Probabilistic Structure Learning for EEG/MEG Source Imaging With Hierarchical Graph Priors

Feng Liu, Li Wang, Yifei Lou, *Member, IEEE*, Ren-Cang Li, and Patrick L. Purdon, *Member, IEEE*

***Abstract*— Brain source imaging is an important method for noninvasively characterizing brain activity using Electroencephalogram (EEG) or Magnetoencephalography (MEG) recordings. Traditional EEG/MEG Source Imaging (ESI) methods usually assume the source activities at different time points are unrelated, and do not utilize the temporal structure in the source activation, making the ESI analysis sensitive to noise. Some methods may encourage very similar activation patterns across the entire time course and may be incapable of accounting the variation along the time course. To effectively deal with noise while maintaining flexibility and continuity among brain activation patterns, we propose a novel probabilistic ESI model based on a hierarchical graph prior. Under our method, a spanning tree constraint ensures that activity patterns have spatiotemporal continuity. An efficient algorithm based on an alternating convex search is presented to solve the resulting problem of the proposed model with guaranteed convergence. Comprehensive numerical studies using synthetic data on a realistic brain model are conducted under different levels of signal-to-noise ratio (SNR) from both sensor and source spaces. We also examine the EEG/MEG datasets in two real applications, in which our ESI reconstructions are neurologically plausible. All the results demonstrate significant improvements of the proposed method over benchmark methods in terms of source localization performance, especially at high noise levels.**

***Index Terms*— EEG/MEG source imaging, source localization, inverse problem, graph structure learning.**

Manuscript received July 12, 2020; revised September 11, 2020; accepted September 15, 2020. Date of publication September 21, 2020; date of current version December 29, 2020. The work of Li Wang was supported by the NSF under Grant DMS-2009689. The work of Yifei Lou was supported in part by the NSF under Award DMS-1522786 and Award CAREER 1846690. The work of Ren-Cang Li was supported in part by the NSF under Grant CCF-1527104, Grant DMS-1719620, and Grant DMS-2009689. The work of Patrick L. Purdon was supported in part by the NIH under Grant R01AG054081 and Grant R01AG056015 and in part by the Tiny Blue Dot Foundation. *(Feng Liu and Li Wang contributed equally to this work.) (Corresponding authors: Li Wang; Patrick L. Purdon.)*

Feng Liu is with the School of Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ 07030 USA (e-mail: fliu22@stevens.edu).

Li Wang and Ren-Cang Li are with the Department of Mathematics, The University of Texas at Arlington, Arlington, TX 76013 USA (e-mail: li.wang@uta.edu; rcli@uta.edu).

Yifei Lou is with the Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: yifei.lou@utdallas.edu).

Patrick L. Purdon is with the Department of Anesthesia, Critical Care, and Pain Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02119 USA, and also with The Picower Institute for Learning and Memory, MIT, Cambridge, MA 02139 USA (e-mail: patrick.purdon@mgh.harvard.edu).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2020.3025608

I. Introduction

ACTIVITY in the human brain is composed of complex firing patterns and interactions among neurons and neuronal circuits. This brain activity generates electromagnetic, hemodynamic, and metabolic signals that can be measured non-invasively [1]. Electromagnetic fields generated by post-synaptic neuronal activity can be measured using the Electroencephalogram (EEG) and Magnetoencephalogram (MEG) [2]. In contrast to brain imaging modalities such as functional magnetic resonance imaging (fMRI), positron emission tomography (PET), and single-photon emission computed tomography (SPECT), EEG and MEG are unique in their ability to measure rapid, millisecond-level brain dynamics [1], [3]–[5]. However, in EEG/MEG, an ill-conditioned inverse problem must be solved in order to localize the cerebral sources underlying the observed electromagnetic fields, referred to as the EEG/MEG source localization or EEG/MEG source imaging (ESI) problem [6]. ESI techniques have been used in many applications such as the study of language, cognition, sensory function, pre-surgical planning, localization of focal drug-resistant epilepsy for drug-resistance patients [7]–[10], and brain-computer interfaces [11]. ESI has also been used as a tool to characterize brain networks in neuroscience studies and to discover pathological network biomarkers for clinical applications [12]–[17].

The number of EEG/MEG sensors is far fewer than the number of candidate brain sources. Hence the ESI problem is highly ill-posed. Neurophysiologically plausible assumptions or priors can be incorporated into ESI algorithms as regularizations [6], [18]. In some event-related experimental designs, it has been argued that only a small fraction of the brain may be consistently activated [19], implying that ESI solutions for such experiments could be sparse. Accordingly, sparsity constraints could significantly improve the spatial resolution of ESI solutions [3], [20], [21]. In addition, brain activity has temporal structure that can be exploited to improve ESI performance [22].

One other area for potential improvement relates to how uncertainty or noise is represented in statistical models for

ESI. Uncertainty or noise can be introduced at the sensor level and also at the space of the source generators. Reconstructed source solutions often include spurious sources that explain the measured noise. In principle, better models of this uncertainty could improve ESI solutions by making distinct the contributions of measurement noise, fluctuations in underlying currents, or variation among latent spatial patterns of activity to the overall observed uncertainty. For instance, Liu *et al.* developed a model in which source activations under the same brain state or stimulus condition were constrained to be similar [23]. However, this assumption could be made less conservative to allow fluctuations in the source signals.

Here we describe a novel ESI method that employs a sophisticated hierarchical prior to describe underlying source activity patterns. This hierarchical prior uses soft-clustering centers, termed "landmarks," to characterize underlying spatial activation patterns. These landmarks have dependencies on one another in terms of a graph structure. The combination of the landmarks and the graph structure forms a manifold that can regularize ESI solutions very effectively even in the presence of significant noise at both the sensor and source levels. Both the landmarks and the graph structure are learned from the observed data.

The main contributions of this paper are:

(i) We formulate a novel probabilistic model for ESI with a hierarchical graph prior, which naturally integrates the landmark learning and the graph structure learning into the inverse problem of ESI (Section III);

(ii) An efficient optimization method is proposed to solve the resulting problem of the proposed model. Its convergence is provably guaranteed (Section IV);

(iii) We perform extensive experiments to examine the performance of the proposed method and to compare it with benchmark methods; Numerical experiments show that our proposed method performs particularly well when the measurement noise is high (Section V).

II. Related Work

There are two main categories of source imaging algorithms: dipole fitting algorithms and distributed source imaging algorithms [1], [24]. Dipole fitting algorithms are the earliest approaches in ESI which empirically solve the MEG/EEG forward and inverse problems by characterizing one or several equivalent current dipoles (ECD) responsible for the electrical potential detected by the scalp sensors. The algorithms typically allow free orientation or varied dipole locations across moments. However, localizing multiple dipoles can be challenging as no prior information is known with regards to the number of dipoles [25].

Recently, numerous algorithms in the category of distributed inverse solvers have been developed. Under this paradigm, a class of beamforming and scanning methods was proposed. Specifically, the linearly constrained minimum variance (LCMV) beamformer is a type of adaptive spatial filter that localizes sources of interest by minimizing the contributions of other uncorrelated sources [26]. Among the scanning strategies, Mosher *et al.* proposed a special case of MUltiple SIgnal Classification (MUSIC) algorithm, where multiple dipoles can be found by scanning potential locations using one dipole model [27]. Mäkelä *et al.* introduced truncated recursively-applied-and-projected MUSIC (TRAP-MUSIC), which builds on the RAP-MUSIC algorithm [28] by applying a sequential dimension reduction to the signal-subspace projection.

Except for the beamforming and scanning methods, under the same category of distributed inverse solvers, many algorithms employ neurophysiologically plausible priors or regularization. As the ESI is an ill-posed problem, certain priori information or regularization about the desired source characteristics or physiological assumptions needs to be incorporated, such as the ℓ_2 norm or the ℓ_p norm regularizers with $p \leq 1$. One seminal approach based on the ℓ_2 norm is the minimum norm estimate (MNE) [29]. Dynamic statistical parametric mapping (dSPM) [30] and standardized low-resolution brain electromagnetic tomography (sLORETA) [31] are two variants of MNE. The ℓ_2-norm based methods tend to produce spatially-diffuse solutions. To overcome this drawback, Uutela *et al.* [32] introduced the minimum current estimate (MCE) using the ℓ_1 norm for sparse source reconstruction. Gorodnitsky *et al.* proposed FOCUSS which is based on recursive, weighted norm minimization and can provide high resolution solutions [33]. To solve the ESI problem with the sparsity inducing ℓ_p norm with $p \leq 1$, numerous algorithms were also proposed. Specifically, Rao and Kreutz-Delgado proposed an affine scaling method [34]. Bore *et al.* proposed to use the ℓ_p-norm regularization ($p < 1$) on the source signal and the ℓ_1 norm on the data fitting error term [35]. An extension of greedy subspace-pursuit algorithms [3] was further developed to jointly identify distributed cortical and subcortical sources [36]. Many approaches treat the source activity as being independent in time. However, neural activity can be highly structured in time and can exhibit different degrees of temporal smoothness depending on the context.

In order to account for this temporal smoothness, a number of methods were developed. The mixed norm estimate (MxNE) [37] uses an $\ell_{p,2}$-norm regularization ($p \leq 1$) to impose ℓ_2 on the temporal direction and ℓ_p ($p \leq 1$) in the source space for temporal smoothness and spatial sparsity. An iteratively reweighted optimization method (irMxNE) is used to obtain more precise and stable MxNE solutions [38]. Huang *et al.* proposed a vector-based spatial–temporal analysis using an ℓ_1-minimum-norm (VESTAL) to address the "spiky-looking" discontinuity in source space [39]. The spatiotemporal unifying tomography (STOUT) algorithm [19] combines Time-Frequency (TF)-MxNE [40] in the time-frequency domain and sparse spatial basis field expansions to obtain temporally and spatially smooth solutions.

Other recent approaches are built on more sophisticated probabilistic models of source activity. Cai *et al.* proposed a hierarchical multiscale Bayesian algorithm called tree-Champagne which enables robust reconstruction of sources with different spatial extents using a probabilistic generative model [41]. The same group later proposed a robust empirical Bayesian algorithm for better reconstruction of distributed brain source activity using kernel smoothing and hyperparameter tiling [42]. Pirondini *et al.* [22] modeled source activity as a spatiotemporal dynamic system and used the Kalman

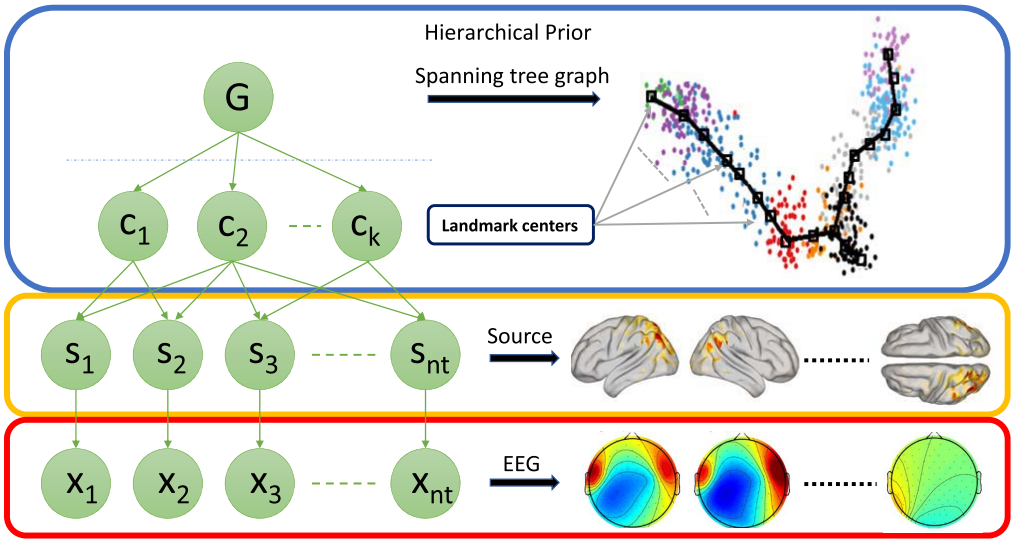

Fig. 1. Illustration of our hierarchical graphical model in the ESI problem. The left part of this figure shows how the variables are dependent of each other in a hierarchical way. The right part of the figure gives an exemplary illustration of what the left components represent in ESI. *G* (marked with black lines) is a spanning tree that is considered as a graph prior to connect landmark centers *C* (marked with black rectangles). *S* is dependent on the landmark centers *C* to explain the EEG data *X*. On the top of right hand side, each dot around the landmark centers characterizes a particular source activation pattern (s_i). Different colors for the dots in the upper right figure represent distinct brain states or conditions.

filter/smoother and the Expectation-Maximization algorithm to obtain source and parameter estimates. Bekhti *et al.* proposed a majorization-minimization approach to solve the EEG/MEG inverse problem, and Markov Chain Monte-Carlo (MCMC) techniques to obtain a multimodal posterior density for the source activity [43].

