0$ and update



[FIG1] **A schematic illustration of Bayesian inference and the BML techniques introduced in this tutorial. Bayesian inference is carried out via two phases: model specification and inference. In the model-specification phase, the likelihood $L(Z) = p(X\,|\,Z)$, which describes how the EEG/MEG data X are related to the unknown variables Z (including parameters and latent variables) is specified. Parametric or nonparametric priors can be imposed on Z. The hierarchy is built up by imposing hyperpriors $p(\gamma)$ on the hyperparameters γ in the priors. In the inference phase, a particular algorithm is chosen to infer Z, based on deterministic or stochastic approximations. Under specific conditions, empirical Bayesian methods, penalized likelihood methods, and VB methods are equivalent to each other. See the section "Bayesian Inference: A Brief Overview" for details.**

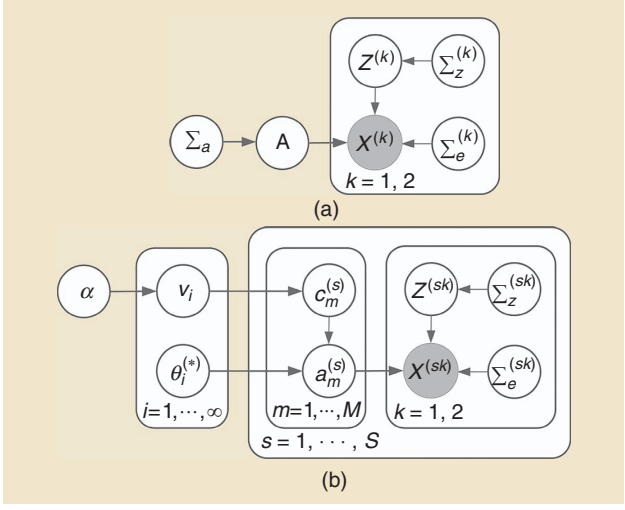
the variational posterior $q(\Theta)$; repeat these two steps until convergence.

VB FOR LEARNING COMMON EEG COMPONENTS

Learning common EEG components that distinguish different conditions is an effective feature-extraction strategy for brain-state classification. Among the various methods, the common spatial patterns (CSP) algorithm [16], which finds common EEG components with the largest variance ratios between two conditions, has attracted considerable attention as being successful in extracting sensorimotor rhythms for BCIs. Nonetheless, as a multivariate algorithm, CSP is known to suffer from overfitting, which may yield poor generalization performance [16]. Although various regularization strategies may be utilized to ameliorate CSP's overfitting, the algorithm was designed primarily for classifying instead of modeling the EEG data and, therefore, not specifically designed for exploring the underlying spatiotemporal dynamics.

A VB-EM algorithm called the *VB-CSP algorithm* was developed in [17] to address the overfitting issue of CSP. The algorithm is based on the following model that is a multicondition extension of (2):

$$\begin{aligned} X^{(k)} &= \mathbf{A}Z^{(k)} + E^{(k)} \\ \mathbf{A} &\sim \prod_{n=1}^{N} \mathcal{N}(\mathbf{a}_n \mid 0, \Sigma_a),\ Z^{(k)} \sim \prod_{t=1}^{T} \mathcal{N}(z_t^{(k)} \mid 0, \Sigma_z^{(k)}),\ E^{(k)} \\ &\sim \prod_{t=1}^{T} \mathcal{N}(e_t^{(k)} \mid 0, \Sigma_e^{(k)}) \\ \sigma_m &\sim \mathcal{IG}(\sigma_m \mid \alpha, \beta),\ \rho_m^{(k)} \sim \mathcal{IG}(\rho_m^{(k)} \mid \alpha, \beta),\ \eta_n^{(k)} \sim \mathcal{IG}(\eta_n^{(k)} \mid \alpha, \beta), \end{aligned} \tag{16}$$



[FIG2] Graphical model representations for two extended CSP models: (a) the VB-CSP model [17] and (b) the Bayesian CSP-DP model [18]. VB-CSP is parametric, while Bayesian CSP-DP is nonparametric. In VB-CSP, $X^{(k)}$ denotes the EEG data for the kth condition. A and $Z^{(k)}$ are the unknown mixing matrix and latent variables, respectively. Σ_a, $\Sigma_z^{(k)}$, and $\Sigma_e^{(k)}$ are the covariance matrices for A, $Z^{(k)}$, and $E^{(k)}$, respectively. All unknown latent variables and parameters are shown as unshaded nodes and estimated from the data using the update rules provided in Algorithm 1. In the Bayesian CSP-DP model, s represents the subject index. It is assumed that $\alpha \sim \mathcal{G}(a_0, b_0)$, $v_i \sim \mathcal{B}(1, \alpha)$, $p(c_m^{(s)}(i) = 1) = v_i \prod_{j=1}^{i-1}(1 - v_j)$, $a_m^{(s)} \mid z_m^{(s)}(i) = 1 \sim \mathcal{N}(\mu_i^*, (\Phi_i^*)^{-1})$ (see the section "DP Modeling" for details).

where $\Sigma_a \triangleq \mathrm{diag}[\sigma_{1:M}]$, $\Sigma_z^{(k)} \triangleq \mathrm{diag}[\rho_{1:M}^{(k)}]$, and $\Sigma_e^{(k)} \triangleq \mathrm{diag}[\eta_{1:N}^{(k)}]$. $\mathcal{IG}$ denotes the inverse-gamma distribution: $\mathcal{IG}(x \mid \alpha, \beta) \triangleq [\beta^\alpha / \Gamma(\alpha)] x^{-\alpha-1} \exp(-\beta/x)$. The corresponding graphical model is shown in Figure 2(a).

Equation (16) provides a principled generative framework for interpreting and improving CSP, since it can be shown that the ML estimation of A in the noiseless and square mixing setup (i.e., $E^{(k)} = \mathbf{0}$ and $M = N$) leads to the same cost function as CSP [19]. The overfitting issue of CSP stems from the square mixing and noiseless assumptions. The noiseless assumption implies that the EEG data are fully characterized by the estimated components and the mixing matrix. This assumption does not take into account noise or interference. The square mixing assumption is closely linked to the noiseless assumption in that, if we relax the square mixing assumption by using a smaller number of components, a model mismatch will automatically arise between the best linear fit and the EEG data.

We further assume in (16) that $\alpha \to 0, \beta \to 0$ to impose noninformative priors on σ_m, $\rho_m^{(k)}$, and $\eta_n^{(k)}$ [11]. This also allows us to leverage Bayes' rule to achieve automatic model size determination in conjunction with parameter estimation. However, exact Bayesian inference is not viable for (16) due to the product coupling of A and $Z^{(k)}$ in the likelihood, as well as the inconvenient form of the hierarchical priors.

Through the VB-MF approximation to $p(\mathbf{A}, Z^{(k)} \mid X^{(k)})$, the variational free energy $\mathcal{L}(q(\mathbf{A}, Z^{(k)}))$ can be derived as in (15), which can be further lower bounded by invoking Fenchel's duality to locally approximate the hierarchical priors on A, $Z^{(k)}$, and $E^{(k)}$, yielding

$$\begin{aligned} \tilde{\mathcal{L}}(q(\mathbf{A}, Z^{(k)})) \triangleq \min_{\substack{\Sigma_a, \Sigma_z^{(k)} \\ \Sigma_e^{(k)}}} & \\ & -(\frac{L}{2} + \alpha) \sum_k \log |\Sigma_e^{(k)}| - (\frac{L}{2} + \alpha) \sum_k \log |\Sigma_z^{(k)}| \\ & -(\frac{N}{2} + \alpha) \log |\Sigma_a| \\ & -\frac{1}{2} \sum_k \mathbb{E}_q[\mathrm{trace}[\Sigma_e^{(k)-1}[(X^{(k)} - \mathbf{A}Z^{(k)})(X^{(k)} - \mathbf{A}Z^{(k)})^\top + 2\beta\mathbf{I}]]] \\ & -\frac{1}{2} \sum_k \mathbb{E}_q[\mathrm{trace}[\Sigma_z^{(k)-1}(Z^{(k)}Z^{(k)\mathrm{T}} + 2\beta\mathbf{I})]] \\ & -\frac{1}{2} \mathbb{E}_q[\mathrm{trace}[\Sigma_e^{(k)-1}(\mathbf{A}^\mathrm{T}\mathbf{A} + 2\beta\mathbf{I})]] \\ & -\mathbb{E}_q[\log q(\mathbf{A})] - \sum_k \mathbb{E}_q[\log q(Z^{(k)})]. \end{aligned}$$

VB-CSP uses the coordinate descent to solve the following optimization problem:

$$\min_{q(\mathbf{A}),\, q(Z^{(k)})} \tilde{\mathcal{L}}(q(\mathbf{A}), q(Z^{(k)})). \tag{17}$$

Detailed derivations of VB-CSP are referred to in [17]. The pseudocode is provided in Algorithm 1. Because of the multiplicative structure of (16), the unknown variables can only be identified up to the permutation ambiguity unless additional constraints are imposed. Hence, the posterior distribution is intrinsically multimodal. This naturally casts doubts on the validity of VB for approximate inference, given that the resulting approximate posterior is unimodal. According to (13), there is a large positive contribution to the KL divergence from regions of $\{\mathbf{A}, Z^{(k)}\}$ space in which

Algorithm 1: The VB-CSP algorithm.

Input: $X^{(k)}$
Output: $\Sigma_a, \Sigma_z^{(k)}, \Sigma_e^{(k)}, q(Z^{(k)}), q(\mathbf{A})$
1: **Initialization**
2: **repeat**
3: $q(Z_k) = \prod_{t=1}^{T} \mathcal{N}(z_t^{(k)} \mid \mathbb{E}[z_t^{(k)}], \mathbb{C}[z_t^{(k)}])$,
where $\mathbb{E}[z_t^{(k)}] \triangleq \mathbb{C}[z_t^{(k)}]\mathbb{E}^{\mathrm{T}}[\mathbf{A}][\Sigma_e^{(k)}]^{-1}x_t^{(k)}$, $\mathbb{C}[z_t^{(k)}] \triangleq [\mathbb{E}^{\mathrm{T}}[\mathbf{A}][\Sigma_e^{(k)}]^{-1}\mathbb{E}[\mathbf{A}] + \sum_n \mathbb{C}[\mathbf{a}_n^{(k)}]/\eta_n^{(k)} + [\Sigma_z^{(k)}]^{-1}]^{-1}$
4: $q(\mathbf{A}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{a}_n \mid \mathbb{E}[\mathbf{a}_n], \mathbb{C}[\mathbf{a}_n])$,
where $\mathbb{E}[\mathbf{a}_n] \triangleq \sum_{t,k} x_{t,n}^{(k)} \mathbb{E}^{\top}[z_t^{(k)}]\mathbb{C}[\mathbf{a}_n]/\eta_n^{(k)}$, $\mathbb{C}[\mathbf{a}_n] \triangleq [\Sigma_a^{-1} + \sum_{t,k}(\mathbb{C}[z_t^{(k)}] + \mathbb{E}[z_t^{(k)}]\mathbb{E}^{\top}[z_t^{(k)}])/\eta_n^{(k)}]^{-1}$
5: $\Sigma_z^{(k)} = \frac{1}{T+2\alpha}\sum_t (\mathrm{diag}[\mathbb{C}[z_t^{(k)}] + \mathbb{E}[z_t^{(k)}]\mathbb{E}^{\top}[z_t^{(k)}]] + 2\beta\mathbf{I})$
6: $\Sigma_e^{(k)} = \frac{1}{T+2\alpha}\sum_t (\mathrm{diag}[X_t^{(k)}[X_t^{(k)}]^{\top} - 2X_t^{(k)}[Z_t^{(k)}]^{\top}\mathbb{E}^{\top}[\mathbf{A}] + \mathbb{E}[\mathbf{A}[\mathbb{C}[z_t^{(k)}] + \mathbb{E}[z_t^{(k)}]\mathbb{E}^{\top}[z_t^{(k)}]]\mathbf{A}^{\top})] + 2\beta\mathbf{I}]$
7: $\Sigma_a = \frac{1}{N+2\alpha}\sum_n (\mathrm{diag}[\mathbb{C}[\mathbf{a}_n] + \mathbb{E}^{\top}[\mathbf{a}_n]\mathbb{E}[\mathbf{a}_n]] + 2\beta\mathbf{I})$
8: **until** Convergence

$p(\mathbf{A}, Z^{(k)} \mid X^{(k)})$ is near zero unless $q(\mathbf{A}, Z^{(k)})$ is also close to zero. Thus, minimizing the KL divergence in (13) leads to distributions $q(\mathbf{A}, Z^{(k)})$ that avoid regions in which $p(\mathbf{A}, Z^{(k)} \mid X^{(k)})$ is small. This means that VB tends to find one of the modes in the posterior distribution. Hence, the unimodality of the VB posterior does not nullify the effectiveness in the inferring (16). Similar arguments apply to the other VB-based spatiotemporal decomposition algorithms discussed in this article.

Our empirical observations indicated that the algorithmic performance is only slightly affected by initialization (see sensitivity analysis in [17]). VB-CSP is particularly suited for the single-trial analysis of motor imagery EEG data, since imagined movements give rise to an attenuation of the sensorimotor rhythms in specific regions of the sensorimotor cortices, a phenomenon known as *event-related desynchronization,* which can be captured by evaluating the variance change of EEG spatial patterns across conditions.

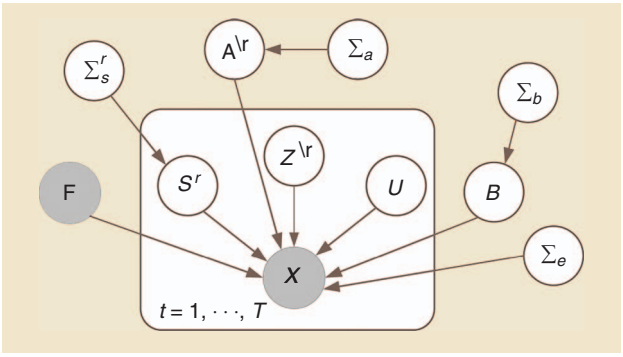
VB FOR LOCALIZING EVOKED SOURCE ACTIVITY WITH INTERFERENCE SUPPRESSION

EEG/MEG sensor measurements are often contaminated by many types of interference, such as background activity from outside the regions of interest, biological and nonbiological artifacts, and sensor noise. Source activity using knowledge of event timing for independence from noise and interference (SAKETINI) is a probabilistic graphical modeling algorithm that localizes and estimates neural source activity measured by EEG/MEG data while suppressing the interference or noise [20]. It is assumed that the recorded EEG/MEG data are separated by the zero time point (stimulus onset or experimental marker). Ongoing brain activity, biological noise, background environmental noise, and sensor noise are present in both the prestimulus (before zero) and poststimulus (after zero) periods; however, the evoked neural sources of interest are only present in the poststimulus time period. It is also assumed that the sensor data can be described as coming from four types of sources: 1) an evoked source at a particular voxel, 2) all other evoked sources not at that voxel, 3) all background noise sources with spatial covariance at the sensors (including the brain, biological, or environmental sources), and 4) sensor noise. The first step is to infer the model describing source types 3 and 4 from the prestimulus data, and the second step is to infer the full model describing the remaining source types 1 and 2 from the poststimulus data. After inference of the model, a map of the source activity is created, as well as a map of the likelihood of activity across voxels.

