Efficient hierarchical Bayesian inference for spatio-temporal regression models in neuroimaging

Ali Hashemi[1,2], **Yijing Gao**[3], **Chang Cai**[3,4], **Sanjay Ghosh**[3],
Klaus-Robert Müller[2,5,6,7], **Srikantan S. Nagarajan**[3], **and Stefan Haufe**[1,8,9,10]
[1]Uncertainty, Inverse Modeling and Machine Learning Group, Technische Universität Berlin, Germany.
[2]Machine Learning Group, Technische Universität Berlin, Germany.
[3]Department of Radiology and Biomedical Imaging, University of California, San Francisco, USA.
[4]National Engineering Research Center for E-Learning, Central China Normal University, China.
[5]BIFOLD – Berlin Institute for the Foundations of Learning and Data, Berlin, Germany.
[6]Department of Artificial Intelligence, Korea University, South Korea.
[7]Max Planck Institute for Informatics, Saarbrücken, Germany.
[8]Physikalisch-Technische Bundesanstalt, Berlin, Germany.
[9]Charité – Universitätsmedizin Berlin, Germany.
[10]Bernstein Center for Computational Neuroscience, Berlin, Germany.

Abstract

Several problems in neuroimaging and beyond require inference on the parameters of multi-task sparse hierarchical regression models. Examples include M/EEG inverse problems, neural encoding models for task-based fMRI analyses, and temperature monitoring of climate or CPU and GPU. In these domains, both the model parameters to be inferred and the measurement noise may exhibit a complex spatio-temporal structure. Existing work either neglects the *temporal* structure or leads to computationally demanding inference schemes. Overcoming these limitations, we devise a novel flexible hierarchical Bayesian framework within which the spatio-temporal dynamics of model parameters and noise are modeled to have Kronecker product covariance structure. Inference in our framework is based on majorization-minimization optimization and has guaranteed convergence properties. Our highly efficient algorithms exploit the intrinsic Riemannian geometry of temporal autocovariance matrices. For stationary dynamics described by Toeplitz matrices, the theory of circulant embeddings is employed. We prove convex bounding properties and derive update rules of the resulting algorithms. On both synthetic and real neural data from M/EEG, we demonstrate that our methods lead to improved performance.

1 Introduction

Probabilistic graphical models for regression problems where both the model parameters and the observation noise can have complex spatio-temporal correlation structure are quite prevalent in neuroscience, neuroimaging and beyond [1]. Example domains where these models are being applied include Gaussian process inference [2], sparse signal recovery and multi-task learning [3–6], array signal processing [7], fMRI analyses [8], temperature monitoring of climate or CPU/GPU [9, 10], and brain computer interfaces (e.g. [11, 12]). The focus in this paper will be brain source imaging (BSI) [13], i.e. the reconstruction of brain source activity from non-invasive magneto- or electroetoencephalography (M/EEG) sensors [14]. Although a number of generative models that encompass spatio-temporal correlations have been proposed within the BSI regression problem, inference solutions have either imposed specific simplifications and model constraints [15–17] or have ignored temporal correlations overall [5], and therefore not fully addressed the inherent spatio-temporal problem structure.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

We propose a novel and efficient algorithmic framework using Type-II Bayesian regression that explicitely considers the spatio-temporal covariance structure in both model coefficients and observation noise. To this end, we focus on the M/EEG inverse problem of brain source reconstruction, for which we adopt a multi-task regression approach and formulate the source reconstruction problem as a probabilistic generative model with Gaussian scale mixture priors for sources with spatio-temporal prior covariance matrices that are expressed as a Kronecker product of separable spatial and temporal covariances. Their solutions are constrained to the Riemannian manifold of positive definite (P.D.) temporal covariance matrices. Exploiting the concept of geodesic convexity on the Riemannian manifold of covariance matrices with Kronecker structure, we are then able to derive robust, fast and efficient *majorization-minimization* (MM) optimization algorithms for model inference with provable convergence guarantees. In addition to deriving update rules for full-structural P.D. temporal covariance matrices, we also show that assuming a Toeplitz structure for the temporal covariance matrix results in computationally efficient update rules within the proposed MM framework.

2 Spatio-temporal generative model

Let us consider a multi-task linear regression problem, mathematically formulated as

$$\mathbf{Y}_g = \mathbf{L}\mathbf{X}_g + \mathbf{E}_g \quad \text{for } g = 1, \ldots, G \,, \tag{1}$$

in which a forward matrix, $\mathbf{L} \in \mathbb{R}^{M \times N}$, maps a set of coefficients or source components $\mathbf{X}_g \in \mathbb{R}^{N \times T}$ to measurements $\mathbf{Y}_g \in \mathbb{R}^{M \times T}$, with independent Gaussian noise $\mathbf{E}_g \in \mathbb{R}^{M \times T}$. G denotes the number of sample blocks in the multi-task problem, while N, M, and T denote the number of sources or coefficients, the number of sensors or observations, and the number of time instants, respectively. The problem of estimating $\{\mathbf{X}_g\}_{g=1}^G$ given $\mathbf{L}$ and $\{\mathbf{Y}_g\}_{g=1}^G$ can represent an inverse problem in physics, a multi-task regression problem in machine learning, or a multiple measurement vector (MMV) recovery problem in signal processing [18].

In the context of BSI, $\{\mathbf{Y}_g\}_{g=1}^G$ refers to M/EEG sensor measurement data, and $\{\mathbf{X}_g\}_{g=1}^G$ refers to brain source activity contributing to the observed sensor data. Here, G can be defined as the number of epochs, trials or experimental tasks. The goal of BSI is to infer the underlying brain activity from the M/EEG measurements given the lead field matrix $\mathbf{L}$. In practice, $\mathbf{L}$ can be computed using discretization methods such as the finite element method for a given head geometry and known electrical conductivities using the quasi-static approximation of Maxwell's equations [19]. As the number of locations of potential brain sources is dramatically larger than the number of sensors, the M/EEG inverse problem is highly ill-posed, which can be dealt with by incorporating prior assumptions. Adopting a Bayesian treatment is useful in this capacity because it allows these assumptions to be made explicitly by specifying prior distributions for the model parameters. Inference can be performed either through Maximum-a-Posteriori (MAP) estimation (*Type-I Bayesian learning*) [20–22] or, when the model has unknown hyperparameters, through Type-II Maximum-Likelihood (ML) estimation (*Type-II Bayesian learning*) [23–25].

Here we focus on full spatio-temporal Type-II Bayesian learning, which assumes a family of prior distributions $p(\mathbf{X}_g|\mathbf{\Theta})$ parameterized by a set of hyperparameters $\mathbf{\Theta}$. We further assume that the spatio-temporal correlation structure of the brain sources can be modeled by a Gaussian $p(\mathbf{x}_g|\mathbf{\Gamma}, \mathbf{B}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_0)$, where $\mathbf{x}_g = \text{vec}(\mathbf{X}_g^\top) \in \mathbb{R}^{NT \times 1}$ and where the covariance matrix $\mathbf{\Sigma}_0 = \mathbf{\Gamma} \otimes \mathbf{B}$ has Kronecker structure, meaning that the temporal and spatial correlations of the sources are modeled independently through matrices $\mathbf{\Gamma}$ and $\mathbf{B}$, where $\otimes$ stands for the Kronecker product. Note that the decoupling of spatial and temporal covariance is a neurophysiologically plausible assumption as neuronal sources originating from different parts of the brain can be assumed to share a similar autocorrelation spectrum. In this paper, we set $\mathbf{\Gamma} = \text{diag}(\boldsymbol{\gamma})$, $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_N]^\top$, and $\mathbf{B} \in \mathbb{S}_{++}^T$, where $\mathbb{S}_{++}^T$ denotes the set of positive definite matrices with size $T \times T$. Note that using a diagonal matrix $\mathbf{\Gamma}$ amounts to assuming that sources in different locations are independent. This assumption may be relaxed in our future work. On the other hand, modeling a full temporal covariance matrix $\mathbf{B}$ amounts to assuming non-stationary dynamics of the sources, which is appropriate in event-related experimental designs. To deal with the more stationary sources (ongoing brain activity), we will also constrain $\mathbf{B}$ to have Toeplitz structure later on. Based on these specifications, the prior distribution of the i-th brain source is modeled as $p((\mathbf{X}_g)_{i.}|\gamma_i, \mathbf{B}) \sim \mathcal{N}(0, \gamma_i \mathbf{B})$, for $i = 1, \ldots, N$, where $(\mathbf{X}_g)_{i.}$ denotes the i-th row of source matrix $\mathbf{X}_g$.

Analogous to the definition of the sources, we here also model the noise to have spatio-temporal structure. The matrix $\mathbf{E}_g \in \mathbb{R}^{M\times T}$ represents T time instances of zero-mean Gaussian noise with full covariance, $\mathbf{\Sigma_e} = \mathbf{\Lambda} \otimes \mathbf{\Upsilon}$, where $\mathbf{e}_g = \text{vec}(\mathbf{E}_g^\top) \in \mathbb{R}^{MT\times 1} \sim \mathcal{N}(0, \mathbf{\Sigma_e})$, and where $\mathbf{\Lambda}$ and $\mathbf{\Upsilon}$ denote spatial and temporal noise covariance matrices, respectively. Here, we assume that the noise and sources share the same temporal structure, i.e., $\mathbf{\Upsilon} = \mathbf{B}$. We later investigate violations of this assumption in the simulation section. Analogous to the sources, we consider spatially independent noise characterized by diagonal spatial covariance $\mathbf{\Lambda} = \text{diag}(\boldsymbol{\lambda})$, $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_M]^\top$. We consider both the general, *heteroscedastic*, case in which the noise level may be different for each sensor as well as the *homoscedastic* case in which the noise level is uniform.

For later use, we also define augmented versions of the source and noise covariance matrices as well as the lead field. Specifically, we set $\mathbf{H}$ and $\mathbf{\Phi}$ so that $\mathbf{H} := [\mathbf{\Gamma}, \mathbf{0}; \mathbf{0}, \mathbf{\Lambda}]$, and $\mathbf{\Phi} := [\mathbf{L}, \mathbf{I}]$. These definitions allow us to unify source and noise covariance parameters within the single variable $\mathbf{H}$, which facilitates the concurrent estimation of both quantities. $\mathbf{\Phi}$ now plays the same role as the lead field $\mathbf{L}$, in the sense that it maps $\boldsymbol{\eta}_g := [\mathbf{x}_g^\top, \mathbf{e}_g^\top]$ to the measurements; i.e., $\mathbf{y}_g = \mathbf{\Phi}\boldsymbol{\eta}_g$. Figure 1-(a) illustrates a probabilistic graphical model summarizing the spatio-temporal generative model of our multi-task linear regression problem.

The MMV model Eq. (1) can be formulated equivalently in *single measurement vector (SMV)* form by vectorizing the spatio-temporal data matrices and using the Kronecker product as follows: $\mathbf{y}_g = \mathbf{D}\mathbf{x}_g + \mathbf{e}_g$, where $\mathbf{y}_g = \text{vec}\left(\mathbf{Y}_g^\top\right) \in \mathbb{R}^{MT\times 1}$, and $\mathbf{D} = \mathbf{L} \otimes \mathbf{I}_T$. In a Type-II Bayesian learning framework, the hyper-parameters of the spatio-temporal source model are optimized jointly with the model parameters $\{\boldsymbol{X}_g\}_{g=1}^G$. In our case, these hyper-parameters comprise the unknown source, noise and temporal covariance matrices, i.e., $\mathbf{\Theta} = \{\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B}\}$.

3 Proposed method — full Dugh

The unknown parameters $\mathbf{\Gamma}$, $\mathbf{\Lambda}$, and $\mathbf{B}$ can be optimized in an alternating iterative process. Given initial estimates, the posterior distribution of the sources is a Gaussian of the form $p(\mathbf{x}_g|\mathbf{y}_g, \mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B}) \sim \mathcal{N}(\bar{\mathbf{x}}_g, \mathbf{\Sigma_x})$, whose mean and covariance are defined as follows:

$$\bar{\mathbf{x}}_g = \mathbf{\Sigma}_0\mathbf{D}^\top(\mathbf{\Lambda} \otimes \mathbf{B} + \mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^\top)^{-1}\mathbf{y}_g = \mathbf{\Sigma}_0\mathbf{D}^\top\tilde{\mathbf{\Sigma}}_\mathbf{y}^{-1}\mathbf{y}_g \ , \tag{2}$$

$$\mathbf{\Sigma_x} = \mathbf{\Sigma}_0 - \mathbf{\Sigma}_0\mathbf{D}^\top\tilde{\mathbf{\Sigma}}_\mathbf{y}^{-1}\mathbf{D}\mathbf{\Sigma}_0 \ , \tag{3}$$

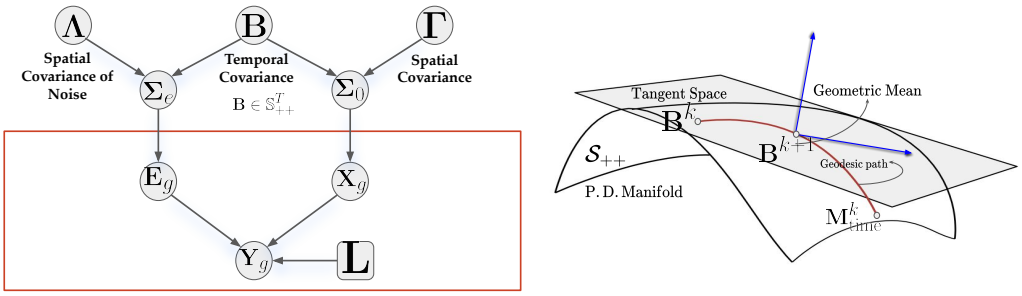
where $\mathbf{\Sigma_y} = \mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top + \mathbf{\Lambda}$, and where $\tilde{\mathbf{\Sigma}}_\mathbf{y} = \mathbf{\Sigma_y} \otimes \mathbf{B}$ denotes the spatio-temporal statistical model covariance matrix. The estimated posterior parameters $\bar{\mathbf{x}}_g$ and $\mathbf{\Sigma_x}$ are then in turn used to update $\mathbf{\Gamma}$, $\mathbf{\Lambda}$, and $\mathbf{B}$ as the minimizers of the negative log of the marginal likelihood $p(\mathbf{Y}|\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$, which is given by:

$$\mathcal{L}_{\text{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B}) = T\log|\mathbf{\Sigma_y}| + M\log|\mathbf{B}| + \frac{1}{G}\sum_{g=1}^{G}\text{tr}(\mathbf{\Sigma}_\mathbf{y}^{-1}\mathbf{Y}_g\mathbf{B}^{-1}\mathbf{Y}_g^\top) \ , \tag{4}$$

where $|\cdot|$ denotes the determinant of a matrix. A detailed derivation is provided in Appendix A. Given the solution of the hyperparameters $\mathbf{\Gamma}$, $\mathbf{\Lambda}$, and $\mathbf{B}$, the posterior source distribution is obtained by plugging these estimates into Eqs. (2)–(3). This process is repeated until convergence.

The challenge in this high-dimensional inference problem is to find (locally) optimal solutions to Eq. (4), which is a non-convex cost function, in adequate time. Here we propose a novel efficient algorithm, which is able to do so. Our algorithm thus learns the full spatio-temporal correlation structure of sources and noise. Hashemi et al. [26] have previously demonstrated that *majorization-minimization* (MM) [27] is a powerful non-linear optimization framework that can be leveraged to solve similar Bayesian Type-II inference problems. Here we extend this work to our spatio-temporal setting. Building on the idea of majorization-minimization, we construct convex surrogate functions that *majorize* $\mathcal{L}_{\text{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$ in each iteration of the proposed algorithm. Then, we show the minimization equivalence between the constructed majorizing functions and Eq. (4). These results are presented in theorems 1 and 3. Theorems 2 and 4 propose an efficient alternating optimization algorithm for solving $\mathcal{L}_{\text{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$, which leads to update rules for the spatial and temporal covariance matrices $\mathbf{\Gamma}$ and $\mathbf{B}$ as well as the noise covariance matrix $\mathbf{\Lambda}$.

Starting with the estimation of the temporal covariance based on the current source estimate, we can state the following theorem:



(a) (b)

Figure 1: (a) Probabilistic graphical model for the multi-task linear regression problem. (b) Geometric representation of the geodesic path between the pair of matrices $\{\mathbf{B}^k, \mathbf{M}^k_{\text{time}}\}$ on the P.D. manifold and the geometric mean between them, which is used to update $\mathbf{B}^{k+1}$.

Theorem 1. *Optimizing the non-convex Type-II cost function in Eq.* (4)*,* $\mathcal{L}_{kron}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$*, with respect to the temporal covariance matrix* $\mathbf{B}$ *is equivalent to optimizing the following convex surrogate function, which* majorizes *Eq.* (4)*:*

$$\mathcal{L}^{\text{time}}_{\text{conv}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B}) = \text{tr}\left((\mathbf{B}^k)^{-1}\mathbf{B}\right) + \text{tr}(\mathbf{M}^k_{\text{time}}\mathbf{B}^{-1}), \tag{5}$$

where $\mathbf{M}^k_{\text{time}}$ *is defined as*

$$\mathbf{M}^k_{\text{time}} := \frac{1}{MG}\sum_{g=1}^{G} \mathbf{Y}_g^\top \left(\mathbf{\Sigma}^k_{\mathbf{y}}\right)^{-1} \mathbf{Y}_g \ , \tag{6}$$

and where $\mathbf{\Gamma}^k$*,* $\mathbf{\Lambda}^k$*, and* $\mathbf{\Sigma}^k_{\mathbf{y}}$ *denote the source, noise and statistical model covariance matrices at the* k*-th iteration, which are treated as constants when optimizing over* $\mathbf{B}$*.*

Proof. A detailed proof is provided in Appendix B. □

The solution of Eq. (5) can be obtained by taking advantage of the Riemannian geometry of the search space. The following theorem holds.

Theorem 2. *The cost function* $\mathcal{L}^{\text{time}}_{\text{conv}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B})$ *is strictly geodesically convex with respect to the P.D. manifold and its minimum with respect to* $\mathbf{B}$ *can be attained by iterating the following update rule until convergence:*

$$\mathbf{B}^{k+1} \leftarrow (\mathbf{B}^k)^{1/2}\left((\mathbf{B}^k)^{-1/2}\mathbf{M}^k_{\text{time}}(\mathbf{B}^k)^{-1/2}\right)^{1/2}(\mathbf{B}^k)^{1/2} \ . \tag{7}$$

Proof. A detailed proof is provided in Appendix C. Moreover, a geometric representation of the geodesic path between the pair of matrices $\{\mathbf{B}^k, \mathbf{M}^k_{\text{time}}\}$ on the P.D. manifold and the geometric mean between them, representing the update for $\mathbf{B}^{k+1}$, is provided in Figure 1-(b). □

Remark 1. *A geometric interpretation of the update rule* (7) *is that it finds the geometric mean between a spatially whitened version of the empirical temporal covariance matrix,* $\mathbf{M}^k_{\text{time}}$*, and the model temporal covariance matrix from the previous iteration,* $\mathbf{B}^k$*.*

We now derive an update rule for the spatial source and noise covariance matrices using an analogous approach. To this end, we first construct a convex surrogate function that *majorizes* $\mathcal{L}_{\text{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$ in each iteration of the optimization algorithm by considering augmented variables comprising both the source and noise covariances, i.e., $\mathbf{H} := [\mathbf{\Gamma}, \mathbf{0}; \mathbf{0}, \mathbf{\Lambda}]$, and $\mathbf{\Phi} := [\mathbf{L}, \mathbf{I}]$. The following theorem holds.