As discussed in Section I, novel methods that can account for different levels of uncertainty at the sensor and source may provide a means for further improvement in ESI performance. In this paper, we propose a new hierarchical probabilistic model that learns a manifold pattern for the source signal to robustly reconstruct source activity in the presence of noise at both the sensor and source.

III. Probabilistic Model With Hierarchical Prior

Our method relies on three key ingredients: sparsity of the source signals, source signal denoising, and manifold relationships among landmarks for the source signals. Here we present the probabilistic formulations of these three ingredients and combine them to form a novel hierarchical probabilistic model. The architecture of this graphical model is illustrated in Fig. 1, followed by separate elaborations on each of the components.

A. The Inverse Problem of ESI

The EEG/MEG data measures the electromagnetic field at a set of N_c sensors (or channels) for N_t time points. We denote the EEG/MEG measurements as $X = [\mathbf{x}_1, \ldots, \mathbf{x}_{N_t}] \in \mathbb{R}^{N_c \times N_t}$. The linear mapping from the brain sources to the sensors on the scalp is often referred as the *lead field gain matrix*, denoted by $L \in \mathbb{R}^{N_c \times N_s}$, obtained from the quasi-static approximation of Maxwell's equations [44] where N_s is the number of distributed sources used to represent the discretized 3D head model. Specifically, each column of L represents the electrical propagation weight vector from a particular source location to the EEG/MEG electrodes. Given X and L, the objective is to find a source activation (source signal), denoted as $S = [\mathbf{s}_1, \ldots, \mathbf{s}_{N_t}] \in \mathbb{R}^{N_s \times N_t}$, where each column corresponds to electrical potentials in N_s source locations for one of the N_t time points. The measured EEG/MEG data X can be described as a linear function of sources S with an additive noise,

$$\mathbf{x}_i = L\mathbf{s}_i + \epsilon_i, \quad \forall i = 1, \cdots, N_t. \tag{1}$$

Here $\mathbf{s}_i$ corresponds to the activation of the source space. The activation pattern is comprised of the true activation which may relate to functional neural activities given external stimuli or tasks, or neurophysiological disorders such as interictal activations in epilepsy patient, and spurious sources, which refer to random or spontaneous activity. With proper whitening [45], the noise term ϵ_i can be assumed to follow the Gaussian distribution with zero mean and an identity matrix I as covariance, i.e., $\epsilon_i \sim \mathcal{N}(0, I)$. Therefore, the probability of $\mathbf{x}_i$, given $\mathbf{s}_i$, is expressed as

$$p(\mathbf{x}_i|\mathbf{s}_i) \propto \mathcal{N}(L\mathbf{s}_i, I), \quad \forall i = 1, \cdots, N_t. \tag{2}$$

To find source estimates with spatially sparse but smooth active regions, we use a spatial smoothness basis matrix which specifies a dictionary of spatial basis vectors [42], [46]. By introducing the spatial basis functions, problem (1) can be formulated as

$$\mathbf{x}_i = L\mathbf{s}_i + \epsilon_i = L\Psi\Psi^{-1}\mathbf{s}_i + \epsilon_i, \tag{3}$$

where $\Psi = [\psi_1, \psi_2, \ldots, \psi_{N_s}]$ is the basis matrix, and the j-th element ψ_{ij} of ψ_i is defined as:

$$\psi_{ij} = \begin{cases} 1 & i = j, \\ \exp(-\|d_{ij}/\varrho\|^2) & j \in \Omega_i, \\ 0 & \text{otherwise}, \end{cases} \quad (4)$$

d_{ij} is the distance between sources i and j, ϱ is a spatial smoothness scalar controlling the smoothness level, and Ω_i is a set of nearest neighboring sources of i based on geodesic distances. Equation (3) can be written as $\mathbf{x}_i = \tilde{L}\tilde{\mathbf{s}}_i + \epsilon_i$ with $\tilde{L} = L\Psi$ and $\tilde{\mathbf{s}}_i = \Psi^{-1}\mathbf{s}_i$. In what follows, we will describe how to recover $\tilde{\mathbf{s}}_i$. For convenience, we will rename $\tilde{L}$ and $\tilde{\mathbf{s}}_i$ back to L and $\mathbf{s}_i$, respectively. The final reconstructed signal should be left-multiplied by Ψ to recover the source signal S.

B. A Sparse Generative Prior Over Latent Landmarks

Generally in ESI, the number of sources is much larger than the number of electrodes, i.e., $N_s \gg N_c$, which implies that (1) is an under-determined system and results in an ill-posed problem for ESI. In order to properly recover $\mathbf{s}_i$ from $\mathbf{x}_i$, prior knowledge is often incorporated as a regularization. A popular choice is the ℓ_1 penalty [47] for a sparse solution, which follows specifically the Laplacian distribution. Each entry of $\mathbf{s}_i$ is i.i.d. drawn from a Laplacian distribution with mean zero and a positive parameter γ_1, i.e.,

$$p(\mathbf{s}_i|\gamma_1) = (\frac{\gamma_1}{2})^{N_s} \exp(-\gamma_1\|\mathbf{s}_i\|_1), \quad \forall i = 1, \cdots, N_t. \quad (5)$$

In this paper, we represent the source signals using a probability distribution conditioned on latent landmarks. The latent landmarks help to denoise the source signal estimates. The denoised signals are then constrained by a graph structure, which will be detailed in Section III-C.

Intuitively, the source signal denosing becomes simple if we know the true distribution of the clean source signals. In addition to the sparse prior in (5), we assume that the true distribution of clean source signals depends on a set of latent variables $C = [\mathbf{c}_1, \cdots, \mathbf{c}_K] \in \mathbb{R}^{N_s \times K}$, which we call *landmarks* for ease of reference. The distribution of the source signal $\mathbf{s}_i$ conditional on these landmarks is denoted by $p(\mathbf{s}_i|C)$,

$$p(\mathbf{s}_i|C) = \frac{1}{K\sigma^{N_s}} \sum_{k=1}^{K} (2\pi)^{-\frac{N_s}{2}} \exp(-\frac{1}{2\sigma^2}\|\mathbf{c}_k - \mathbf{s}_i\|^2). \quad (6)$$

Given K landmarks in C and the sparsity prior parameter γ_1, we represent $p(\mathbf{s}_i|C, \gamma_1)$ as a mixture of Gaussian distributions with the sparsity prior. We can combine the priors expressed in (5) and (6) to arrive at a sparse generative prior over latent landmarks:

$$p(\mathbf{s}_i|C, \gamma_1) \propto \left[\frac{1}{K\sigma^{N_s}} \sum_{k=1}^{K} (2\pi)^{-\frac{N_s}{2}} \exp(-\frac{1}{2\sigma^2}\|\mathbf{c}_k - \mathbf{s}_i\|^2)\right] \times \left[(\frac{\gamma_1}{2})^{N_s} \exp(-\gamma_1\|\mathbf{s}_i\|_1)\right], \quad (7)$$

where the first part is the generative prior for $\mathbf{s}_i$ conditioned on the landmarks C characterized by the mixture of Gaussian distributions, the second part is the sparsity prior on $\mathbf{s}_i$ to address the ill-posedness.

Under this model, the denoised source signal S is sparse and centered around the landmarks C. In this paper C is also an unknown variable that needs to be estimated. The denoising property of this prior will be further discussed in Section III-D.

C. Graph Structure Sparse Priors for Landmarks

The sources $\mathbf{s}_i$ are sparse given their Laplacian prior, yet they also depend on the landmarks. As such, it would be reasonable and consistent if the landmarks $\mathbf{c}_k$ themselves also followed a sparsity prior, for instance, given by a Laplacian distribution similar to $\mathbf{s}_i$ but with a possibly different parameter γ_2, i.e.,

$$p(\mathbf{c}_k|\gamma_2) = (\frac{\gamma_2}{2})^{N_s} \exp(-\gamma_2\|\mathbf{c}_k\|_1), \quad \forall k = 1, \cdots, K. \quad (8)$$

Recall that the landmarks are introduced to model the true distribution of the source signals. Instead of assuming independence among these landmarks, we attempt to model the pairwise relationships between landmarks in terms of a constrained graph.

We assume the landmarks have an underlying graph structure, denoted by $G \in \mathbb{R}^{K \times K}$, where each vertex of G corresponds to one landmark. We obtain the joint distribution

$$p(C|G) \propto \exp\left(-\frac{\beta}{2} \sum_{k=1}^{K} \sum_{k'=1}^{K} g_{k,k'}\|\mathbf{c}_k - \mathbf{c}_{k'}\|^2\right), \quad (9)$$

where β is a positive parameter. We can then combine the priors expressed in (8) and (9) to arrive at a sparse graph structure prior for landmarks:

$$p(C|G, \gamma_2) \propto \exp\left(-\frac{\beta}{2} \sum_{k=1}^{K} \sum_{k'=1}^{K} g_{k,k'}\|\mathbf{c}_k - \mathbf{c}_{k'}\|^2\right) \times (\frac{\gamma_2}{2})^{KN_s} \exp(-\gamma_2 \sum_{k=1}^{K} \|\mathbf{c}_k\|_1), \quad (10)$$

where the first part is the graph structure prior over the landmarks, and the second part is the sparsity prior. This form of distribution (10) enforces smoothness among landmarks [48] in the sense that a larger $g_{k,k'}$ results in a smaller distance of $\|\mathbf{c}_k - \mathbf{c}_{k'}\|^2$. G is required to be non-negative and symmetric, i.e., $g_{k,k'} \geq 0$ and $g_{k,k'} = g_{k',k}, \forall k, k'$, respectively.