Let the time-ranges for prestimulus and poststimulus data be T_{pre} and T_{post}, respectively. The generative model for data $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T] \in \mathbb{R}^{N\times T}$ from N sensors is assumed as follows:

$$\boldsymbol{x}_t = \begin{cases} \mathbf{B}\boldsymbol{u}_t + e_t & t \in T_{\mathrm{pre}} \\ \mathbf{F}^r\boldsymbol{s}_t^r + \mathbf{A}^{\setminus r}\boldsymbol{z}_t^{\setminus r} + \mathbf{B}\boldsymbol{u}_t + e_t & t \in T_{\mathrm{post}}. \end{cases} \tag{18}$$

The lead-field matrix $\mathbf{F}^r \in \mathbb{R}^{N\times q}$ represents the physical (and linear) relationship between a dipole source at voxel r for each dipole orientation q along a coordinate basis and its influence on sensors [21]. The source activity $S^r = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T] \in \mathbb{R}^{q\times T}$ matrix of dipole time course for the voxel r. The $\mathbf{A}^{\setminus r} \in \mathbb{R}^{N\times L}$ and $Z^{\setminus r} = [\boldsymbol{z}_1^{\setminus r}, \ldots, \boldsymbol{z}_T^{\setminus r}] \in \mathbb{R}^{L\times T}$ represent the poststimulus mixing matrix and evoked nonlocalized factors, respectively, corresponding to source type 2 discussed previously, where the superscript $\setminus r$ indicates for all voxels not at voxel r. $\mathbf{B} \in \mathbb{R}^{N\times M}$ and $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T] \in \mathbb{R}^{M\times T}$ represent the background mixing matrix and background factors, respectively, corresponding to source type 3. Representing the sensor-level noise is e_t, which is assumed to be drawn from a Gaussian distribution with zero-mean and precision (inverse covariance) defined by the diagonal matrix Σ_e. All quantities depend on r in the poststimulus period, except for $\mathbf{B}$, U, and Σ_e, which will be learned from the prestimulus data. Note, however, that the posterior update for U (represented by $\overline{U}$) does depend on the voxel r. The graphical model is shown in Figure 3.



[FIG3] A graphical model representation for SAKETINI. The observed data $\boldsymbol{X}$ are assumed to be generated from a linear mixing of several latent sources. Sources $\boldsymbol{S^r}$ correspond to brain activity that contributes to sensor data through a known mixing matrix F. Brain activity not located at the scanning location r is characterized by a set of factors $Z^{\backslash r}$ with an unknown mixing matrix $\mathbf{A}^{\backslash r}$. Additionally, background brain activity is assumed to arise from a set of factors U, associated with an unknown mixing matrix B. The mixing matrices A and B are assumed to be drawn from a Gaussian prior with covariance Σ_a and Σ_b, respectively. Finally, sensor data is assumed to have sensor noise that is also drawn from a Gaussian prior with covariance Σ_e. All unknown latent variables and parameters are shown as unshaded nodes and estimated from the data using update rules provided in Algorithm 2.

LEARNING BACKGROUND BRAIN ACTIVITY FROM PRESTIMULUS DATA

We use a VB factor analysis (VBFA) approach to describe the part of the sensor signals contributed by background factors, as arising from L underlying factors, via an unknown mixing matrix. In mathematical terms, let $\boldsymbol{u}_t$ denote the M-dimensional vector corresponding to the background factors. We assume Gaussian prior distributions on the background factors and sensor noise and a flat prior on the sensor noise precision. We further assume the background factors are independent and identically distributed (i.i.d.) across time. Therefore, the following is assumed for $t \in T_{\text{pre}}$:

$$p(U) = \prod_t p(\boldsymbol{u}_t); \quad p(\boldsymbol{u}_t) = \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{19}$$

$$p(E) = \prod_t p(\boldsymbol{e}_t); \quad p(\boldsymbol{e}_t) = \mathcal{N}(\mathbf{0}, \Sigma_e), \quad p(\Sigma_e) = \text{const.} \tag{20}$$

$$p(X \mid U, \mathrm{B}, \Sigma_e) = \prod_t p(\boldsymbol{x}_t \mid \boldsymbol{u}_t, \mathrm{B}, \Sigma_e) = \prod_t \mathcal{N}(\mathrm{B}\boldsymbol{u}_t, \Sigma_e). \tag{21}$$

We use a conjugate prior for the background mixing matrix B, as follows:

$$p(\mathrm{B}) = \prod_{n=1}^{N} \prod_{m=1}^{M} p(b_{nm}) = \prod_{n=1}^{N} \prod_{m=1}^{M} \mathcal{N}(0, \lambda_n \beta_m), \tag{22}$$

where β_m is a hyperparameter over the mth column of B, and λ_n is the precision of the nth sensor. The matrix $\Sigma_b = \text{diag}(\beta_1, ..., \beta_M)$ provides a robust mechanism for automatic model order selection (see the "Introduction" section), so that the optimal size of B is inferred from the data through Σ_b.

Exact inference on this model is also intractable, and we choose to factorize the marginal conditional posterior distribution, assuming conditional independence of the background factors U and mixing matrix B

$$p(U, \mathrm{B} \mid X) \approx q(U, \mathrm{B}) = q(U)q(\mathrm{B}). \tag{23}$$

The VB-EM algorithm is used to iteratively maximize the free-energy with respect to $q(U)$, $q(\mathrm{B})$, Σ_b, and Σ_e.

LOCALIZATION OF EVOKED SOURCES LEARNED FROM POSTSTIMULUS DATA

In the stimulus-evoked paradigm, the source strength at each voxel is learned from the poststimulus data. The background mixing matrix B and sensor noise precision Σ_e are assumed to be fixed after prestimulus estimation. We assume those quantities remain constant through the poststimulus period and are independent of source location.

The source factors have Gaussian distribution with zero mean and covariance $\Sigma_s^r \in \mathbb{R}^{q \times q}$, which relates to the strength of the dipole in each of q directions

$$p(S^r) = \prod_t p(\boldsymbol{s}_t^r); \quad p(\boldsymbol{s}_t^r) = \mathcal{N}(\mathbf{0}, \Sigma_s^r). \tag{24}$$

The interference and background factors are assumed to be Gaussian with zero mean and identity covariance matrix (i.e., non-informative prior).

$$p(Z^{\backslash r}) = \prod_t p(\boldsymbol{z}_t^{\backslash r}); \quad p(\boldsymbol{z}_t^{\backslash r}) = \mathcal{N}(0, \mathbf{I}),$$
$$p(U) = \prod_t p(\boldsymbol{u}_t); \quad p(\boldsymbol{u}_t) = \mathcal{N}(0, \mathbf{I}). \tag{25}$$

We also use a conjugate prior for the interference mixing matrix $\mathbf{A}^{\backslash r}$, where $\Sigma_a = \text{diag}(\alpha_1, ..., \alpha_L)$ is a hyperparameter that helps determine the size of $\mathbf{A}^{\backslash r}$ (in a similar fashion for matrix B)

$$p(\mathbf{A}^{\backslash r}) = \prod_{n=1}^{N} \prod_{l=1}^{L} p(a_{nl}) = \prod_{n=1}^{N} \prod_{l=1}^{L} \mathcal{N}(0, \lambda_n \alpha_l). \tag{26}$$

We now specify the full model

$$p(X \mid S^r, Z^{\backslash r}, U, \mathbf{A}^{\backslash r}, \mathrm{B}, \Sigma_e) = \prod_t p(\boldsymbol{x}_t \mid \boldsymbol{s}_t^r, \boldsymbol{z}_t^{\backslash r}, \boldsymbol{u}_t, \mathbf{A}^{\backslash r}, \mathrm{B}, \Sigma_e)$$
$$= \prod_t \mathcal{N}(\mathrm{F}^r \boldsymbol{s}_t^r + \mathbf{A}^{\backslash r} \boldsymbol{z}_t^{\backslash r} + \mathrm{B}\boldsymbol{u}_t, \Sigma). \tag{27}$$

Exact inference on this model is also intractable, and we similarly use VB approximation for evoked nonlocalized factors $Z^{\backslash r}$ and mixing matrix $\mathbf{A}^{\backslash r}$

$$p(S^r, Z^{\backslash r}, U, \mathbf{A}^{\backslash r} \mid X) \approx q(S^r, Z^{\backslash r}, U, \mathbf{A}^{\backslash r} \mid X)$$
$$= q(S^r, Z^{\backslash r}, U \mid X) q(\mathbf{A}^{\backslash r} \mid X). \tag{28}$$

All variables, parameters and hyperparameters are unknown and learned from the data using the VB-EM algorithm (Algorithm 2).

Extensions of this approach, whose performance benefits are yet to be worked out or determined, are to assume non-Gaussian prior distributions for the evoked or background factors using either mixture-of-Gaussian or Gaussian-scale mixture distributions [22], [15].

VB FOR GROUP EEG/MEG ANALYSIS

EEG/MEG recordings often arise from a multilevel structure (e.g., multiple subjects/sessions/trials), and group analysis at each level entails building statistical models that characterize the homogeneity and variability within the group. One way to impose the group structure is to use HB models [11], with the variations at

Algorithm 2: The SAKETINI algorithm.

Input: X for T_{pre} and T_{post}

Output: $q(U), q(S^r), q(Z^{\setminus r}), q(\mathrm{A}^{\setminus r}), q(\mathrm{B}), \Sigma_e, \Sigma_b, \Sigma_a$

1: **Initialization**

2: **repeat**

3: $q(\mathrm{B}\,|\,X) = \mathcal{N}(\overline{\mathrm{B}}, \psi)$, where $\overline{\mathrm{B}} = \mathrm{R}_{\mathrm{XU}}\psi, \psi = (\mathrm{R}_{\mathrm{UU}} + \Sigma_b)^{-1}, \Sigma_b^{-1} = \mathrm{diag}(\frac{1}{N}\overline{\mathrm{B}}^{\top}\Sigma_e\overline{\mathrm{B}} + \psi), \Sigma_e^{-1} = \frac{1}{T}\mathrm{diag}(\mathrm{R}_{\mathrm{XX}} - \overline{\mathrm{B}}\mathrm{R}_{\mathrm{XU}}^{\top}), \mathrm{R}_{\mathrm{XX}} = \sum_t \boldsymbol{x}_t\,\overline{\boldsymbol{x}}_t^{\top}$ and $\mathrm{R}_{\mathrm{XU}} = \sum_t \boldsymbol{x}_t\,\overline{\boldsymbol{u}}_t^{\top}$.

4: $q(U\,|\,X) = \prod_t q(\boldsymbol{u}_t\,|\,\boldsymbol{x}_t) = \prod_t \mathcal{N}(\overline{\boldsymbol{u}}_t, \gamma)$, where $\overline{\boldsymbol{u}}_t = \gamma^{-1}\overline{\mathrm{B}}^{\top}\Sigma_e\,\boldsymbol{x}_t, \gamma = \overline{\mathrm{B}}^{\top}\Sigma_e\,\overline{\mathrm{B}} + N\psi^{-1} + \mathrm{I}$

5: Let $V = (S^r\ Z^{\setminus r}\ U)$, then $q(V\,|\,X) = \prod_t q(\boldsymbol{v}_t\,|\,\boldsymbol{x}_t) = \prod_t \mathcal{N}(\overline{\boldsymbol{v}}_t, \Gamma)$, where $\overline{\boldsymbol{v}}_t = \Gamma^{-1}\overline{\mathrm{A}}_{\mathrm{aug}}^{\top}\,\Sigma_e\,\boldsymbol{x}_t, \Gamma = \overline{\mathrm{A}}_{\mathrm{aug}}^{\top}\Sigma_e\,\overline{\mathrm{A}}_{\mathrm{aug}} + N\Psi + \mathrm{I}_{\mathrm{aug}}$, where $\overline{\mathrm{A}}_{\mathrm{aug}} = (\mathrm{F}^r\overline{\mathrm{A}}^{\setminus r}\overline{\mathrm{B}}), \mathrm{I}_{\mathrm{aug}} = \begin{pmatrix}\Phi & 0 & 0\\ 0 & \mathrm{I} & 0\\ 0 & 0 & \mathrm{I}\end{pmatrix}, \Psi = \begin{pmatrix}0 & 0 & 0\\ 0 & \Psi_{\mathrm{AA}} & 0\\ 0 & 0 & 0\end{pmatrix}$. From marginalization, we also obtain $q(S^r\,|\,X)$ and $q(Z^{\setminus r}\,|\,X)$.

6: $q(\mathrm{A}^{\setminus r}\,|\,X) = \mathcal{N}(\overline{\mathrm{A}}, \Psi_{\mathrm{AA}})$, where $\overline{\mathrm{A}} = (\mathrm{R}_{\mathrm{XZ}} - \mathrm{F}^r\mathrm{R}_{\mathrm{SZ}} - \overline{\mathrm{B}}\mathrm{R}_{\mathrm{UZ}})\,\Psi_{\mathrm{AA}}, \Psi_{\mathrm{AA}} = (\mathrm{R}_{\mathrm{ZZ}} + \Sigma_a)^{-1}, \Sigma_a^{-1} = \mathrm{diag}(\frac{1}{N}\overline{\mathrm{A}}^{\top}\Sigma_e\overline{\mathrm{A}} + \Psi_{\mathrm{AA}})$, where the covariance matrices can be defined, for instance: $\mathrm{R}_{\mathrm{SS}} = \sum_t \overline{\boldsymbol{s}}_t\,\overline{\boldsymbol{s}}_t^{\top} + N\,\Sigma_{\mathrm{SS}}$, and similarly for $\mathrm{R}_{\mathrm{SZ}}, \mathrm{R}_{\mathrm{SU}}, \mathrm{R}_{\mathrm{UU}}, \mathrm{R}_{\mathrm{ZU}}, \mathrm{R}_{\mathrm{UU}}$, where $\Sigma = \Gamma^{-1}$ is specified as: $\Sigma = \begin{pmatrix}\Sigma_{\mathrm{SS}} & \Sigma_{\mathrm{SZ}} & \Sigma_{\mathrm{SU}}\\ \Sigma_{\mathrm{SZ}}^{\top} & \Sigma_{\mathrm{ZZ}} & \Sigma_{\mathrm{ZU}}\\ \Sigma_{\mathrm{SU}}^{\top} & \Sigma_{\mathrm{ZU}}^{\top} & \Sigma_{\mathrm{UU}}\end{pmatrix}$

7: **until** Convergence

each low level described by a submodel and variations in the hyperparameters of the submodels described by an upper-level model. For instance, let $X^{(s)}(s = 1, \cdots, S)$ denote a group of S-subject EEG/MEG data sets, with s being the subject index; we may have

$$X^{(s)} \sim p(X^{(s)}\,|\,\Theta^{(s)}), \quad \Theta^{(s)} \sim p(\Theta^{(s)}\,|\,\Omega). \tag{29}$$

The upper-level prior acts as the group constraint on the parameters, $\Theta^{(s)}$, of each subject's model. The strength of HB modeling is that the hyperparameters in an upper-level model can be effectively estimated by pooling information across the data in the lower levels. As a benefit, HB models have proven to be robust, with the posterior distribution being less sensitive to the hierarchical priors. A two-level HB model was proposed to characterize the intertrial amplitude variability in the EEG signals during motor imagery tasks [19]. In this article, we use an example taken from [23] to illustrate an alternative group modeling approach to achieve multisubject electromagnetic brain imaging.