Theorem 3. *Minimizing the non-convex Type-II cost function in Eq.* (4)*,* $\mathcal{L}_{kron}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$*, with respect to the spatial covariance matrix* $\mathbf{H}$ *is equivalent to minimizing the following convex surrogate function, which* majorizes *Eq.* (4)*:*

$$\mathcal{L}^{\text{space}}_{\text{conv}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B}^k) = \mathcal{L}^{\text{space}}_{\text{conv}}(\mathbf{H}, \mathbf{B}^k) = \text{tr}\left(\mathbf{\Phi}^\top(\mathbf{\Sigma}^k_{\mathbf{y}})^{-1}\mathbf{\Phi}\mathbf{H}\right) + \text{tr}\left(\mathbf{M}^k_{\text{SN}}\mathbf{H}^{-1}\right) \ , \tag{8}$$

where $\mathbf{M}_{\mathrm{SN}}^k$ *is defined as*

$$\mathbf{M}_{\mathrm{SN}}^k := \mathbf{H}^k \mathbf{\Phi}^\top (\mathbf{\Sigma}_{\mathbf{y}}^k)^{-1} \mathbf{M}_{\mathrm{space}}^k (\mathbf{\Sigma}_{\mathbf{y}}^k)^{-1} \mathbf{\Phi} \mathbf{H}^k \text{ , with}$$

$$\mathbf{M}_{\mathrm{space}}^k := \frac{1}{TG} \sum_{g=1}^{G} \mathbf{Y}_g (\mathbf{B}^k)^{-1} \mathbf{Y}_g^\top \ , \tag{9}$$

and where $\mathbf{B}^k$ *denotes the temporal model covariance matrix estimated in the* k*-th iteration.*

Proof. A detailed proof is provided in Appendix D. □

While, in principle, update rules for full-structured spatial source and noise covariances may be conceived, in analogy to the ones presented for the temporal covariance, we here restrict ourselves to the discussion of diagonal spatial covariances. Note, though, that such a choice of prior does not prohibit the reconstruction of parameters with more complex correlation structure. We have $\mathbf{H} = \mathrm{diag}(\mathbf{h})$, $\mathbf{h} = [\gamma_1, \ldots, \gamma_N, \sigma_1^2, \ldots, \sigma_M^2]^\top$, and $\mathbf{\Sigma}_{\mathbf{y}} = \mathbf{\Phi}\mathbf{H}\mathbf{\Phi}^\top$. The update rule for $\mathbf{H}$ then takes a simple form, as stated in the following theorem:

Theorem 4. *The cost function* $\mathcal{L}_{\mathrm{conv}}^{\mathrm{space}}(\mathbf{H}, \mathbf{B}^k)$ *is convex in* $\mathbf{h}$*, and its minimum with respect to* $\mathbf{h}$ *can be obtained according to the following closed-form update rule, which concurrently estimates the scalar source variances and heteroscedastic noise variances:*

$$\mathbf{H}^{k+1} = \mathrm{diag}(\mathbf{h}^{k+1}),\ h_i^{k+1} \leftarrow \sqrt{\frac{g_i^k}{z_i^k}} \quad \textit{for } i = 1, \ldots, N+M \textit{ , where} \tag{10}$$

$$\mathbf{g} := \mathrm{diag}(\mathbf{M}_{\mathrm{SN}}^k) \tag{11}$$

$$\mathbf{z} := \mathrm{diag}(\mathbf{\Phi}^\top (\mathbf{\Sigma}_{\mathbf{y}}^k)^{-1} \mathbf{\Phi}) \tag{12}$$

Proof. A detailed proof can be found in Appendix E. □

The final estimate of the posterior distribution (2) can be obtained by starting from any initial guess for the model parameters, $\mathbf{H}^0 = \{\mathbf{\Gamma}^0, \mathbf{\Lambda}^0\}$ and $\mathbf{B}^0$, and iterating between update rules (7), (10), and (2) until convergence. We call the resulting algorithm *full Dugh*. Convergence is shown in the following theorem:

Theorem 5. *Optimizing the non-convex ML cost function Eq.* (4) *with alternating update rules for* $\mathbf{B}$ *and* $\mathbf{H} = \{\mathbf{\Gamma}, \mathbf{\Lambda}\}$ *in Eq.* (7) *and Eq.* (10) *defines an MM algorithm, which is guaranteed to converge to a stationary point.*

Proof. A detailed proof can be found in Appendix G. □

4 Efficient inference for stationary temporal dynamics — thin Dugh

With full Dugh, we could present a general inference method for data with full temporal covariance structure. This algorithm may be used for non-stationary data, for which the covariance between two time samples depends on the absolute position of the two samples within the studied time window. An example is the reconstruction of brain responses to repeated external stimuli from M/EEG (event-related electrical potential or magnetic field) data, where data blocks are aligned to the stimulus onset. In other realistic settings, however, stationarity can be assumed in the sense that the covariance between two samples only depends on their relative position (distance) to each other but not to any external trigger. An example is the reconstruction of ongoing brain dynamics in absence of known external triggers. This situation can be adequately modeled using temporal covariance matrices with Toeplitz structure. In the following, we devise an efficient extension of full Dugh for that setting.

Temporal correlation has been incorporated in various brain source imaging models through different priors [see 13, 28, and references therein]. According to [29], the temporal dynamics of neuronal populations can be approximated by a first order auto-regressive (AR(1)) model of the form $x_n(t+1) = \beta x_n(t) + \sqrt{1-\beta^2}\xi_n(t),\ n = 1, \ldots, N; t = 1, \ldots, T$ with AR coefficient $\beta \in (-1, 1)$ and innovation noise $\xi_n(t)$. It can be shown that the temporal correlation matrix corresponding to this AR

model has Toeplitz structure, $\mathbf{B}_{i,j} = \beta^{|i-j|}$ [30]. Consequently, we now constrain the cost function Eq. (5) to the set of Toeplitz matrices:

$$\mathbf{B}^{k+1} = \underset{\mathbf{B}\in\mathcal{B},\ \mathbf{H}=\mathbf{H}^k}{\arg\min}\ \mathrm{tr}((\mathbf{B}^k)^{-1}\mathbf{B}) + \mathrm{tr}(\mathbf{M}^k_{\mathrm{time}}\mathbf{B}^{-1})\ , \tag{13}$$

where set $\mathcal{B}$ denotes the set of real-valued positive-definite Toeplitz matrices of size $T \times T$.

In order to be able to derive efficient update rules for this setting, we bring $\mathbf{B}$ into a diagonal form. The following proposition holds.

Proposition 1. *Let $\mathbf{P} \in \mathbb{R}^{L\times L}$ with $L > T$, be the circulant embedding of matrix $\mathbf{B} \in \mathcal{B}^L$, where $\mathcal{B}^L$ is assumed to be the subset of $\mathcal{B}$ that guarantees that $\mathbf{P}$ is a real-valued circulant P.D. matrix. Then the Toeplitz matrix $\mathbf{B}$ can be diagonalized as follows:*

$$\mathbf{B} = \mathbf{Q}\mathbf{P}\mathbf{Q}^H \quad \textit{with} \quad \mathbf{Q} = [\mathbf{I}_M, \mathbf{0}]\mathbf{F}_L,\ \textit{and} \tag{14}$$

$$[\mathbf{F}_L]_{m,l} = \frac{1}{\sqrt{L}} e^{\frac{i2\pi(l-1)}{L}(m-1)}\ , \tag{15}$$

where $\mathbf{P} = \mathrm{diag}(\mathbf{p}) = \mathrm{diag}(p_0, p_1, \ldots, p_{L-1})$ is a diagonal matrix. The main diagonal coefficients of $\mathbf{P}$ are given as the normalized discrete Fourier transform (DFT), represented by the linear operator $\mathbf{F}_L$, of the first row of $\mathbf{B}$: $\mathbf{p} = \mathbf{F}_L\mathbf{B}_{1.}$ (Note that a Toeplitz matrix can be represented by its first row or column, $\mathbf{B}_{1.}$).

Proof. This is a direct implication of the fact that Toeplitz matrices can be embedded into circulant matrices of larger size [31] and the result that circulant matrices can be approximately diagonalized using the Fourier transform [32]. Further details can be found in Appendix H. □

The solution of Eq. (13) can be obtained by direct application of Proposition 1. The results is summarized in the following theorem:

Theorem 6. *The cost function Eq.* (13) *is convex in $\mathbf{p}$, and its minimum with respect to $\mathbf{p}$ can be obtained by iterating the following closed-form update rule until convergence:*

$$p_l^{k+1} \leftarrow \sqrt{\frac{\hat{g}_l^k}{\hat{z}_l^k}}\ \textit{for}\ l = 1, \ldots, L\ , \textit{where} \tag{16}$$

$$\hat{\mathbf{g}} := \mathrm{diag}(\mathbf{P}^k\mathbf{Q}^H(\mathbf{B}^k)^{-1}\mathbf{M}^k_{\mathrm{time}}(\mathbf{B}^k)^{-1}\mathbf{Q}\mathbf{P}^k) \tag{17}$$

$$\hat{\mathbf{z}} := \mathrm{diag}(\mathbf{Q}^H(\mathbf{B}^k)^{-1}\mathbf{Q})\ . \tag{18}$$

Proof. A detailed proof is provided in Appendix I. □

4.1 Efficient computation of the posterior

Given estimates of the spatial and temporal covariance matrices, we can efficiently compute the posterior mean by exploiting their intrinsic diagonal structure.

Theorem 7. *Given the diagonalization $\mathbf{B} = \mathbf{Q}\mathbf{P}\mathbf{Q}^H$ of the temporal correlation matrix and the eigenvalue decomposition $\mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top = \mathbf{U}_\mathbf{x}\mathbf{D}_\mathbf{x}\mathbf{U}_\mathbf{x}^\top$ of $\mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top$, where $\mathbf{D}_\mathbf{x} = \mathrm{diag}(d_1, \ldots, d_M)$, the posterior mean is efficiently computed as*

$$\bar{\mathbf{x}}_g = (\mathbf{\Gamma}\otimes\mathbf{B})\,\mathbf{D}^\top\tilde{\mathbf{\Sigma}}_\mathbf{y}^{-1}\mathbf{y}_g = \mathrm{tr}\left(\mathbf{Q}\mathbf{P}\left(\mathbf{\Pi}\odot\mathbf{Q}^H\mathbf{Y}_g^\top\mathbf{U}_\mathbf{x}\right)\left(\mathbf{U}_\mathbf{x}^\top\mathbf{L}\mathbf{\Gamma}^\top\right)\right)\ , \tag{19}$$

where $\odot$ denotes the Hadamard product between the corresponding elements of two matrices of equal size. In addition, $\mathbf{\Pi}$ is defined as follows: $[\mathbf{\Pi}]_{l,m} = \frac{1}{\sigma_m^2 + p_l d_m}$ for $l = 1, \ldots, L$ and $m = 1, \ldots, M$.

Proof. A detailed proof is provided in Appendix J. □

The resulting algorithm, obtained by iterating between update rules (16), (10) and (19), is called *thin Dugh* (as opposed to *full Dugh* introduced above).

5 Simulations

We present two sets of experiments to assess the performance of our proposed methods. In the first experiment, we compare the reconstruction performance of the proposed Dugh algorithm variants to that of Champagne [33] and two other competitive methods – eLORETA [34] and S-FLEX [35] – for a range of SNR levels, numbers of time samples, and orders of AR coefficients. In the second experiment, we test the impact of model violations on the temporal covariance estimation. All experiments are performed using `Matlab` on a machine with a 2.50 GHz Intel(R) Xeon(R) Platinum 8160 CPU. The codes are publicly available at https://github.com/AliHashemi-ai/Dugh-NeurIPS-2021.

5.1 Pseudo-EEG signal generation and benchmark comparison

We simulate a sparse set of $N_0 = 3$ active sources placed at random locations on the cortex. To simulate the electrical neural activity, we sample time series of length $T \in \{10, 20, 50, 100\}$ from a univariate linear autoregressive AR(P) process. We use stable AR systems of order $P \in \{1, 2, 5, 7\}$. The resulting source distribution is then projected to the EEG sensors, denoted by $\mathbf{Y}^{\text{signal}}$, using a realistic lead field matrix, $\mathbf{L} \in \mathbb{R}^{58\times 2004}$. We generate $\mathbf{L}$ using the New York Head model [36] taking into account the realistic anatomy and electrical tissue conductivities of an average human head. Finally, we add Gaussian white noise to the sensor space signal. Note that the simulated noise and source time courses do not share a similar temporal structure here – sources are modeled with a univariate autoregressive AR(P) process while a temporally white Gaussian distribution is used for modeling noise. Thus, we could assess the robustness of our proposed method under violation of the model assumption that the temporal structure of sources and noise is similar. The resulting noise matrix $\mathbf{E} = [\mathbf{e}(1), \ldots, \mathbf{e}(T)]$ is first normalized and then added to the signal matrix $\mathbf{Y}^{\text{signal}}$ as follows: $\mathbf{Y} = \mathbf{Y}^{\text{signal}} + \frac{(1-\alpha)\|\mathbf{Y}^{\text{signal}}\|_F}{\alpha\|\mathbf{E}\|_F}\mathbf{E}$, where α determines the signal-to-noise ratio (SNR) in sensor space. Precisely, SNR is defined as $\text{SNR} = 20\log_{10}\left(\alpha/_{1-\alpha}\right)$. In this experiment the following values of α are used: $\alpha \in \{0.55, 0.65, 0.7, 0.8\}$, which correspond to the following SNRs: SNR $\in \{1.7, 5.4, 7.4, 12\}$ (dB). Interested readers can refer to Appendix K and [37] for further details on the simulation framework. We quantify the performance of all algorithms using the *earth mover's distance* (EMD) [21, 38] and the maximal correlation between the time courses of the simulated and the reconstructed sources (TCE). Each simulation is carried out 100 times using different instances of $\mathbf{X}$ and $\mathbf{E}$, and the mean and standard error of the mean (SEM) of each performance measure across repetitions is calculated.

In order to investigate the impact of model violations on the temporal covariance estimation, we generate a random Gaussian source matrix, $\mathbf{X} \in \mathbb{R}^{2004\times 30\times G}$ representing the brain activity of 2004 brain sources at 30 time instances for different numbers of trials $G \in 10, 20, 30, 40, 50$. In all trials, sources are randomly sampled from a zero-mean normal distribution with spatio-temporal covariance matrix $\mathbf{\Gamma} \otimes \mathbf{B}$, where $\mathbf{B} \in \mathbb{R}^{30\times 30}$ is either a full-structural random PSD matrix or a Toeplitz matrix with $\beta = 0.8$. Gaussian noise $\mathbf{E}$ sharing the same temporal covariance with the sources is added to the measurements $\mathbf{Y} = \mathbf{LX}$, so that the overall SNR is 0 dB. We evaluate the accuracy of the temporal covariance reconstruction using Thin and full Dugh. The performance is evaluated using two measures: Pearson correlation between the original and reconstructed temporal covariance matrices, $\mathbf{B}$ and $\hat{\mathbf{B}}$, denoted by $r(\mathbf{B}, \hat{\mathbf{B}})$, and the normalized mean squared error (NMSE) defined as: NMSE $= ||\hat{\mathbf{B}} - \mathbf{B}||_F^2 / ||\mathbf{B}||_F^2$. The similarity error is defined as: $1 - r(\mathbf{B}, \hat{\mathbf{B}})$. Note that NMSE measures the reconstruction at the true scale of the temporal covariance; while $r(\mathbf{B}, \hat{\mathbf{B}})$ is scale-invariant and hence only quantifies the overall structural similarity between simulated and estimated noise covariance matrices.

In Figure 2, we show the source reconstruction performance (mean $\pm$ SEM) of the four different brain source imaging schemes, namely thin Dugh, Champagne, eLORETA and S-FLEX. We notice that Dugh achieves superior performance in terms of EMD metric, whereas it is competitive in terms of TCE. Note that since thin Dugh incorporates the temporal structure of the sources into the inference scheme, its performance with respect to EMD and TCE can be significantly improved by increasing the number of time samples. Figure 3 demonstrates the estimated temporal covariance matrix obtained from our two proposed spatio-temporal learning schemes, namely thin and full Dugh, indicated by cyan and magenta colors, respectively. The upper panel illustrates the reconstruction results for a

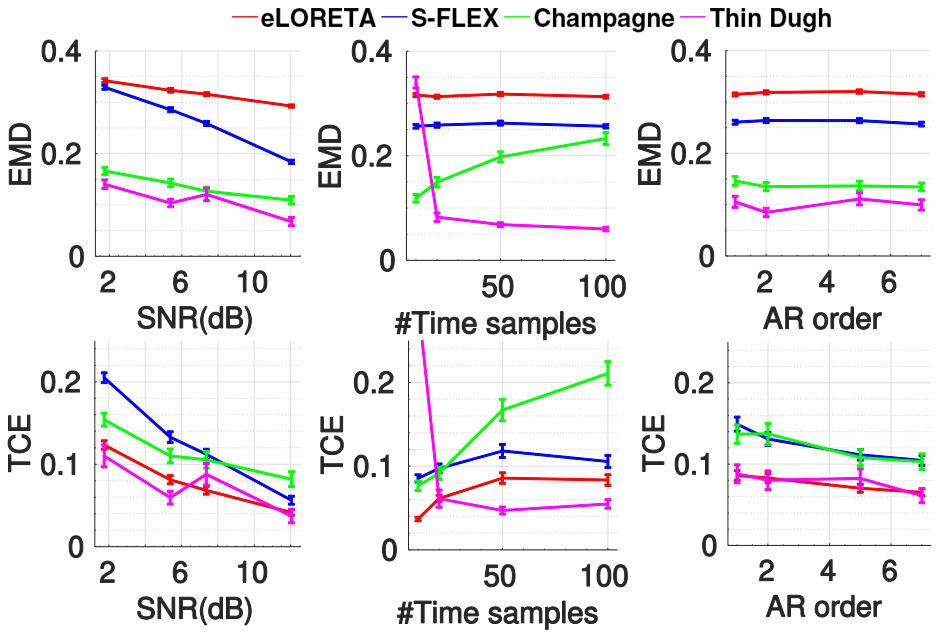


Figure 2: Source reconstruction performance (mean ± SEM) of the four different brain source imaging schemes, namely Thin Dugh, Champagne, eLORETA, and S-FLEX, for data generated by a realistic lead field matrix. Performance is assessed for different settings including a wide range of SNRs, different numbers of time samples, and different AR model orders. Performance is evaluated in terms of the earth mover's distance (EMD) and time-course correlation error (TCE) between each simulated source and the reconstructed source with highest maximum absolute correlation.

setting where the ground-truth temporal covariance matrix has full structure, while the lower panel shows a case with Toeplitz temporal covariance matrix structure. It can be seen that *full Dugh* can better capture the overall structure of ground truth full-structure temporal covariance as evidenced by lower NMSE and similarity errors compared to *thin Dugh* that is only able to recover a Toeplitz matrix. As can be expected, the behavior is reversed in the lower-panel of Figure 3, where the ground truth temporal covariance matrix is indeed Toeplitz. Full Dugh, however, still provides reasonable performance in terms of NMSE and similarity error, even though it estimates a full-structure temporal covariance matrix. Finally, the performance of both learning schemes are significantly improved by increasing the number of trials.