In general, G is unknown, so it is challenging to model the density function over all graphs that have C as their vertices. To make the structure learning pragmatic, constraints on the set of graphs are required. To overcome this challenge, we directly model $\log p(G)$ instead of modeling $p(G)$ for learning a specific type of graphs from data such as spanning trees [49]. Informally, given a connected undirected graph $(\mathcal{V}, \mathcal{E})$ with edges $(V_k, V_{k'}) \in \mathcal{E}, \forall k, k'$, let $\mathcal{T}$ be a set of all spanning trees with vertices $\mathcal{V}$ and $\mathcal{E}$ be the edges forming a tree. In order to represent and learn a tree structure, $\{g_{k,k'}\}$ are formulated as binary variables where $g_{k,k'} = 1$ if $(V_k, V_{k'}) \in \mathcal{E}$ and 0 otherwise, i.e., $G = [g_{k,k'}] \in \{0, 1\}^{K \times K}$. To this end, we express the parametric formulation of $\log p(G)$ as an indicator function of the set of trees given by

$$\log p(G) \propto \begin{cases} 0, & G \in \mathcal{T}, \\ -\infty, & \text{otherwise}, \end{cases} \quad (11)$$

where $\mathcal{T}$ satisfies four conditions to be a set of spanning trees: (I) binary: $\{G \in \{0,1\}^{K\times K}\}$; (II) symmetric $\{G = G^T\}$; (III) acyclic $\{\frac{1}{2}\sum_{k,k'} g_{k,k'} = |\mathcal{V}| - 1\}$; and (IV) connected: $\{\frac{1}{2}\sum_{V_k\in\mathcal{S},V_{k'}\in\mathcal{S}} g_{k,k'} \leq |\mathcal{S}| - 1, \forall \mathcal{S} \subseteq \mathcal{V}\}$ [49]. The third constraint states that a spanning tree only has $|\mathcal{V}|-1$ edges and the fourth constraint imposes the connectivity properties of a tree. Equation (11) states that any graph as a tree is uniformly sampled from $p(G)$, and the probability of $G \notin \mathcal{T}$ is zero.

D. The Proposed Probabilistic Hierarchical Model

Given the likelihood function (2) and priors (7), (10), and (11), we are ready to formulate the joint conditional distribution $p(S, C, G|X, \gamma_1, \gamma_2)$. By applying the Bayesian network from the perspective of directed graphical model, the joint likelihood function $p(S, C, G|X, \gamma_1, \gamma_2)$ is written as

$$\begin{aligned} &p(S, C, G|X, \gamma_1, \gamma_2) \\ &\quad\propto\ p(X|S)p(S|C, \gamma_1)p(C|G, \gamma_2)p(G) \\ &\quad\propto \prod_{i=1}^{N_t}\Big[p(\mathbf{x}_i|\mathbf{s}_i)p(\mathbf{s}_i|C, \gamma_1)\Big]p(C|G, \gamma_2)p(G). \end{aligned} \tag{12}$$

To obtain estimates S, C and G from data, we propose to employ the maximum a posterior estimation by minimizing the negative logarithm of the conditional probability $p(S, C, G|X, \gamma_1, \gamma_2)$, given by,

$$\begin{aligned} \min_{S,C,G\in\mathcal{T}} &\|X - LS\|_F^2 + \gamma_1 \sum_{i=1}^{N_t} \|\mathbf{s}_i\|_1 + \gamma_2 \sum_{k=1}^{K} \|\mathbf{c}_k\|_1 \\ &- 2\lambda\sigma^2 \sum_{i=1}^{N_t} \log \sum_{k=1}^{K} \exp(-\frac{\|\mathbf{s}_i - \mathbf{c}_k\|^2}{2\sigma^2}) \\ &+ \frac{\beta}{2}\sum_{k=1}^{K}\sum_{k'=1}^{K} \|\mathbf{c}_k - \mathbf{c}_{k'}\|^2 g_{k,k'} - \log p(G). \end{aligned} \tag{13}$$

To better understand the proposed model (13), we elaborate its two important properties as follows:

1) Graph Structure Learning via a Minimum-Cost Spanning Tree: Problem (13) with respect to G, namely the graph structure learning, is equivalent to the minimum-cost spanning tree (MST) [49]. By substituting (11) into (13), we have the following optimization problem with respect to G:

$$\min_{G\in\mathcal{T}} \frac{\beta}{2}\sum_{k=1}^{K}\sum_{k'=1}^{K} \|\mathbf{c}_k - \mathbf{c}_{k'}\|^2 g_{k,k'}, \tag{14}$$

which is the problem of the minimum-cost spanning tree with the cost for the edge $(V_k, V_{k'})$ defined as $\frac{\beta}{2}\|\mathbf{c}_k - \mathbf{c}_{k'}\|^2$. It is clear that (i) the landmarks are not independent; (ii) minimizing the objective function of (14) with respect to G given C can help improve the smoothness over the landmarks connected by the tree G. Specifically, if $g_{k,k'}$ is one, we expect $\|\mathbf{c}_k - \mathbf{c}_{k'}\|^2$ to be small, so that $\mathbf{c}_k$ should be close to $\mathbf{c}_{k'}$. Hence, the connected vertices on the tree will enforce similarity between any two connected landmarks.

2) Source Signal Denoising via Landmarks: Based on the density function (6), we can obtain the probability of any point that is not in the set of landmarks. In particular, we can introduce an assignment matrix $R \in \mathbb{R}^{N_t\times K}$ such that each entry $r_{i,k}$ represents the probability of assigning $\mathbf{s}_i$ to $\mathbf{c}_k$. With the help of Proposition 1 below, we can equivalently rewrite the maximization problem of the logarithm of the density function (6) over $\{\mathbf{s}_i\}_{i=1}^{N_t}$ as a joint minimization problem with respect to R and $\{\mathbf{s}_i\}_{i=1}^{N_t}$.

Proposition 1: Let α be a positive number and the feasible set $\mathcal{R} = \{r_{i,k}|r_{i,k} \geq 0, \sum_{k=1}^{K} r_{i,k} = 1\,\forall i\}$. We have

$$\tilde{g}(S, C) = \min_{R\in\mathcal{R}} g(S, C, R), \tag{15}$$

where

$$\tilde{g}(S, C) = -\alpha \sum_{i=1}^{N_t} \log \sum_{k=1}^{K} \exp(-\frac{\|\mathbf{s}_i - \mathbf{c}_k\|^2}{\alpha}), \tag{16}$$

$$g(S, C, R) = \sum_{i=1}^{N_t}\sum_{k=1}^{K} [r_{i,k}\|\mathbf{s}_i - \mathbf{c}_k\|^2 + \alpha r_{i,k}\log r_{i,k}]. \tag{17}$$

The proof of Proposition 1 is given in A-A of the supplemental material. Proposition 2 spells out the optimal R in (15).

Proposition 2: Given C and S, minimizing $g(S, C, R)$ with respect to $R \in \mathcal{R}$ has the closed form solution

$$r_{i,k} = \frac{\exp\left(-\frac{\|\mathbf{s}_i - \mathbf{c}_k\|^2}{\alpha}\right)}{\sum_{k'=1}^{K} \exp\left(-\frac{\|\mathbf{s}_i - \mathbf{c}_{k'}\|^2}{\alpha}\right)}, \quad \forall i, k. \tag{18}$$

Proof: This is obtained directly according to the KKT conditions in the proof of Proposition 1. □

It is worth noting that introducing the function $g(S, C, R)$ can be helpful to uncover the underlying properties of our proposed method, as claimed by Proposition 3.

Proposition 3: Given $R \in \mathcal{R}$ and C, $\min_S g(S, C, R)$ has the closed form solution

$$\mathbf{s}_i = \frac{\sum_{k=1}^{K} r_{i,k}\mathbf{c}_k}{\sum_{k=1}^{K} r_{i,k}}, \quad \forall i. \tag{19}$$

Proof: The function $g(S, C, R)$ with respect to S is convex and quadratic. The optimal solution is obtained by setting the first derivative to zero. □

Next, we explain how the proposed model can help denoise the source signal.

Source signal denoising is achieved by the proposed method based on the intuition that points in the high density regions are less noisy than those in low density regions. If the source signals are optimized to be points in the high density regions of the true density of source signals, we could argue that they are less noisy.

It is worth noting that both the source signals and the true density of source signals are unknown. Hence, we model them as in Section III-B, where the true density is parameterized by landmarks depending on the unknown graph structure as presented in Section III-C. For ease of discussion, let us assume the true density of source signals is available. This is

reasonable since our model optimizes all unknown variables jointly. Below, we will explain how the true distribution on landmarks helps to denoise source signals by optimizing source signals to be points in high density regions.

In Section III-B, we model the density of source signals S in terms of the conditional distribution on landmarks C. To explain how S becomes the points residing in high density regions formed by C, we first show the equivalence of two minimization problems for objective functions (16) and (17) with respect to S in Proposition 1, and then prove the crucially important property in Proposition 3. Specifically, (16) is the negative logarithmic function of density (6), so it is a mixture of Gaussians, but can also be interpreted as the nonparametrically estimated density via kernel density estimation for any source signal built on a set of landmarks. According to Proposition 1 and Proposition 3, minimizing (16) with respect to each $\mathbf{s}_i$ is equivalent to finding its corresponding maximum point of the density function by applying the update rule (19), that is, $\{\mathbf{s}_i\}$ are finally convergent to points in the high density regions. Intuitively, since the convergent source signals are in the high density regions, they become less noisy.

The denoising process is further explained by the convergent process of the optimal source signals by combining (18) in Proposition 2 and (19) in Proposition 3: i) Suppose that $\mathbf{s}_j$ is one noisy source signal. Proposition 2 tells us the weighting $r_{i,k}$ will be very small if $\mathbf{s}_i$ is far away from $\mathbf{c}_k$ since the distance between the source signal and the landmark $\mathbf{c}_k$ is large. ii) According to Proposition 3, the source signal $\mathbf{s}_i$ will continuously move to the landmarks that are closer to the source signal and finally converge to a high density region near the closest landmarks with large weights.

The representation of the noisy source signal to a robust point inside the high density region of landmarks is preferred for denoising source signals since landmarks are the representative points of the true distribution of clean source signals. Thus, the source signals are viewed in reference to a robust point centered around landmarks representing the true distribution of clean source signals.

We further show that K-means [50] is a special case of minimizing $g(S, C, R)$ with respect to $R \in \mathcal{R}$ as illustrated in Proposition 4. The proof of Proposition 4 is given in A-B of the supplemental material.

Proposition 4: Minimizing $g(S, C, R)$ with respect to $R \in \mathcal{R}$ is equivalent to K-means as $\alpha \to 0$.

The parameter α provides the flexibility for our method to take the density of points into account. As a result, our soft-assignment weighting strategy (18) becomes more robust than the hard-assignment in K-means.

IV. The Proposed Optimization Approach

According to Proposition 1 and with $\alpha = 2\sigma^2$, the estimates S, C and G of (13) can be equivalently obtained by introducing $R \in \mathcal{R}$ and solving the following problem

$$\min_{S,C,G\in\mathcal{T},R\in\mathcal{R}} h(S, C, G, R), \tag{20}$$

where $h(\cdot)$ is defined as

$$\begin{aligned} h(S, C, G, R) := \|X - LS\|_F^2 + \frac{\beta}{2}\sum_{k,k'=1}^{K} g_{k,k'}\|\mathbf{c}_k - \mathbf{c}_{k'}\|^2 \\ + \lambda\sum_{i=1}^{N_t}\sum_{k=1}^{K}[r_{i,k}\|\mathbf{s}_i - \mathbf{c}_k\|^2 + \alpha r_{i,k}\log r_{i,k}] \\ + \gamma_1\sum_{i=1}^{N_t}\|\mathbf{s}_i\|_1 + \gamma_2\sum_{i=1}^{K}\|\mathbf{c}_k\|_1. \end{aligned} \tag{21}$$

Next, we will describe an efficient algorithm to solve (20), followed by the analysis of its theoretical convergence and computational complexity.