In the multisubject setup, the generative model (2) can be compactly written in the following form:

$$\begin{bmatrix}X^{(1)}\\ \vdots\\ X^{(S)}\end{bmatrix} = \begin{bmatrix}\mathrm{A}^{(1)} & & \\ & \ddots & \\ & & \mathrm{A}^{(S)}\end{bmatrix}\begin{bmatrix}Z^{(1)}\\ \vdots\\ Z^{(S)}\end{bmatrix} + \begin{bmatrix}E^{(1)}\\ \vdots\\ E^{(S)}\end{bmatrix}, \tag{30}$$

or in a compact form $\overline{X} = \overline{\mathrm{A}}\,\overline{Z} + \overline{E}$, where the overline symbol denotes the concatenated matrix form. Here, for simplicity of description, we assume that time samples are i.i.d., i.e., $E^{(s)} \sim \prod_{t=1}^{T}\mathcal{N}(e_t^{(s)}\,|\,0, \Sigma_e^{(s)})$. It is also assumed that the noise levels are proportional across subjects, i.e., $\Sigma_e^{(s)} = \sigma^{(s)}\mathrm{Q}_e$, where Q_e is a known covariance component over channels. To facilitate group analysis, a canonical mesh is warped to each individual subject's cortical sheet such that the reconstructed source activity can be assigned to the same sources over subjects. The group structure is then enforced by assuming that the source activity for each subject can be factorized into a source-specific component and a subject-specific component

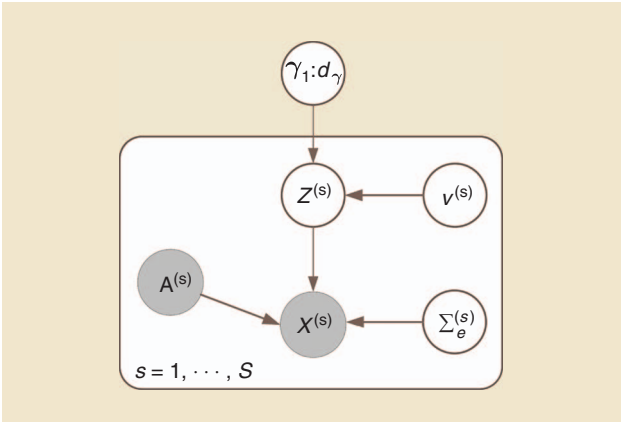
$$p(Z^{(s)}) = \prod_{t=1}^{T} p(z_t^{(s)}) = \prod_{t=1}^{T}\mathcal{N}\left(0, \upsilon^{(s)}\sum_{i=1}^{d_\gamma}\gamma_i\,\mathrm{C}_i\right), \tag{31}$$

where the source-specific scale hyperparameter γ_i represents the group constraint over the subject, and the subject-specific scale hyperparameter $\upsilon^{(s)}$ encodes the additional variability unique to each subject. This prior is a generalization of the source prior (7). In addition, log-normal hyperpriors are placed on the hyperparameters:

$$\lambda \triangleq \log([\gamma_{1:d_\gamma}, \upsilon^{(1:S)}, \sigma^{(1:S)}]^{\top}) \sim \mathcal{N}(\eta, \mathrm{R}).$$

Weakly informative hyperpriors of η and R were used in [23] to allow automatic model selection (see Figure 4 for a graphical model illustration).

For model inference, a computationally efficient two-stage source reconstruction algorithm [termed *gMSP* henceforth since multiple sparse priors akin to (7) are imposed on the source activity] was developed to estimate the hyperparameters $\boldsymbol{\lambda}$ and source activity [23]. The idea is to decompose the covariance function of the group data into a sparse set of source components and a noise component. The sparse set of source components is then used as empirical priors to estimate the subject-specific hyperparameters. More specifically, the



[FIG4] A graphical model representation of the group source imaging algorithm, gMSP, in the presence of S subjects. $X^{(s)}$ and $Z^{(s)}$ are the observed EEG data and unobserved source activity for the sth subject. $A^{(s)}$ and $\Sigma_e^{(s)}$ are the associated lead-field matrix and noise covariance matrix, respectively. $v^{(s)}$ is a subject-specific scale hyperparameter, and $\gamma_{1:d_\gamma}$ are source-specific scale hyperparameters that act as the group constraint. All unknown latent variables and parameters are shown as unshaded nodes and estimated from the data. See the section "VB for Group EEG/MEG Analysis" for details.

mean-field approximation is combined with the Laplace approximation by assuming that the posterior distribution can be factorized into Gaussian marginals (i.e., a combination of VB-MF and VB-LA; see the section "SBL Methods"):

$$q(\overline{Z},\lambda) = q(\overline{Z})\,q(\boldsymbol{\lambda}) \triangleq \mathcal{N}(\boldsymbol{\mu}_{\bar{z}},\Sigma_{\bar{z}})\,\mathcal{N}(\mu_\lambda,\Sigma_\lambda).$$

This leads to variational free energies at both the individual subject and group levels. The gMSP algorithm then proceeds as follows: in the first stage, $q(\lambda_{1:d_\gamma})$ is estimated by maximizing the group-level variational free energy; in the second stage, $q(\lambda_{(d_\gamma+1):(d_\gamma+2S+1)})$ is estimated by maximizing the subject-level variational free energy. With all the hyperparameters estimated, $q(\overline{Z})$ can be obtained as the MAP estimates of the source activity. Note that, to achieve sparsity on the source covariance components, both ARD and a greedy search approach were developed to optimize the source-specific hyperparameters. Refer to [23] and [24] for algorithmic details. Only gMSP with ARD is illustrated in this article (see the section "Group Electromagnetic Brain Imaging Using gMSP"). This hierarchical model approach allows one to place within- and between-subject constraints on the reconstructed source activity in the context of group studies.

SBL METHODS

Sparse learning, also known as *compressed sensing* in signal processing [25], is referred to as a *collection of learning methods* that seek a tradeoff between certain goodness-of-fit measure and sparsity of the solution, the latter of which allows better interpretability and enhanced generalization ability of the model. Sparse learning is particularly suited for analyzing EEG/MEG signals with high dimensionality, small sample size, and low SNR. See [26] for a recent review on the applications of sparse learning to brain signal processing.

Compared with its most common counterpart, two advantages of SBL [9], [27] are noteworthy:

- SBL allows automatic model selection. This can be achieved by both EB (see the "Introduction" section), and fully Bayesian methods. In EB, maximizing $p(X\,|\,\gamma)$ provides a natural regularizing mechanism that yields sparse solutions. In fully Bayesian methods, as suggested by the VB methods presented in the sections "VB for Learning Common EEG Components" and "VB for Group EEG/MEG Analysis," automatic sparse learning can be achieved by imposing noninformative priors on γ, which leads to sparsity since the hierarchical priors on the parameters are typically sparse priors. For instance, by marginalizing the variance, the normal-inverse-Gamma prior for each brain source in (16) amounts to a Student t-distribution:

$$p(z_{m,t}^{(k)}) = \int p(z_{m,t}^{(k)}\,|\,\rho_m^{(k)})\,p(\rho_m^{(k)})\,\mathrm{d}\rho_m^{(k)} = \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)\sqrt{2\pi}}\beta^\alpha[\beta+\frac{[z_{m,t}^{(k)}]^2}{2}]^{-(\alpha+\frac{1}{2})}, \quad (32)$$

  which is sharply peaked at zero in the noninformative case (i.e., when $\alpha \to 0, \beta \to 0$).
- SBL is more capable of finding sparse solutions than conventional methods. In typical electromagnetic brain-imaging problem setups, many columns of the lead-field matrix are highly correlated; in this case, the convex l_1-norm-based MAP (sparse) solution performs poorly since the restricted isometry property is violated. In spatiotemporal decomposition problems where $\mathbf{A}$ is unknown, the MAP estimation may suffer from too many local minima in the solution space due to the multiplicative structure of the spatiotemporal model (2). Consider the multiple covariance parameterization in (7). According to [28], if $\mathbf{C}_i = \mathbf{e}_i\,\mathbf{e}_i^\top$, where $\mathbf{e}_i$ is an indexing vector with one for the ith element and zero otherwise, the ARD solution can equivalently be viewed as the solution to the following MAP problem:

$$\max_{z_t} \|\boldsymbol{x}_t - \mathbf{A}z_t\|_2^2 + \lambda h^*(z_t), \quad (33)$$

  where λ denotes the regularization parameter, and $h^*(z_t) \triangleq \min_\gamma [z_t^2]^\top\gamma + \log|\Sigma_x|$ is the concave conjugate of $-\log|\Sigma_x|$. The prior distribution associated with $h^*(z_t)$ is generally nonfactorial, i.e., $p(z_t) \neq \prod_m p(z_{m,t})$, and is dependent on both $\mathbf{A}$ and λ. It has been shown that $h(z_t)$ provides a tighter approximation that promotes greater sparsity than the l_1-norm while conveniently producing many fewer local minima than using the l_0-norm [8]. In [17], we establish a similar result for VB-CSP in the spatiotemporal decomposition setup.

In this section, we review two recent works to illustrate SBL in EEG/MEG signal processing.

SBL FOR LEARNING ERPs

ERPs are stereotyped electrical activities of the brain that occur in response to specific events or stimuli. Accurate estimation of the amplitude and latency of ERPs is pivotal in delineating the successive

stages of mental processes and differentiating among various events or stimuli. In multichannel EEG recordings, since the number of ERP components can be considerably smaller than the number of the sensors, SBL is suited to automatically infer the sparse ERP components.

In [29], a Bayesian model was proposed to estimate the components specific to each experimental condition for ERPs from multicondition and multitrial EEG data. More specifically, we adapted (2) to the following linear latent variable model to accommodate the ERP estimation:

$$\boldsymbol{x}_t^{(\mathrm{kl})} = \sum_{m=1}^{M} c_m^{(k)} \mathbf{a}_m z_{m,t+\tau_m^{(k)}} + e_t^{(\mathrm{kl})}, \tag{34}$$

where $\boldsymbol{x}_t^{(\mathrm{kl})} \in \mathbb{R}^N$ and $e_t^{(\mathrm{kl})} \in \mathbb{R}^N$ denote the EEG signal and the noise-plus-interference term from the lth ($l \in \{1, \cdots, L\}$) trial of condition $k \in \{1, \cdots, K\}$, respectively; $z_{m,t}$ denotes the waveform of the stereotyped mth ERP component; $\mathbf{a}_m \in \mathbb{R}^N$ denotes the scalp map of the mth ERP component; and $c_m^{(k)}$ and $\tau_m^{(k)}$ denote the amplitude factor and the latency of the mth ERP component for condition k, respectively. We further assume the noise terms are independent and identically Gaussian distributed across time and trials $e_t^{(\mathrm{kl})} \sim \mathcal{N}(0, \Sigma_e^{(k)})$, where $\Sigma_e^{(k)}$ are the spatial covariance matrices. Note that no particular structure is assumed for $\Sigma_e^{(k)}$ (e.g., $\Sigma_e^{(k)}$ can be nondiagonal).

Two key assumptions are made in (34): 1) the waveform and spatial pattern of each ERP component are invariant across trials and conditions, and 2) intercondition ERPs differ only in their amplitudes and latencies. These assumptions are motivated by the existing experimental observations that ERPs are approximately time- and phase-locked across trials and that ERP variability is typically far less within conditions than across conditions. Furthermore, unlike fixed-effect ERP components, the noise-plus-interference activities $e_t^{(\mathrm{kl})}$ are modeled as random effects in (34), which is also in agreement with existing experimental observations. Crucially, $e_t^{(\mathrm{kl})}$ is allowed to be correlated and nonisotropic among sensors, with the covariance matrices $\Sigma_e^{(k)}$ following noninformative inverse Wishart distributions for $\nu^{-1} \to 0$:

$$[\Sigma_e^{(k)}]^{-1} \sim \mathcal{W}(\nu\mathbf{I}, M) \triangleq \frac{1}{2^{\frac{M^2}{2}} |\, \nu\mathbf{I} \,|^{\frac{M}{2}} \Gamma_M\left(\frac{M}{2}\right)} |\, \Sigma_e^{(k)} \,|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\mathrm{trace}\left[\nu^{-1}[\Sigma_e^{(k)}]^{-1}\right]\right). \tag{35}$$

To address the inherent scaling ambiguity in (34), z_m and $\mathbf{a}_m$ are endowed with standard Gaussian priors:

$$\mathbf{a}_m \sim \mathcal{N}(0, \mathbf{I}), z_{m,t} \sim \mathcal{N}(0,1). \tag{36}$$

To allow for automatic determination of the component number M, a noninformative hierarchical prior akin to (32) can be imposed on $c_m^{(k)}$. Here, at the first level, a half-Gaussian is assumed to account for a fixed polarity of each ERP component across conditions

$$\begin{aligned} &c_m^{(k)} \sim \mathcal{N}^+(0, \alpha_m^{-1}) \triangleq 2\sqrt{\frac{\alpha_m}{2\pi}} \exp\left(-\frac{1}{2}\alpha_m [c_m^{(k)}]^2\right), c_m^{(k)} \geq 0 \\ &\alpha_m \sim \mathcal{G}(u, v) \triangleq \frac{v^u}{\Gamma(u)} \alpha_m^{u-1} \exp(-v\alpha_m), u \to 0, v \to 0. \end{aligned} \tag{37}$$

Moreover, the latency shifts $\tau_m^{(k)}$ are integers within a preset interval $[t_1, t_2]$. The probabilistic graphical model is shown in Figure 5.

In [29], a VB-EM algorithm termed *Bayesian estimation of ERPs* (*BEEP*) was developed for inferring model (34) based on the following mean-field approximation:

Algorithm 3: The BEEP algorithm.

Input: $\{\mathbf{X}^{(\mathrm{kl})}\}$

Output: $q(\mathbf{a}_m), q(z_{m,t}), q(c_m^{(k)}), q(\alpha_m), q(\Sigma_e^{(k)}), \tau_m^{(k)}$

1: **Initialization**

2: **repeat**

3: $q(\mathbf{a}_m) = \mathcal{N}(\mathbb{E}[\mathbf{a}_m], C[\mathbf{a}_m])$, where $\mathbb{C}[\mathbf{a}_m] = [L\sum_{k,t} \mathbb{E}[[c_m^{(k)}]^2] \mathbb{E}[z_{m,t}^2] \mathbb{E}[[\Sigma_e^{(k)}]^{-1}] + \mathbf{I}]^{-1}$,
$\mathbb{E}[a_m] = R\sum_{k,t} \mathbb{C}[a_m] \mathbb{E}[c_m^{(k)}] \mathbb{E}[z_{m,t}] \mathbb{E}[[\Sigma_e^{(k)}]^{-1}] \bar{s}_{m,t}^{(k)}$, $\bar{s}_{m,t}^{(k)} \triangleq \sum_r [x_{m,t}^{(k)} - \sum_{i \neq m} \mathbf{a}_i c_i^{(k)} z_{i,t-\tau_{li}^*}]$

4: $q(z_{m,t}) = \mathcal{N}(\mathbb{E}(z_{m,t}), \mathbb{C}(z_{m,t}))$, where $\mathbb{C}[z_{m,t}] = [L\sum_k \mathbb{E}([c_m^k]^2)(\mathrm{trace}(\mathbb{C}[\mathbf{a}_m] \mathbb{E}[[\Sigma_e^{(k)}]^{-1}]) + \mathbb{E}^\top[\mathbf{a}_m] \mathbb{E}[[\Sigma_e^{(k)}]^{-1}] \mathbb{E}[\mathbf{a}_m]) + K]^{-1}$,
$\mathbb{E}[z_{m,t}] = [L\sum_k [\bar{s}_{m,t-\tau_{lj}}^{(k)}]^\top \mathbb{E}[[\Sigma_e^{(k)}]^{-1}] \mathbb{E}[\mathbf{a}_m] \mathbb{E}[c_m^{(k)}]] \mathbb{C}[z_{m,t}]$

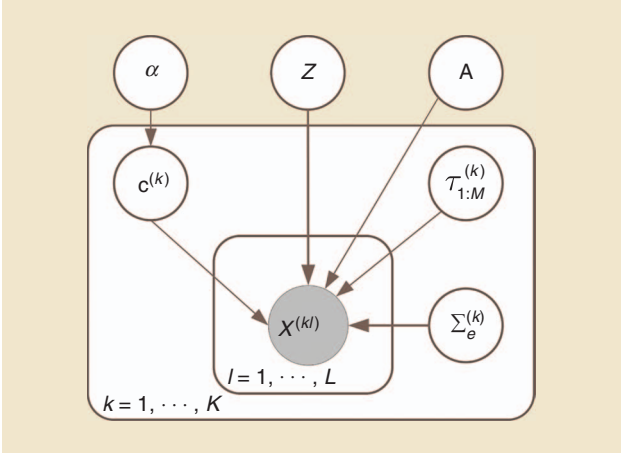
5: $q(c_m^{(k)}) = \kappa \cdot \mathcal{N}(\mu_m^{(k)}, \sigma_m^{(k)})$, $c_m^{(k)} \geq 0$, where $\sigma_m^{(k)} \triangleq \left[\alpha_m + L\sum_t \mathbb{E}([z_{m,t}]^2)(\mathrm{trace}(\mathbb{C}[\mathbf{a}_m] \mathbb{E}[[\Sigma_e^{(k)}]^{-1}]) + \mathbb{E}^\top[\mathbf{a}_m] \mathbb{E}[[\Sigma_e^{(k)}]^{-1}] \mathbb{E}[\mathbf{a}_m])\right]^{-1}$,
$\mu_m^{(k)} \triangleq [L\sum_t [\bar{s}_{m,t-\tau_{lj}}^{(k)}]^\top \mathbb{E}[[\Sigma_e^{(k)}]^{-1}] \mathbb{E}[a_m] \mathbb{E}[z_{m,t}]] \sigma_m^{(k)}$, $\kappa \triangleq 1/(1 - \Phi(-\frac{\mu_m^{(k)}}{\sqrt{\sigma_m^{(k)}}}))$.