6 Analysis of real MEG and EEG data recordings

We now demonstrate the performance of the novel algorithms on two MEG datasets (see Figure 4 and 5) and one EEG dataset (see Appendix L). All participants provided informed written consent prior to study participation and received monetary compensation for their participation. The studies were approved by the University of California, San Francisco Committee on Human Research. The MEG datasets from two female subjects comprised five trials of visual or auditory stimulus presentation, where the goal was to reconstruct cortical activity reflecting auditory and visual processing. No prior work has shown success in reconstruction in such extreme low SNR data.

Figure 4 shows the reconstructed sources for auditory evoked fields (AEF) from a representative subject using eLORETA, MCE, thin and full Dugh. In this case, we tested the reconstruction performance of all algorithms with the number of trials limited to 5. As Figure 4 demonstrates, the reconstructions of thin and full Dugh both show focal sources at the expected locations of the auditory cortex. Limiting the number of trials to as few as 5 does not negatively influence the reconstruction result of Dugh methods, while it severely affects the reconstruction performance of competing methods. For the visual evoked field (VEF) data in Figure 5, thin Dugh was able to reconstruct two sources in the visual cortex with their corresponding time-courses that demonstrate an early peak at 100 ms and a later peak around 200 ms (similar performance was observed for full Dugh).

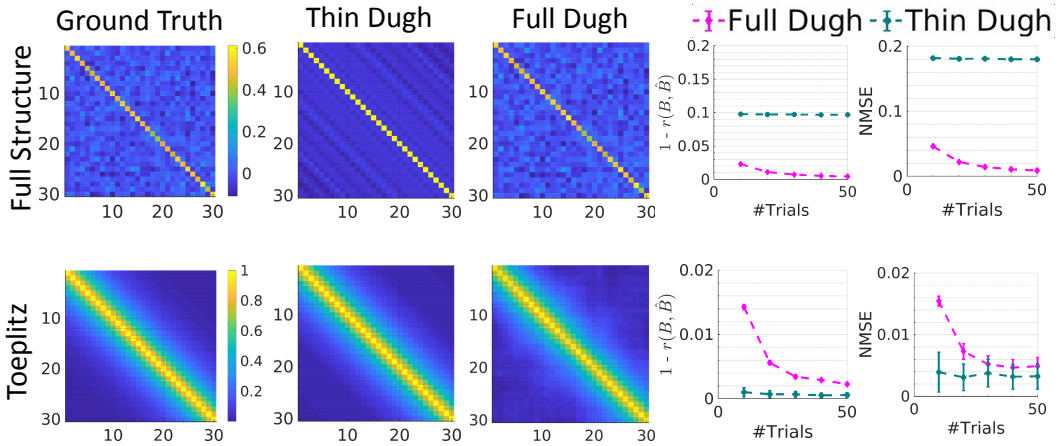


Figure 3: Accuracy of the temporal covariance matrix reconstruction incurred by two different temporal learning approaches, namely thin (cyan) and full (magenta) Dugh, assuming Toeplitz and full temporal covariance structure, respectively. The ground-truth temporal covariance matrix (first column) has full covariance structure in the first row, and Toeplitz structure in the second row. Performance is assessed in terms of the Pearson correlation between the entries of the original and reconstructed temporal covariance matrices, $\mathbf{B}$ and $\hat{\mathbf{B}}$, denoted by $r(\mathbf{B}, \hat{\mathbf{B}})$. Shown is the similarity error $1 - r(\mathbf{B}, \hat{\mathbf{B}})$ (forth column). Further, the normalized mean squared error (NMSE) between $\mathbf{B}$ and $\hat{\mathbf{B}}$, defined as NMSE $= ||\hat{\mathbf{B}} - \mathbf{B}||_F^2 / ||\mathbf{B}||_F^2$ is reported (fifth column).

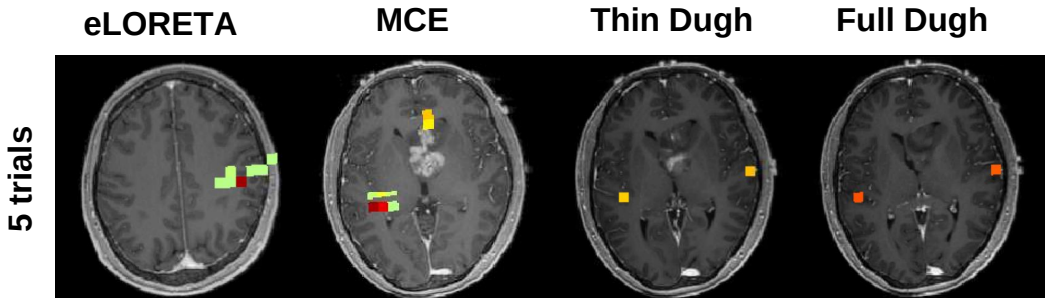


Figure 4: Analysis of auditory evoked fields (AEF) of one representative subject using eLORETA, MCE, thin and full Dugh. As it can be seen, thin and full Dugh can correctly localize bilateral auditory activity to Heschl's gyrus, which is the characteristic location of the primary auditory cortex, with as few as 5 trials. In this challenging setting, all competing methods show inferior performance.

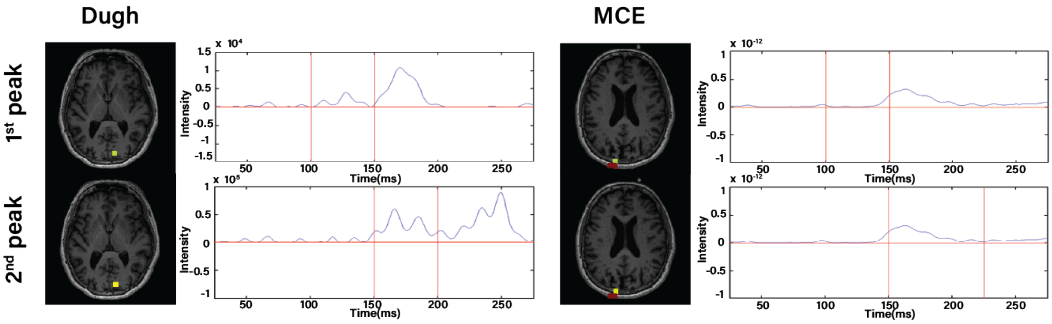


Figure 5: Spatial maps and corresponding time-courses reconstructed from five trials of visual stimulus presentation using Dugh and MCE. Distinct activity in visual cortex was reconstructed.

7 Discussion

Inverse modeling is a challenging problem that can be cast as a multi-task regression problem. Machine learning-based methods have contributed by systematically incorporating prior information into such models (e.g. [13, 25, 39–41]). While, so far, explicit models of spatio-temporal dynamics are rare, we contribute in this paper by deriving efficient inference algorithms for regression models with spatio-temporal dynamics in model parameters and noise. Specifically, we employ separable Gaussian distributions using Kronecker products of temporal and spatial covariance matrices. We assume sparsity and independent activations in the spatial domain, while the temporal covariance is modeled to have either full or Toeplitz structure. The proposed Dugh framework is encompassing efficient optimization algorithms for jointly estimating these distributions within a hierarchical Bayesian inference and MM optimization framework. Interestingly, we could theoretically prove convergence for the proposed update rules yielding estimates of the spatial and temporal covariances. In careful simulation studies, we have demonstrated that the inclusion of both spatial and temporal model parameters in the model indeed leads to significantly improved reconstruction performance. Finally, the utility of our algorithms is showcased in challenging real-world applications by reconstructing expected sources from real M/EEG data based on very few experimental trials.

The M/EEG BSI problem has been previously framed as a multi-task regression problem in the context of Type-I learning methods [5, 42]. Bertrand et al. [5] proposed a method to extend the group Lasso class of algorithms to the multi-task learning case. Their method shows promising results for large numbers of trials but is not competitive for smaller sample sizes, presumably due to not modeling the temporal structure of the sources. In contrast, Dugh is a Type-II method that learns the full spatio-temporal prior source distribution as part of the model fitting. Note that the assumption of spatially independent noise made here can be readily dropped, as in [43]. A number of alternative approaches have been proposed to estimate the spatio-temporal correlation structure of the *noise* [44–48]. These works, however, do not estimate the noise characteristics as part of the source reconstruction problem on the same data but require separate noise recordings. Our proposed algorithm substantially differs in this respect, as it learns the full noise covariance jointly with the brain source distribution. This joint estimation perspective is very much in line with the end-to-end learning philosophy as opposed to a step-wise independent estimation process that can give rise to error accumulation.

With respect to the limits of our Dugh framework, we would like to note that using the same temporal correlation prior for noise and sources is a potentially restricting assumption in our modeling. It is made here to achieve tractable inference due to the fact that the spatio-temporal statistical model covariance matrix can be formulated in a separable form, i.e., $\tilde{\mathbf{\Sigma}}_{\mathbf{y}} = \mathbf{\Sigma}_{\mathbf{y}} \otimes \mathbf{B}$. Although we have demonstrated empirically that the reconstruction results of our proposed learning schemes are fairly robust against violating this assumption, this constraint may be further relaxed by exploiting eigenvalue decomposition techniques presented in [2, 49]. Another potentially limiting assumption in our model is to assume Gaussian distributions for the sources and noise. Although Gaussian priors are commonly justified, the extension of our framework to heavy-tailed noise distributions, which are more robust to outliers, is a direction of future work. Regarding the societal impact of this work, we note that we intend to solve inverse problems that have non-unique solutions, which depend on explicit or implicit assumptions of priors. These have applications in basic, clinical, and translational neuroscience imaging studies. Our use of a hierarchical empirical Bayesian framework allows for explicit specifications of the priors in our model that are learned from data, and a more clear interpretation of the reconstructions with reduced bias. Nevertheless, in our algorithms, we assume sparse spatial priors, and in scenarios where this assumption embodied in the priors is incorrect, the resulting reconstructions will be inaccurate. Users and neuroscientists will need to be cognizant of these issues.

In conclusion, we could derive novel flexible hierarchical Bayesian algorithms for multi-task regression problems with spatio-temporal covariances. Incorporating prior knowledge, namely, constraining the solutions to Riemannian manifolds, and using ideas of geodesic convexity, circulant embeddings, and majorization-minimization, we derive inference update rules and prove convergence guarantees. The proposed Dugh algorithms show robust and competitive performance both on synthetic and real neural data from M/EEG recordings and thus contribute to a well-founded solution to complex inverse problems in neuroscience.

Broader Impact

In this paper we focus on sparse multi-task linear regression within the hierarchical Bayesian regression framework and its application in EEG/MEG brain source imaging. Our algorithm, however, is suitable for a wider range of applications. The same concepts used here for full-structural spatio-temporal covariance learning could be employed in other contexts where hyperparameters like kernel widths in Gaussian process regression [50] or dictionary elements in the dictionary learning problem [51] need to be inferred from data. Our proposed method, Dugh, may also prove useful for practical scenarios in which model residuals and signals are expected to be correlated, e.g., probabilistic canonical correlation analysis (CCA) [52], spectral independent component analysis (ICA) [53], direction of arrival (DoA) and channel estimation in massive Multiple Input Multiple Output (MIMO) systems [54–56], robust portfolio optimization in finance [57], covariance matching and estimation [58–62], graph learning [63], thermal field reconstruction [64–66], and brain functional imaging [67].

Acknowledgments and Disclosure of Funding

This result is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 758985). AH acknowledges scholarship support from the Machine Learning/Intelligent Data Analysis research group at Technische Universität Berlin. He further wishes to thank the Charité – Universitätsmedizin Berlin, the Berlin Mathematical School (BMS), and the Berlin Mathematics Research Center MATH+ for partial support. CC was supported by the National Natural Science Foundation of China under Grant 62007013. KRM was funded by the German Ministry for Education and Research as BIFOLD – Berlin Institute for the Foundations of Learning and Data (ref. 01IS18025A and ref. 01IS18037A), and the German Research Foundation (DFG) as Math+: Berlin Mathematics Research Center (EXC 2046/1, project-ID: 390685689), Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea Government (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University). SSN was funded in part by National Institutes of Health grants (R01DC004855, R01EB022717, R01DC176960, R01DC010145, R01NS100440, R01AG062196, and R01DC013979), University of California MRPI MRP-17–454755, the US Department of Defense grant (W81XWH-13-1-0494).

References

[1] Richard M Shiffrin, Michael D Lee, Woojae Kim, and Eric-Jan Wagenmakers. A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, 32(8):1248–1284, 2008.

[2] Barbara Rakitsch, Christoph Lippert, Karsten Borgwardt, and Oliver Stegle. It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 1*, pages 1466–1474, 2013.

[3] Emmanuel J Candès, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[4] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4): 1289–1306, 2006.

[5] Quentin Bertrand, Mathurin Massias, Alexandre Gramfort, and Joseph Salmon. Handling correlated and repeated measurements with the smoothed multivariate square-root Lasso. In *Advances in Neural Information Processing Systems*, pages 3959–3970, 2019.

[6] Jérôme-Alexis Chevalier, Alexandre Gramfort, Joseph Salmon, and Bertrand Thirion. Statistical control for spatio-temporal MEG/EEG source imaging with desparsified multi-task lasso. *arXiv preprint arXiv:2009.14310*, 2020.

[7] Dmitry Malioutov, Müjdat Cetin, and Alan S Willsky. A sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Transactions on Signal Processing*, 53(8): 3010–3022, 2005.

[8] Ming Bo Cai, Michael Shvartsman, Anqi Wu, Hejia Zhang, and Xia Zhu. Incorporating structured assumptions with probabilistic graphical models in fMRI data analysis. *Neuropsychologia*, page 107500, 2020.

[9] S Beirle, U Platt, Mr Wenig, and T Wagner. Weekly cycle of NO2 by GOME measurements: a signature of anthropogenic sources. *Atmospheric Chemistry and Physics*, 3(6):2225–2232, 2003.

[10] Juri Ranieri, Alessandro Vincenzi, Amina Chebira, David Atienza, and Martin Vetterli. Near-optimal thermal monitoring framework for many-core systems-on-chip. *IEEE Transactions on Computers*, 64(11):3197–3209, 2015.

[11] Steven Lemm, Benjamin Blankertz, Gabriel Curio, and K-R Muller. Spatio-spectral filters for improving the classification of single trial eeg. *IEEE transactions on biomedical engineering*, 52(9):1541–1548, 2005.

[12] Daniel Bartz and Klaus-Robert Müller. Covariance shrinkage for autocorrelated data. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[13] Wei Wu, Srikantan Nagarajan, and Zhe Chen. Bayesian machine learning: EEG\MEG signal processing measurements. *IEEE Signal Processing Magazine*, 33(1):14–36, 2016.

[14] Sylvain Baillet, John C Mosher, and Richard M Leahy. Electromagnetic brain mapping. *IEEE Signal Processing Magazine*, 18(6):14–30, 2001.

[15] Mingbo Cai, Nicolas W Schuck, Jonathan W Pillow, and Yael Niv. A Bayesian method for reducing bias in neural representational similarity analysis. In *Advances in Neural Information Processing Systems*, pages 4951–4959, 2016.

[16] Michael Shvartsman, Narayanan Sundaram, Mikio Aoi, Adam Charles, Theodore Willke, and Jonathan Cohen. Matrix-normal models for fMRI analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 1914–1923. PMLR, 2018.

[17] Ming Bo Cai, Nicolas W Schuck, Jonathan W Pillow, and Yael Niv. Representational structure or task structure? Bias in neural representational similarity analysis and a Bayesian method for reducing bias. *PLoS computational biology*, 15(5):e1006299, 2019.

[18] Shane F Cotter, Bhaskar D Rao, Kjersti Engan, and Kenneth Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Transactions on signal processing*, 53(7):2477–2488, 2005.

[19] Matti Hämäläinen, Riitta Hari, Risto J Ilmoniemi, Jukka Knuutila, and Olli V Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of modern Physics*, 65(2):413, 1993.

[20] Roberto D Pascual-Marqui, Christoph M Michel, and Dietrich Lehmann. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. *International Journal of psychophysiology*, 18(1):49–65, 1994.

[21] Stefan Haufe, Vadim V Nikulin, Andreas Ziehe, Klaus-Robert Müller, and Guido Nolte. Combining sparsity and rotational invariance in EEG/MEG source reconstruction. *NeuroImage*, 42(2):726–738, 2008.

[22] Alexandre Gramfort, Matthieu Kowalski, and Matti Hämäläinen. Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods. *Physics in medicine and biology*, 57(7):1937, 2012.

[23] Sebastian Mika, Gunnar Rätsch, and Klaus-Robert Müller. A mathematical programming approach to the kernel fisher algorithm. *Advances in neural information processing systems*, 13:591–597, 2000.

[24] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1(June):211–244, 2001.

[25] David Wipf and Srikantan Nagarajan. A unified Bayesian framework for MEG/EEG source imaging. *NeuroImage*, 44(3):947–966, 2009.

[26] Ali Hashemi, Chang Cai, Gitta Kutyniok, Klaus-Robert Müller, Srikantan Nagarajan, and Stefan Haufe. Unification of sparse Bayesian learning algorithms for electromagnetic brain imaging with the majorization minimization framework. *NeuroImage*, 239:118309, 2021.

[27] Ying Sun, Prabhu Babu, and Daniel P Palomar. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing*, 65(3):794–816, 2017.

[28] Elvira Pirondini, Behtash Babadi, Gabriel Obregon-Henao, Camilo Lamus, Wasim Q Malik, Matti S Hamalainen, and Patrick L Purdon. Computationally-efficient algorithms for sparse, dynamic solutions to the eeg source localization problem. *IEEE Transactions on Biomedical Engineering*, 2017.

[29] Camilo Lamus, Matti S Hämäläinen, Simona Temereanca, Emery N Brown, and Patrick L Purdon. A spatiotemporal dynamic distributed solution to the MEG inverse problem. *NeuroImage*, 63(2):894–909, 2012.

[30] Zhilin Zhang and Bhaskar D Rao. Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning. *IEEE Journal of Selected Topics in Signal Processing*, 5(5):912–926, 2011.

[31] Amir Dembo, Colin L Mallows, and Lawrence A Shepp. Embedding nonnegative definite Toeplitz matrices in nonnegative definite circulant matrices, with application to covariance estimation. *IEEE Transactions on Information Theory*, 35(6):1206–1212, 1989.