A. Optimization Method

We consider an alternating convex search (ACS) method [51] to solve the proposed model (20). The ACS algorithm iterates as follows,

$$\begin{cases} S^{(n+1)} = \arg\min_{S} h(S, C^{(n)}, G^{(n)}, R^{(n)}), \\ C^{(n+1)} = \arg\min_{C} h(S^{(n+1)}, C, G^{(n)}, R^{(n)}), \\ G^{(n+1)} = \arg\min_{G\in\mathcal{T}} h(S^{(n+1)}, C^{(n+1)}, G, R^{(n)}), \\ R^{(n+1)} = \arg\min_{R\in\mathcal{R}} h(S^{(n+1)}, C^{(n+1)}, G^{(n+1)}, R), \end{cases} \tag{22}$$

where n is the iteration index. We omit the iteration index when the context is clear. Below we will describe how to solve each subproblem in details.

We rewrite the S-subproblem in (22) as

$$\begin{aligned} &\arg\min_{S}\sum_{i=1}^{N_t}\left(\|\mathbf{x}_i - L\mathbf{s}_i\|_2^2 + \lambda\mathbf{s}_i^T\mathbf{s}_i - 2\lambda(\sum_{k=1}^{K} r_{i,k}\mathbf{c}_k^T)\mathbf{s}_i + \gamma_1\|\mathbf{s}_i\|_1\right) \\ &= \arg\min_{S}\sum_{i=1}^{N_t}\left(\mathbf{s}_i^T U^T U\mathbf{s}_i - 2\mathbf{b}_i^T\mathbf{s}_i + \gamma_1\|\mathbf{s}_i\|_1\right) \\ &= \arg\min_{S}\sum_{i=1}^{N_t}\|U\mathbf{s}_i - U^{-T}\mathbf{b}_i\|_2^2 + \gamma_1\|\mathbf{s}_i\|_1, \end{aligned} \tag{23}$$

where U is the Cholesky factor of $L^TL + \lambda I_{N_s} = U^TU$, I_{N_s} is the $N_s \times N_s$ identity matrix and

$$\mathbf{b}_i = \left(\mathbf{x}_i^T L + \lambda\sum_{k=1}^{K} r_{i,k}\mathbf{c}_k^T\right)^T = L^T\mathbf{x}_i + \lambda\sum_{k=1}^{K} r_{i,k}\mathbf{c}_k.$$

It is straightforward to see that U is invertible. Let $Y = U^{-T}[\mathbf{b}_1, \ldots, \mathbf{b}_{N_t}] = U^{-T}\left(L^TX + \lambda CR^T\right)$, we see that the S-subproblem is equivalent to solving N_t independently strictly convex subproblems:

$$\mathbf{s}_t := \arg\min_{\mathbf{s}} \|U\mathbf{s} - U^{-T}\mathbf{b}_t\|_2^2 + \gamma_1\|\mathbf{s}\|_1. \tag{24}$$

The C-subproblem in (22) can be expressed as

$$\begin{aligned} \arg\min_{C}\ &\frac{\beta}{2}\sum_{k,k'=1}^{K}\|\mathbf{c}_k - \mathbf{c}_{k'}\|^2 g_{k,k'} \\ &+ \lambda\sum_{i=1}^{N_t}\sum_{k=1}^{K} r_{i,k}\|\mathbf{s}_i - \mathbf{c}_k\|^2 + \gamma_2\sum_{k=1}^{K}\|\mathbf{c}_k\|_1. \end{aligned} \tag{25}$$

Algorithm 1 The Proposed ACS Algorithm

Input: Data $X \in \mathbb{R}^{N_c \times N_t}$, lead field gain matrix $L \in \mathbb{R}^{N_c \times N_s}$, parameters $\beta, \lambda, \alpha, \gamma_1, \gamma_2, K$.
Initialize S, C, G, R.
repeat
  update S by solving (24);
  update C by solving (27);
  update G by solving (14) using Kruskal's algorithm;
  update R using (18);
until convergent.
Output: S, C, G, R.

Denote by $P = \text{diag}(G\mathbf{1}) - G$, the graph Laplacian matrix of G, and $\Lambda = \text{diag}(\mathbf{1}^T R)$, where $\mathbf{1}$ is a vector of all ones. Let $\|C\|_{1,1} = \sum_{k=1}^{K} \|\mathbf{c}_k\|_1$. We obtain an equivalent form of (25) as,

$$\min_{C} \text{trace}\left(C\left(\beta P + \lambda\Lambda\right)C^T - 2\lambda S R C^T\right) + \gamma_2\|C\|_{1,1}. \quad (26)$$

This is an (unconstrained) strictly convex problem, which admits a unique solution. To solve (26), we use the Cholesky decomposition of the matrix $\beta P + \lambda\Lambda := VV^T$, where $V \in \mathbb{R}^{K\times K}$. Since the Laplacian matrix P is positive semidefinite and the diagonal matrix Λ is positive definite according to $r_{i,k} > 0$ in Proposition 2, the matrix $\beta P + \lambda\Lambda$ is positive definite, which guarantees that V is invertible. As a result, we have the solution C given by:

$$C := \arg\min_{C} \|V^T C^T - \lambda V^{-1} R^T S^T\|_F^2 + \gamma_2\|C\|_{1,1}. \quad (27)$$

Both (24) and (27) are ℓ_1 regularized strictly convex quadratic programming problems and hence there exists a unique solution for each subproblem. Furthermore, it can be efficiently solved by many well developed methods such as Homotopy [52], ADMM LASSO [53], and FISTA [54]. In this paper, we adopt the Homotopy for solving both (24) and (27).

The G-subproblem in (22) boils down to an MST problem as shown in problem (14). It can be solved efficiently by the Kruskal algorithm [55]. Finally, there is a closed-form solution given by (18) for the R-subproblem.

The proposed ACS algorithm for solving problem (20) is outlined in Algorithm 1.

B. Initialization

Due to the nonconvexity of problem (20), we consider to have proper initializations of optimizing variables: S is solved by Homotopy [52] for the inverse problem, C is obtained by applying the K-means method [50] on the initialized S, G is computed by the Kruskal algorithm and R is updated using (18).

C. Theoretical Convergence Analysis

The convergence of Algorithm 1 is established in Theorem 1. Its proof is given in A-C of the supplemental material.

Theorem 1: Suppose $\{S^{(n)}, C^{(n)}, G^{(n)}, R^{(n)}\}$ *is the optimal solution of problem (20) in the n-th iteration with each subproblem solved exactly. Let* $h^{(n)} = h(S^{(n)}, C^{(n)}, G^{(n)}, R^{(n)})$ *be the corresponding objective function value. We have*

(a) *the function* $h(S, C, G, R)$ *is coercive;*
(b) $h^{(n+1)} \le h^{(n)}$*, i.e., objective function value is monotonically decreasing;*
(c) *the sequence* $\{h^{(n)}\}$ *converges as* $n \to \infty$*;*
(d) *the sequence* $\{S^{(n)}, C^{(n)}, G^{(n)}, R^{(n)}\}$ *has a convergent subsequence.*

It is worth noting that our convergent point is not guaranteed to be a global optimal solution due to the nonconvexity of the objective function. However, we observe that the solution obtained from Algorithm 1 can achieve good performance. Our explanation is that the nonconvexity is incurred from the multiplicity of C and G, and the source space can have different graph structures, however due to the existence of graph structure and landmarks, the reconstructed source signal is still robust to noise.

D. Computational Complexity Analysis

The computational complexity of Algorithm 1 can be summarized by adding the complexities of multiple key modules. First, some calculations can be precomputed, such as the Cholesky factor of $L^T L + \lambda I_{N_s} = U^T U$ and the inverse of the upper triangular matrix U. The complexity is $O(N_s^3)$. To solve (24), we employ the Homotopy method [52]. The computation cost of each homotopy step takes about $N_s^2 + N_s\tau_1 + 3\tau_1^2 + O(N_s)$ flops by assuming τ_1 elements in the support ($\tau_1 \le N_s$). As problem (23) is decomposable in terms of N_t columns of S, we need to solve N_t subproblem (24) independently. As a result, it takes $O(N_t(N_s^2 + N_s\tau_1 + 3\tau_1^2))$ for each Homotopy step of problem (23). Second, for a given spanning tree G, the Laplacian matrix P maintains the sparsity with only $2(K-1)$ nonzero values. The Cholesky decomposition of $\beta P + \lambda\Lambda = VV^T$ only takes $O(K)$. Again, we employ the Homotopy method to solve problem (27) with variable C^T, which is decomposable in terms of N_s columns of C^T, and we can solve each column in parallel. So the total cost is $O(N_s(K^2 + K\tau_2 + 3\tau_2^2))$ by assuming τ_2 elements in the support ($\tau_2 \le K$). Third, the complexity of Kruskal's algorithm on the fully connected graph with weights on the landmarks C can be calculated as follows. It requires $O(K^2\ N_s)$ for computing the fully connected graph and $O(K^2\log(K))$ for finding the spanning tree. Finally, computing R takes $O(N_s N_t K)$. As we will show in Section V-B, Algorithm 1 converges very fast with usually fewer than 10 iterations.

V. Numerical Experiments

We evaluate our proposed algorithm using one synthetic dataset and two real EEG/MEG datasets to illustrate its effectiveness. The simulation study employs a realistic MRI-based head model. Although it is known that noise is introduced at both the sensor and brain source levels, it is unclear the extent to which noise in the sensor and source spaces impacts ESI solutions. Therefore we consider various levels of signal-to-noise ratio (SNR) in both the sensor and source spaces to validate the proposed algorithm in comparison with well-established algorithms in the ESI literature

[29], [31], [32], [37]. The SNR is defined by SNR $= 10\log_{10}(P_s/P_n)$, where P_s is the power of signal and P_n is the power of noise signal at either the sensor or source level.

A. Benchmark Algorithms and Error Metrics

We compare our method with four benchmark algorithms, namely MNE [29], MCE [32], sLORETA [31], and MxNE [37]. For MCE, we employ the Homotopy algorithm due to its superior performance observed in our previous work on ESI [23]. In addition, we find that incorporating spatial smoothness with the ℓ_1 approach in MCE improves the reconstruction results, so we adopt this improved version of MCE in our experiments. As for MxNE [37], we use the ℓ_1 norm on the source space and the ℓ_2 norm across time, a version of MxNE, leading to the $\ell_{2,1}$ regularization with an assumption on fixed source orientations and spatial smoothness.

We quantitatively evaluate the performance of each competing algorithm based on the following metrics:

- Localization error (LE) measures the geodesic distance between two source locations on the cortex meshes using the Dijkstra shortest path algorithm. Since the two hemispheres of a brain are disconnected, we calculate LE separately for each hemisphere.
- Area under curve (AUC) [56] is particularly useful to characterize the overlap of an extended source activation pattern [57]–[59].
- Data fitting (DF) metric is defined as $r^2 = 1 - \frac{E_{res}}{E_{tot}}$, where

$$E_{tot} = \sum_{i=1} \|\mathbf{x}_i - \bar{\mathbf{x}}\|_2^2 \quad \text{and} \quad E_{res} = \sum_{i=1} \|\mathbf{x}_i - \hat{\mathbf{x}}\|_2^2,$$

  $\mathbf{x}_i$ is the i-th column in the EEG data X, $\bar{\mathbf{x}}$ is the mean of X along the time axis, and $\hat{\mathbf{x}}_i$ is the fitted value defined as $\hat{\mathbf{x}}_i = L\hat{\mathbf{s}}_i$ for any reconstructed source signal at time i.
- Reconstructed error (RE) in source location is defined as $\text{RE} = \frac{\|\hat{S}-S\|_F^2}{\|S\|_F^2}$.