6: $q(\alpha_m) = \mathcal{G}(\tilde{u}, \tilde{v})$, where $\tilde{u} \triangleq \frac{K}{2}$, $\tilde{v} \triangleq \frac{1}{2}\sum_k \mathbb{E}[[c_m^{(k)}]^2]$.

7: $q([\Sigma_e^{(k)}]^{-1}) = \mathcal{W}([\Gamma^{(k)}]^{-1}, LT + M)$, where
$\Gamma^{(k)} \triangleq \sum_l \sum_t (\boldsymbol{x}_t^{(\mathrm{kl})} [\boldsymbol{x}_t^{(\mathrm{kl})}]^\top) - 2\sum_l \sum_t \left(\boldsymbol{x}_t^{(\mathrm{kl})} \sum_{j=1}^M [\mathbf{h}_{j,t}^{(k)}]^\top\right) + L\sum_t \sum_{i=1}^M \sum_{j=1, j \neq i}^M \mathbf{h}_{i,t}^{(k)} [\mathbf{h}_{j,t}^{(k)}]^\top + L\sum_t \sum_{j=1}^M \mathbb{E}[\mathbf{a}_j \mathbf{a}_j^\top] \mathbb{E}[[c_j^{(k)}]^2] \mathbb{E}[[z_{j,t}^{(k)}]^2]$,
$\mathbf{h}_{j,t}^{(k)} \triangleq \mathbb{E}[\mathbf{a}_j] \mathbb{E}[c_j^{(k)}] \mathbb{E}[z_{j,t}^{(k)}]$.

8: $\tau_m^k = \arg\max_{\tau_m^k \in [t_1, t_2]} f(\bar{\tau}_m^k)$, where $f(\bar{\tau}_m^k) \triangleq \sum_t \mathbb{E}[c_m^{(k)}] \mathbb{E}[a_m] \mathbb{E}^\top[z_{m,t-\bar{\tau}_m^{(k)}}] \mathbb{E}[[\Gamma^{(k)}]^{-1}] \bar{s}_{m,t}^{(k)}$.

9: **until** Convergence



[FIG5] **A graphical model representation for BEEP in the presence of K conditions and L trials. $X^{(kl)}$ denotes the EEG data for the lth trial of condition k. $Z = [z_1, \cdots, z_M]^\top$ contains the time courses of the M ERP components. A denotes the mixing matrix. $\Sigma_e^{(k)}$ is the noise spatial covariance matrix for condition k. $c^{(k)} = [c_1^{(k)}, \cdots, c_M^{(k)}]$ consists of the amplitude factors of the ERP components for condition k, and $\alpha = [\alpha_1, \cdots, \alpha_M]$ are the corresponding hyperparameters. $\tau_m^{(k)}$ denotes the latency shift of the mth ERP component for condition k. All unknown latent variables and parameters are shown as unshaded nodes and estimated from the data using update rules provided in Algorithm 3.**

$$q(\Theta) = \prod_{k,m,t} q(\mathrm{a}_m)\, q(c_m^{(k)})\, q(z_{m,t})\, q(\alpha_m)\, q(\Sigma_e^{(k)}), \tag{38}$$

where $q(\mathrm{a}_m), q(c_m^{(k)}), q(z_{m,t}), q(\alpha_m), q(\Sigma_e^{(k)})$ are updated alternately, and $\tau_m^{(k)}$ is updated by maximizing the variational free energy over integer interval $[t_1, t_2]$. BEEP is summarized in Algorithm 3. Results on synthetic data and 13 subjects' EEG recordings collected in a face-inversion experiment demonstrated that BEEP was superior to several state-of-the-art ERP analysis algorithms, yielding neurophysiologically meaningful ERP components [29].

SBL FOR BRAIN IMAGING OF CORRELATED SOURCES

This section briefly describes another SBL algorithm called *Champagne* for electromagnetic brain imaging. The Champagne algorithm relies on segmenting data into pre- and poststimulus periods, learning the statistics of the background activity from the prestimulus period, and then applying the statistics of the background activity to the poststimulus data to uncover the stimulus-evoked activity. The underlying assumption is that noise and nonstimulus-locked brain activity in the prestimulus period remain unchanged in the poststimulus period, whereas the stimulus-evoked activity is linearly superimposed on the prestimulus activity. In contrast to SAKETINI (which is a scanning method and estimates the statistics of noise and interference at each voxel), in the Champagne algorithm, all voxel activity is simultaneously estimated (such as the variance parameters for each voxel). This formulation leads to implicit sparse estimation whereby many voxel variance parameters are guaranteed to be zero, if the noise and interference statistics are accurate. Therefore, the Champagne algorithm is a considerably faster method for whole-brain imaging in the presence of sparse sources.

The poststimulus sensor data X_{post} is modeled as

$$X_{\text{post}} = \sum_{r=1}^{d_z} \mathrm{A}_r\, Z_r + E, \tag{39}$$

where $X_{\text{post}} \in \mathbb{R}^{N\times T}$ represents T sample points in the poststimulus period from N sensors. $\mathrm{A}_r \in \mathbb{R}^{N\times M}$ is the lead-field matrix in M orientations for the rth voxel. Each unknown source $Z_r \in \mathbb{R}^{M\times T}$ is an M-dimensional neural current-dipole source at T time points, projecting from the ith voxel. There are d_z voxels under consideration. $E \in \mathbb{R}^{N\times T}$ is a noise plus interference term estimated from the prestimulus period using partitioned factor analysis models [30], which assumes independence over time. The first step of the Champagne algorithm is to estimate Σ_e by source localization. The second step of the Champagne algorithm is to estimate hyperparameters of Γ that govern the statistical model of the poststimulus data. We can fully define the probability distribution of the data conditioned on the sources

$$p(X_{\text{post}} \mid Z) \propto \exp\left(-\frac{1}{2}\left\| X_{\text{post}} - \sum_{r=1}^{d_z} \mathrm{A}_r Z_r \right\|^2_{\Sigma_e^{-1}}\right), \tag{40}$$

where $\| X \|_{\mathrm{W}}$ denotes the weighted matrix norm $\sqrt{\mathrm{trace}[X\mathrm{W}X^\top]}$.

The following source prior is assumed for Z:

$$p(Z \mid \Gamma) \propto \exp\left(-\frac{1}{2}\mathrm{trace}\left[\sum_{r=1}^{d_z} Z_r^\top \Gamma_r^{-1} Z_r\right]\right). \tag{41}$$

This is equivalent to applying independently, at each time point, a zero-mean Gaussian distribution with covariance $\mathbf{\Gamma}_r$ to each source Z_r. We define $\mathbf{\Gamma} = \mathrm{diag}(\mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_{d_z})$ as the $d_z M \times d_z M$ block-diagonal matrix, formed by ordering each $\mathbf{\Gamma}_r$ along the diagonal of an otherwise zero-valued matrix. If the lead-field has only one orientation (scalar/orientation-constrained lead-field), $\mathbf{\Gamma}$ reduces to a diagonal matrix. Since $\mathbf{\Gamma}$ is unknown, we use the approximation $\hat{\mathbf{\Gamma}}$ by integrating out the sources Z of the joint distribution $p(Z, X_{\text{post}} \mid \mathbf{\Gamma}) \propto p(X_{\text{post}} \mid Z) p(Z \mid \mathbf{\Gamma})$, i.e.,

$$p(X_{\text{post}} \mid \mathbf{\Gamma}) = \int p(Z, X_{\text{post}} \mid \mathbf{\Gamma})\, \mathrm{d}Z, \tag{42}$$

and then minimizing the cost function

$$\mathcal{L}(\mathbf{\Gamma}) \triangleq -2 \log p(X_{\text{post}} \mid \mathbf{\Gamma}) \equiv \mathrm{trace}[\mathrm{C}_x\, \Sigma_x^{-1}] + \log|\Sigma_x|, \tag{43}$$

where $\mathrm{C}_x \triangleq (1/T) X_{\text{post}} X_{\text{post}}^\top$ is the empirical covariance and $\Sigma_x = \Sigma_e + \sum_{r=1}^{dz} \mathrm{A}_r \mathbf{\Gamma}_r \mathrm{A}_r^\top$.

Minimizing the cost function (43) with respect to $\mathbf{\Gamma}$ can be done in a variety of ways, including gradient descent or the EM algorithm, but these and other generic methods are exceedingly slow when d_z is large. Instead, we use an alternative optimization procedure that resorts to convex bounding techniques [9]. This method expands on ideas from [9], [10], [28], and [31], handles arbitrary/unknown dipole-source orientations, and converges quickly due to a tighter upper bound [10]. This optimization procedure yields a modified cost function:

$$\mathcal{L}(\{\mathbf{\Gamma}_r\}, \{Z_r\}, \{Y_r\}) = \left\| \tilde{X} - \sum_{r=1}^{d_z} \mathrm{A}_r Z_r \right\|^2_{\Sigma_e^{-1}} + \sum_{r=1}^{d_z} [\| Z_r \|^2_{\mathbf{\Gamma}_r^{-1}} + \mathrm{trace}(Y_r^\top \mathbf{\Gamma}_r)] - h^*(Y_r), \tag{44}$$

Algorithm 4: The Champagne algorithm.

Input: $X_{\text{post}}, \mathbf{A}_r, \Sigma_e$
Output: $Z_r, \mathbf{\Gamma}_r ((r = 1, \ldots, d_z))$
1: **Initialization**
2: **repeat**
3: $\Sigma_x = \Sigma_e + \sum_r \mathbf{A}_r \mathbf{\Gamma}_r \mathbf{A}_r^\top$
4: $Z_r = \mathbf{\Gamma}_r \mathbf{A}_r^\top \Sigma_x^{-1} \tilde{X}$
5: $Y_r = \nabla_{\mathbf{\Gamma}_r} \log|\Sigma_x| = \mathbf{A}_r^\top \Sigma_x^{-1} \mathbf{A}_r$
6: $\mathbf{\Gamma}_r = Y_r^{-1/2} (Y_r^{1/2} Z_r Z_r^\top Y_r^{1/2})^{1/2} Y_r^{-1/2}$.
7: **until** Convergence

where $h^*(Y_r)$ is the concave conjugate of $\log|\Sigma_x|$ for auxiliary variables $Y_r = \mathbf{A}_r^\top \Sigma_x^{-1} \mathbf{A}_r$. By construction $\mathcal{L}(\mathbf{\Gamma}) = \min_{\{Z_r\}} \min_{\{Y_r\}} \mathcal{L}(\{\mathbf{\Gamma}_r\}, \{Z_r\}, \{Y_r\})$, the matrix $\tilde{X} \in \mathbb{R}^{N \times \text{rank}(X_{\text{post}})}$, $\tilde{Z}\tilde{Z}^\top = \mathbf{C}_x$. Minimizing this modified cost function yields three update rules (summarized in Algorithm 4) [31]. In summary, the Champagne algorithm estimates $\mathbf{\Gamma}$ by iterating between $\{Z_r\}$, $\{Y_r\}$, $\{\mathbf{\Gamma}_r\}$, and with each pass monotonically decreasing $\mathcal{L}(\mathbf{\Gamma})$.

NB METHODS

The historic roots of Bayesian nonparametrics date back to the late 1960s, but its applications in machine learning have not been widespread until recently [32]. In contrast to parametric models, NB models accommodate a large number of degrees of freedom (infinite dimensional parameter space) to exhibit a rich class of probabilistic structure, which make them more appealing, flexible, and powerful in data representation. The fundamental building blocks of Bayesian nonparametrics are two stochastic processes: the GP and DP. Two excellent introductory books on GP and DP are [33] and [34]. This section presents a few examples of these nonparametric modeling tools for EEG/MEG data analysis.

GP MODELING

Any finite set of random variables is a GP if it has a joint Gaussian distribution. Unlike the fixed finite-dimensional parametric model, the GP defines priors for the mean function and covariance function, where the covariance kernel determines the smoothness and (non)-stationarity between the time-series data points. To construct a GP model, given an input–output training sample set $(\boldsymbol{x}_{1:T}, \boldsymbol{y}_{1:T})$ of size T (where $\boldsymbol{x}_i \in \mathbb{R}^N$, $y_i \in \mathbb{R}$,, without loss of generality, assuming the time series has zero mean), the goal is to estimate a distribution from the data and predict the outcome of an unseen input data vector $\boldsymbol{x}_{T+1}$

$$P(y_{T+1} | \boldsymbol{x}_{T+1}, \boldsymbol{x}_{1:T}, \boldsymbol{y}_{1:T}) = \frac{1}{Z} \exp\left(-\frac{1}{2} \boldsymbol{y}_{1:T+1}^\top \mathbf{C}_{T+1}^{-1} \boldsymbol{y}_{1:T+1}\right), \quad (45)$$

where Z is a normalizing constant and $\mathbf{C}$ defines the covariance matrix for input data, characterized by an unknown hyperparameter vector θ. The covariance kernel can be either stationary or nonstationary [35], and different choices of the covariance kernel result in various types of GPs.

GP FOR EEG SEIZURE DETECTION

Important tasks of GP modeling are prediction or outlier detection [36], as well as EEG signal classification [37]. Specifically, the predictive distribution can be computed by integrating out the hyperparameters θ:

$$P(y_{T+1} | \boldsymbol{x}_{T+1}, \boldsymbol{x}_{1:T}, \boldsymbol{y}_{1:T}) = \int P(y_{T+1} | \boldsymbol{x}_{T+1}, \mathbf{C}, \theta) p(\mathbf{C}, \theta) \, \mathrm{d}\theta, \quad (46)$$

and the covariance kernel has the form $\mathbf{C}(\boldsymbol{x}_t, \boldsymbol{x}_{t'}) = \theta_0 e^{-\frac{1}{2}\sum_{i=1}^N \theta_i (x_{i,t} - x_{i,t'})^2} + \theta_\eta \delta_{tt'}$, where $\theta = [\theta_0, \theta_1, \ldots, \theta_N, \theta_\eta] \in \mathbb{R}^{N+2}$ denotes the hyperparameters, and $\delta_{tt'}$ is a Kronecker delta function (which is equal to one if and only if $t' = t$).