[32] Ulf Grenander and Gabor Szegö. *Toeplitz forms and their applications*. Univ of California Press, 1958.

[33] David P Wipf, Julia P Owen, Hagai T Attias, Kensuke Sekihara, and Srikantan S Nagarajan. Robust Bayesian estimation of the location, orientation, and time course of multiple correlated neural sources using MEG. *NeuroImage*, 49(1):641–655, 2010.

[34] Roberto D Pascual-Marqui. Discrete, 3D distributed, linear imaging methods of electric neuronal activity. part 1: exact, zero error localization. *arXiv preprint arXiv:0710.3341*, 2007.

[35] Stefan Haufe, Ryota Tomioka, Thorsten Dickhaus, Claudia Sannelli, Benjamin Blankertz, Guido Nolte, and Klaus-Robert Müller. Large-scale EEG/MEG source localization with spatial flexibility. *NeuroImage*, 54(2):851–859, 2011.

[36] Yu Huang, Lucas C Parra, and Stefan Haufe. The New York head—a precise standardized volume conductor model for EEG source localization and tes targeting. *NeuroImage*, 140: 150–162, 2016.

[37] Stefan Haufe and Arne Ewald. A simulation framework for benchmarking EEG-based brain connectivity estimation methodologies. *Brain topography*, pages 1–18, 2016.

[38] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *Int J Comput Vision*, 40:99–121, 2000.

[39] Michael T McCann, Kyong Hwan Jin, and Michael Unser. Convolutional neural networks for inverse problems in imaging: A review. *IEEE Signal Processing Magazine*, 34(6):85–95, 2017.

[40] Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

[41] Tatiana A Bubba, Gitta Kutyniok, Matti Lassas, Maximilian Maerz, Wojciech Samek, Samuli Siltanen, and Vignesh Srinivasan. Learning the invisible: a hybrid deep learning-shearlet framework for limited angle computed tomography. *Inverse Problems*, 35(6):064002, 2019.

[42] Mathurin Massias, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Generalized concomitant multi-task lasso for sparse multimodal regression. In *International Conference on Artificial Intelligence and Statistics*, pages 998–1007, 2018.

[43] Ali Hashemi, Chang Cai, Klaus Robert Muller, Srikantan Nagarajan, and Stefan Haufe. Joint learning of full-structure noise in hierarchical Bayesian regression models, 2021. URL `https://openreview.net/forum?id=yRP4_BOxdu`.

[44] Hilde M Huizenga, Jan C De Munck, Lourens J Waldorp, and Raoul PPP Grasman. Spatiotemporal EEG/MEG source analysis based on a parametric noise covariance model. *IEEE Transactions on Biomedical Engineering*, 49(6):533–539, 2002.

[45] Jan Casper De Munck, Hilde M Huizenga, Lourens J Waldorp, and RA Heethaar. Estimating stationary dipoles from MEG/EEG data contaminated with spatially and temporally correlated background noise. *IEEE Transactions on Signal Processing*, 50(7):1565–1572, 2002.

[46] Fetsje Bijma, Jan C De Munck, Hilde M Huizenga, and Rob M Heethaar. A mathematical approach to the temporal stationarity of background noise in MEG/EEG measurements. *NeuroImage*, 20(1):233–243, 2003.

[47] Jan Casper De Munck, Fetsje Bijma, Pawel Gaura, Cezary Andrzej Sieluzycki, Maria Inês Branco, and Rob M Heethaar. A maximum-likelihood estimator for trial-to-trial variations in noisy MEG/EEG data sets. *IEEE Transactions on Biomedical Engineering*, 51(12):2123–2128, 2004.

[48] Sung C Jun, Sergey M Plis, Doug M Ranken, and David M Schmidt. Spatiotemporal noise covariance estimation from limited empirical magnetoencephalographic data. *Physics in Medicine & Biology*, 51(21):5549, 2006.

[49] Anqi Wu, Stan Pashkovski, Sandeep R Datta, and Jonathan W Pillow. Learning a latent manifold of odor representations from neural responses in piriform cortex. *Advances in Neural Information Processing Systems*, 31:5378–5388, 2018.

[50] Anqi Wu, Oluwasanmi Koyejo, and Jonathan W Pillow. Dependent relevance determination for smooth and structured sparse regression. *J. Mach. Learn. Res.*, 20(89):1–43, 2019.

[51] Onur Dikmen and Cédric Févotte. Maximum marginal likelihood estimation for nonnegative dictionary learning in the Gamma-Poisson model. *IEEE Transactions on Signal Processing*, 60(10):5163–5175, 2012.

[52] Francis R Bach and Michael I Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

[53] Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. Spectral independent component analysis with noise modeling for M/EEG source separation. *arXiv preprint arXiv:2008.09693*, 2020.

[54] Ranjitha Prasad, Chandra R Murthy, and Bhaskar D Rao. Joint channel estimation and data detection in MIMO-OFDM systems: A sparse Bayesian learning approach. *IEEE Transactions on Signal Processing*, 63(20):5369–5382, 2015.

[55] Peter Gerstoft, Christoph F Mecklenbräuker, Angeliki Xenaki, and Santosh Nannuru. Multisnapshot sparse Bayesian learning for DOA. *IEEE Signal Processing Letters*, 23(10): 1469–1473, 2016.

[56] Saeid Haghighatshoar and Giuseppe Caire. Massive MIMO channel subspace estimation from low-dimensional projections. *IEEE Transactions on Signal Processing*, 65(2):303–318, 2017.

[57] Yiyong Feng, Daniel P Palomar, et al. A signal processing perspective on financial engineering. *Foundations and Trends® in Signal Processing*, 9(1–2):1–231, 2016.

[58] Karl Werner, Magnus Jansson, and Petre Stoica. On estimation of covariance matrices with Kronecker product structure. *IEEE Transactions on Signal Processing*, 56(2):478–491, 2008.

[59] Theodoros Tsiligkaridis and Alfred O Hero. Covariance estimation in high dimensions via Kronecker product expansions. *IEEE Transactions on Signal Processing*, 61(21):5347–5360, 2013.

[60] Abdelhak M Zoubir, Visa Koivunen, Esa Ollila, and Michael Muma. *Robust statistics for signal processing*. Cambridge University Press, 2018.

[61] Alessandro Benfenati, Emilie Chouzenoux, and J-C Pesquet. Proximal approaches for matrix optimization problems: Application to robust precision matrix estimation. *Signal Processing*, 169:107417, 2020.

[62] Esa Ollila, Daniel P Palomar, and Frédéric Pascal. Shrinking the eigenvalues of M-estimators of covariance matrix. *arXiv preprint arXiv:2006.10005*, 2020.

[63] Sandeep Kumar, Jiaxi Ying, José Vinícius de Miranda Cardoso, and Daniel P Palomar. A unified framework for structured graph learning via spectral constraints. *Journal of Machine Learning Research*, 21(22):1–60, 2020.

[64] Ali Hashemi, Mohammad Rostami, and Ngai-Man Cheung. Efficient environmental temperature monitoring using compressed sensing. In *2016 Data Compression Conference (DCC)*, pages 602–602. IEEE, 2016.

[65] Axel Flinth and Ali Hashemi. Thermal source localization through infinite-dimensional compressed sensing. *arXiv preprint arXiv:1710.02016*, 2017.

[66] Axel Flinth and Ali Hashemi. Approximate recovery of initial point-like and instantaneous sources from coarsely sampled thermal fields via infinite-dimensional compressed sensing. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1720–1724. IEEE, 2018.

[67] Huilin Wei, Amirhossein Jafarian, Peter Zeidman, Vladimir Litvak, Adeel Razi, Dewen Hu, and Karl J Friston. Bayesian fusion and multimodal DCM for EEG and fMRI. *NeuroImage*, 211:116595, 2020.

[68] Yonina C Eldar and Moshe Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on information theory*, 55(11):5302–5316, 2009.

[69] Yuzhe Jin and Bhaskar D Rao. Support recovery of sparse signals in the presence of multiple measurement vectors. *IEEE Transactions on Information Theory*, 59(5):3139–3157, 2013.

[70] Piya Pal and PP Vaidyanathan. Pushing the limits of sparse support recovery using correlation information. *IEEE Transactions on Signal Processing*, 63(3):711–726, 2014.

[71] Rajendra Bhatia. *Positive definite matrices*, volume 24. Princeton University Press, 2009.

[72] Pourya Zadeh, Reshad Hosseini, and Suvrit Sra. Geometric mean metric learning. In *International conference on machine learning*, pages 2464–2471, 2016.

[73] Ami Wiesel, Teng Zhang, et al. Structured robust covariance estimation. *Foundations and Trends® in Signal Processing*, 8(3):127–216, 2015.

[74] Athanase Papadopoulos. *Metric spaces, convexity and nonpositive curvature*, volume 6. European Mathematical Society, 2005.

[75] Tamas Rapcsak. Geodesic convexity in nonlinear optimization. *Journal of Optimization Theory and Applications*, 69(1):169–183, 1991.

[76] Aharon Ben-Tal. On generalized means and generalized convex functions. *Journal of Optimization Theory and Applications*, 21(1):1–13, 1977.

[77] Leo Liberti. On a class of nonconvex problems where all local minima are global. *Publications de l'Institut Mathémathique*, 76(90):101–109, 2004.

[78] Diethard Ernst Pallaschke and Stefan Rolewicz. *Foundations of mathematical optimization: convex analysis without linearity*, volume 388. Springer Science & Business Media, 2013.

[79] Silvere Bonnabel and Rodolphe Sepulchre. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31(3): 1055–1070, 2009.

[80] Maher Moakher. A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 26(3):735–747, 2005.

[81] Suvrit Sra and Reshad Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, 25(1):713–739, 2015.

[82] Nisheeth K Vishnoi. Geodesic convex optimization: Differentiation on manifolds, geodesics, and convexity. *arXiv preprint arXiv:1806.06373*, 2018.

[83] Constantin Niculescu and Lars-Erik Persson. *Convex functions and their applications*. Springer, 2006.

[84] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007.

[85] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[86] Chang Cai, Ali Hashemi, Mithun Diwakar, Stefan Haufe, Kensuke Sekihara, and Srikantan S Nagarajan. Robust estimation of noise for electromagnetic brain imaging with the Champagne algorithm. *NeuroImage*, 225:117411, 2021.

[87] David R Hunter and Kenneth Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37, 2004.

[88] Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

[89] Matthew W Jacobson and Jeffrey A Fessler. An expanded theoretical treatment of iteration-dependent majorize-minimize algorithms. *IEEE Transactions on Image Processing*, 16(10): 2411–2422, 2007.

[90] Tong Tong Wu, Kenneth Lange, et al. The MM alternative to EM. *Statistical Science*, 25(4): 492–505, 2010.

[91] Eric Mjolsness and Charles Garrett. Algebraic transformations of objective functions. *Neural Networks*, 3(6):651–669, 1990.

[92] Alan L Yuille and Anand Rangarajan. The concave-convex procedure. *Neural computation*, 15(4):915–936, 2003.

[93] Thomas Lipp and Stephen Boyd. Variations and extension of the convex–concave procedure. *Optimization and Engineering*, 17(2):263–287, 2016.

[94] Dylan Fagot, Herwig Wendt, Cédric Févotte, and Paris Smaragdis. Majorization-minimization algorithms for convolutive NMF with the beta-divergence. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8202–8206. IEEE, 2019.

[95] Mahdi Barzegar Khalilsarai, Tianyu Yang, Saeid Haghighatshoar, and Giuseppe Caire. Structured channel covariance estimation from limited samples in Massive MIMO. In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pages 1–7. IEEE, 2020.

[96] Alexander Fengler, Giuseppe Caire, Peter Jung, and Saeid Haghighatshoar. Massive MIMO unsourced random access. *arXiv preprint arXiv:1901.00828*, 2019.

[97] Alexander Fengler, Saeid Haghighatshoar, Peter Jung, and Giuseppe Caire. Non-Bayesian activity detection, large-scale fading coefficient estimation, and unsourced random access with a massive MIMO receiver. *arXiv preprint arXiv:1910.11266*, 2019.

[98] Martin Luessi, Matti S Hämäläinen, and Victor Solo. Sparse component selection with application to MEG source localization. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 556–559. IEEE, 2013.

[99] Ali Hashemi and Stefan Haufe. Improving EEG source localization through spatio-temporal sparse Bayesian learning. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1935–1939. IEEE, 2018.

[100] Yousra Bekhti, Felix Lucka, Joseph Salmon, and Alexandre Gramfort. A hierarchical Bayesian perspective on majorization-minimization for non-convex sparse regression: application to M/EEG source imaging. *Inverse Problems*, 34(8):085010, 2018.

[101] Prabhu Babu. MELT—maximum-likelihood estimation of low-rank Toeplitz covariance matrix. *IEEE Signal Processing Letters*, 23(11):1587–1591, 2016.

[102] Petre Stoica, Randolph L Moses, et al. *Spectral analysis of signals*, volume 1. Pearson Prentice Hall Upper Saddle River, NJ, 2005.

[103] Robert M Gray et al. Toeplitz and circulant matrices: A review. *Foundations and Trends® in Communications and Information Theory*, 2(3):155–239, 2006.

[104] Ying Sun, Prabhu Babu, and Daniel P Palomar. Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions. *IEEE Transactions on Signal Processing*, 64 (14):3576–3590, 2016.

[105] Yunus Saatçi. Scalable inference for structured gaussian process models. *Ph.D. thesis, University of Cambridge Cambridge, UK.*, 2012.

[106] Arno Solin, Pasi Jylänki, Jaakko Kauramäki, Tom Heskes, Marcel AJ van Gerven, and Simo Särkkä. Regularizing solutions to the MEG inverse problem using space-time separable covariance functions. *arXiv preprint arXiv:1604.04931*, 2016.

[107] R. Oostenveld and P. Praamstra. The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, 112(4):713–719, 2001. ISSN 1388-2457. doi: 10.1016/s1388-2457(00)00527-7.

[108] Alexandre Gramfort. *Mapping, timing and tracking cortical activations with MEG and EEG: Methods and application to human vision*. PhD thesis, Ecole nationale supérieure des telecommunications-ENST, 2009.

[109] Kanta Matsuura and Yoichi Okabe. Selective minimum-norm solution of the biomagnetic inverse problem. *IEEE Transactions on Biomedical Engineering*, 42(6):608–615, 1995.

[110] Sarang S Dalal, JM Zumer, V Agrawal, KE Hild, Kensuke Sekihara, and SS Nagarajan. NUTMEG: a neuromagnetic source reconstruction toolbox. *Neurology & Clinical Neurophysiology: NCN*, 2004:52, 2004.

[111] Sarang S Dalal, Johanna M Zumer, Adrian G Guggisberg, Michael Trumpis, Daniel DE Wong, Kensuke Sekihara, and Srikantan S Nagarajan. MEG/EEG source reconstruction, statistical evaluation, and visualization with NUTMEG. *Computational Intelligence and Neuroscience*, 2011, 2011.

[112] Julia P Owen, David P Wipf, Hagai T Attias, Kensuke Sekihara, and Srikantan S Nagarajan. Performance evaluation of the Champagne source reconstruction algorithm on simulated and real M/EEG data. *Neuroimage*, 60(1):305–323, 2012.

[113] Chang Cai, Mithun Diwakar, Dan Chen, Kensuke Sekihara, and Srikantan S Nagarajan. Robust empirical Bayesian reconstruction of distributed sources for electromagnetic brain imaging. *IEEE Transactions on Medical Imaging*, 39(3):567–577, 2019.

Supplementary Material: Efficient hierarchical Bayesian inference for spatio-temporal regression models in neuroimaging

Summary of the proposed algorithms and derived update rules

We proposed two algorithms in the main paper, namely full and thin Dugh, which are summarized in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1: Full Dugh

Input : The lead field matrix $\mathbf{L} \in \mathbb{R}^{M \times N}$ and G trials of measurement vectors $\{\mathbf{Y}_g\}_{g=1}^{G}$, where $\mathbf{Y}_g \in \mathbb{R}^{M \times T}$.

Result: Estimates of the source and noise variances $\mathbf{h} = [\gamma_1, \ldots, \gamma_N, \sigma_1^2, \ldots, \sigma_M^2]^\top$, the temporal covariance $\mathbf{B}$, and the posterior mean $\{\bar{\mathbf{x}}_g\}_{g=1}^{G}$ and covariance $\mathbf{\Sigma_x}$ of the sources.

Choose a random initial value for $\mathbf{B}$ as well as $\mathbf{h} = [\gamma_1, \ldots, \gamma_N, \sigma_1^2, \ldots, \sigma_M^2]^\top$, and construct $\mathbf{H} = \mathrm{diag}(\mathbf{h})$ and $\mathbf{\Gamma} = \mathrm{diag}([\gamma_1, \ldots, \gamma_N]^\top)$.

Construct the augmented lead field $\mathbf{\Phi} = [\mathbf{L}, \mathbf{I}_M]$.

Calculate the lead field $\mathbf{D} = \mathbf{L} \otimes \mathbf{I}_T$ for vectorized sources.

Calculate the prior spatio-temporal covariance for the sources as $\mathbf{\Sigma}_0 = \mathbf{\Gamma} \otimes \mathbf{B}$.

Calculate the spatial statistical covariance $\mathbf{\Sigma_y} = \mathbf{\Phi H \Phi}^\top$.

Calculate the spatio-temporal statistical covariance $\tilde{\mathbf{\Sigma}}_\mathbf{y} = \mathbf{\Sigma_y} \otimes \mathbf{B}$.

Initialize $k \leftarrow 1$

repeat

- Calculate the posterior mean as $\bar{\mathbf{x}}_g = \mathbf{\Sigma}_0 \mathbf{D}^\top \tilde{\mathbf{\Sigma}}_\mathbf{y}^{-1} \mathbf{y}_g$, for $g = 1, \ldots, G$, where $\mathbf{y}_g = \mathrm{vec}\left(\mathbf{Y}_g^\top\right) \in \mathbb{R}^{MT \times 1}$.
- Calculate $\mathbf{M}_{\mathrm{time}}^k$ based on Eq. (6), and update $\mathbf{B}$ based on Eq. (7) according to the Riemannian update on the manifold of P.D. matrices.
- Calculate $\mathbf{M}_{\mathrm{SN}}^k$ based on Eq. (9), and update $\mathbf{H}$ based on Eq. (10).
- $k \leftarrow k + 1$

until *stopping condition is satisfied:* $\left\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\right\|_2^2 \leq \epsilon$ *or* $k = k_{max}$;

Calculate the posterior covariance as $\mathbf{\Sigma_x} = \mathbf{\Sigma}_0 - \mathbf{\Sigma}_0 \mathbf{D}^\top \tilde{\mathbf{\Sigma}}_\mathbf{y}^{-1} \mathbf{D} \mathbf{\Sigma}_0$.

Algorithm 2: Thin Dugh

Input : The lead field matrix $\mathbf{L} \in \mathbb{R}^{M \times N}$, and G trials of measurement vectors $\{\mathbf{Y}_g\}_{g=1}^{G}$, where $\mathbf{Y}_g \in \mathbb{R}^{M \times T}$.