Better accuracy for localization are expected if LE are close to 0 and AUC is close to 1. Although the regularization will compromise DF and RE, we still use those two as the performance reference for all the algorithms.

B. Simulation Study

As in many brain imaging problems, the underlying ground truth is generally unknown. As a result, we rely on numerical simulations on synthetic data to measure the performance of various algorithms.

1) Simulated Data Using a Real Head Model: We consider a realistic head model with simulated brain source signals. The head model was reconstructed from T1-MRI images of a 26-year old male subject scanned at Massachusetts General Hospital. We use a 128-channel BioSemi EEG cap layout and co-register it with the subject's head surfaces (Electrodes layout and coregistration can be found in Fig. G.1 of supplemental material). Brain structure segmentation and cortical surface reconstruction were conducted using FreeSurfer. Coregistration of the head surface and EEG electrodes were conducted using Brainstorm [60] and then verified using the GUI of coregistration function in MNE-Python [61]. The conductivity of brain, skull, and scalp were set to the default from MNE-Python, which are 0.3 S/m, 0.006 S/m and 0.3 S/m, respectively. The source space contains 1026 sources in each hemisphere, with 2052 sources in total.

In the simulation, we randomly select two activation locations over the entire source space. We then generate two active source signals via a 5th-order autoregressive (AR) model at these locations (see Fig. B.1 of supplemental material). In the simulated experiments, we consider three states with different activation locations, the sampling frequency as 100 Hz, and the time window as 2 s. A detailed description of the simulated signals can be found in [62]. Each of the three states corresponds to a multi-channel time series of 200 data points. As a result, we obtain a time series with length of 600 in total. We consider additive white noise at various SNR levels in both channel and source spaces, denoted as SNR_C and SNR_S, respectively. We set 4 noise levels for EEG channels: $\text{SNR}_C = 30$ dB, 20 dB, 10 dB, 0 dB, and 3 noise levels in source space $\text{SNR}_S = \infty$ (noiseless), 30 dB, and 10 dB. Examples of how these noise sources impact the true noise-free signal are shown in Fig. B.2 of the supplemental material. As you can see from Fig. B.2, when SNR_C= 10 dB, the EEG signal is significantly distorted. To further examine the performance of all the algorithms under extremely high level of channel noise, we include additional test for the experiments with $\text{SNR}_C = -5$. We also validate our algorithm with 3 active sources, and we observe the superiority of our proposed method over the benchmark algorithms, similar to the situations with 2 active sources. More details can be found in C-B of the supplemental material.

2) Parameter Sensitivity Analysis: We investigate the impacts of the four parameters on both AUC and LE in Algorithm 1. Since the sparsity levels of signals and landmarks are expected to follow the same patterns, it is natural to assume their regularization parameters are similar. Hence, we set $\gamma_1 = \gamma_2 = \gamma$. In order to investigate the sensitivity of our method with respect to these parameters, we evaluate Algorithm 1 by varying parameters in the following ranges: $\lambda \in \{1, 5, 10, 15, 20\}$, $K \in \{30, 60, 90, 120\}$, $\gamma \in \{1\text{e-}4, 1\text{e-}3, 0.01, 0.1, 1\}$, $\beta \in \{1, 3, 10, 20\}$, in the setting of $\text{SNR}_C = 20$ dB and $\text{SNR}_S =$ 10 dB and $\alpha = 0.01$. We used AUC and LE to evaluate the sensitivity for the varying parameters.

As usual, performance improves if AUC increases and LE decreases. We demonstrate the parameter sensitivity in Fig. 2 by showing the best AUC and LE over two of the four parameters as the other two vary. As can be seen, AUC is not sensitive to K, γ and β. We observe that $\lambda \in [0.5, 5]$ can generally lead to quite satisfactory results for both AUC and LE, even though our method is sensitive to the parameter λ varying in a wide range.

Based on the above sensitivity analysis, we employ the following settings for the synthetic data experiment with $\lambda = 3$, $\beta = 3$, $\alpha = 0.01$, while tuning $K \in \{60, 120\}$ and $\gamma \in \{0.001, 0.01\}$, for Algorithm 1. We also tune the regularization parameters for the compared algorithms, including MNE, MCE, sLORETA and $\ell_{2,1}$, over a wide range

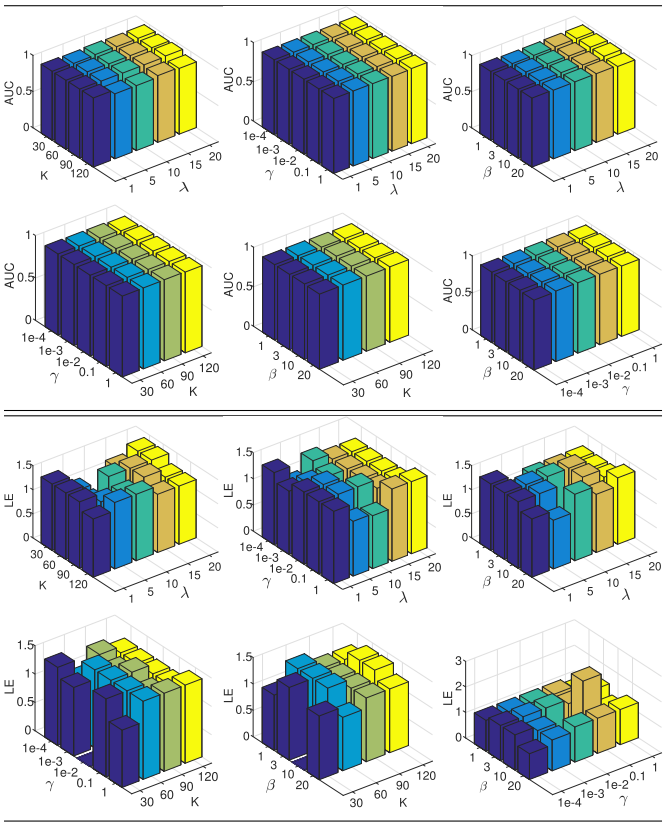

Fig. 2. AUC and LE by the proposed algorithm with respect to any two of the parameters at the noise level $\text{SNR}_C = 20$ dB and $\text{SNR}_S = 10$ dB.

$\{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 500, 1000\}$ in order to ensure that they can achieve the best possible performance. After solving the ESI problem using different algorithms, we use AUC to choose the best solution by each algorithm for synthetic data.

3) Convergence Analysis and Computational Analysis: We conduct experiments using the settings $\text{SNR}_C = 20$ dB, $\text{SNR}_C = 10$ dB and $\text{SNR}_C = 0$ dB, while $\text{SNR}_S = 30$ dB. We observe that the objective function values can usually converge within 5-10 steps as shown in Fig. 3. In Fig. E.1 of supplemental material, it is shown that the empirical computational complexity measured in seconds for each iteration using the same settings as in Fig. 3 for $\text{SNR}_S = 30$ dB and $\text{SNR}_C = 10$ dB. It takes less than 3 minutes per iteration on a Linux machine (Intel Xeon W-2133 CPU @3.60GHz with 12 cores). In Section IV-D, we show that the computational complexity for each iteration is almost constant, which is also verified in Fig. E.1. It is worth noting that steps 3, 4, and 6 of Algorithm 1 can be efficiently executed in parallel as shown in section IV-A.

4) Performance Evaluation: To compare against benchmark methods, we conduct both numerical and real human subject experiments to demonstrate the advantage of the proposed method. We first report the numerical comparisons in terms of the four metrics in Table I with respect to different sensor noise levels. Each reported value is the mean from 10 repeated experiments. The results show that the proposed approach outperforms the benchmark algorithms in most cases,

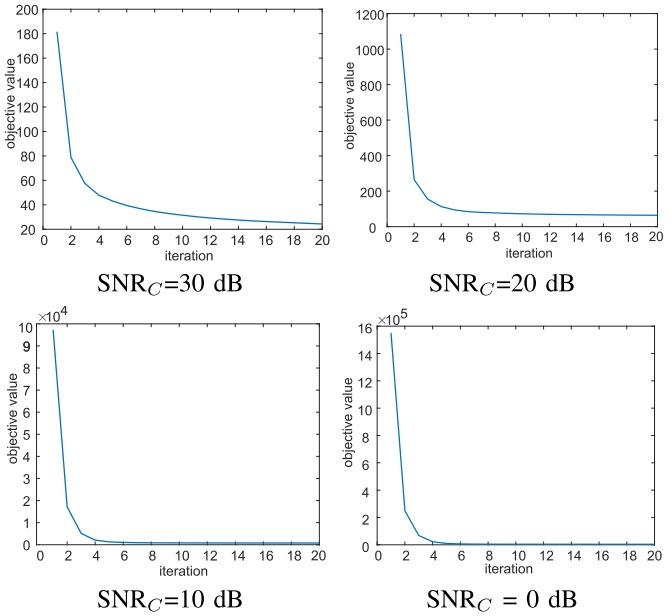

Fig. 3. Convergence analysis of the proposed algorithm under different SNR settings with $\lambda = 3, \beta = 3, K = 120, \gamma = 0.01$ and $\text{SNR}_S = 30$ dB.

particularly excelling in terms of LE and AUC. Specifically, for higher $\text{SNR}_C = 30$ dB (less noise in the channel space), all the algorithms perform very well in terms of LE. But for both $\text{SNR}_C = 10$ dB and $\text{SNR}_C = 0$ dB, the proposed approach stands out in terms of LE and AUC. The performance result when $\text{SNR}_C = -5$ dB is given in Table C.1 of the supplemental material. When $\text{SNR}_C = -5$ dB, all the algorithms performed poorly, with very large LE and small value of AUC.

Based on the above results, we have the following observations: i) when there is little noise, our algorithm performs slightly better; ii) when the noise intensity becomes slightly larger, the performances of all algorithms except ours degrade; iii) when the channel noise level is -5 dB, all algorithms perform very poorly.