In most of cases, the integral of (46) is analytically intractable. Two possible solutions can be considered: a Monte Carlo method or a likelihood method that replaces the distribution of $\boldsymbol{\theta}$ with a maximum likelihood (point) estimate. For EEG outlier prediction, the variance approach and hyperparameter approach have been proposed [36]. In the variance approach, the variance of predicted output is monitored: if the variance is outside the range of the normal training samples, it will indicate the change of EEG structure or the presence of an outlier. In the hyperparameter approach, the ratio of two hyperparameters (e.g., $|\theta_0/\theta_\eta|$, which reflects the level of determinism in the EEG signal) is monitored.

MEG MODELING AND DIMENSIONALITY REDUCTION

For high-dimensional, structured EEG/EEG data, dimensionality reduction is useful for data denoising and interpretation. It is also insufficient to assume that the correlations between the elements of observations vector are static in many practical applications. Recently, GP has been used for heteroscedastic modeling of noisy high-dimensional MEG data [38]. Specifically, the N-dimensional MEG time series at the lth trial can be modeled as a non-Markovian dynamic linear factor analysis model

$$z_t^{(l)} = \psi^{(l)}(\tau_t) + v_t^{(l)} \quad (47)$$
$$\boldsymbol{x}_t^{(l)} = \mathbf{A}(\tau_t) z_t^{(l)} + e_t^{(l)}, \quad (48)$$

where $\mathbf{A}(\tau_t) \in \mathbb{R}^{N \times M} ((M \ll N)$ denotes a time-evolving factor loading matrix to account for nonstationarity, and $v_t^{(l)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $e_t^{(l)} \sim \mathcal{N}(\mathbf{0}, \Sigma_0)$ denote the zero-mean Gaussian dynamic and measurement noise, respectively. The noise covariance matrix Σ_0 is assigned with an inverse gamma prior.

This model differs from (3) and (4) in that the latent process is drawn from GPs. In the state equation, the low-dimensional latent process is constructed by a hierarchy, in which the evolution of the latent factor z_t, at each trial, is governed by a collection of M dictionary functions. Each element itself is a GP [38]: $\psi^{(l)}(\tau) = [\psi_1^{(l)}(\tau), \ldots, \psi_M^{(l)}(\tau)] \in \mathbb{R}^M$, where $\psi_j^{(0)}(\cdot) \backsim \text{GP}(0, \mathbf{C}_0)$, $\psi_j^{(l)}(\cdot) \backsim \text{GP}(\psi_j^{(0)}, \mathbf{C}_1)$ are two GPs with squared exponential correlation function, with $\mathbf{C}_k(\boldsymbol{\xi}, \boldsymbol{\xi}) = d_k \exp(-\kappa \|\boldsymbol{\xi} - \boldsymbol{\xi}'\|_2^2)$ for $k = 0, 1$. Therefore, the child processes $\psi_j^{(l)}$ are centered around the parent process $\psi_j^{(0)}$, and they all share information through the parent process.

In the observation equation, the factor loading matrix $\mathbf{A} \in \mathbb{R}^{N \times M}$ is modeled by a weighted combination of R-dimensional latent covariance dictionary functions [38]

$$\mathbf{A}(\tau) = \mathbf{B}\mathbf{\Phi}(\tau),$$

where $\mathbf{B} \in \mathbb{R}^{N\times R}$ $(R \ll M \ll N)$, and $\Phi \in \mathbb{R}^{R\times M} = \{\phi_{jk}(\cdot)\} \sim \mathrm{GP}(0, \mathbf{C}_0)$ consists of a covariance dictionary. Such GP modeling of dictionary elements allows the MEG signals' long-range dependencies to be captured. Each row of $\mathbf{B}$ will be assigned with a sparse shrinkage prior (that penalizes a large coefficient). The proposed hierarchical model is able to characterize nonstationarity (via the time-varying factor loading matrix) and to share information (via coupling) between single trials.

The goal of Bayesian modeling is to infer the latent variable and parameters of the hierarchical factor analysis model. First, the componentwise observation $x_{n,t}^{(l)}$ can be written as

$$x_{n,t}^{(l)} = \sum_{m=1}^{M} z_{m,t}^{(l)} \sum_{r=1}^{R} b_{\mathrm{nr}} \phi_{\mathrm{rm}}(\tau_t) + e_{i,t}^{(l)}.$$

Marginalizing the dynamic and measurement noise $v_t^{(l)}$ and $e_t^{(l)}$ induces the following time-varying mean and covariance structure for the observed signal $x_t^{(l)} \sim \mathcal{N}(\mu^{(l)}(\tau_t), \Sigma(\tau_t))$, where

$$\begin{aligned}\mu^{(l)}(\tau_t) &= \mathbf{B}\Phi(\tau_t)\psi^{(l)}(\tau_t)\\ \Sigma(\tau_t) &= \mathbf{B}\Phi(\tau_t)\Phi(\tau_t)^{\top}\mathbf{B}^{\top} + \Sigma_0.\end{aligned}$$

The time-varying covariance structure captures the heteroscedasticity of time series, and the overcomplete representation provides flexibility and computational advantages. Bayesian inference (posterior computation and predictive likelihood) of this hierarchical GP model is achieved by MCMC sampling methods of $\{\psi_j^{(l)}, v_{1:T}^{(l)}, \psi^{(0)}, \Phi, \mathbf{B}, \Sigma_0\}$ [38]. As shown in [38], such hierarchical NB modeling provides a powerful framework to characterize a signal noisy MEG recording that allows word category classification.

DP MODELING

Mixture modeling has been commonly used time-series data analysis. Unlike finite mixture models, NB models define a prior distribution over the set of all possible partitions, where the number of clusters or partitions may grow as the data samples increase. This is particularly useful for EEG/MEG applications in clustering, partition, segmentation, and classification. Examples of static or dynamic mixture models include DP mixtures, the infinite hidden Markov model, and hierarchical DP. As an illustration, we consider an example of multisubject EEG classification in the context of NB CSP, which defines the Bayesian CSP model with either a DP prior or an Indian-buffet-process prior [18], [39].

The NB CSP model extends the probabilistic CSP (see [17]) and further introduces the DP prior, which was referred to as *BCSP-DP* [18]. The BCSP-DP uses a DP mixture model to learn the number of spatial patterns among multiple subjects. The spatial patterns with the same hyperparameter are grouped in the same cluster [see the graphical model illustration in Figure 2(b)], thereby facilitating the information transfer between subjects with similar spatial patterns. Essentially, BCSP-DP is a nonparametric counterpart of VB-CSP.

For condition $k \in \{1, 2\}$ and subject $s \in \{1, \ldots, S\}$, the multisubject CSP model is rewritten as

$$X^{(\mathrm{sk})} = \mathrm{A}^{(s)} Z^{(\mathrm{sk})} + E^{(\mathrm{sk})}. \tag{49}$$

As seen in Figure 2(b), spatial patterns $\{\mathrm{a}_m^{(s)}\}$ are drawn from Gaussian distributions $\mathcal{N}(\mathrm{a}_m^{(s)} \mid \mu_m^{(s)}, (\Phi_m^{(s)})^{-1})$, where the Gaussian parameters $(\theta_m^{(s)} = \{\mu_m^{(s)}, \Phi_m^{(s)}\})$ are further drawn from a random measure G in a DP, $\theta_m^{(s)} \sim G, G \sim \mathrm{DP}(\alpha, G_0)$ for $m = 1, \ldots, M$ and $s = 1, \ldots, S$, where $\alpha > 0$ is a positive concentration parameter (which specifies how strong this discretization is: when $\alpha \to 0$, the realizations are all concentrated on a single value, when $\alpha \to \infty$, the realizations become continuous), and the base measure $G_0 = p(\mu, \Phi)$ is set to be a Gaussian–Wishart distribution, which is the conjugate prior for Gaussian likelihood

$$p(\mu, \Phi) = \mathcal{N}(\mu \mid m_0, (\beta_0\Phi)^{-1})\mathcal{W}(\Phi \mid \nu_0, \mathbf{W}_0), \tag{50}$$

where $\mathcal{W}(\Phi \mid \nu_0, \mathbf{W}_0)$ denotes a Wishart distribution with the degree of freedom ν_0 and the scale matrix $\mathbf{W}_0$. The random measure G has the following stick-breaking representation

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i^*}, \quad ;\pi_i = v_i \prod_{j=1}^{i-1}(1 - v_i), \tag{51}$$

where $v_i \sim \mathcal{B}(v_i \mid 1, \alpha)$ and $\theta_i^* \in G_0$ are independent and random variables drawn from a beta distribution and the beta measure G_0, respectively. The mixing proportions $\{\pi_i\}$ are given by successively breaking a unit-length stick into an infinite number of pieces.

Assuming hierarchical priors for $z_t^{(\mathrm{sk})}$ and $x_t^{(\mathrm{sk})}$ [see legend of Figure 2(b) for the priors for α, v_i, $\mathrm{A}^{(s)}$]

$$\begin{aligned}z_t^{(\mathrm{sk})} &\sim \mathcal{N}(0, \Sigma_z^{(\mathrm{sk})})\\ x_t^{(\mathrm{sk})} &\sim \mathcal{N}(A^{(s)} z_t^{(\mathrm{sk})}, \Sigma_e^{(\mathrm{sk})}),\end{aligned}$$

where Σ_z and Σ_e denote two diagonal covariance matrices with diagonal elements drawn from inverse-Gamma distributions. Finally, let Θ define the collective set of unknown variables of parameters and hyperparameters

$$\Theta = \{\{\mathrm{A}^{(s)}\}, c_m^{(s)}, \{Z^{(\mathrm{sk})}\}, \{v_i\}, \alpha, \{\Sigma_z^{(\mathrm{sk})}\}, \{\Sigma_e^{(\mathrm{sk})}\}, \{\mu_i^*, \Phi_i^*\}\}.$$

The authors of [18] employed VB inference by assuming a factorized variational distribution $q(\Theta)$

$$\begin{aligned}q(\Theta) = q(\{\mathrm{A}^{(s)}\})q(c_m^{(s)})q(\{Z^{(\mathrm{sk})}\})q(\{v_k\})\\ \times q(\alpha)q(\{\Sigma_z^{(\mathrm{sk})}\})q(\{\Sigma_e^{(\mathrm{sk})}\})q(\{\mu_i^*, \Phi_i^*\}).\end{aligned}$$

The variational posterior $q(Z^{(\mathrm{sk})}) = \prod_t q(z_t^{(\mathrm{sk})})$, $q(z_t^{(\mathrm{sk})}) = \mathcal{N}(\mu_t^{(\mathrm{sk})}, \Sigma_*^{(\mathrm{sk})})$ has an analytic form, with

$$\begin{aligned}(\Sigma_*^{(\mathrm{sk})})^{-1} &= \mathbb{E}_q[(\Sigma_z^{(\mathrm{sk})})^{-1}] + \sum_{i=1}^{N} \mathbb{E}_q[(\Sigma_e^{(\mathrm{sk})})^{-1}]_i \, \mathbb{E}_q[[A^{(s)}]_{i,:}^{\top}[A^{(s)}]_{i,:}]\\ \mu_t^{(\mathrm{sk})} &= \Sigma_*^{(\mathrm{sk})} \mathbb{E}_q[A^{(s)}]^{\top} \mathbb{E}_q[(\Sigma_e^{(\mathrm{sk})})^{-1}] x_t^{(\mathrm{sk})}.\end{aligned}$$

For inference details, see [18]. At the cost of increasing computational complexity, the BCSP-DP model achieved improved EEG classification performance compared to state-of-the-art CSP models [18].

APPLICATIONS

BML has been successfully used in a variety of EEG/MEG applications. As mentioned in the "Introduction" section, the two main applications of BML algorithms for EEG/MEG data are brain imaging and brain-state classification. Electromagnetic brain imaging has diverse applications in basic and clinical neuroscience. Brain-state classification may include development of BCIs and methods for diagnosis of abnormal brain activity. In this section, we briefly describe a few representative examples of both applications.

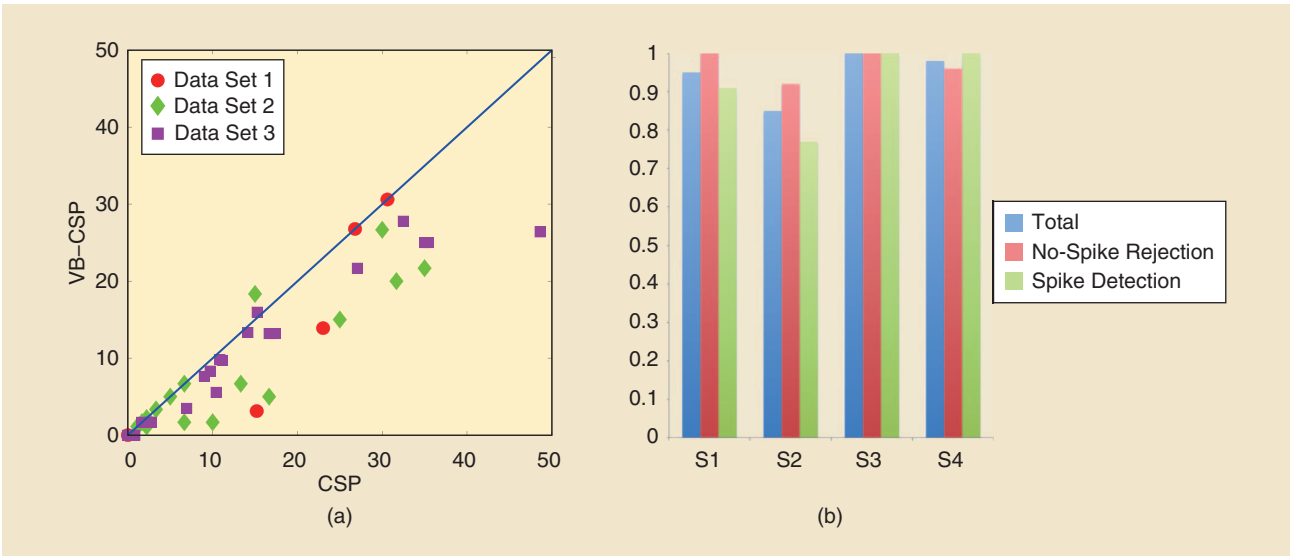
BRAIN-STATE CLASSIFICATION USING VB-CSP

The VB-CSP algorithm can be used for brain-state classification, such as the motor imagery task [17]. Imagined limb or tongue movements give rise to an attenuation of the EEG rhythms in different regions of the sensorimotor cortices, and CSP-based methods have proven to be effective in distinguishing imagined movements by extracting their distinctive EEG spatiotemporal patterns. The standard CSP methods have the overfitting problem. In our previous work, the efficacy of VB-CSP was demonstrated through applications to single-trial classifications of three motor imagery EEG data sets, two of which were taken from BCI Competition III (Data Set IIIa as data set 1, and Data Set IVa as data set 2). A total of 43 binary-class data subsets were generated from the three data sets by pairing MI tasks. The overall test error for VB-CSP is 10.36%, compared with 14.22% for CSP. Paired t-tests indicate that VB-CSP significantly outperforms CSP ($p = 1.61 \times 10^{-5}$). For data sets with a small number of training trials, which CSP has a tendency to overfit, VB-CSP alleviates the overfitting by using substantially less but automatically determined components than the available EEG channels.

The VB-CSP algorithm was also employed to identify periods of interictal epileptiform activity in four different patients with epilepsy. Generally, the challenges in analyzing EEG and/or MEG data from epilepsy patients are twofold. First, it is important to identify events in the data that represent epileptiform activity and this is often done manually by domain experts. After identification of such abnormal epileptiform activity events, then imaging algorithms are run on these events to determine origin in the brain to identify the epileptogenic zone. Algorithms like VB-CSP can be used to automatically segment long MEG or EEG traces into short pieces that can be classified as representing epileptiform activity, and such segmented data can subsequently be used with imaging algorithms. Alternative non-Bayesian approaches in the literature have included engineering solutions [40]. In this particular data set, a registered EEG/evoked potential technologist marked segments of the continuously collected data set that contained spontaneous epileptiform spikes, as well as segments that clearly contained no spikes. All of these data were split into training and test data sets. The VB-CSP was first applied to learn a generative model for spike and "nonspike" training data, and then the classification performance was evaluated in independent test data. Results from the four subjects, shown in Figure 6, demonstrated a brain-state classification accuracy of above 85% for spikes and nonspikes in all subjects.