Result: Estimates of the source and noise variances $\mathbf{h} = [\gamma_1, \ldots, \gamma_N, \sigma_1^2, \ldots, \sigma_M^2]^\top$, the temporal covariance $\mathbf{B}$, and the posterior mean $\{\bar{\mathbf{x}}_g\}_{g=1}^{G}$.

Choose a random initial value for $\mathbf{p}$ as well as $\mathbf{h}$, and construct $\mathbf{H} = \mathrm{diag}(\mathbf{h})$ and $\mathbf{P} = \mathrm{diag}(\mathbf{p})$.

Construct $\mathbf{B} = \mathbf{Q}\mathbf{P}\mathbf{Q}^H$, where $\mathbf{Q} = [\mathbf{I}_M, \mathbf{0}]\mathbf{F}_L$ with $L = 2T + 1$ and $\mathbf{F}_L$ as DFT.

Construct the augmented lead field $\mathbf{\Phi} := [\mathbf{L}, \mathbf{I}_M]$.

Calculate the lead field $\mathbf{D} = \mathbf{L} \otimes \mathbf{I}_T$ for vectorized sources.

Calculate the statistical covariance $\mathbf{\Sigma_y} = \mathbf{\Phi}\mathbf{H}\mathbf{\Phi}^\top$.

Calculate the statistical covariance $\mathbf{\Sigma_y} = \mathbf{\Phi}\mathbf{H}\mathbf{\Phi}^\top$.

Calculate the spatio-temporal statistical covariance $\tilde{\mathbf{\Sigma}}_\mathbf{y} = \mathbf{\Sigma_y} \otimes \mathbf{B}$.

Initialize $k \leftarrow 1$

repeat

Calculate the posterior mean efficiently based on Eq. (19) as $\bar{\mathbf{x}}_g = \mathrm{tr}\left(\mathbf{Q}\mathbf{P}\left(\mathbf{\Pi} \odot \mathbf{Q}^H \mathbf{Y}_g^\top \mathbf{U_x}\right)\left(\mathbf{U_x}^\top \mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top\right)\right)$, where $\mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top = \mathbf{U_x}\mathbf{D_x}\mathbf{U_x}^\top$ and $[\mathbf{\Pi}]_{l,m} = \frac{1}{\sigma_m^2 + p_l d_m}$ for $l = 1, \ldots, L$ and $m = 1, \ldots, M$.

Calculate $\mathbf{M}_{\mathrm{time}}^k$ based on Eq. (6), and update $\mathbf{B}$ based on Eq. (16) according to Riemannian update for Toeplitz matrices using circulant embedding.

Calculate $\mathbf{M}_{\mathrm{SN}}^k$ based on Eq. (9), and update $\mathbf{H}$ based on Eq. (10).

$k \leftarrow k + 1$

until *stopping condition is satisfied:* $\left\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\right\|_2^2 \leq \epsilon$ *or* $k = k_{max}$;

Calculate the posterior covariance as $\mathbf{\Sigma_x} = \mathbf{\Sigma}_0 - \mathbf{\Sigma}_0 \mathbf{D}^\top \tilde{\mathbf{\Sigma}}_\mathbf{y}^{-1} \mathbf{D} \mathbf{\Sigma}_0$.

A Derivation of Type-II Bayesian cost function for full-structural spatio-temporal models

In this section, we provide a detailed derivation of Type-II Bayesian learning for full-structural spatio-temporal models. To this end, we first briefly explain the *multiple measurement vector* (MMV) model and then formulate Type-II Bayesian learning with full-structural spatio-temporal covariance structure for this setting. Note that, to simplify the problem, we first present the derivations of the MMV model only for a single trial. We later extend this simplified setting to the multi-trials case.

A.1 Multiple measurement vector (MMV) model

In M/EEG brain source imaging, a sequence of measurement vectors are often available. Thus, the following *multiple measurement vector (MMV)* model can be formulated1:

$$\mathbf{Y} = \mathbf{L}\mathbf{X} + \mathbf{E}\ ,$$

where $\mathbf{Y} = [\mathbf{y}(1), \ldots, \mathbf{y}(T)] \in \mathbb{R}^{M \times T}$ consists of T measurement vectors for a sequence of T time samples. $\mathbf{X} = [\mathbf{x}(1), \ldots, \mathbf{x}(T)] \in \mathbb{R}^{N \times T}$ is the desired solution matrix (the amplitude of N brain sources during T time samples in our setting), and $\mathbf{E}$ is an unknown noise matrix. A key assumption in the MMV model is that the support (i.e., the indices of the nonzero entries) of every column in $\mathbf{X}$ is identical (referred to as the *common sparsity assumption* in the literature). The number of nonzero rows in $\mathbf{X}$ needs to be below a threshold to ensure unique and global solution. This implies that $\mathbf{X}$ has only a small number of non-zero rows. It has been shown that the recovery of the support can be greatly improved by increasing the number of measurements [68–70].

A.2 Type-II Bayesian cost function for full-structural spatio-temporal models

To exploit temporal correlations between measurements, we first assume that the voxels are mutually independent. Given the column vector $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_N]^\top$ and a Gaussian probability density for each brain source, the prior distribution with time correlation is modeled as follows:

$$p(\mathbf{X}_i | \gamma_i, \mathbf{B}) \sim \mathcal{N}(0, \gamma_i \mathbf{B}), \qquad i = 1, \ldots, N\ . \tag{20}$$

$\mathbf{X}_i$ denotes the i-th row of source matrix $\mathbf{X}$ and models the probability distribution of the i-th brain source. Note that γ_i is a non-negative hyper-parameter that controls the row sparsity of $\mathbf{X}$; i.e., the values of source $\mathbf{X}_i$ become all zero if $\gamma_i = 0$. Finally, $\mathbf{B}$ is a positive definite matrix that captures the time correlation structure, which is assumed to be shared across all sources. The goal is to obtain the prior distribution of sources, $p(\mathbf{X}|\boldsymbol{\gamma}, \mathbf{B})$, by estimating the hyper-parameters, $\{\boldsymbol{\gamma}, \mathbf{B}\}$. Next, we reformulate the joint MMV model of all sources using vectorization of matrices and Kronecker product operations:

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{e} \ ,$$

where $\mathbf{y} = \text{vec}\left(\mathbf{Y}^\top\right) \in \mathbb{R}^{MT\times 1}$, $\mathbf{x} = \text{vec}\left(\mathbf{X}^\top\right) \in \mathbb{R}^{NT\times 1}$, $\mathbf{e} = \text{vec}\left(\mathbf{E}^\top\right) \in \mathbb{R}^{MT\times 1}$ and $\mathbf{D}$=$\mathbf{L} \otimes \mathbf{I}_T$.

The prior distribution of $\mathbf{x}$ is given as

$$p(\mathbf{x}|\boldsymbol{\gamma}, \mathbf{B}) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_0)$$

where $\boldsymbol{\Sigma}_0$ is defined as

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} \gamma_1 \mathbf{B} & & \\ & \ddots & \\ & & \gamma_N \mathbf{B} \end{bmatrix} = \boldsymbol{\Gamma} \otimes \mathbf{B} \ ,$$

in which $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\gamma}) = \text{diag}(\gamma_1, \ldots, \gamma_N)$.

Similarly, we may assume zero-mean Gaussian noise with covariance $\boldsymbol{\Sigma}_\mathbf{e} = \boldsymbol{\Lambda} \otimes \boldsymbol{\Upsilon}$, where $\mathbf{e} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\mathbf{e})$, and $\boldsymbol{\Lambda}$ and $\boldsymbol{\Upsilon}$ denote the spatial and temporal noise covariance matrices, respectively. Here, we use the same prior for the temporal structure of noise and sources, i.e., $\boldsymbol{\Upsilon} = \mathbf{B}$.

The parameters of the spatio-temporal Type-II model are the unknown source, noise and temporal covariance matrices, i.e., $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}$, and $\mathbf{B}$. The unknown parameters $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}$, and $\mathbf{B}$ are optimized based on the current estimates of the source, noise and temporal covariances in an alternating iterative process. Given initial estimates of $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}$, and $\mathbf{B}$, the posterior distribution of the sources is a Gaussian of the form $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B}) \sim \mathcal{N}(\bar{\mathbf{x}}, \boldsymbol{\Sigma}_\mathbf{x})$, whose mean and covariance are obtained as follows:

$$\bar{\mathbf{x}} = \boldsymbol{\Sigma}_0 \mathbf{D}^\top (\boldsymbol{\Lambda} \otimes \mathbf{B} + \mathbf{D}\boldsymbol{\Sigma}_0\mathbf{D}^\top)^{-1}\mathbf{y} = \boldsymbol{\Sigma}_0 \mathbf{D}^\top \tilde{\boldsymbol{\Sigma}}_\mathbf{y}^{-1}\mathbf{y} \ , \tag{21}$$

$$\boldsymbol{\Sigma}_\mathbf{x} = \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0 \mathbf{D}^\top \tilde{\boldsymbol{\Sigma}}_\mathbf{y}^{-1} \mathbf{D}\boldsymbol{\Sigma}_0 \ , \tag{22}$$

where $\boldsymbol{\Sigma}_\mathbf{y} = \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top + \boldsymbol{\Lambda}$, and where $\tilde{\boldsymbol{\Sigma}}_\mathbf{y} = \boldsymbol{\Sigma}_\mathbf{y} \otimes \mathbf{B}$ denotes the spatio-temporal variant of statistical model covariance matrix. The estimated posterior parameters $\bar{\mathbf{x}}$ and $\boldsymbol{\Sigma}_\mathbf{x}$ are then in turn used to update $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}$, and $\mathbf{B}$ as the minimizers of the negative log of the marginal likelihood $p(\mathbf{Y}|\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B})$, which is given by

$$\mathcal{L}_{\text{kron}}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B}) = \log|\tilde{\boldsymbol{\Sigma}}_\mathbf{y}| + \text{tr}\left(\mathbf{y}^\top \tilde{\boldsymbol{\Sigma}}_\mathbf{y}^{-1}\mathbf{y}\right) \ . \tag{23}$$

Using the same temporal covariance prior for noise and sources, i.e., $\boldsymbol{\Upsilon} = \mathbf{B}$, the statistical model covariance matrix, $\tilde{\boldsymbol{\Sigma}}_\mathbf{y}$, can be written as:

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}}_\mathbf{y} &= \boldsymbol{\Lambda} \otimes \boldsymbol{\Upsilon} + (\mathbf{D}\boldsymbol{\Sigma}_0\mathbf{D}^\top) = \boldsymbol{\Lambda} \otimes \boldsymbol{\Upsilon} + \left(\left(\mathbf{L} \otimes \mathbf{I}^\top\right)(\boldsymbol{\Gamma} \otimes \mathbf{B})\left(\mathbf{L} \otimes \mathbf{I}^\top\right)^\top\right) \\ &= \boldsymbol{\Lambda} \otimes \boldsymbol{\Upsilon} + \left(\mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top \otimes \mathbf{B}\right) \overset{(\boldsymbol{\Upsilon}=\mathbf{B})}{=} (\boldsymbol{\Lambda} + \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top) \otimes \mathbf{B} \\ &= \boldsymbol{\Sigma}_\mathbf{y} \otimes \mathbf{B} \ , \ . \end{aligned} \tag{24}$$

which leads to the following spatio-temporal Type-II Bayesian learning cost function:

$$\begin{aligned} \mathcal{L}_{\text{kron}}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B}) &= \log|\tilde{\boldsymbol{\Sigma}}_\mathbf{y}| + \text{tr}\left(\mathbf{y}^\top \tilde{\boldsymbol{\Sigma}}_\mathbf{y}^{-1}\mathbf{y}\right) \\ &= \log|\boldsymbol{\Sigma}_\mathbf{y} \otimes \mathbf{B}| + \text{tr}\left(\mathbf{y}^\top \left(\boldsymbol{\Sigma}_\mathbf{y} \otimes \mathbf{B}\right)^{-1}\mathbf{y}\right) \\ &= \log\left(|\boldsymbol{\Sigma}_\mathbf{y}|^T |\mathbf{B}|^M\right) + \text{tr}\left(\mathbf{y}^\top \left(\boldsymbol{\Sigma}_\mathbf{y} \otimes \mathbf{B}\right)^{-1}\mathbf{y}\right) \ . \end{aligned} \tag{25}$$

Here, we assume the presence of G sample blocks $\mathbf{Y}_g \in \mathbb{R}^{M\times T}$, for $g = 1, \ldots, G$. These block samples can be obtained by segmenting a time series into smaller parts that are assumed to independent

and identically distributed. These blocks may represent epochs, trials or experimentl tasks depending on the applications. $\mathcal{L}_{\text{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$ can then be reformulated as

$$\mathcal{L}_{\text{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B}) = T \log |\mathbf{\Sigma_y}| + M \log |\mathbf{B}| + \frac{1}{G} \sum_{g=1}^{G} \text{tr}(\mathbf{\Sigma_y}^{-1} \mathbf{Y}_g \mathbf{B}^{-1} \mathbf{Y}_g^\top) \tag{26}$$

by applying the following matrix equality to Eq. (25):

$$\text{tr}(\mathbf{A}^\top \mathbf{B} \mathbf{C} \mathbf{D}^\top) = \text{vec}(\mathbf{A})^\top (\mathbf{D} \otimes \mathbf{B}) \text{vec}(\mathbf{C}) \ .$$

B Proof of Theorem 1

Proof. We start by recalling $\mathcal{L}_{\text{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$ in Eq. (4):

$$\mathcal{L}_{\text{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B}) = T \log |\mathbf{\Sigma_y}| + M \log |\mathbf{B}| + \frac{1}{G} \sum_{g=1}^{G} \text{tr}(\mathbf{\Sigma_y}^{-1} \mathbf{Y}_g \mathbf{B}^{-1} \mathbf{Y}_g^\top).$$

Let $\mathbf{\Sigma}_\mathbf{y}^k$, $\mathbf{\Gamma}^k$, and $\mathbf{\Lambda}^k$ be the values of statistical model covariance and the source and noise covariances at the k-th iteration, respectively. By ignoring terms that do not depend on $\mathbf{B}$, $\mathcal{L}_{\text{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$ can be written as follows:

$$\begin{aligned}
\mathcal{L}_{\text{kron}}^{\text{time}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B}) &= M \log |\mathbf{B}| + \frac{1}{G} \sum_{g=1}^{G} \text{tr} \left(\left(\mathbf{\Sigma}_\mathbf{y}^k \right)^{-1} \mathbf{Y}_g \mathbf{B}^{-1} \mathbf{Y}_g^\top \right) \\
&= \log |\mathbf{B}| + \frac{1}{MG} \sum_{g=1}^{G} \text{tr} \left(\left(\mathbf{\Sigma}_\mathbf{y}^k \right)^{-1} \mathbf{Y}_g \mathbf{B}^{-1} \mathbf{Y}_g^\top \right) \\
&= \log |\mathbf{B}| + \text{tr} \left(\mathbf{B}^{-1} \frac{1}{MG} \sum_{g=1}^{G} \mathbf{Y}_g^\top \left(\mathbf{\Sigma}_\mathbf{y}^k \right)^{-1} \mathbf{Y}_g \right) \\
&= \log |\mathbf{B}| + \text{tr} \left(\mathbf{B}^{-1} \mathbf{M}_{\text{time}}^k \right) \ ,
\end{aligned} \tag{27}$$

where $\mathbf{M}_{\text{time}}^k := \frac{1}{MG} \sum_{g=1}^{G} \mathbf{Y}_g^\top \left(\mathbf{\Sigma}_\mathbf{y}^k \right)^{-1} \mathbf{Y}_g$.

By virtue of the concavity of the log-determinant function and its first order Taylor expansion around $\mathbf{B}^k$, the following inequality holds:

$$\begin{aligned}
\mathcal{L}_{\text{kron}}^{\text{time}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B}) &= \log |\mathbf{B}| + \text{tr} \left(\mathbf{B}^{-1} \mathbf{M}_{\text{time}}^k \right) \\
&\leq \log \left| \mathbf{B}^k \right| + \text{tr} \left(\left(\mathbf{B}^k \right)^{-1} \left(\mathbf{B} - \mathbf{B}^k \right) \right) + \text{tr} \left(\mathbf{B}^{-1} \mathbf{M}_{\text{time}}^k \right) \\
&= \log \left| \mathbf{B}^k \right| + \text{tr} \left(\left(\mathbf{B}^k \right)^{-1} \mathbf{B} \right) - \text{tr} \left(\left(\mathbf{B}^k \right)^{-1} \mathbf{B}^k \right) + \text{tr} \left(\mathbf{B}^{-1} \mathbf{M}_{\text{time}}^k \right) \\
&= \text{tr} \left((\mathbf{B}^k)^{-1} \mathbf{B} \right) + \text{tr}(\mathbf{B}^{-1} \mathbf{M}_{\text{time}}^k) + \text{const} \\
&= \mathcal{L}_{\text{conv}}^{\text{time}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B}) + \text{const} \ .
\end{aligned} \tag{28}$$

Note that constant values in (28) do not depend on $\mathbf{B}$; hence, they can be ignored in the optimization procedure. Hence, we have shown that minimizing Eq. (4) with respect to $\mathbf{B}$ is equivalent to minimizing $\mathcal{L}_{\text{conv}}^{\text{time}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B})$, which concludes the proof. □

C Proof of Theorem 2

Before presenting the proof, the subsequent definitions and propositions are required:

Definition 1 (Geodesic path)**.** *Let $\mathcal{M}$ be a Riemannian manifold, i.e., a differentiable manifold whose tangent space is endowed with an inner product that defines local Euclidean structures. Then, a geodesic between two points on $\mathcal{M}$, denoted by $\mathbf{p}_0, \mathbf{p}_1 \in \mathcal{M}$, is defined as the shortest connecting path between those two points along the manifold, $\zeta_l(\mathbf{p}_0, \mathbf{p}_1) \in \mathcal{M}$ for $l \in [0, 1]$, where $l = 0$ and $l = 1$ defines the starting and end points of the path, respectively.*

In the current context, $\zeta_l(\mathbf{p}_0, \mathbf{p}_1)$ defines a geodesic curve on the P.D. manifold joining two P.D. matrices, $\mathbf{P}_0, \mathbf{P}_1 > 0$. The specific pair of matrices we will deal with is $\{\mathbf{B}^k, \mathbf{M}^k_{\text{time}}\}$.