We illustrate the results of source reconstruction under different noise levels and locations in Fig. 4. We normalize the final reconstructed source and scale the signal between 0 and 1. The threshold value is set to 1/3. The top row of Fig. 4 is for $\text{SNR}_C = 20$ dB, $\text{SNR}_S = 30$ dB with the true activation in the rostral middle frontal area. In this case of high SNR, all methods achieve satisfactory results, except for the spatially-diffuse solutions by MNE and sLORETA. We then increase the noise to $\text{SNR}_C = 10$ dB, $\text{SNR}_S = 30$ dB and generate the ground truth source activation as illustrated in the second row of Fig. 4, which shows that the activation patterns by the proposed algorithm are the closest to the ground-truth. MCE and $\ell_{2,1}$ can correctly identify some locations but also produce many spurious activations on the cortex surface. On the other hand, MNE and sLORETA again render highly diffuse solutions, but the locations with the largest magnitude on both hemispheres align well with the true activation locations. The third row in Fig. 4 shows a case when $\text{SNR}_C = 10$ dB, $\text{SNR}_S = 10$ dB with two sources that are located close to each other, and

TABLE I

PERFORMANCE (AVERAGE ± STANDARD DEVIATION) OF ALL ALGORITHMS USING THE FOUR METRICS OVER 10 RANDOMLY REPEATED EXPERIMENTS BASED ON SYNTHETIC DATA FOR $\text{SNR}_C = \{30, 20, 10, 0\}$ dB AND $\text{SNR}_S = \{\infty, 30, 10\}$ dB. AS THE SOLUTION FROM sLORETA MAINLY EXPLAINS THE SIGNAL VARIANCE, THE DF FIELD IS NOT APPLICABLE TO sLORETA AND IS LEFT BLANK

SNR_S	Methods	SNR_c=30 dB				SNR_c = 20 dB			
		DF	RE	LE	AUC	DF	RE	LE	AUC
∞ dB	MNE	0.926±0.030	0.560 ±0.068	2.613±3.803	0.951±0.012	0.871±0.015	0.659±0.069	5.508±6.511	0.927±0.021
	MCE	0.973±0.003	0.076±0.054	0.425±2.286	0.965±0.014	0.914±0.001	0.144±0.078	0.694±1.562	0.939±0.024
	sLORETA	–	0.843±0.099	1.796±3.098	0.939±0.017	–	0.919±0.072	5.754±7.366	0.914±0.015
	$\ell_{2,1}$	0.911±0.012	0.307±0.071	1.275±2.936	0.939±0.035	0.866±0.012	0.363±0.152	3.962±8.482	0.939±0.047
	Proposed	0.968±0.002	0.119±0.048	0.353±1.898	0.985±0.008	0.912±0.001	0.185±0.075	0.694±1.562	0.970±0.013
30 dB	MNE	0.933±0.011	0.557±0.065	4.310±4.194	0.956±0.014	0.872±0.015	0.635±0.041	4.348±3.862	0.936±0.014
	MCE	0.972±0.003	0.086±0.066	0.598±1.978	0.968±0.019	0.914±0.001	0.175±0.077	0.591±1.333	0.931±0.021
	sLORETA	–	0.860±0.089	3.300±3.444	0.942±0.013	–	0.973± 0.137	3.364±3.919	0.929±0.017
	$\ell_{2,1}$	0.918±0.014	0.327±0.160	2.517±4.814	0.929±0.051	0.872±0.008	0.3578±0.090	2.111±4.687	0.918±0.041
	Proposed	0.968±0.001	0.135±0.076	0.646±2.772	0.985±0.010	0.912±0.001	0.199±0.064	0.241±0.924	0.970±0.010
10 dB	MNE	0.944±0.012	0.529±0.057	2.433±3.507	0.938±0.012	0.880±0.010	0.628±0.054	3.948±4.099	0.930±0.017
	MCE	0.979±0.001	0.154±0.051	0.355±1.076	0.937±0.017	0.917±0.001	0.219±0.055	0.722±1.827	0.922±0.024
	sLORETA	–	0.964±0.096	2.063±3.090	0.925±0.012	–	1.180±0.111	3.279±3.463	0.907±0.014
	$\ell_{2,1}$	0.888±0.006	0.356±0.072	2.267±3.561	0.927±0.024	0.863±0.007	0.320±0.089	1.927±2.811	0.937±0.041
	Proposed	0.963±0.001	0.167±0.056	0.672±1.669	0.962±0.008	0.913±0.001	0.193±0.058	0.731±1.800	0.959±0.014

SNR_S	Methods	SNR_c=10 dB				SNR_c = 0 dB			
		DF	RE	LE	AUC	DF	RE	LE	AUC
∞ dB	MNE	0.666±0.006	0.771±0.044	7.689±7.317	0.893±0.033	0.285±0.102	0.886±0.026	35.011±10.257	0.835±0.022
	MCE	0.718±0.002	0.306±0.107	2.433±3.709	0.862±0.042	0.389±0.105	2.451±4.783	41.653±15.815	0.746±0.019
	sLORETA	–	1.379±0.196	6.383±16.04	0.875±0.026	–	2.842±0.356	24.972±18.747	0.806±0.02
	$\ell_{2,1}$	0.706±0.004	0.387±0.128	3.160±4.700	0.915±0.061	0.354±0.111	0.731±0.063	29.093±9.309	0.823±0.036
	Proposed	0.739±0.013	0.402±0.122	3.306±4.060	0.930±0.031	0.398±0.093	0.736±0.082	20.715±13.955	0.829±0.019
30 dB	MNE	0.674±0.004	0.791±0.041	8.856±5.066	0.893±0.028	0.293±0.104	0.887±0.033	37.333±10.243	0.834±0.017
	MCE	0.723±0.016	0.445±0.292	5.207±4.821	0.844±0.043	0.388±0.14	2.387±5.778	45.745±14.853	0.738±0.024
	sLORETA	–	1.25±0.176	7.297±15.241	0.876±0.019	–	2.759±0.665	27.226±17.099	0.813±0.019
	$\ell_{2,1}$	0.704±0.004	0.421±0.124	4.881±4.938	0.909±0.065	0.365±0.118	0.782±0.083	34.543±14.257	0.821±0.049
	Proposed	0.735±0.015	0.434±0.127	4.740±4.335	0.926±0.03	0.394±0.123	0.818±0.095	23.649±14.04	0.832±0.022
10 dB	MNE	0.672±0.007	0.807±0.04	8.62±5.982	0.882±0.025	0.325±0.088	0.891±0.028	43.838±10.329	0.815±0.026
	MCE	0.718±0.002	0.393±0.077	4.298±4.256	0.838±0.036	0.388±0.122	2.613±7.632	48.724±13.258	0.732±0.021
	sLORETA	–	1.484±0.133	7.892±11.89	0.862±0.020	–	3.326±0.682	35.248±19.388	0.789±0.024
	$\ell_{2,1}$	0.702±0.005	0.393±0.105	4.218±3.728	0.921±0.042	0.365±0.101	0.780±0.099	39.734±10.205	0.829±0.044
	Proposed	0.743±0.001	0.455±0.054	3.575±3.134	0.915±0.021	0.401±0.102	0.904±0.105	25.695±15.514	0.832±0.028

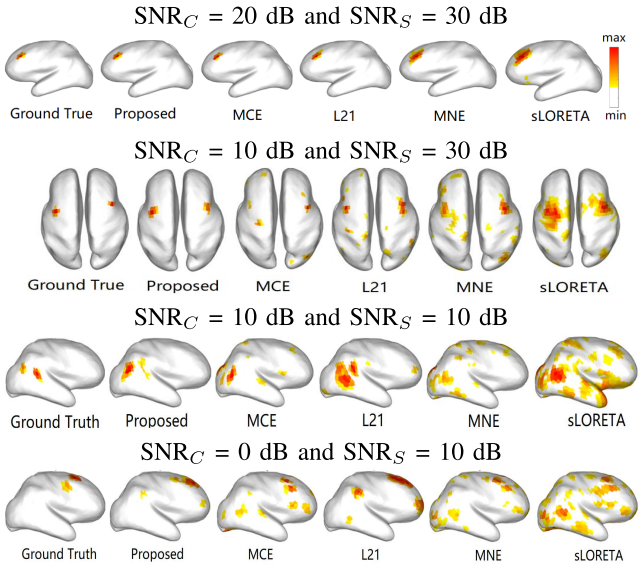

Fig. 4. Source reconstruction comparison for all algorithms under different noise levels.

the fourth row shows the reconstructed results of all algorithms when SNR_C = 0 dB and SNR_S = 10 dB, where the proposed algorithm provides a better reconstruction than any other algorithms. We also illustrated the reconstructed time series in Fig. F.2 of the supplemental material and the corresponding activation pattern in Fig. F.1 of the supplemental material.

In order to better illustrate the performances of the different algorithms, we decompose the true signal using PCA into its top 3 principal components and visualized the reconstructed signals by all methods on this basis. The tree structure learned by the proposed method is drawn based on the projected points from landmarks C as the vertices. The results obtained by the compared methods are illustrated in Fig. 5, from which we observe how well the solutions by different algorithms are aligned with the ground truth signals. It is clear that when the noise level is low (SNR_C = 30 dB), both solutions by MCE and our proposed algorithm align well with the ground truth signals, which explains why they have good LE and AUC performance. When the noise is large (SNR_C = 10 dB), our algorithm aligns well with the ground truth signals but the ones by MCE and sLORETA fail to align well, which explains the deteriorating performances by MCE and sLORETA when the noise is large. We can also see the landmarks C are a faithful representation of the ground truth signals, and the tree structure can correctly capture the manifold structure of data over various noise levels.

In summary, numerical results on the synthetic EEG data show that the proposed approach is comparable to the benchmark algorithms at lower noise levels, but at high noise levels, our algorithm significantly improves the benchmark algorithms. The observation confirms that source signal denoising and graph structure can identify activation patterns even at a high noise level in both sensor and source spaces.

C. Real Data Experiment I

We analyze a real dataset that is publicly accessible through the MNE-Python package [61]. The EEG/MEG data was collected when the subject was given auditory and visual stimuli.

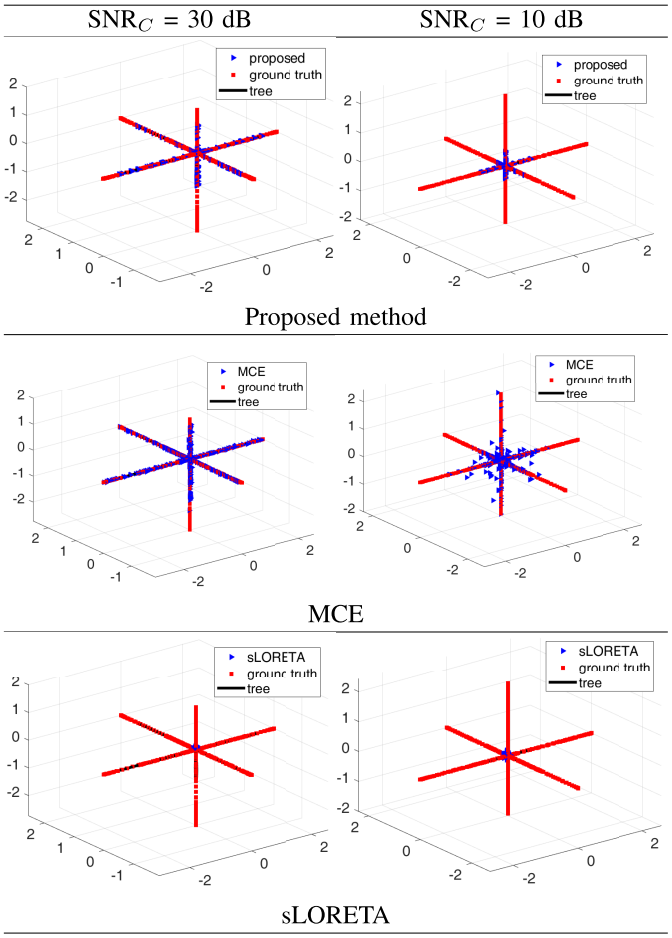

Fig. 5. Source solutions from three algorithms projected on the top-3 principal components of the ground truth signal S matrix with $\text{SNR}_C \in \{10, 30\}$ dB, and $\text{SNR}_S = 30$ dB. This figure shows how close all solutions are to the ground truth S at different levels of noise.

In the experiment, the subject was presented with checkerboard patterns to the left or right eye, interspersed by high frequency noise to the left or right ear. The recording device was a whole-head Elekta Neuromag Vector View 306 MEG system with 102 triple-sensor elements (two orthogonal planar gradiometers and one magnetometer per location). The EEG data from a 60-channel cap were also recorded simultaneously. Standard pre-processing steps including bandpass filtering, bad epoch rejection, and bad channel selection were conducted before applying ESI algorithms [40]. The lead field matrix was constructed using MNE-Python. There are 7498 sources distributed over the cortical surfaces.