LOCALIZATION OF INTERICTAL EPILEPTIC SPIKES WITH SAKETINI

As mentioned previously, once data segments can be classified as being epileptiform, it is important to determine the origin of such data in the brain, thereby identifying the epileptogenic zone. With regard to imaging, determining the epileptogenic source location and its time course while removing the measurement noise and artifacts is challenging. The SAKETINI algorithm is one algorithm that can be used to localize interictal spikes in patients [20]. One segment of data, identified as a spike marked at 400 milliseconds, as well as

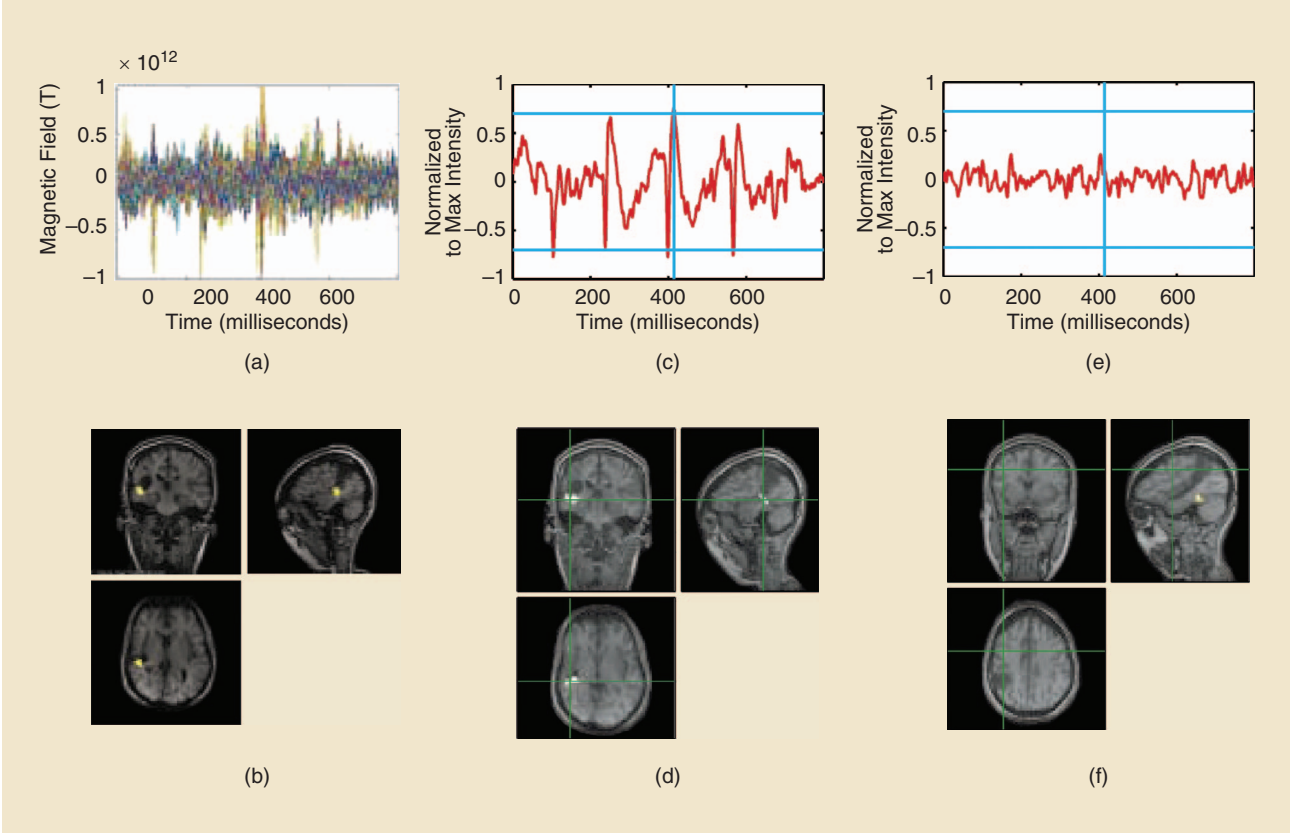


[FIG6] **The brain-state classification performance of VB-CSP. (a) Test errors (%) on three motor imagery BCI data sets (43 subsets) for VB-CSP against CSP. (b) The accuracy of VB-CSP classification of spike and nonspike segments in four epilepsy patients (S1–S4). The overall classification accuracy is well above 80% for all patients. (Figure adapted with permission from [17].)**

three additional spikes in the 800-millisecond segment, was used here as the poststimulus period, and a separate, spike-free segment of equal length was used as the prestimulus period. Figure 7 shows SAKETINI's performance estimating single spikes relative to a spike-free prestimulus period. Figure 7(a) shows the raw sensor data for the segment containing the marked spike. Figure 7(b) shows the location of the equivalent-current dipole (ECD) fit to 20 spikes from this patient. Figure 7(d) shows the SAKETINI likelihood map based on the data in Figure 7(a); the peak is in clear agreement with the standard ECD localization. Figure 7(c) shows the time course estimated for the likelihood spatial peak. The spike at 400 milliseconds is clearly visible; this cleaned waveform can be used by a clinician for analyzing the peak shape. Figure 7(e) and (f) shows a source signal's time course from a randomly selected location far from the epileptic spike source, to show the low noise level and to show the absence of cross talk onto source estimates elsewhere.

RECONSTRUCTION OF LOW SNR MEG DATA WITH SAKETINI

The SAKETINI algorithm can be used to reconstruct MEG data under a low-SNR condition [20]. In one MEG data set for localization of the primary somatosensory cortex (S1), a small diaphragm was placed on the subject's right index finger and was driven by compressed air, and the stimulus was delivered 256 times every 500 milliseconds. If we limit the available data to only a small subset of trials, the lower SNR issue became limiting for most benchmark reconstruction algorithms. We first applied SAKETINI to the average of all 256 trials to assess performance for the high-SNR case. We then applied it to the average of only the first five trials and across other sets of five-trial averages. Figure 8(a) shows typical somatosensory evoked MEG data, with the largest peak at 50 milliseconds, expected to be coming from S1 in the posterior wall of the central sulcus. Figure 8(b) shows performance of SAKETINI for the high-SNR data, which can be accurately localized to the contralateral S1. Figure 8(c) shows the sensor data averaged over the first five trials of the same data set. Therefore, SAKETINI consistently localizes S1 even in these low-SNR data, whereas the benchmark algorithms [minimum variance adaptive beamforming (MVAB) and standardized low-resolution brain electromagnetic tomography (sLORETA); results not shown—see [20] for details] perform poorly in low-SNR regimes.



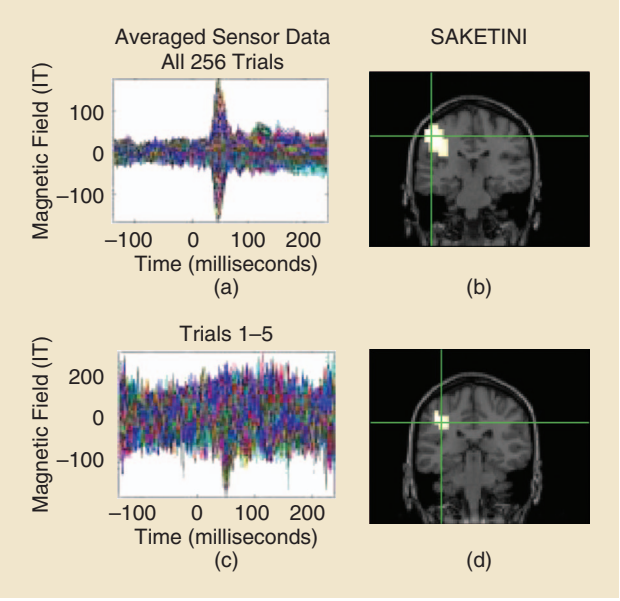
[FIG7] Localization of interictal spike MEG data with SAKETINI. (a) MEG sensor waveforms for a 2-second segment of data from an epilepsy patient with interictal spikes. (b) Three orthogonal views of the MR imaging (MRI) for this patient, and a cluster of dipoles fit by traditional dipole fitting methods. (c) The time course of a voxel showing interictal spiking activity. (d) Three orthogonal views of the MRI and the overlay of source activity estimated by SAKETINI for this data. The source localization cluster is consistent with the more traditional methods of localization. This interictal source localizes to the margins of a prior resection cavity seen in the MRI. (e) Time-course of a voxel farther away from the interictal source. The voxel does not show clear spiking activity. (f) The location of this "control" voxel on the patient's MRI. These results demonstrate the sensitivity of SAKETINI in spike localization as well as the source time course estimation. (Figure adapted with permission from [20].)

GROUP ELECTROMAGNETIC BRAIN IMAGING USING gMSP

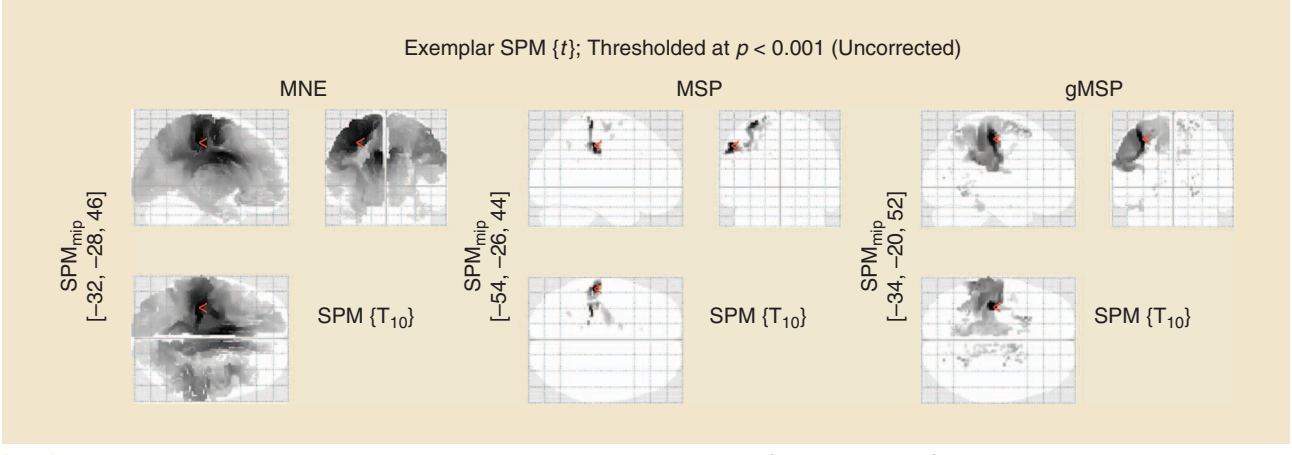
Conventional source imaging algorithms are mainly designed for EEG/MEG data from a single subject or session and are not optimal for detecting group effects. The gMSP algorithm was used to reconstruct the somatosensory evoked source activity from a median nerve-stimulation EEG data set [23]. The data set consists of recordings of eleven subjects and five session recordings per subject. The reconstruction was performed on the 10–40-millisecond poststimulus data segments from 100 groups of experiments (each with $S = 11$ subjects), which were randomly sampled from the 5^{11} possible groups. The median nerve somatosensory sources have been extensively studied in humans and are known to largely reside in the hand area of the S1. The reconstruction results of gMSP were compared with those of classical minimum norm estimation (MNE) and a multiple sparse prior (MSP) modeling approach to individual subjects without the group constraint. To assess the group-level reconstructed source activity, for each method, paired t-tests were performed between the reconstructed source activities and same images flipped across the midsagittal plane to generate statistical parametric maps (SPMs). Figure 9 shows the SPMs for one randomly selected group. All SPMs exhibit maximal activation around the left sensorimotor area. However, despite MSP's regional specificity, the t-values are generally low because of the few overlaps of sparse solutions across subjects. By contrast, large t-values are dispersed over a wide range of brain regions for MNE due to its low spatial resolution. The gMSP solution attains t-values close to those of MNE and preserves sparsity consistent across subjects due to the group constraint.

RECONSTRUCTION OF COMPLEX BRAIN SOURCE CONFIGURATIONS WITH THE CHAMPAGNE ALGORITHM

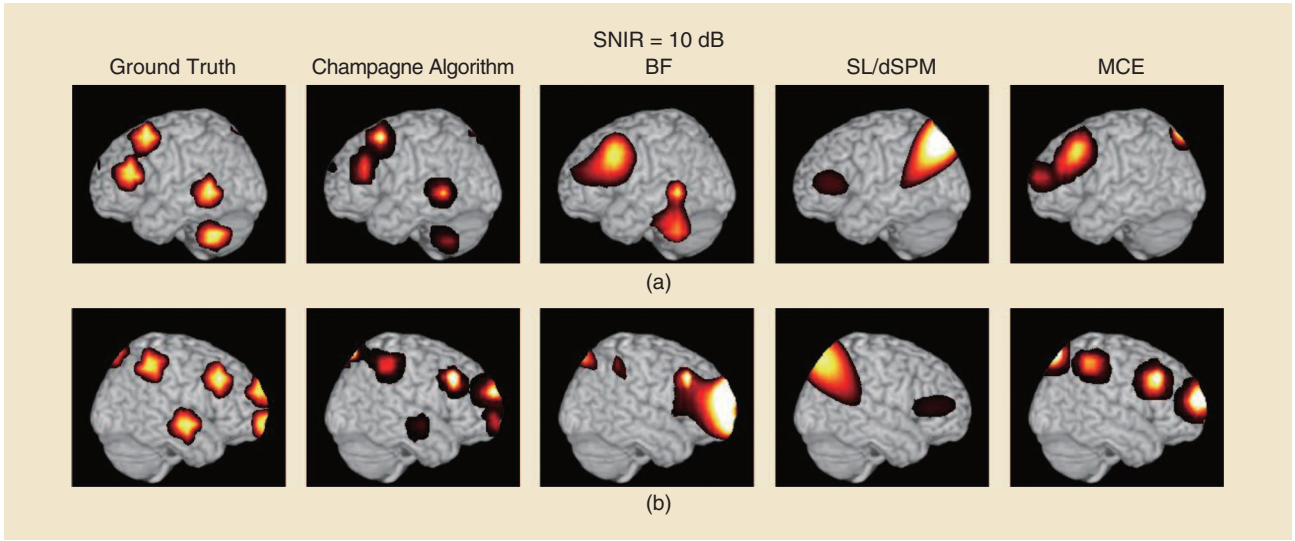
Typical brain-imaging algorithms are designed to reconstruct either isolated dipoles or extended sources with very large spatial extents. However, it is clear from functional MRI (fMRI) studies that typical source activations occur in spatially clustered sources with complex geometries of activations. Very few computational algorithms have attempted reconstructing such complex source configurations, especially from noisy sensor data corrupted by biological and nonbiological artifacts. Here, we apply the Champagne algorithm to localize complex distributed activity from simulated clusters of brain sources. Ground truth from one such simulation consists of ten clusters, with ten dipolar sources within each cluster. The placement of the cluster center is random, and the clusters consist of sources seeded in the nine nearest neighboring voxels. The source time courses within each cluster have an interdipole correlation coefficient of 0.8 and an intradipole correlation coefficient of 0.25. The source simulation is constructed with realistic



[FIG8] Localization performance of SAKETINI on high- and low-SNR MEG data. (a) High-SNR sensor waveforms for tactile stimulation obtained by averaging 256 trials. (b) Localization of the waveforms in (a) shows a clear peak in S1. (c) Low-SNR sensor waveforms for tactile stimulation obtained by averaging only five trials. (d) Localization of low-SNR data is robust in showing S1 activity. (Figure adapted with permission from [20].)