Definition 2 (Geodesic on the P.D. manifold)**.** *Geodesics on the manifold of P.D. matrices can be shown to form a cone within the embedding space. We denote this manifold by $\mathcal{S}_{++}$. Assume two P.D. matrices $\mathbf{P}_0, \mathbf{P}_1 \in \mathcal{S}_{++}$. Then, for $l \in [0, 1]$, the geodesic curve joining $\mathbf{P}_0$ to $\mathbf{P}_1$ is defined as [71, Chapter. 6]:*

$$\xi_l(\mathbf{P}_0, \mathbf{P}_1) = (\mathbf{P}_0)^{\frac{1}{2}} \left((\mathbf{P}_0)^{-1/2} \mathbf{P}_1 (\mathbf{P}_0)^{-1/2} \right)^l (\mathbf{P}_0)^{\frac{1}{2}} \quad l \in [0, 1] \ . \tag{29}$$

Note that $\mathbf{P}_0$ and $\mathbf{P}_1$ are obtained as the starting and end points of the geodesic path by choosing $l = 0$ and $l = 1$, respectively. The midpoint of the geodesic, obtained by setting $l = \frac{1}{2}$, is called the *geometric mean*. Note that, according to Definition 2, the following equality holds :

$$\begin{aligned} \xi_l(\mathbf{B}_0, \mathbf{B}_1)^{-1} &= \left((\mathbf{B}_0)^{1/2} \left((\mathbf{B}_0)^{-1/2} \mathbf{B}_1 (\mathbf{B}_0)^{-1/2} \right)^l (\mathbf{B}_0)^{1/2} \right)^{-1} \\ &= \left((\mathbf{B}_0)^{-1/2} \left((\mathbf{B}_0)^{1/2} (\mathbf{B}_1)^{-1} (\mathbf{B}_0)^{1/2} \right)^l (\mathbf{B}_0)^{-1/2} \right) = \xi_l(\mathbf{B}_0^{-1}, \mathbf{B}_1^{-1}) \ . \end{aligned} \tag{30}$$

Definition 3 (Geodesic convexity)**.** *Let $\mathbf{p}_0$ and $\mathbf{p}_1$ be two arbitrary points on a subset $\mathcal{A}$ of a Riemannian manifold $\mathcal{M}$. Then, a real-valued function $f : \mathcal{A} \to \mathbb{R}$ with domain $\mathcal{A} \subset \mathcal{M}$ is called* geodesic convex *(g-convex) if the following relation holds:*

$$f\left(\zeta_l(\mathbf{p}_0, \mathbf{p}_1)\right) \leq l f(\mathbf{p}_0) + (1 - l) f(\mathbf{p}_1) \ , \tag{31}$$

where $l \in [0, 1]$ and $\zeta(\mathbf{p}_0, \mathbf{p}_1)$ denotes the geodesic path connecting two points $\mathbf{p}_0$ and $\mathbf{p}_1$ as defined in Definition 1. Thus, in analogy to classical convexity, the function f is g-convex if every geodesic $\zeta(\mathbf{p}_0, \mathbf{p}_1)$ of $\mathcal{M}$ between $\mathbf{p}_0, \mathbf{p}_1 \in \mathcal{A}$, lies in the g-convex set $\mathcal{A}$. Note that the set $\mathcal{A} \subset \mathcal{M}$ is called g-convex, if any geodesics joining an arbitrary pair of points lies completely in $\mathcal{A}$.

Remark 2. *Note that g-convexity is a generalization of classical (linear) convexity to non-Euclidean (non-linear) geometry and metric spaces. Therefore, it is straightforward to show that all convex functions in Euclidean geometry are also g-convex, where the geodesics between pairs of matrices are simply line segments:*

$$\zeta_l(\mathbf{p}_0, \mathbf{p}_1) = l\mathbf{p}_0 + (1 - l)\mathbf{p}_1 \ . \tag{32}$$

For the sake of brevity, we omit a detailed theoretical introduction of g-convexity, and only borrow a result from [72]. Interested readers are referred to [73, Chapter 1] for a gentle introduction to this topic, and [74, Chapter. 2]; [75–82] for more in-depth technical details. Now we are ready to state the proof, which parallels the one provided in Zadeh et al. [72, Theorem. 3].

Proof. We proceed in two steps. First, we consider P.D. manifolds and express (31) in terms of geodesic paths and functions that lie on this particular space. We then show that $\mathcal{L}^{\text{time}}_{\text{conv}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B})$ is strictly g-convex on this specific domain. In the second step, we then derive the update rule proposed in (7).

C.1 Part I: G-convexity of the majorizing cost function

We consider geodesics along the P.D. manifold by setting $\zeta_l(\mathbf{p}_0, \mathbf{p}_1)$ to $\xi_l(\mathbf{B}_0, \mathbf{B}_1)$ as presented in Definition 2, and define $f(.)$ to be $f(\mathbf{B}) = \text{tr}\left(\left(\mathbf{B}^k\right)^{-1} \mathbf{B} \right) + \text{tr}(\mathbf{M}^k_{\text{time}} \mathbf{B}^{-1})$, representing the cost function $\mathcal{L}^{\text{time}}_{\text{conv}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B})$.

We now show that $f(\mathbf{B})$ is strictly g-convex on this specific domain. For continuous functions as considered in this paper, fulfilling (31) for $f(\mathbf{B})$ and $\xi_l(\mathbf{B}_0, \mathbf{B}_1)$ with $l = 1/2$ is sufficient for strict g-convexity according to *mid-point convexity* [83]:

$$\begin{aligned} &\text{tr}\left(\left(\mathbf{B}^k\right)^{-1} \xi_{1/2}(\mathbf{B}_0, \mathbf{B}_1) \right) + \text{tr}\left(\mathbf{M}^k_{\text{time}} \xi_{1/2}(\mathbf{B}_0, \mathbf{B}_1)^{-1} \right) \\ &< \frac{1}{2}\text{tr}\left(\left(\mathbf{B}^k\right)^{-1} \mathbf{B}_0 \right) + \frac{1}{2}\text{tr}\left(\mathbf{M}^k_{\text{time}} {\mathbf{B}_0}^{-1} \right) \\ &\quad + \frac{1}{2}\text{tr}\left(\left(\mathbf{B}^k\right)^{-1} \mathbf{B}_1 \right) + \frac{1}{2}\text{tr}\left(\mathbf{M}^k_{\text{time}} {\mathbf{B}_1}^{-1} \right) \ . \end{aligned} \tag{33}$$

Given $\left(\mathbf{B}^k\right)^{-1} \in \mathcal{S}_{++}$, i.e., $\left(\mathbf{B}^k\right)^{-1} > 0$ and the operator inequality [71, Chapter. 4]

$$\xi_{1/2}(\mathbf{B}_0, \mathbf{B}_1) \prec \frac{1}{2}\mathbf{B}_0 + \frac{1}{2}\mathbf{B}_1 \ , \tag{34}$$

we have:

$$\mathrm{tr}\left(\left(\mathbf{B}^k\right)^{-1} \xi_{1/2}(\mathbf{B}_0, \mathbf{B}_1)\right) < \frac{1}{2}\mathrm{tr}\left(\left(\mathbf{B}^k\right)^{-1} \mathbf{B}_0\right) + \frac{1}{2}\mathrm{tr}\left(\left(\mathbf{B}^k\right)^{-1} \mathbf{B}_1\right) \ , \tag{35}$$

which is derived by multiplying both sides of Eq. (34) with $\left(\mathbf{B}^k\right)^{-1}$ followed by taking the trace on both sides.

Similarly, we can write the operator inequality for $\{\mathbf{B}_0^{-1}, \mathbf{B}_1^{-1}\}$ using Eq. (30) as:

$$\xi_{1/2}(\mathbf{B}_0, \mathbf{B}_1)^{-1} = \xi_{1/2}(\mathbf{B}_0^{-1}, \mathbf{B}_1^{-1}) \prec \frac{1}{2}\mathbf{B}_0^{-1} + \frac{1}{2}\mathbf{B}_1^{-1} \ . \tag{36}$$

Multiplying both sides of Eq. (36) by $\mathbf{M}_{\text{time}}^k \in \mathcal{S}_{++}$ and applying the trace operator on both sides leads to:

$$\mathrm{tr}\left(\mathbf{M}_{\text{time}}^k \xi_{1/2}(\mathbf{B}_0, \mathbf{B}_1)^{-1}\right) < \frac{1}{2}\mathrm{tr}\left(\mathbf{M}_{\text{time}}^k \mathbf{B}_0{}^{-1}\right) + \frac{1}{2}\mathrm{tr}\left(\mathbf{M}_{\text{time}}^k \mathbf{B}_1{}^{-1}\right) \ . \tag{37}$$

Summing up (35) and (37) proves inequality (33) and concludes the first part of the proof.

C.2 Part II: Derivation of the update rule in Eq. (7)

We now present the second part of the proof by deriving the update rule in Eq. (7). Since the cost function $\mathcal{L}_{\text{conv}}^{\text{time}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B})$ is strictly g-convex, its optimal solution in the k-th iteration is unique. More concretely, the optimum can be analytically derived by taking the derivative of Eq. (7) and setting the result to zero as follows:

$$\nabla \mathcal{L}_{\text{conv}}^{\text{time}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B}) = \left(\mathbf{B}^k\right)^{-1} - \mathbf{B}^{-1}\mathbf{M}_{\text{time}}^k\mathbf{B}^{-1} = 0 \ , \tag{38}$$

which results in

$$\mathbf{B}\left(\mathbf{B}^k\right)^{-1}\mathbf{B} = \mathbf{M}_{\text{time}}^k \ . \tag{39}$$

This solution is known as the *Riccati equation* and is the geometric mean between $\mathbf{B}^k$ and $\mathbf{M}_{\text{time}}^k$ [79, 84]:

$$\mathbf{B}^{k+1} \leftarrow (\mathbf{B}^k)^{\frac{1}{2}} \left((\mathbf{B}^k)^{-1/2}\mathbf{M}_{\text{time}}^k(\mathbf{B}^k)^{-1/2}\right)^{\frac{1}{2}} (\mathbf{B}^k)^{\frac{1}{2}} \ .$$

Deriving the update rule in Eq. (7) concludes the second part of the proof of Theorem 2. □

D Proof of Theorem 3

Proof. Analogous to the proof of Theorem 1 in Appendix B, we start by recalling $\mathcal{L}_{\text{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$ in Eq. (4):

$$\mathcal{L}_{\text{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B}) = T\log|\mathbf{\Sigma}_{\mathbf{y}}| + M\log|\mathbf{B}| + \frac{1}{G}\sum_{g=1}^{G}\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{y}}^{-1}\mathbf{Y}_g\mathbf{B}^{-1}\mathbf{Y}_g^\top).$$

Let $\mathbf{B}^k$ be the value of the temporal covariance matrix learned using Eq. (7) at k-th iteration. Then, by ignoring the term $M\log|\mathbf{B}^k|$ that is only a function of $\mathbf{B}^k$, $\mathcal{L}_{\text{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$ can be written as follows:

$$\begin{aligned}
\mathcal{L}_{\text{kron}}^{\text{space}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B}^k) &= T\log|\mathbf{\Sigma}_{\mathbf{y}}| + \frac{1}{G}\sum_{g=1}^{G}\mathrm{tr}\left(\mathbf{\Sigma}_{\mathbf{y}}^{-1}\mathbf{Y}_g(\mathbf{B}^k)^{-1}\mathbf{Y}_g^\top\right) \\
&= \log|\mathbf{\Sigma}_{\mathbf{y}}| + \frac{1}{TG}\sum_{g=1}^{G}\mathrm{tr}\left(\mathbf{\Sigma}_{\mathbf{y}}^{-1}\mathbf{Y}_g(\mathbf{B}^k)^{-1}\mathbf{Y}_g^\top\right) \\
&= \log|\mathbf{\Sigma}_{\mathbf{y}}| + \mathrm{tr}\left(\mathbf{\Sigma}_{\mathbf{y}}^{-1}\frac{1}{TG}\sum_{g=1}^{G}\mathbf{Y}_g(\mathbf{B}^k)^{-1}\mathbf{Y}_g^\top\right) \\
&= \log|\mathbf{\Sigma}_{\mathbf{y}}| + \mathrm{tr}\left(\mathbf{\Sigma}_{\mathbf{y}}^{-1}\mathbf{M}_{\text{space}}^k\right) \ ,
\end{aligned} \tag{40}$$

where $\mathbf{M}^k_{\text{space}} := \frac{1}{TG}\sum_{g=1}^{G} \mathbf{Y}_g(\mathbf{B}^k)^{-1}\mathbf{Y}_g^\top$.

Similar to the argument made in Appendix B, a first order Taylor expansion of the log-determinant function around $\mathbf{\Sigma_y}$ provides the following inequality:

$$\begin{aligned}\log|\mathbf{\Sigma_y}| &\leq \log\left|\mathbf{\Sigma}^k_{\mathbf{y}}\right| + \text{tr}\left(\left(\mathbf{\Sigma}^k_{\mathbf{y}}\right)^{-1}\left(\mathbf{\Sigma_y} - \mathbf{\Sigma}^k_{\mathbf{y}}\right)\right)\\ &= \log\left|\mathbf{\Sigma}^k_{\mathbf{y}}\right| + \text{tr}\left(\left(\mathbf{\Sigma}^k_{\mathbf{y}}\right)^{-1}\mathbf{\Sigma_y}\right) - \text{tr}\left(\left(\mathbf{\Sigma}^k_{\mathbf{y}}\right)^{-1}\mathbf{\Sigma}^k_{\mathbf{y}}\right)\\ &= \text{tr}(\mathbf{\Phi}^\top(\mathbf{\Sigma}^k_{\mathbf{y}})^{-1}\mathbf{\Phi H}) + \text{const}\ ,\end{aligned} \tag{41}$$

where the last step is derived using the augmented source and noise covariances, $\mathbf{H} := [\mathbf{\Gamma}, \mathbf{0}; \mathbf{0}, \mathbf{\Lambda}]$, $\mathbf{\Phi} := [\mathbf{L}, \mathbf{I}]$ and $\mathbf{\Sigma_y} = \mathbf{\Phi H \Phi}^\top$.

By inserting Eq. (40) into Eq. (41), the first term of Eq. (8), $\text{tr}(\mathbf{\Phi}^\top(\mathbf{\Sigma}^k_{\mathbf{y}})^{-1}\mathbf{\Phi H})$, can be directly inferred:

$$\begin{aligned}\mathcal{L}^{\text{space}}_{\text{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B}^k) = \mathcal{L}^{\text{space}}_{\text{kron}}(\mathbf{H}, \mathbf{B}^k) &= \log|\mathbf{\Sigma_y}| + \text{tr}\left(\mathbf{\Sigma_y}^{-1}\mathbf{M}^k_{\text{space}}\right)\\ &\leq \text{tr}(\mathbf{\Phi}^\top(\mathbf{\Sigma}^k_{\mathbf{y}})^{-1}\mathbf{\Phi H}) + \text{tr}\left(\mathbf{\Sigma_y}^{-1}\mathbf{M}^k_{\text{space}}\right) + \text{const}\ ,\end{aligned} \tag{42}$$

We further show how the second term in Eq. (8) can be derived. To this end, we construct an upper bound on $\text{tr}\left(\mathbf{\Sigma_y}^{-1}\mathbf{M}^k_{\text{space}}\right)$ using an inequality derived from the Schur complement of $\mathbf{\Sigma_y}$. Before presenting this inequality, the subsequent definition of the Schur complement of matrix $\mathbf{\Sigma_y}$ is required:

Definition 4. *For a positive semidefinite (PSD) matrix $\mathbf{\Sigma_y}$, and a partitioning*

$$\mathbf{X} = \begin{bmatrix} \mathbf{D} & \mathbf{G} \\ \mathbf{G}^\top & \mathbf{B} \end{bmatrix}\ , \tag{43}$$

its Schur complement is defined as

$$\mathbf{S} := \mathbf{D} - \mathbf{G}\mathbf{\Sigma_y}^{-1}\mathbf{G}^\top \tag{44}$$

$$\tag{45}$$

The Schur complement condition states that the matrix $\mathbf{X}$ is PSD, $\mathbf{X} \geq \mathbf{0}$, if and only if the Schur complement of $\mathbf{\Sigma_y}$ is PSD, $\mathbf{S} \geq \mathbf{0}$.

Now we are ready to construct an upper bound on $\text{tr}\left(\mathbf{\Sigma_y}^{-1}\mathbf{M}^k_{\text{space}}\right)$. To this end, we show that $\text{tr}\left(\mathbf{\Sigma_y}^{-1}\mathbf{M}^k_{\text{space}}\right)$ can be majorized as follows:

$$\text{tr}\left(\mathbf{\Sigma_y}^{-1}\mathbf{M}^k_{\text{space}}\right) \leq \text{tr}(\mathbf{H}^k\mathbf{\Phi}^\top(\mathbf{\Sigma}^k_{\mathbf{y}})^{-1}\mathbf{M}^k_{\text{space}}(\mathbf{\Sigma}^k_{\mathbf{y}})^{-1}\mathbf{\Phi H}^k\mathbf{H}^{-1})\ . \tag{46}$$

By defining $\mathbf{V}$ as:

$$\mathbf{V} = \begin{bmatrix} (\mathbf{\Sigma}^k_{\mathbf{y}})^{-1}\mathbf{\Phi H}^k\mathbf{H}^{\frac{-1}{2}} \\ \mathbf{\Phi H}^{\frac{1}{2}} \end{bmatrix}\ , \tag{47}$$

the PSD property of $\mathbf{S}$ can be inferred as:

$$\mathbf{S} = \begin{bmatrix} (\mathbf{\Sigma}^k_{\mathbf{y}})^{-1}\mathbf{\Phi H}^k\mathbf{H}^{-1}\mathbf{H}^k\mathbf{\Phi}^\top(\mathbf{\Sigma}^k_{\mathbf{y}})^{-1} & \mathbf{I} \\ \mathbf{I} & \mathbf{\Phi H \Phi}^\top \end{bmatrix} = \mathbf{V}\mathbf{V}^\top \geq 0\ . \tag{48}$$

By employing the definition of the Schur complement with $\mathbf{D} = (\mathbf{\Sigma}^k_{\mathbf{y}})^{-1}\mathbf{\Phi H}^k\mathbf{H}^{-1}\mathbf{H}^k\mathbf{\Phi}^\top(\mathbf{\Sigma}^k_{\mathbf{y}})^{-1}$, $\mathbf{G} = \mathbf{I}$ and $\mathbf{\Sigma_y} = \mathbf{\Phi H \Phi}^\top$, we have:

$$(\mathbf{\Sigma}^k_{\mathbf{y}})^{-1}\mathbf{\Phi H}^k\mathbf{H}^{-1}\mathbf{H}^k\mathbf{\Phi}^\top(\mathbf{\Sigma}^k_{\mathbf{y}})^{-1} \geq \left(\mathbf{\Phi H \Phi}^\top\right)^{-1}\ . \tag{49}$$

The inequality in Eq. (46) can be directly inferred by multiplying $\mathbf{M}^k_{\text{space}}$ to both sides of Eq. (49), applying trace operator, and rearranging the arguments in the trace operator:

$$\begin{aligned}\text{tr}(\mathbf{M}^k_{\text{space}}\mathbf{\Sigma_y}^{-1}] &\leq \text{tr}(\mathbf{M}^k_{\text{space}}(\mathbf{\Sigma}^k_{\mathbf{y}})^{-1}\mathbf{\Phi H}^k\mathbf{H}^{-1}\mathbf{H}^k\mathbf{\Phi}^\top(\mathbf{\Sigma}^k_{\mathbf{y}})^{-1})\\ &= \text{tr}(\mathbf{H}^k\mathbf{\Phi}^\top(\mathbf{\Sigma}^k_{\mathbf{y}})^{-1}\mathbf{M}^k_{\text{space}}(\mathbf{\Sigma}^k_{\mathbf{y}})^{-1}\mathbf{\Phi H}^k\mathbf{H}^{-1})\\ &= \text{tr}\left(\mathbf{M}^k_{\text{SN}}\mathbf{H}^{-1}\right)\ .\end{aligned} \tag{50}$$

By inserting Eq. (50) into Eq. (42), we have

$$\begin{aligned}\mathcal{L}_{\text{kron}}^{\text{space}}(\boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \mathbf{B}^k) = \mathcal{L}_{\text{kron}}^{\text{space}}(\mathbf{H}, \mathbf{B}^k) &\leq \text{tr}(\boldsymbol{\Phi}^\top (\boldsymbol{\Sigma}_{\mathbf{y}}^k)^{-1} \boldsymbol{\Phi}\mathbf{H}) + \text{tr}\left(\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{M}_{\text{space}}^k\right) + \text{const} \\ &\leq \text{tr}(\boldsymbol{\Phi}^\top (\boldsymbol{\Sigma}_{\mathbf{y}}^k)^{-1} \boldsymbol{\Phi}\mathbf{H}) + \text{tr}\left(\mathbf{M}_{\text{SN}}^k \mathbf{H}^{-1}\right) + \text{const} \\ &= \mathcal{L}_{\text{conv}}^{\text{space}}(\mathbf{H}, \mathbf{B}^k) + \text{const} \ . \end{aligned} \tag{51}$$

Note that constant values in (51) do not depend on $\mathbf{H}$; hence, they can be ignored in the optimization procedure. We have shown that minimizing Eq. (4) with respect to $\mathbf{H}$ is equivalent to minimizing $\mathcal{L}_{\text{conv}}^{\text{space}}(\mathbf{H}, \mathbf{B}^k)$, which concludes the proof. □

E Proof of Theorem 4

Proof. We proceed in two steps. First, we show that $\mathcal{L}_{\text{conv}}^{\text{space}}(\mathbf{H}, \mathbf{B}^k)$ is convex in $\mathbf{h}$. Then, we derive the update rule proposed in Eq. (10).