The epochs under study are from left auditory stimuli (LAS) and left visual stimuli (LVS). There are 66 epochs for LAS and 73 for LVS. The time range used for all the epochs of LVS and LAS is from 0.03 s to 0.16 s after the stimuli events. The average time course aligned with stimuli events for EEG, Gradiometers and Gagnetometers are illustrated in Fig. G.3 for LAS and Fig. G.6 for LVS in supplemental material. By checking the time series plots in Fig. G.3, the event related field (ERF) activation pattern for LAS is very clear from time 0.08 s to 0.12 s and the related topomaps are illustrated on the rightmost plot in Fig. 6. Following the same parameter setting as the simulation study, we perform the source localization on an averaged epoch in a moving window with 15 epochs (overlap=5) for LAS and LVS and the final activation patterns on the cortex for LAS and LVS are illustrated in Fig. 6 for $t = 0.1$ s, and the reconstruction results for $t = 0.08$ s and $t = 0.12$ s are illustrated in Fig. G.5 the in the supplemental material.

The other competing methods show highly diffuse or spurious activations in their reconstructions, similar to the results from the synthetic data.

To visualize and validate the existence of a spanning tree in this real data, we project the source solutions of all averaged epochs from different moving windows for LAS and LVS onto a three-dimensional space in Fig. 7 using the discriminative dimensionality reduction tree algorithm [63]. The tree structure is shown in the source space where the blue dots and red dots are the sources under LAS and LVS, respectively, and the intersection of blue and red points is the preceding part of ERF when the evoked response is not very strong, so both conditions have much similar background activities. The activations under different stimuli conditions progress on the branches of the spanning tree. Fig. 7 also shows the landmarks activation patterns reside in a manifold structure.

Compared to the benchmark algorithms, our approach provides a sparse source reconstruction with apparently fewer spuriously activated sources with activation patterns that are more consistent across time.

D. Real Data Experiment II

For this real data experiment, we use a publicly-accessible epilepsy dataset available from the Brainstorm tutorial webpage [60]. The data were collected at the Epilepsy Center, University Hospital Freiburg, Germany. The patient suffered from focal epilepsy since the age of 8 years. The patient's 3T MRI and PET scans were both normal. The scalp EEG showed sharp waves and epileptiform spikes in the left frontocentral area (FC1, Cz channels have the maximum amplitude). The patient had an invasive EEG recording followed by a left frontal resection surgery. The patient's health condition was followed-up for 5 years after surgery and the patient was found to be seizure free. The patient provided written informed consent for this study, obtained by Dr. Andreas Schulze-Bonhage and Dr. Marcel Heers at the Epilepsy Center Freiburg. We obtained permission from the Epilepsy Center Feiburg to use this dataset in this work.

The EEG data were recorded using a Neurofile NT digital video-EEG system at 256 Hz and was filtered with a high-pass filter followed by a low-pass filter with cutoff frequencies of 0.16 Hz and 344 Hz, respectively. After discarding bad channels, there were 29 channels used for testing ESI algorithms. The spike locations were manually marked by the epileptologists at the Epilepsy Center in Freiburg. There were 58 time points identified as the epileptiform spikes. We selected [-100 ms, 100 ms] around the time stamp of each spike as one epoch. Eventually we had 58 epochs each with a length of 51 time points. We concatenated all the epochs to obtain a multivariate time series with 2958 data points.

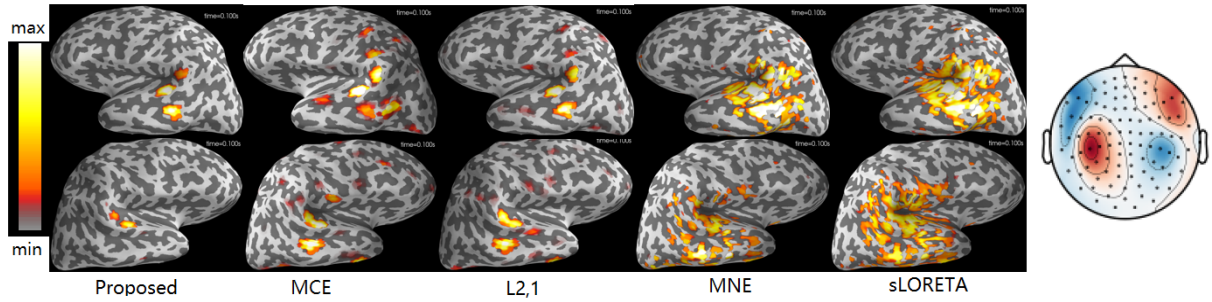

Fig. 6. Source activation patterns from MEG data at $t = 0.1$ s.

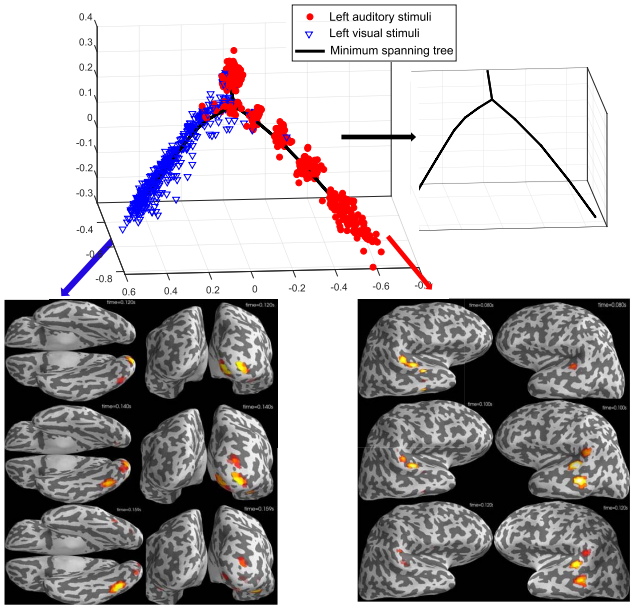

Fig. 7. Reconstructed sources and the spanning tree over landmarks projected onto a three-dimensional space for LVS (in blue) and LAS (in red), and the corresponding cortex activation for different time points for LVS (left) and LAS (right). The tree is extracted out for better viewing.

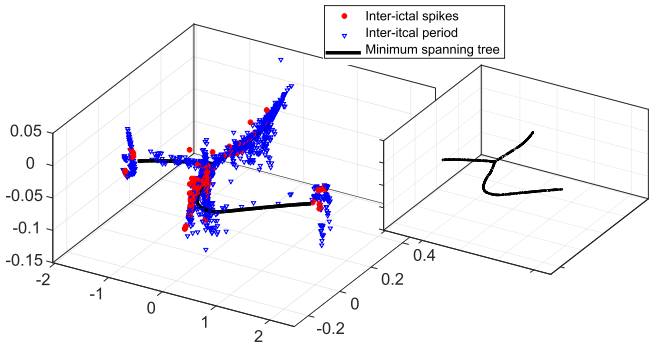

Fig. 8. Reconstructed sources and the spanning tree over landmarks projected onto a three-dimensional space. The tree is extracted out on the right side of the figure for better viewing.

We set the parameters for our algorithm as $\lambda = 3, \beta = 3, K = 500, \gamma_1 = \gamma_2 = 0.01$. Similar to Section V-C, we project the solution onto a three-dimensional space in Fig. 8. The marked spike data points and 2 data points before and after that are illustrated in red. The other data points are illustrated in blue. As we see in Fig. 8, most sources

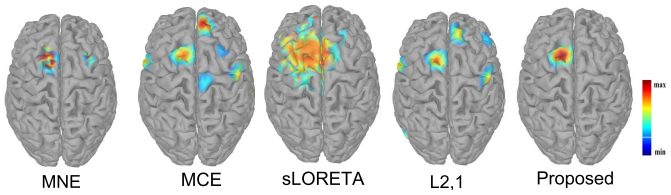

Fig. 9. Averaged inter-ictal spike source localization results. From left to right, the cortex activations are from MNE, MCE, sLORETA, $\ell_{2,1}$ and the proposed method.

reconstructed from the spikes and adjacent time points are aligned in one branch of the tree. We see that the spikes vary in amplitude, but form similar patterns connected by the spanning tree. The spanning tree structure shows the signal variations around the inter-ictal spikes. The averaged results are shown in Fig. 9 for the proposed algorithm and MNE, MCE, sLORETA, and the $\ell_{2,1}$ method. The scale of the colorbar in the figure matches the range of the max and min values for all solutions. Our results show consistency with the clinical results reported in [64] and other recent source localization results [65] for this same dataset.

VI. Conclusion

We have presented a novel probabilistic ESI model employing a hierarchical graph prior that identifies consistent patterns across time spanned from a sparse set of representative landmark activation patterns. An efficient algorithm based on an alternating convex search is proposed with provable convergence. We conduct extensive numerical experiments, including on both synthetic and real datasets. Our numerical experiments demonstrate that the proposed algorithm can robustly localize the activated sources with satisfactory precision despite high levels of noise at both the channel and source levels. On the other hand, we see that well-established benchmark algorithms are highly sensitive to noise at both levels, and particularly at the sensor level. For the synthetic data experiments, our algorithm outperforms all benchmark algorithms and yields significant improvements especially when the noise levels in channel space and source space are high. In the examples of real data that we studied, our algorithm renders more consistent reconstructions, while allowing variations across time. Compared with the benchmark algorithms, the source signals reconstructed by our proposed method are less contaminated by spurious sources originating from spontaneous brain fluctuations or measurement noise.

ACKNOWLEDGMENT

The authors are grateful for Dr. Andreas Schulze-Bonhage and Dr. Marcel Heers at the Epilepsy Center Freiburg for their permission to use the epilepsy dataset in this work.

REFERENCES

[1] B. He, A. Sohrabpour, E. Brown, and Z. Liu, "Electrophysiological source imaging: A noninvasive window to brain dynamics," *Annu. Rev. Biomed. Eng.*, vol. 20, no. 1, pp. 171–196, Jun. 2018.

[2] C. Phillips, M. D. Rugg, and K. J. Friston, "Anatomically informed basis functions for EEG source localization: Combining functional and anatomical constraints," *NeuroImage*, vol. 16, no. 3, pp. 678–695, Jul. 2002.

[3] B. Babadi, 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.

[4] A. Custo, S. Vulliemoz, F. Grouiller, D. Van De Ville, and C. Michel, "EEG source imaging of brain states using spatiotemporal regression," *NeuroImage*, vol. 96, pp. 106–116, Aug. 2014.

[5] C. Li *et al.*, "Epileptogenic source imaging using cross-frequency coupled signals from scalp EEG," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 12, pp. 2607–2618, Dec. 2016.

[6] C. M. Michel, M. M. Murray, G. Lantz, S. Gonzalez, L. Spinelli, and R. G. de Peralta, "EEG source imaging," *Clin. Neurophysiol.*, vol. 115, no. 10, pp. 2195–2222, 2004.

[7] P. Mégevand *et al.*, "Electric source imaging of interictal activity accurately localises the seizure onset zone," *J. Neurol. Neurosurg. Psychiatry*, vol. 85, no. 1, pp. 38–43, 2014.

[8] B. Erem, D. E. Hyde, J. M. Peters, F. H. Duffy, and S. K. Warfield, "Dynamic electrical source imaging (Des.) of seizures and interictal epileptic discharges without ensemble averaging," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 98–110, Jan. 2017.

[9] R. B. Willemse, A. Hillebrand, H. E. Ronner, W. P. Vandertop, and C. J. Stam, "Magnetoencephalographic study of hand and foot sensorimotor organization in 325 consecutive patients evaluated for tumor or epilepsy surgery," *NeuroImage Clin.*, vol. 10, pp. 46–53, Jan. 2016.