[FIG9] SPMs comparing the reconstructed source activities with the midsagittal-flipped version of the same images. Results are overlaid on a glass brain. The location of the maximum t-value in each SPM is marked by a red pointer. (Figure adapted with permission from [23]).

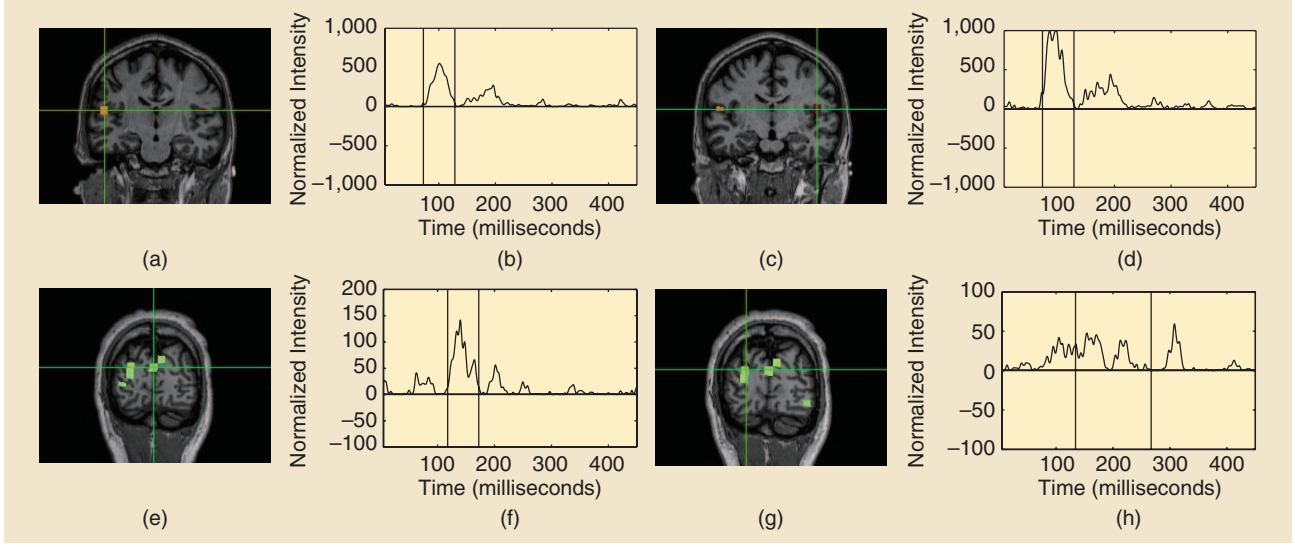


[FIG10] **The performance of the Champagne algorithm in computer simulations. The ground truth has activity in ten dipole clusters spreading throughout the brain. Reconstruction comparison using the Champagne algorithm, beamforming (BF), sLORETA (SL), dynamic statistical parameter mapping (dSPM), and MCE methods. (Figure adapted with permission from [41].)**

brain noise and 10 dB signal-to-noise-plus-interference-ratio (SNIR). Reconstruction results from this simulation with clusters are shown in Figure 10. In this example, the result derived from the Champagne algorithm is superior to other benchmark algorithms, such as MVAB, generalized minimum current estimation (MCE), and minimum norm methods (sLORETA/dSPM) [41], [31]. It can be seen that the standard benchmark algorithms are not able to reconstruct such complex brain source activity configurations. More details of benchmark algorithms and Champagne algorithm performance can be found in [41] and [31].

RECONSTRUCTION OF AUDIOVISUAL NETWORK ACTIVITY FROM MEG DATA WITH CHAMPAGNE

One example of a complex source activity configuration is the network of brain regions recruited during passive viewing of simultaneously presented audiovisual stimuli. This network consists of partially overlapping and partially segregated spatial and temporal activity in auditory and visual cortices as well as brain regions involved in integration of audiovisual information. The Champagne algorithm was used to reconstruct this audiovisual network activity from MEG data to examine the integration of auditory and visual



[FIG11] **Results of audiovisual data localization from the Champagne algorithm. The Champagne algorithm localizes a bilateral auditory response at 100 milliseconds after the simultaneous presentation of tones and a visual stimulus. (a) and (c) Localized bilateral auditory activity from the Champagne algorithm, with time courses shown in (b) and (d). The Champagne algorithm localizes an early visual response at 150 milliseconds after the simultaneous presentation of tones and a visual stimulus. The time course in (f) corresponds to the location indicated by the crosshairs in (e) the coronal sections. The Champagne algorithm localizes a later visual response later than 150 milliseconds after simultaneous presentation of tones and a visual stimulus. The time course in (h) corresponds to the location indicated by the crosshairs in (g) the coronal sections. (Figure adapted with permission from [41].)**

information [41]. In one MEG data set, a healthy control participant was presented single 35-millisecond-duration tones (1 kHz) simultaneously with a visual stimulus. The visual stimulus consisted of a white cross at the center of a black monitor screen. The data were averaged across 100 trials (after the trials were time-aligned to the stimulus). The prestimulus window was selected to be –100 to 5 milliseconds, and the poststimulus time window was selected to be 5–450 milliseconds, where zero marks the onset of the simultaneous auditory and visual stimulation. Champagne's results are presented in Figure 11. In the first and second rows, the brain activations associated with the auditory stimulus are shown. The Champagne algorithm is able to localize bilateral auditory activity in Heschel's gyrus in the window around the M100 peak. The time courses for the left and right auditory sources are shown in Figure 11(b) and (d), respectively, along with the window used around the M100 peak. The two auditory sources had the maximum power in the window around the M100 peak. The Champagne algorithm is also able to localize a visual source in medial, occipital gyrus with a peak around 150 milliseconds. The power in the window around this peak is shown in Figure 11(e), and the time course of the source marked with the crosshairs is shown in Figure 11(f).

RECONSTRUCTION OF FACE PROCESSING NETWORK ACTIVITY FROM EEG DATA WITH THE CHAMPAGNE ALGORITHM

Another example of overlapping spatial and temporal activity in a variety of brain regions occurs during the early brain response to face visual stimuli. The Champagne algorithm was used to reconstruct the face-processing network activity from EEG data. In Figure 12, we present the results from using the Champagne algorithm on a face-processing EEG data set. The 126-channel EEG data was downloaded from a public website (http://www.fil.ion.ucl.ac.uk/spm/data/mmfaces). A three-orientation component lead-field matrix was calculated in SPM8 using the coarse resolution. The EEG data paradigm involved randomized presentation of at least 86 faces, and the average across trials time-aligned to the presentation of the face was used for source reconstruction. The prestimulus window was selected to be –200 to 5 milliseconds, and the poststimulus time window was selected to be 5–250 milliseconds. Reconstructed power was plotted on a three-dimensional brain, and the time courses for the peak voxels are plotted (the arrows point from a particular voxel to its time course). In Figure 12, it is shown that the Champagne algorithm is able to localize early visual areas that have a peak around 100 milliseconds as well as activations in and around fusiform gyrus that peaks around 170 milliseconds, corresponding to the N170 seen in the sensor data. These results are consistent with those obtained in [42] using the same EEG data set but show greater spatial resolution than previously published methods.

DISCUSSION AND CONCLUSIONS

EEG and MEG signals offer a unique and noninvasive way to record brain activity at a fine temporal resolution, providing many opportunities to address important neuroscience questions. However, EEG/MEG analysis is also confronted with a multitude of challenges, such as high dimensionality, nonstationarity, and very large sample size (e.g., continuous EEG recordings) or very sparse sample size (e.g., single trial analysis, disease diagnosis). BML is an emerging research area that integrates Bayesian inference, optimization, Monte Carlo sampling, and machine-learning techniques for data analysis. The Bayesian framework enables us to capture various sources of uncertainties in the data or parameters. By taking into account (hierarchical) priors, Bayesian inference provides an optimal estimate of the model or model parameters. In addition, hierarchical modeling brings the advantages in extra flexibility and the insensitivity to priors.

Although this article focuses on brain-state classification and electromagnetic brain imaging as the two main applications, BML methods have also been successfully developed in some other EEG/MEG signal processing applications, e.g., functional brain connectivity analysis [43], multimodal brain data fusion [42], and data compression [44]. The topics introduced in this article are theoretical underpinnings of these methods and provide the key to understanding their mechanisms. Table 3 lists some useful resources (especially on software) related to the topics reviewed in this article.

CHALLENGES AND FUTURE RESEARCH DIRECTIONS

ELECTROMAGNETIC BRAIN IMAGING FOR SPATIALLY EXTENDED SOURCES

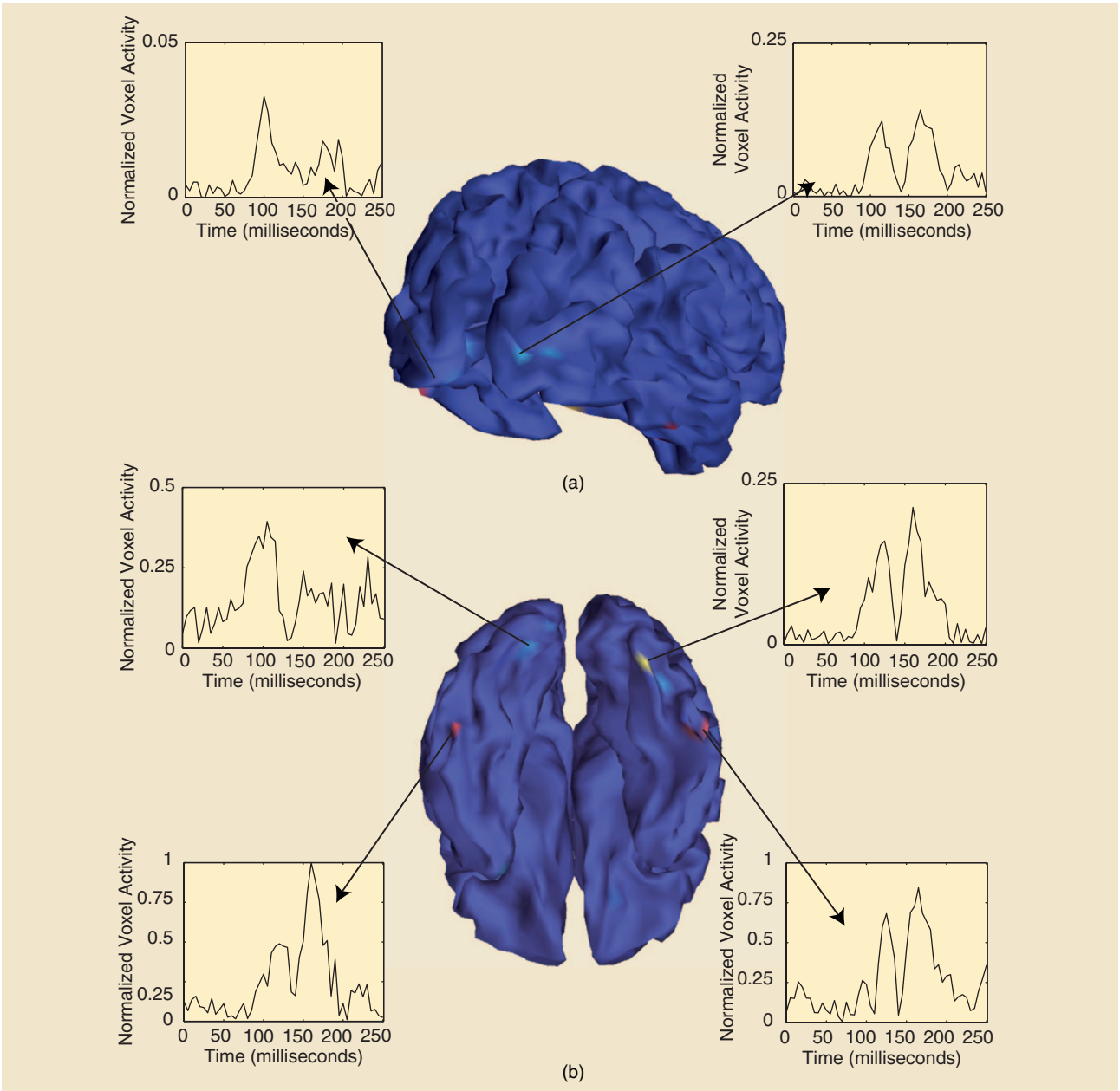
Accurate estimation of locations and spatial extents of brain sources remains a challenge for EEG/MEG source imaging. Conventional approaches yield source estimates that are either too sparse or blurry. Proper determination of the source extents requires modeling the functional interactions between sources within local patches [3]. Moreover, the majority of existing source imaging methods assume temporal independence (i.e., the source estimation is applied to each time point separately), which ignores the temporal dynamics structure clearly present in many EEG/MEG measurements. Naturally, the information afforded by this structure could be employed to further regularize the solution space and lead to performance improvement. A promising future research direction is to develop new Bayesian methods that integrate spatial and temporal modeling at the local patch level and leverage SBL techniques to automatically infer the patch size.

JOINT IMAGING OF ACTIVITY AND FUNCTIONAL CONNECTIVITY

Current approaches to functional connectivity imaging with EEG/MEG data comprise first performing imaging estimates and, subsequently, conducting functional connectivity analyses. In principle, these two steps could be integrated within a single framework. However, few efforts have been undertaken to integrate such estimates for functional connectivity due to the large number of parameters and model complexity.

LEARNING MODEL STRUCTURE

The essence of Bayesian modeling is to capture the inherent data structure. In a probabilistic graphical model, the inference task is



[FIG12] **Results for EEG face processing data from the Champagne algorithm: (a) two early visual responses in the occipital cortex with the time courses and (b) four ventral activations in (or near) the face fusiform area with time courses showing peaks around 170 milliseconds [ventral side of brain shown in (b), with the right hemisphere on the right]. (Figure adapted with permission from [41].)**

[TABLE 3] RESOURCES RELATED TO BAYESIAN INFERENCE AND RELATED TOPICS.

TOPIC	WEB RESOURCES
VB METHODS	HTTP://WWW.VARIATIONAL-BAYES.ORG
GIBBS SAMPLING (BUGS)	HTTP://WWW.MRC-BSU.CAM.AC.UK/SOFTWARE/BUGS/
SPARSE LEARNING	HTTP://WWW.MIKETIPPING.COM/DOWNLOADS.HTM
	HTTP://SPAMS-DEVEL.GFORGE.INRIA.FR
	HTTP://WWW.YELAB.NET/SOFTWARE/SLEP
NB (GP)	HTTP://WWW.GAUSSIANPROCESS.ORG
NB (DP)	HTTPS://CRAN.R-PROJECT.ORG/PACKAGE=DPPACKAGE
DEEP LEARNING	HTTP://DEEPLEARNING.NET/SOFTWARE_LINKS/

to identify the parameters or hyperparmeters specified by the model. In EEG/MEG data analysis, model specification is often empirically determined by researchers, according to certain hypotheses. This is not viable in exploratory analyses where the knowledge regarding the model structure is weak or even unavailable (e.g., in the functional connectivity analysis where the involved brain regions are unknown a priori). How to infer the graphical model structure remains an active research topic [45], [46]. For other observed input variables, it is also important to identify the causal or statistical dependency (including temporal delay) between the input and observed EEG/MEG signals.