E.1 Part I: Convexity of the majorizing cost function

We start the proof by constraining $\mathbf{H}$ to the set of diagonal matrices with non-negative entries $\mathcal{S}$, i.e., $\mathcal{S} = \{\mathbf{H} \mid \mathbf{H} = \text{diag}(\mathbf{h}) = \text{diag}([h_1, \ldots, h_{N+M}]^\top),\ h_n \geq 0,\ \text{for } i = 1, \ldots, N+M\}$. We continue by reformulating the constrained optimization with respect to the source covariance matrix,

$$\mathbf{H}^{k+1} = \underset{\mathbf{H} \in \mathcal{S},\ \mathbf{B} = \mathbf{B}^k}{\arg\min}\ \text{tr}\left(\boldsymbol{\Phi}^\top \left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1} \boldsymbol{\Phi}\mathbf{H}\right) + \text{tr}(\mathbf{M}_{\text{SN}}^k \mathbf{H}^{-1}) \ , \tag{52}$$

as follows:

$$\mathbf{h}^{k+1} = \underset{\mathbf{h} \geq 0,\ \mathbf{B} = \mathbf{B}^k}{\arg\min}\ \underbrace{\text{diag}\left(\boldsymbol{\Phi}^\top \left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1} \boldsymbol{\Phi}\right) \mathbf{h} + \text{diag}\left(\mathbf{M}_{\text{SN}}^k\right) \mathbf{h}^{-1}}_{\mathcal{L}_{\text{diag}}^{\text{space}}(\mathbf{h}|\mathbf{h}^k)} \ , \tag{53}$$

where $\mathbf{h}^{-1} = [h_1^{-1}, \ldots, h_N^{-1}]^\top$ is defined as the element-wise inversion of $\mathbf{h}$. Let $\mathbf{V}^k := \boldsymbol{\Phi}^\top \left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1} \boldsymbol{\Phi}$. Then, we rewrite $\mathcal{L}_{\text{diag}}^{\text{space}}(\mathbf{h}|\mathbf{h}^k)$ as

$$\mathcal{L}_{\text{diag}}^{\text{space}}(\mathbf{h}|\mathbf{h}^k) = \text{diag}\left(\mathbf{V}^k\right) \mathbf{h} + \text{diag}\left(\mathbf{M}_{\text{SN}}^k\right) \mathbf{h}^{-1} \ . \tag{54}$$

The convexity of $\mathcal{L}_{\text{diag}}^{\text{space}}(\mathbf{h}|\mathbf{h}^k)$ can be directly inferred from the convexity of $\text{diag}\left[\mathbf{V}^k\right] \mathbf{h}$ and $\text{diag}\left[\mathbf{M}_{\text{SN}}^k\right] \mathbf{h}^{-1}$ with respect to $\mathbf{h}$ [85, Chapter. 3].

E.2 Part II: Derivation of the update rule in Eq. (10)

We now present the second part of the proof by deriving the update rule in Eq. (10). Since the cost function $\mathcal{L}_{\text{diag}}^{\text{space}}(\mathbf{h}|\mathbf{h}^k)$ is convex, its optimal solution in the k-th iteration is unique. Therefore, the optimization with respect to heteroscedastic source and noise variances is carried out by taking the derivative of (53) with respect to $h_i,$ for $n = 1, \ldots, M+N$, and setting it to zero:

$$\begin{aligned}&\frac{\partial}{\partial h_i}\left(\left[\boldsymbol{\Phi}^\top \left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1} \boldsymbol{\Phi}\right] h_i + \left[\mathbf{M}_{\text{SN}}^k\right] h_i^{-1}\right) \\ &\quad = \left[\boldsymbol{\Phi}^\top \left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1} \boldsymbol{\Phi}\right]_{i,i} - \frac{1}{(h_i)^2}\left[\mathbf{M}_{\text{SN}}^k\right]_{i,i} \\ &\quad = 0 \quad \text{for } i = 1, \ldots, N+M \ , \end{aligned}$$

where $\boldsymbol{\Phi}_i$ denotes the n-th column of the augmented lead field matrix. This yields the following update rule:

$$\mathbf{H}^{k+1} = \text{diag}(\mathbf{h}^{k+1}),\ \text{where,} \quad h_i^{k+1} \leftarrow \sqrt{\frac{\left[\mathbf{M}_{\text{SN}}^k\right]_{i,i}}{\left[\boldsymbol{\Phi}^\top \left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1} \boldsymbol{\Phi}\right]_{i,i}}} = \sqrt{\frac{\frac{1}{T}\sum_{t=1}^{T}(\bar{\boldsymbol{\eta}}_n^k(t))^2}{\boldsymbol{\Phi}_n^\top \left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1} \boldsymbol{\Phi}_i}}$$
$$\text{for } i = 1, \ldots, N+M \ . \tag{55}$$

The updates rule in Eq. (10) can be directly inferred by defining $\mathbf{g} := \text{diag}(\mathbf{M}_{\text{SN}}^k)$ and $\mathbf{z} := \text{diag}(\boldsymbol{\Phi}^\top (\boldsymbol{\Sigma}_{\mathbf{y}}^k)^{-1} \boldsymbol{\Phi})$, which leads to: $g_i^k = \left[\mathbf{M}_{\text{SN}}^k\right]_{i,i}$ and $z_i^k = \left[\boldsymbol{\Phi}^\top \left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1} \boldsymbol{\Phi}\right]_{i,i}$. This concludes the proof. □

F Champagne with heteroscedastic noise learning

Interestingly, identical update rules as those proposed in Champagne [33] and heteroscedastic noise learning [86] can be derived for source and noise variances, respectively, by selecting the corresponding indices of matrix $\mathbf{H}$ associated to noise and source covariances.

F.1 Update rule for source variances

Given $[\mathbf{\Phi}]_{1:M,1:N} = \mathbf{L}$, $[\mathbf{H}]_{1:N,1:N} = \mathbf{\Gamma}$, and $[\bar{\boldsymbol{\eta}}(t)]_{1:N} = \bar{\mathbf{x}}(t)$, the update rule for $\mathbf{\Gamma}^{k+1} = \mathrm{diag}(\boldsymbol{\gamma}^{k+1})$ is derived by replacing $\mathbf{H}$, $\mathbf{\Phi}$ and $\bar{\boldsymbol{\eta}}_n^k(t)$ in Eq. (55) with $\mathbf{\Gamma}$, $\mathbf{L}$ and $\bar{\mathbf{x}}_n^k(t)$, respectively, and defining the counterpart of $\mathbf{M}_{\mathrm{SN}}^k$ for sources accordingly as $\mathbf{M}_{\mathrm{S}}^k := \boldsymbol{\omega}_{\mathrm{S}}^k \mathbf{M}_{\mathrm{space}}^k (\boldsymbol{\omega}_{\mathrm{S}}^k)^\top$, where $\boldsymbol{\omega}_{\mathrm{S}}^k := \mathbf{\Gamma}^k \mathbf{L}^\top (\mathbf{\Sigma}_{\mathbf{y}}^k)^{-1}$. The update rule for the source variances is then obtained as follows:

$$\gamma_n^{k+1} \leftarrow \sqrt{\frac{\left[\mathbf{M}_{\mathrm{S}}^k\right]_{n,n}}{\left[\mathbf{L}^\top \left(\mathbf{\Sigma}_{\mathbf{y}}^k\right)^{-1} \mathbf{L}\right]_{n,n}}} = \sqrt{\frac{\frac{1}{T}\sum_{t=1}^T (\bar{\mathbf{x}}_n^k(t))^2}{\mathbf{L}_n^\top \left(\mathbf{\Sigma}_{\mathbf{y}}^k\right)^{-1} \mathbf{L}_n}} \quad \text{for } n = 1, \ldots, N \,, \tag{56}$$

where $\mathbf{L}_n$ denotes the n-th column of the lead field matrix.

F.2 Update rule for noise variances

Similarly, given $[\mathbf{\Phi}]_{1:M,N+1:N+M} = \mathbf{I}$, $[\mathbf{H}]_{N+1:N+M,N+1:N+M} = \mathbf{\Lambda}$, and $[\bar{\boldsymbol{\eta}}(t)]_{N+1:N+M} = \bar{\mathbf{e}}(t) := \mathbf{y}(t) - \mathbf{L}\bar{\mathbf{x}}(t)$, the update rule for $\mathbf{\Lambda}^{k+1} = \mathrm{diag}(\boldsymbol{\lambda}^{k+1})$ is derived by replacing $\mathbf{H}$, $\mathbf{\Phi}$ and $\bar{\boldsymbol{\eta}}_n^k(t)$ in Eq. (55) with $\mathbf{\Lambda}$, $\mathbf{I}$ and $\bar{\mathbf{e}}_n^k(t)$, respectively, and defining the counterpart of $\mathbf{M}_{\mathrm{SN}}^k$ for the noise accordingly as $\mathbf{M}_{\mathrm{N}}^k := \boldsymbol{\omega}_{\mathrm{N}}^k \mathbf{M}_{\mathrm{space}}^k (\boldsymbol{\omega}_{\mathrm{N}}^k)^\top$ with $\boldsymbol{\omega}_{\mathrm{N}}^k = \mathbf{\Lambda}^k (\mathbf{\Sigma}_{\mathbf{y}}^k)^{-1}$. The update rule for the noise variances is then derived as follows:

$$\lambda_m^{k+1} \leftarrow \sqrt{\frac{\left[\mathbf{M}_{\mathrm{N}}^k\right]_{m,m}}{\left[\left(\mathbf{\Sigma}_{\mathbf{y}}^k\right)^{-1}\right]_{m,m}}} = \sqrt{\frac{\sum_{t=1}^T (\bar{\mathbf{e}}_n^k(t))^2}{\left[\left(\mathbf{\Sigma}_{\mathbf{y}}^k\right)^{-1}\right]_{m,m}}} \quad \text{for } m = 1, \ldots, M \,, \tag{57}$$

which is identical to the update rule of the Champagne with heteroscedastic noise learning as presented in Cai et al. [86].

G Proof of Theorem 5

We prove Theorem 5 by showing that the alternating update rules for $\mathbf{B}$ and $\mathbf{H}$, Eqs. (7) and (10), are guaranteed to converge to a stationary point of the Bayesian Type-II likelihood $\mathcal{L}_{\text{kron}}(\mathbf{\Gamma}, \mathbf{\Lambda}, \mathbf{B})$ Eq. (4). More generally, we prove that full Dugh is an instance of the general class of majorization-minimization (MM) algorithms, for which this property follows by construction. To this end, we first briefly review theoretical concepts behind the majorization-minimization (MM) algorithmic framework [87–90].

G.1 Required conditions for majorization-minimization algorithms

MM is a versatile framework for optimizing general non-linear optimization programs. The main idea behind MM is to replace the original cost function in each iteration by an upper bound, also known as majorizing function, whose minimum is easy to find. Compared to other popular optimization paradigms such as (quasi)-Newton methods, MM algorithms enjoy guaranteed convergence to a stationary point [27]. The MM class covers a broad range of common optimization algorithms such as *convex-concave procedures (CCCP)* and *proximal methods* [27, Section IV], [91–93]. Such algorithms have been applied in various domains such as non-negative matrix factorization [94], graph learning [63], robust portfolio optimization in finance [57], direction of arrival (DoA) and channel estimation in wireless communications [54–56, 95], internet of things (IoT) [96, 97], and brain source

imaging [26, 86, 98–100]. Interested readers are referred to Sun et al. [27] for an extensive list of applications on MM.

We define an original optimization problem with the objective of minimizing a continuous function $f(\mathbf{u})$ within a closed convex set $\mathcal{U} \subset \mathbb{R}^n$:

$$\min_{\mathbf{u}} f(\mathbf{u}) \quad \text{subject to } \mathbf{u} \in \mathcal{U} \ . \tag{58}$$

Then, the idea of MM can be summarized as follows. First, construct a continuous *surrogate function* $g(\mathbf{u}|\mathbf{u}^k)$ that *majorizes*, or upper-bounds, the original function $f(\mathbf{u})$ and coincides with $f(\mathbf{u})$ at a given point $\mathbf{u}^k$:

$$\begin{aligned} &[\text{A1}] \qquad && g(\mathbf{u}^k|\mathbf{u}^k) = f(\mathbf{u}^k) && \forall\ \mathbf{u}^k \in \mathcal{U} \\ &[\text{A2}] \qquad && g(\mathbf{u}|\mathbf{u}^k) \geq f(\mathbf{u}) && \forall\ \mathbf{u}, \mathbf{u}^k \in \mathcal{U} \ . \end{aligned}$$

Second, starting from an initial value $\mathbf{u}^0$, generate a sequence of feasible points $\mathbf{u}^1, \mathbf{u}^2, \ldots, \mathbf{u}^k, \mathbf{u}^{k+1}$ as solutions of a series of successive simple optimization problems, where

$$[\text{A3}] \qquad \mathbf{u}^{k+1} := \arg\min_{\mathbf{u} \in \mathcal{U}} g(\mathbf{u}|\mathbf{u}^k) \ .$$

Definition 5. *Any algorithm fulfilling conditions [A1]–[A3] is called a Majorization Minimization (MM) algorithm.*

If a surrogate function fulfills conditions [A1]–[A3], then the value of the cost function f decreases in each iteration:

Corollary 1. *An MM algorithm has a* descending trend *property, whereby the value of the cost function f decreases in each iteration: $f(\mathbf{u}^{k+1}) \leq f(\mathbf{u}^k)$.*

Proof. To verify the descending trend in the MM framework, it is sufficient to show that $f(\mathbf{u}^{k+1}) \leq f(\mathbf{u}^k)$. To this end, we have $f(\mathbf{u}^{k+1}) \leq g(\mathbf{u}^{k+1}|\mathbf{u}^k)$ from condition [A2]. Condition [A3] further states that $g(\mathbf{u}^{k+1}|\mathbf{u}^k) \leq g(\mathbf{u}^k|\mathbf{u}^k)$, while $g(\mathbf{u}^k|\mathbf{u}^k) = f(\mathbf{u}^k)$ holds according to [A1]. Putting everything together, we have:

$$f(\mathbf{u}^{k+1}) \overset{[\text{A2}]}{\leq} g(\mathbf{u}^{k+1}|\mathbf{u}^k) \overset{[\text{A3}]}{\leq} g(\mathbf{u}^k|\mathbf{u}^k) \overset{[\text{A1}]}{=} f(\mathbf{u}^k) \ ,$$

which concludes the proof. □

While Corollary 1 guarantees a descending trend, convergence requires additional assumptions on particular properties of f and g [88, 89]. For the smooth functions considered in this paper, we require that the derivatives of the original and surrogate functions coincide at $\mathbf{u}^k$:

$$[\text{A4}] \qquad \nabla g(\mathbf{u}^k|\mathbf{u}^k) = \nabla f(\mathbf{u}^k) \qquad \forall\ \mathbf{u}^k \in \mathcal{U} \ .$$

We can then formulate the following, stronger, theorem:

Theorem 8. *For an MM algorithm that additionally satisfies [A4], every limit point of the sequence of minimizers generated through [A3] is a stationary point of the original optimization problem Eq.* (58)*.*

Proof. A detailed proof is provided in Razaviyayn et al. [88, Theorem 1]. □

Note that since we are working with smooth functions, conditions [A1]–[A4] are sufficient to prove convergence to a stationary point according to Theorem 8.

G.2 Detailed derivation of the proof of Theorem 5

We now show that full Dugh is an instance of majorization-minimization as defined above, which fulfills Theorem 8.

Proof. We need to prove that conditions [A1]–[A4] are fulfilled for full Dugh. To this end, we first prove conditions [A1]–[A4] for the optimization with respect to $\mathbf{B}$ based on the convex surrogate function in Eq. (5), $\mathcal{L}_{\text{conv}}^{\text{time}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B})$. For this purpose, we recall the upper bound on $\log|\mathbf{B}|$ in Eq. (28), which fulfills condition [A2] since it majorizes $\log|\mathbf{B}|$ as a result of the concavity of the log-determinant function and its first-order Taylor expansion around $\mathbf{B}^k$. Besides, it automatically satisfies conditions [A1] and [A4] by construction, because the majorizing function in Eq. (28) is obtained through a Taylor expansion around $\mathbf{B}^k$. Concretely, [A1] is satisfied because the equality in Eq. (28) holds for $\mathbf{B} = \mathbf{B}^k$. Similarly, [A4] is satisfied because the gradient of $\log|\mathbf{B}|$ at point $\mathbf{B}^k$, $\left(\mathbf{B}^k\right)^{-1}$ defines the linear Taylor approximation $\log\left|\mathbf{B}^k\right| + \text{tr}\left(\left(\mathbf{B}^k\right)^{-1}\left(\mathbf{B} - \mathbf{B}^k\right)\right)$. Thus, both gradients coincide in $\mathbf{B}^k$ by construction. We can further prove that [A3] can be satisfied by showing that $\mathcal{L}_{\text{conv}}^{\text{time}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B})$ reaches its global minimum in each MM iteration. This is guaranteed if $\mathcal{L}_{\text{conv}}^{\text{time}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}^k, \mathbf{B})$ can be shown to be convex or g-convex with respect to $\mathbf{B}$. To this end, we first require the subsequent proposition:

Proposition 2. *Any local minimum of a g-convex function over a g-convex set is a global minimum.*

Proof. A detailed proof is presented in Rapcsak [75, Theorem 2.1]. □

Given the proof presented in Appendix C.1, we can conclude that $\mathcal{L}_{\text{conv}}^{\text{time}}(\mathbf{H}^k, \mathbf{B})$ is g-convex; hence, any local minimum of $\mathcal{L}_{\text{conv}}^{\text{time}}(\mathbf{H}^k, \mathbf{B})$ is a global minimum according to Proposition 2. This proves that condition [A3] is fulfilled and completes the proof that the optimization of Eq. (4) with respect to $\mathbf{B}$ using the convex surrogate cost function Eq. (5) leads to an MM algorithm.