[10] A. G. Baroumand *et al.*, "Automated EEG source imaging: A retrospective, blinded clinical validation study," *Clin. Neurophysiol.*, vol. 129, no. 11, pp. 2403–2410, Nov. 2018.

[11] B. J. Edelman, B. Baxter, and B. He, "EEG source imaging enhances the decoding of complex right-hand motor imagery tasks," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 1, pp. 4–14, Jan. 2016.

[12] S. A. H. Hosseini, A. Sohrabpour, and B. He, "Electromagnetic source imaging using simultaneous scalp EEG and intracranial EEG: An emerging tool for interacting with pathological brain networks," *Clin. Neurophysiol.*, vol. 129, no. 1, pp. 168–187, Jan. 2018.

[13] P. van Mierlo, Y. Höller, N. K. Focke, and S. Vulliemoz, "Network perspectives on epilepsy using EEG/MEG source connectivity," *Frontiers Neurol.*, vol. 10, p. 721, Jul. 2019.

[14] S. Olbrich, A. Tränkner, T. Chittka, U. Hegerl, and P. Schönknecht, "Functional connectivity in major depression: Increased phase synchronization between frontal cortical EEG-source estimates," *Psychiatry Research: Neuroimaging*, vol. 222, nos. 1–2, pp. 91–99, Apr. 2014.

[15] G. Di Lorenzo *et al.*, "Altered resting-state EEG source functional connectivity in schizophrenia: The effect of illness duration," *Frontiers Human Neurosci.*, vol. 9, p. 234, May 2015.

[16] Y. Chen *et al.*, "Multimodal imaging of repetitive transcranial magnetic stimulation effect on brain network: A combined electroencephalogram and functional magnetic resonance imaging study," *Brain Connectivity*, vol. 9, no. 4, pp. 311–321, 2019.

[17] Y. Aoki *et al.*, "EEG resting-state networks responsible for gait disturbance features in idiopathic normal pressure hydrocephalus," *Clin. EEG Neurosci.*, vol. 50, no. 3, pp. 210–218, 2019.

[18] R. Grech *et al.*, "Review on solving the inverse problem in EEG source analysis," *J. neuroengineering Rehabil.*, vol. 5, no. 1, p. 1, 2008.

[19] S. Casta no-Candamil, J. Höhne, J.-D. Martínez-Vargas, X.-W. An, G. Castellanos-Domínguez, and S. Haufe, "Solving the EEG inverse problem based on space–time–frequency structured sparsity constraints," *NeuroImage*, vol. 118, pp. 598–612, Sep. 2015.

[20] B. Cai, P. Zille, J. M. Stephen, T. W. Wilson, V. D. Calhoun, and Y. P. Wang, "Estimation of dynamic sparse connectivity patterns from resting state fMRI," *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1224–1234, May 2018.

[21] F. Costa, H. Batatia, L. Chaari, and J.-Y. Tourneret, "Sparse EEG source localization using Bernoulli Laplacian priors," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 12, pp. 2888–2898, Dec. 2015.

[22] E. Pirondini *et al.*, "Computationally efficient algorithms for sparse, dynamic solutions to the EEG source localization problem," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 6, pp. 1359–1372, Jun. 2018.

[23] F. Liu, J. Rosenberger, Y. Lou, R. Hosseini, J. Su, and S. Wang, "Graph regularized EEG source imaging with in-class consistency and out-class discrimination," *IEEE Trans. Big Data*, vol. 3, no. 4, pp. 378–391, Dec. 2017.

[24] W. Ou, M. S. Hämäläinen, and P. Golland, "A distributed spatio-temporal EEG/MEG inverse solver," *NeuroImage*, vol. 44, no. 3, pp. 932–946, Feb. 2009.

[25] B. He, L. Yang, C. Wilke, and H. Yuan, "Electrophysiological imaging of brain activity and connectivity—Challenges and opportunities," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 7, pp. 1918–1931, 2011.

[26] B. D. Van Veen, W. Van Drongelen, M. Yuchtman, and A. Suzuki, "Localization of brain electrical activity via linearly constrained minimum variance spatial filtering," *IEEE Trans. Biomed. Eng.*, vol. 44, no. 9, pp. 867–880, Sep. 1997.

[27] J. C. Mosher, P. S. Lewis, and R. M. Leahy, "Multiple dipole modeling and localization from spatio-temporal MEG data," *IEEE Trans. Biomed. Eng.*, vol. 39, no. 6, pp. 541–557, Jun. 1992.

[28] J. C. Mosher and R. M. Leahy, "Source localization using recursively applied and projected (RAP) MUSIC," *IEEE Trans. Signal Process.*, vol. 47, no. 2, pp. 332–340, Feb. 1999.

[29] M. S. Hämäläinen and R. J. Ilmoniemi, "Interpreting magnetic fields of the brain: Minimum norm estimates," *Med. Biol. Eng. Comput.*, vol. 32, no. 1, pp. 35–42, Jan. 1994.

[30] A. M. Dale *et al.*, "Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity," *Neuron*, vol. 26, no. 1, pp. 55–67, 2000.

[31] R. D. Pascual-Marqui, "Standardized low-resolution brain electromagnetic tomography (sLORETA): Technical details," *Methods Find Exp. Clin. Pharmacol.*, vol. 24, pp. 5–12, Jan. 2002.

[32] K. Uutela, M. Hämäläinen, and E. Somersalo, "Visualization of magnetoencephalographic data using minimum current estimates," *NeuroImage*, vol. 10, no. 2, pp. 173–180, Aug. 1999.

[33] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic source imaging with FOCUSS: A recursive weighted minimum norm algorithm," *Electroencephalogr. Clin. Neurophysiol.*, vol. 95, no. 4, pp. 231–251, Oct. 1995.

[34] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. Signal Process.*, vol. 47, no. 1, pp. 187–200, Jan. 1999.

[35] J. C. Bore *et al.*, "Sparse EEG source localization using LAPPS: Least absolute I-P ($0 < p < 1$) penalized solution," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 7, pp. 1927–1939, Jul. 2018.

[36] P. Krishnaswamy *et al.*, "Sparsity enables estimation of both subcortical and cortical activity from MEG and EEG," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 48, pp. E10465–E10474, 2017.

[37] A. Gramfort, M. Kowalski, and M. Hämäläinen, "Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods," *Phys. Med. Biol.*, vol. 57, no. 7, p. 1937, 2012.

[38] D. Strohmeier, Y. Bekhti, J. Haueisen, and A. Gramfort, "The iterative reweighted mixed-norm estimate for spatio-temporal MEG/EEG source reconstruction," *IEEE Trans. Med. Imag.*, vol. 35, no. 10, pp. 2218–2228, Oct. 2016.

[39] M.-X. Huang *et al.*, "Vector-based spatial–temporal minimum L1-norm solution for MEG," *NeuroImage*, vol. 31, no. 3, pp. 1025–1037, 2006.

[40] A. Gramfort, D. Strohmeier, J. Haueisen, M. S. Hämäläinen, and M. Kowalski, "Time-frequency mixed-norm estimates: Sparse M/EEG imaging with non-stationary source activations," *NeuroImage*, vol. 70, pp. 410–422, Apr. 2013.

[41] C. Cai, K. Sekihara, and S. S. Nagarajan, "Hierarchical multiscale Bayesian algorithm for robust MEG/EEG source reconstruction," *NeuroImage*, vol. 183, pp. 698–715, Dec. 2018.

[42] C. Cai, M. Diwakar, D. Chen, K. Sekihara, and S. S. Nagarajan, "Robust empirical Bayesian reconstruction of distributed sources for electromagnetic brain imaging," *IEEE Trans. Med. Imag.*, vol. 39, no. 3, pp. 567–577, Mar. 2020.

[43] Y. Bekhti, F. Lucka, J. Salmon, and A. Gramfort, "A hierarchical Bayesian perspective on majorization-minimization for non-convex sparse regression: Application to M/EEG source imaging," *Inverse Problems*, vol. 34, no. 8, Aug. 2018, Art. no. 085010.

[44] C. H. Wolters, L. Grasedyck, and W. Hackbusch, "Efficient computation of lead field bases and influence matrix for the FEM-based EEG and MEG inverse problem," *Inverse problems*, vol. 20, no. 4, p. 1099, 2004.
[45] D. A. Engemann and A. Gramfort, "Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals," *NeuroImage*, vol. 108, pp. 328–342, Mar. 2015.
[46] S. Haufe *et al.*, "Large-scale EEG/MEG source localization with spatial flexibility," *NeuroImage*, vol. 54, no. 2, pp. 851–859, 2011.
[47] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 689–696.
[48] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. 20th Int. Conf. Mach. Learn. (ICML)*, 2003, pp. 912–919.
[49] M. Cheung. (2008). *Minimum-Cost Spanning Trees*. [Online]. Available: http://people.orie.cornell.edu/dpw/orie6300/fall2008/Recitations/rec09.pdf
[50] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.
[51] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: A survey and extensions," *Math. Methods Oper. Res.*, vol. 66, no. 3, pp. 373–407, Nov. 2007.
[52] M. S. Asif and J. Romberg, "Sparse recovery of streaming signals using L1-homotopy," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4209–4223, Aug. 2014.
[53] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.
[54] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, Jan. 2009.
[55] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proc. Amer. Math. Soc.*, vol. 7, no. 1, pp. 48–50, 1956.
[56] A. Molins, S. M. Stufflebeam, E. N. Brown, and M. S. Hämäläinen, "Quantification of the benefit from integrating MEG and EEG data in minimum l_2-norm estimation," *NeuroImage*, vol. 42, no. 3, pp. 1069–1077, 2008.
[57] L. Ding, "Reconstructing cortical current density by exploring sparseness in the transform domain," *Phys. Med. Biol.*, vol. 54, no. 9, p. 2683, 2009.
[58] M. Zhu, W. Zhang, D. L. Dickens, and L. Ding, "Reconstructing spatially extended brain sources via enforcing multiple transform sparseness," *NeuroImage*, vol. 86, pp. 280–293, Feb. 2014.
[59] A. Sohrabpour, Y. Lu, G. Worrell, and B. He, "Imaging brain source extent from EEG/MEG by means of an iteratively reweighted edge sparsity minimization (IRES) strategy," *NeuroImage*, vol. 142, pp. 27–42, Nov. 2016.
[60] F. Tadel, S. Baillet, J. C. Mosher, D. Pantazis, and R. M. Leahy, "Brainstorm: A user-friendly application for MEG/EEG analysis," *Comput. Intell. Neurosci.*, vol. 2011, p. 8, Oct. 2011.
[61] A. Gramfort *et al.*, "MNE software for processing MEG and EEG data," *NeuroImage*, vol. 86, pp. 446–460, Feb. 2014.
[62] S. Haufe and A. Ewald, "A simulation framework for benchmarking EEG-based brain connectivity estimation methodologies," *Brain topography*, pp. 1–18, 2016.
[63] L. Wang and Q. Mao, "Probabilistic dimensionality reduction via structure learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 205–219, Jan. 2017.
[64] M. Dümpelmann, T. Ball, and A. Schulze-Bonhage, "SLORETA allows reliable distributed source reconstruction based on subdural strip and grid recordings," *Human Brain Mapping*, vol. 33, no. 5, pp. 1172–1188, May 2012.
[65] K. Liu, Z. L. Yu, W. Wu, Z. Gu, Y. Li, and S. Nagarajan, "Variation sparse source imaging based on conditional mean for electromagnetic extended sources," *Neurocomputing*, vol. 313, pp. 96–110, Nov. 2018.
[66] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.