DEVELOPMENT OF EFFICIENT INFERENCE AND OPTIMIZATION ENGINES

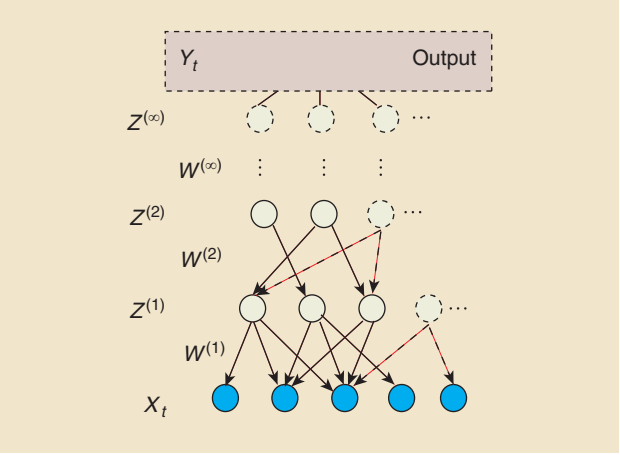
It is our belief that BML will play an important role in modern EEG/MEG signal processing applications. In comparison to the standard signal processing or likelihood-based inference, BML comes at the cost of increasing computational complexity, which is particularly problematic for EEG/MEG brain imaging in the high-resolution source space. Therefore, development of approximate, computationally efficient Bayesian inference algorithms (including parallel or distributed modules) remains an active task in the research field.

One direction is to exploit the data structure or to impose constraints (e.g., sparsity in space or time) in the model and restrict the parameter search space, in which compressed sensing idea can be further explored [47]. Another interesting direction is to employ the so-called Bayesian optimization technique, which aims to automatically tune the parameters and hyperparameters in HB models.

BAYESIAN DEEP LEARNING METHODS

Nonlinear features have been investigated and proven useful in EEG/MEG applications such as BCIs and epileptic seizure prediction. Commonly, spatiotemporal or spectral-temporal features are manually selected, which are mostly based on second-order statistics (e.g., power spectra, eigenvectors of covariance). To overcome these limitations, it is desirable to automatically extract nonlinear or high-order features that are tailored to the properties of signals of interest [48]. Deep learning is an emerging feature learning method by composing multiple nonlinear transformations of the data to produce more abstract yet potentially more useful representations [49]. It typically combines bottom-up (unsupervised, step one) inference and top-down (supervised, step two) learning procedures. Although deep learning was primarily developed for discriminative learning (which is in contrast to generative models and probabilistic inference), Bayesian analysis can be integrated with deep learning to extract optimal features.

Recently, deep learning has been demonstrated with a superior performance in an EEG-based BCI application [48]. However, to the best of our knowledge, Bayesian deep learning has not been explored in EEG/MEG applications. As deep learning often involves a multilayer network architecture (Figure 13), identifying the optimal network architecture is an important task. With the optimal or suboptimal network architecture inferred from



[FIG13] An infinite generative model for Bayesian deep learning. The shaded nodes (X_t) represent the observed random variables, the open nodes represent the latent variables ($Z^{(1)}, Z^{(2)}, \ldots, Z^{(\infty)}$), and $W^{(1)}, W^{(2)}, \ldots, W^{(\infty)}$ denote the connection weights between layers. The output layer consists of one node (Y_t) for logistic regression. (Figure adapted with permission from [50].)

Bayesian analysis, Bayesian deep learning will have great potential for EEG/MEG feature extraction.

From a Bayesian modeling viewpoint, the number of hidden layers and the number of hidden factors (i.e., latent variables) for each layer need to be determined, and both of them can be potentially infinite [49]. Each hidden factor in hidden layers may correspond to certain higher-order features extracted from the EEG time series. Hierarchical NB model provides a principled approach to this solution. In [50], the authors proposed a generative deep network architecture and imposed an infinity structure both latently and hierarchically. At the stage of unsupervised learning, Gibbs sampling can be used to infer the hidden factors and connection weights between layers. At the stage of supervised learning, supervisory signals y_t (e.g., target labels) can be fed to the output layer, followed by a discriminative learning procedure [49].

In conclusion, the BML framework provides an integrated framework for EEG/MEG data analysis. Despite many signal processing challenges, more advanced BML algorithmic development and successful EEG/MEG applications are anticipated in the forthcoming years.

ACKNOWLEDGMENTS

We would like to thank the following former and current colleagues for their contributions, cited in this article: Hagai Attias, Kensuke Sekihara, David Wipf, Julia Owen, Sarang Dalal, and Johanna Zumer. We also thank the cited publishers for permission to use previously published figures. We are grateful to Karl Friston, Vladimir Litvak, and Jose David López for helpful discussions. Wei Wu acknowledges support from the 973 Program of China (2015CB351703), the 863 Program of China (2012AA011601), the National Natural Science Foundation of China (61403114), the Guangdong Natural Science Foundation (2014A030312005 and S2013010013445), and the Steven and Alexandra Cohen Foundation. Srikantan Nagarajan received support from the U.S. National Institutes of Health (R01DC010145 and R01DC013979), the U.S.

National Science Foundation (NSF) (BCS 1262297), and the U.S. Department of Defense (W81XWH-13-1-0494). Zhe Chen was supported by a Collaborative Research in Computational Neuroscience Award (1307645) from the U.S. NSF.

AUTHORS

Wei Wu (wwumed@stanford.edu) is with the School of Automation Science and Engineering, South China University of Technology, Guangzhou, and the Department of Psychiatry and Behavioral Sciences, Stanford University, California. He is a Member of the IEEE.

Srikantan Nagarajan (sri@ucsf.edu) is with the Department of Radiology and Biomedical Imaging, University of California at San Francisco. He is a Senior Member of the IEEE.

Zhe Chen (zhe.chen3@nyumc.org) is with the Department of Psychiatry, Neuroscience, and Physiology, School of Medicine, New York University. He is a Senior Member of the IEEE.

REFERENCES

[1] S. Baillet, J. C. Mosher, and R. M. Leahy, "Electromagnetic brain mapping," *IEEE Signal Processing. Mag*., vol. 18, no. 6, pp. 14–30, 2001.

[2] S. Sanei and J. A. Chambers, *EEG Signal Processing*. Hoboken, NJ: Wiley, 2007.

[3] C. Lamus, M. S. Hämäläinen, S. Temereanca, E. N. Brown, and P. L. Purdon, "A spatiotemporal dynamic distributed solution to the MEG inverse problem," *Neuroimage*, vol. 63, no. 2, pp. 894–909, 2012.

[4] K. Friston, L. Harrison, J. Daunizeau, S. Kiebel, C. Phillips, N. Trujillo-Barreto, R. Henson, G. Flandin et al., "Multiple sparse priors for the M/EEG inverse problem," *Neuroimage*, vol. 39, no. 3, pp. 1104–1120, 2008.

[5] A. Bolstad, B. V. Veen, and R. Nowak, "Space-time event sparse penalization for magneto-/electro-encephalography," *Neuroimage*, vol. 46, no. 4, pp. 1066–1081, 2009.

[6] K. J. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, and J. Ashburner, "Classical and Bayesian inference in neuroimaging: Theory," *Neuroimage*, vol. 16, no. 2, pp. 465–483, 2002.

[7] J. Mattout, C. Phillips, W. D. Penny, M. D. Rugg, and K. J. Friston, "MEG source localization under multiple constraints: an extended Bayesian framework," *Neuroimage*, vol. 30, no. 3, pp. 753–767, 2006.

[8] M. W. Seeger and D. P. Wipf, "Variational Bayesian inference techniques," *IEEE Signal Processing. Mag*., vol. 27, no. 6, pp. 81–91, 2010.

[9] D. P. Wipf, B. D. Rao, and S. Nagarajan, "Latent variable Bayesian models for promoting sparsity," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 6236–6255, 2011.

[10] D. P. Wipf and S. Nagarajan, "A unified Bayesian framework for MEG/EEG source imaging," *Neuroimage*, vol. 44, no. 3, pp. 947–966, 2009.

[11] C. P. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, 2nd ed. New York: Springer-Verlag, 2007.

[12] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall/CRC, 1995.

[13] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 6, no. 6, pp. 721–741, 1984.

[14] T. Marshall and G. Roberts, "An adaptive approach to Langevin MCMC," *Stat. Comput*., vol. 22, no. 5, pp. 1041–1057, 2012.

[15] H. Attias, "Independent factor analysis," *Neural Comput*., vol. 11, no. 4, pp. 803–851, 1999.

[16] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K. R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Processing. Mag*., vol. 25, no. 1, pp. 41–56, 2008.

[17] W. Wu, Z. Chen, X. Gao, Y. Li, E. Brown, and S. Gao, "Probabilistic common spatial patterns for multichannel EEG analysis," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 37, no. 3, pp. 639–653, 2015.

[18] H. Kang and S. Choi, "Bayesian common spatial patterns with Dirichlet process priors for multi-subject EEG classification," in *Proc. Int. Joint Conf. Neural Networks (IJCNN),* 2012, pp. 1–6.

[19] W. Wu, Z. Chen, S. Gao, and E. N. Brown, "A hierarchical Bayesian approach for learning sparse spatio-temporal decompositions of multichannel EEG," *Neuroimage*, vol. 56, no. 4, pp. 1929–1945, 2011.

[20] J. M. Zumer, H. T. Attias, K. Sekihara, and S. S. Nagarajan, "A probabilistic algorithm integrating source localization and noise suppression for MEG and EEG data," *Neuroimage*, vol. 37, no. 1, pp. 102–115, 2007.

[21] J. Sarvas, "Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem," *Phys. Med. Biol*., vol. 32, no. 1, pp. 11–22, 1987.

[22] S. S. Nagarajan, H. T. Attias, K. E. H. II, and K. Sekihara, "A graphical model for estimating stimulus-evoked brain responses from magnetoencephalography data with large background brain activity," *Neuroimage*, vol. 30, no. 2, pp. 400–416, 2006.

[23] V. Litvak and K. Friston, "Electromagnetic source reconstruction for group studies," *Neuroimage*, vol. 42, no. 4, pp. 1490–1498, 2008.

[24] J. D. López, V. Litvak, J. J. Espinosa, K. Friston, and G. R. Barnes, "Algorithmic procedures for Bayesian MEG/EEG source reconstruction in SPM," *Neuroimage*, vol. 84, pp. 476–487, Jan., 2014.

[25] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing. Mag*., vol. 25, no. 2, pp. 21–30, 2008.

[26] Y. Li, Z. Yu, N. Bi, Y. Xu, Z. Gu, and S. I. Amari, "Sparse representation for brain signal processing: A tutorial on methods and applications," *IEEE Signal Processing. Mag*., vol. 31, no. 3, pp. 96–106, 2014.

[27] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Processing*, vol. 56, no. 6, pp. 2346–2356, 2008.

[28] D. P. Wipf and S. Nagarajan, "A new view of automatic relevance determination," in *Proc. Advances in Neural Information Processing Systems*, 2008, vol. 20, pp. 1625–1632.

[29] W. Wu, C. Wu, S. Gao, B. Liu, Y. Li, and X. Gao, "Bayesian estimation of ERP components from multicondition and multichannel EEG," *Neuroimage*, vol. 88, pp. 319–339, Mar., 2014.

[30] S. S. Nagarajan, H. T. Attias, K. E. Hild, and K. Sekihara, "A probabilistic algorithm for robust interference suppression in bioelectromagnetic sensor data," *Stat. Med*., vol. 26, no. 21, pp. 3886–910, 2007.

[31] D. P. Wipf, J. P. Owen, H. T. Attias, K. Sekihara, and S. S. Nagarajan, "Robust Bayesian estimation of the location, orientation, and time course of multiple correlated neural sources using MEG," *Neuroimage*, vol. 49, no. 1, pp. 641–655, 2010.

[32] Z. Ghahramani, "Bayesian non-parametrics and the probabilistic approach to modelling," *Philos. Trans. R. Soc. A*, vol. 371, no. 1984, p. 20110553, 2013.

[33] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2005.

[34] N. L. Hjort, C. Holmes, P. Mülcer, and S. G. Walker, Eds., *Bayesian Nonparametrics*. Cambridge, UK: Cambridge Univ. Press, 2010.

[35] C. J. Paciorek and M. J. Schervish, "Nonstationary covariance functions for Gaussian process regression," in *Proc. Advances in Neural Information Processing Systems,* 2004, vol. 16, pp. 273–280.

[36] S. Faul, G. Gregorcic, G. Boylan, W. Marnane, G. Lightbody, and S. Connolly, "Gaussian process modeling of EEG for the detection of neonatal seizure," *IEEE Trans. Biomed. Eng*., vol. 54, no. 12, pp. 2151–2162, 2007.

[37] M. Zhong, F. Lotte, M. Girolami, and A. Lecuyer, "Classifying EEG for brain computer interfaces using Gaussian process," *Pattern Recognit. Lett*., vol. 29, no. 3, pp. 354–359, 2008.

[38] A. Fyshe, E. Fox, D. Dunson, and T. Mitchell, "Hierarchical latent dictionaries for models of brain activation," in *Proc. 15th Int. Conf. Artificial Intelligence and Statistics (AISTATS),* 2012, pp. 409–421.

[39] H. Kang and S. Choi, "Bayesian common spatial patterns for multi-subject EEG classification," *Neural Netw*., vol. 57, pp. 39–50, Sept. 2014.

[40] A. Ossadtchi, S. Baillet, J. C. Mosher, D. Thyerlei, W. Sutherling, and R. M. Leahy, "Automated interictal spike detection and source localization in magnetoencephalography using independent components analysis and spatio-temporal clustering," *Clin. Neurophysiol*., vol. 115, no. 3, pp. 508–522, 2004.

[41] J. P. Owen, D. P. Wipf, H. T. Attias, K. Sekihara, and S. S. Nagarajan, "Performance evaluation of the CHAMPAGNE source reconstruction algorithm on simulated and real M/EEG data," *Neuroimage*, vol. 60, no. 1, pp. 305–323, 2012.

[42] R. N. Henson, G. Flandin, K. J. Friston, and J. Mattout, "A parametric empirical Bayesian framework for fMRI-constrained MEG/EEG source reconstruction," *Hum. Brain Mapp*., vol. 31, no. 10, pp. 1512–1531, 2010.

[43] S. J. Kiebel, M. I. Garrido, R. Moran, C. Chen, and K. J. Friston, "Dynamic causal modeling for EEG and MEG," *Hum. Brain Mapp*., vol. 30, no. 6, pp. 1866–1876, 2009.

[44] Z. Zhang, T. Jung, S. Makeig, and B. D. Rao, "Compressed sensing of EEG for wireless telemonitoring with low energy consumption and inexpensive hardware," *IEEE Trans. Biomed. Eng*., vol. 60, no. 1, pp. 221–224, 2013.

[45] D. Heckerman, "A tutorial on learning with Bayesian networks," in *Learning in Graphical Models*, M. I. Jordan, Ed. Cambridge, MA: MIT Press, 1999, pp. 301–354.

[46] J. Lee and T. Hastie, "Structure learning of mixed graphical models," in *Proc. 16th Int. Conf. Artificial Intelligence and Statistics*, 2013, pp. 388–396.

[47] B. Babdi, G. Obregon-Henao, C. Lamus, M. S. Hämäläinen, E. N. Brown, and P. L. Purdon, "A subspace pursuit-based iterative greedy hierarchical solution to the neuromagnetic inverse problem." *Neuroimage*, vol. 87, pp. 427–443, Feb., 2014.

[48] H. Cecotti and A. Graser, "Convolutional neural networks for P300 detection with application to brain-computer interfaces," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 33, no. 3, pp. 433–445, 2011.

[49] L. Deng and D. Yu, "Deep learning: methods and applications," *Found. Trends Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[50] E. Pan and Z. Han. (2014). Non-parametric Bayesian learning with deep learning structure and its application in wireless network. [Online]. Available: https://arxiv.org/pdf/1410.4599v2.pdf

[SP]