The proof of conditions [A1], [A2] and [A4] for the optimization with respect to $\mathbf{H}$ based on the convex surrogate function in Eq. (8), $\mathcal{L}_{\text{conv}}^{\text{space}}(\mathbf{H}, \mathbf{B}^k)$, can be presented analogously. To this end, we recall the upper bound on $\log|\mathbf{\Sigma_y}|$ in Eq. (41), which fulfills condition [A2] since it majorizes $\log|\mathbf{\Sigma_y}|$ as a result of the concavity of the log-determinant function and its first-order Taylor expansion around $\mathbf{\Sigma}_\mathbf{y}^k$. Besides, it automatically satisfies conditions [A1] and [A4] by construction, because the majorizing function in Eq. (41) is obtained through a Taylor expansion around $\mathbf{\Sigma}_\mathbf{y}^k$. Concretely, [A1] is satisfied because the equality in Eq. (41) holds for $\mathbf{\Sigma_y} = \mathbf{\Sigma}_\mathbf{y}^k$. Similarly, [A4] is satisfied because the gradient of $\log|\mathbf{\Sigma_y}|$ at point $\mathbf{\Sigma}_\mathbf{y}^k$, $\left(\mathbf{\Sigma}_\mathbf{y}^k\right)^{-1}$ defines the linear Taylor approximation $\log\left|\mathbf{\Sigma}_\mathbf{y}^k\right| + \text{tr}\left[\left(\mathbf{\Sigma}_\mathbf{y}^k\right)^{-1}\left(\mathbf{\Sigma_y} - \mathbf{\Sigma}_\mathbf{y}^k\right)\right]$. Thus, both gradients coincide in $\mathbf{\Sigma}_\mathbf{y}^k$ by construction. We can further prove that [A3] can be satisfied by showing that $\mathcal{L}_{\text{conv}}^{\text{time}}(\mathbf{H}^k, \mathbf{B})$ reaches its global minimum in each MM iteration. This is guaranteed if $\mathcal{L}_{\text{conv}}^{\text{time}}(\mathbf{H}^k, \mathbf{B})$ can be shown to be convex with respect to $\mathbf{H} = \text{diag}(\mathbf{h})$. Given the proof presented in Appendix E.2, we can show that [A3] is also satisfied since $\mathcal{L}_{\text{conv}}^{\text{space}}(\mathbf{H}, \mathbf{B}^k)$ in Eq. (8) is a convex function with respect to $\mathbf{h}$. The convexity of $\mathcal{L}_{\text{conv}}^{\text{space}}(\mathbf{H}, \mathbf{B}^k)$, which ensures that condition [A3] can be satisfied using standard optimization, along with the fulfillment of conditions [A1], [A2] and [A4], ensure that Theorem 8 holds for $\mathcal{L}_{\text{conv}}^{\text{space}}(\mathbf{H}, \mathbf{B}^k)$. This completes the proof that the optimization of Eq. (4) with respect to $\mathbf{H}$ using the convex surrogate cost function Eq. (8) leads to an MM algorithm that is guaranteed to converge. □

H Proof of Proposition 1

This is a well-established classical results of the signal processing literature. Therefore, we only provide two remarks highlighting important connections between Preposition 1 and our proposed method, and refer the interested reader to Grenander and Szegö [32, Chapter 5] for a detailed proof of the theorem.

Remark 3. *As indicated in Babu [101], the set of embedded circulant matrices with size $L \times L$, $\mathcal{B}^L$, does not constist exclusively of positive definite Toeplitz matrices. Therefore, we restrict ourselves to embedded circulant matrices $\mathbf{P}$ that are also positive definite. This restriction indeed makes the diagonalization inaccurate, but we can improve the accuracy by choosing a large L. Practical evaluations have shown that choosing $L \geq 2T - 1$ provides sufficient approximation result.*

Remark 4. *The* Carathéodory parametrization *of a Toeplitz matrix [102, Section 4.9.2] states that any PD matrix can be represented as $\mathbf{B} = \mathbf{A}\mathbf{P}^{'}\mathbf{A}^H$ with $[\mathbf{A}]_{m,l} = e^{i(w_l)(m-1)}$ and $\mathbf{P}^{'} = \text{diag}(p_0^{'}, p_1^{'}, \ldots, p_{L-1}^{'})$ where w_l and p_l are specific frequencies and their corresponding*

amplitudes. By comparing the Carathéodory parametrization *of* $\mathbf{B}$ *with its Fourier diagonalization (Eq.* (15)*), it can be seen that Fourier diagonalization force the frequencies to lie on the Fourier grid, i.e.* $w_l = \frac{2\pi(l-1)}{L}$ *, which indeed makes the diagonalization slightly inaccurate. The approximation accuracy can, however, be improved by increasing* L*. The Szegö theorem [32, 103] states that a Toeplitz matrix is asymptotically (*$L \to \infty$*) diagonalized by the DFT matrix.*

I Proof of Theorem 6

Proof. We proceed in two steps. First, we show that the cost function in Eq. (13) is convex with respect to $\mathbf{p}$. In the second step, we then derive the update rule proposed in (16).

I.1 Part I: Convexity of the majorizing cost function

The proof of this section parallels the one provided in [104, Proposition 4]. We start by recalling Eq. (13):

$$\mathbf{B}^* = \underset{\mathbf{B}\in\mathcal{B},\ \mathbf{H}=\mathbf{H}^k}{\arg\min}\ \mathrm{tr}((\mathbf{B}^k)^{-1}\mathbf{B}) + \mathrm{tr}(\mathbf{M}^k_{\mathrm{time}}\mathbf{B}^{-1})\ . \tag{59}$$

We then show that the second term in Eq. (59) can be upper-bounded as follows:

$$\mathrm{tr}(\mathbf{M}^k_{\mathrm{time}}\mathbf{B}^{-1}) \leq \mathrm{tr}(\mathbf{P}^k\mathbf{Q}^H(\mathbf{B}^k)^{-1}\mathbf{M}^k_{\mathrm{time}}(\mathbf{B}^k)^{-1}\mathbf{Q}\mathbf{P}^k\mathbf{P}^{-1})\ . \tag{60}$$

By defining $\mathbf{V}$ as

$$\mathbf{V} = \begin{bmatrix} (\mathbf{B}^k)^{-1}\mathbf{Q}\mathbf{P}^k\mathbf{P}^{\frac{-1}{2}} \\ \mathbf{Q}\mathbf{P}^{\frac{1}{2}} \end{bmatrix}\ , \tag{61}$$

the PSD property of $\mathbf{S}$ can be inferred as

$$\mathbf{S} = \begin{bmatrix} (\mathbf{B}^k)^{-1}\mathbf{Q}\mathbf{P}^k\mathbf{P}^{-1}\mathbf{P}^k\mathbf{Q}^H(\mathbf{B}^k)^{-1} & \mathbf{I} \\ \mathbf{I} & \mathbf{Q}\mathbf{P}\mathbf{Q}^H \end{bmatrix} = \mathbf{V}\mathbf{V}^H \geq 0\ . \tag{62}$$

Therefore by virtue of the Schur complement with $\mathbf{D} = (\mathbf{B}^k)^{-1}\mathbf{Q}\mathbf{P}^k\mathbf{P}^{-1}\mathbf{P}^k\mathbf{Q}^H(\mathbf{B}^k)^{-1}$, $\mathbf{G} = \mathbf{I}$ and $\mathbf{B} = \mathbf{Q}\mathbf{P}\mathbf{Q}^H$, we have:

$$(\mathbf{B}^k)^{-1}\mathbf{Q}\mathbf{P}^k\mathbf{P}^{-1}\mathbf{P}^k\mathbf{Q}^H(\mathbf{B}^k)^{-1} \geq \left(\mathbf{Q}\mathbf{P}\mathbf{Q}^H\right)^{-1}\ . \tag{63}$$

The inequality (Eq. (60)) can be directly obtained by multiplying $\mathbf{M}^k_{\mathrm{time}}$ to both sides of Eq. (63), applying the trace operator, using Eq. (14) and finally rearranging the terms within the trace operator:

$$\mathrm{tr}(\mathbf{M}^k_{\mathrm{time}}(\mathbf{B}^k)^{-1}\mathbf{Q}\mathbf{P}^k\mathbf{P}^{-1}\mathbf{P}^k\mathbf{Q}^H(\mathbf{B}^k)^{-1}) \geq \mathrm{tr}(\mathbf{M}^k_{\mathrm{time}}\mathbf{B}^{-1})\ . \tag{64}$$

Let $\mathbf{B}_k = \mathbf{Q}\mathbf{P}^k\mathbf{Q}^H$ be the Fourier diagonalization of a fixed matrix $\mathbf{B}_k$ in the k-th iteration, one can derive an efficient update rule for the temporal covariance by rewriting Eq. (13) and exploiting Propositions 1 and Eq. (60):

$$\begin{aligned}
&\mathrm{tr}((\mathbf{B}^k)^{-1}\mathbf{B}) + \mathrm{tr}(\mathbf{B}^{-1}\mathbf{M}^k_{\mathrm{time}}) \\
&\qquad \leq \mathrm{tr}((\mathbf{B}^k)^{-1}\mathbf{Q}\mathbf{P}\mathbf{Q}^H) + \mathrm{tr}(\mathbf{P}^k\mathbf{Q}^H(\mathbf{B}^k)^{-1}\mathbf{M}^k_{\mathrm{time}}(\mathbf{B}^k)^{-1}\mathbf{Q}\mathbf{P}^k\mathbf{P}^{-1}) \\
&\qquad = \mathrm{diag}(\mathbf{Q}^H(\mathbf{B}^k)^{-1}\mathbf{Q})\mathbf{p} + \mathrm{diag}(\mathbf{P}^k\mathbf{Q}^H(\mathbf{B}^k)^{-1}\mathbf{M}^k_{\mathrm{time}}(\mathbf{B}^k)^{-1}\mathbf{Q}\mathbf{P}^k)\mathbf{p}^{-1}\ ,
\end{aligned} \tag{65}$$

where $\mathbf{p} = \mathrm{vec}(\mathbf{P})$, and $\mathbf{p}^{-1}$ is defined as the element-wise inversion of $\mathbf{p}$.

We formulate the optimization problem as follows:

$$\begin{aligned}
\mathcal{L}^{\mathrm{time}}_{\mathrm{toeplitz}}(\mathbf{p}) &= \mathrm{diag}(\mathbf{Q}^H(\mathbf{B}^k)^{-1}\mathbf{Q})\mathbf{p} \\
&\quad + \mathrm{diag}(\mathbf{P}^k\mathbf{Q}^H(\mathbf{B}^k)^{-1}\mathbf{M}^k_{\mathrm{time}}(\mathbf{B}^k)^{-1}\mathbf{Q}\mathbf{P}^k)\mathbf{p}^{-1}\ .
\end{aligned} \tag{66}$$

Let $\mathbf{W}^k := \mathbf{Q}^H(\mathbf{B}^k)^{-1}\mathbf{Q}$ and $\mathbf{O}^k := \mathbf{P}^k\mathbf{Q}^H(\mathbf{B}^k)^{-1}\mathbf{M}^k_{\mathrm{time}}(\mathbf{B}^k)^{-1}\mathbf{Q}\mathbf{P}^k$. Then, we rewrite $\mathcal{L}^{\mathrm{time}}_{\mathrm{toeplitz}}(\mathbf{p})$ as

$$\mathcal{L}^{\mathrm{time}}_{\mathrm{toeplitz}}(\mathbf{p}) = \mathrm{diag}(\mathbf{W}^k)\mathbf{p} + \mathrm{diag}(\mathbf{O}^k)\mathbf{p}^{-1}\ . \tag{67}$$

The convexity of $\mathcal{L}^{\mathrm{time}}_{\mathrm{toeplitz}}(\mathbf{p})$ can be directly inferred from the convexity of $\mathrm{diag}\left[\mathbf{W}^k\right]\mathbf{p}$ and $\mathrm{diag}\left[\mathbf{O}^k\right]\mathbf{p}^{-1}$ with respect to $\mathbf{p}$ [85, Chapter. 3].

I.2 Part II: Derivation of the update rule in Eq. (16)

We now present the second part of the proof by deriving the update rule in Eq. (16). Since the cost function $\mathcal{L}^{\text{time}}_{\text{toeplitz}}(\mathbf{p})$ is convex, its optimal solution in the k-th iteration is unique. More concretely, a closed-form solution of the final update rule can be obtained by taking the derivative of Eq. (67) with respect to $\mathbf{p}$ and setting it to zero:

$$p_l^{k+1} \leftarrow \sqrt{\frac{\hat{g}_l^k}{\hat{z}_l^k}} \text{ for } l = 0, \ldots, L-1 \text{ , where} \tag{68}$$

$$\hat{\mathbf{g}} = \text{diag}(\mathbf{P}^k \mathbf{Q}^H (\mathbf{B}^k)^{-1} \mathbf{M}^k_{\text{time}} (\mathbf{B}^k)^{-1} \mathbf{Q} \mathbf{P}^k) \tag{69}$$

$$\hat{\mathbf{z}} = \text{diag}(\mathbf{Q}^H (\mathbf{B}^k)^{-1} \mathbf{Q}) \text{ ,} \tag{70}$$

which concludes the proof. □

J Proof of Theorem 7

Proof. The proof is inspired by ideas presented in Rakitsch et al. [2], Wu et al. [49], Saatçi [105] for spatio-temporal Gaussian process inference, and parallels the one proposed in Solin et al. [106]. Rakitsch et al. [2], Wu et al. [49] provide an efficient method for computing the non-convex spatio-temporal ML cost function by exploiting the compatibility between diagonalization and the Kronecker product. Here we use similar ideas to obtain the posterior mean in an efficient way.

Recalling the diagonalization of the temporal correlation matrix as $\mathbf{B} = \mathbf{Q}\mathbf{P}\mathbf{Q}^H$ and considering the eigenvalue decomposition of $\mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top$ as $\mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top = \mathbf{U}_\mathbf{x}\mathbf{D}_\mathbf{x}\mathbf{U}_\mathbf{x}^\top$ with $\mathbf{D}_\mathbf{x} = \text{diag}(d_1, \ldots, d_M)$, we have:

$$\begin{aligned}
\bar{\mathbf{x}}_g &= (\boldsymbol{\Gamma} \otimes \mathbf{B})\, \mathbf{D}^\top \tilde{\boldsymbol{\Sigma}}_\mathbf{y}^{-1} \mathbf{y}_g \\
&= (\boldsymbol{\Gamma} \otimes \mathbf{B})\, (\mathbf{L} \otimes \mathbf{I})^\top \left(\boldsymbol{\Lambda} \otimes \mathbf{B} + \mathbf{D}\boldsymbol{\Sigma}_0\mathbf{D}^\top\right)^{-1} \text{vec}(\mathbf{Y}_g^\top) \\
&= (\boldsymbol{\Gamma} \otimes \mathbf{B})\, (\mathbf{L}^\top \otimes \mathbf{I}) \left((\boldsymbol{\Lambda} + \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top) \otimes \mathbf{B}\right)^{-1} \text{vec}(\mathbf{Y}_g^\top) \\
&= \left(\boldsymbol{\Gamma}\mathbf{L}^\top \otimes \mathbf{B}\right) \left((\boldsymbol{\Lambda} + \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top) \otimes \mathbf{B}\right)^{-1} \text{vec}(\mathbf{Y}_g^\top) \\
&= \left(\boldsymbol{\Gamma}\mathbf{L}^\top \mathbf{U}_\mathbf{x} \otimes \mathbf{Q}\mathbf{P}\right) (\boldsymbol{\Omega})^{-1} (\mathbf{U}_\mathbf{x}^\top \otimes \mathbf{Q}^H) \text{vec}(\mathbf{Y}_g^\top) \\
&= \left(\boldsymbol{\Gamma}\mathbf{L}^\top \mathbf{U}_\mathbf{x} \otimes \mathbf{Q}\mathbf{P}\right) (\boldsymbol{\Omega})^{-1} \text{tr}\left(\mathbf{Q}^H \mathbf{Y}_g^\top \mathbf{U}_\mathbf{x}\right) \\
&= \left(\boldsymbol{\Gamma}\mathbf{L}^\top \mathbf{U}_\mathbf{x} \otimes \mathbf{Q}\mathbf{P}\right) \text{tr}\left(\boldsymbol{\Pi} \odot \mathbf{Q}^H \mathbf{Y}_g^\top \mathbf{U}_\mathbf{x}\right) \\
&= \text{tr}\left(\mathbf{Q}\mathbf{P}\left(\boldsymbol{\Pi} \odot \mathbf{Q}^H \mathbf{Y}_g^\top \mathbf{U}_\mathbf{x}\right)\left(\mathbf{U}_\mathbf{x}^\top \mathbf{L}\boldsymbol{\Gamma}^\top\right)\right) \text{ ,}
\end{aligned} \tag{71}$$

where $\odot$ denotes the Hadamard product between corresponding elements of two matrices. $\boldsymbol{\Omega}$ and $\boldsymbol{\Pi}$ are defined as follows: $\boldsymbol{\Omega} = \boldsymbol{\Lambda} + \mathbf{D}_\mathbf{x} \otimes \mathbf{P}$ and $[\boldsymbol{\Pi}]_{l,m} = \frac{1}{\sigma_m^2 + p_l d_m}$ for $l = 1, \ldots, L;\ m = 1, \ldots, M$. Note that the last four lines are derived based on the following matrix equality:

$$\text{tr}(\mathbf{A}^\top \mathbf{B}\mathbf{C}\mathbf{D}^\top) = \text{vec}(\mathbf{A})^\top (\mathbf{D} \otimes \mathbf{B}) \text{vec}(\mathbf{C}). \tag{72}$$

Together with the update rule in Eq. (19), this concludes the proof of Theorem 7. □

K Details on the simulation set-up

K.1 Forward modeling

Populations of pyramidal neurons in the cortical gray matter are known to be the main drivers of the EEG signal [14, 19]. Here, we use a realistic volume conductor model of the human head to model the linear relationship between primary electrical source currents generated within these populations and the resulting scalp surface potentials captured by EEG electrodes. The lead field matrix, $\mathbf{L} \in \mathbb{R}^{58 \times 2004}$, which serves as the forward model in our simulations, was generated using the New York Head model [36]. The New York Head model provides a segmentation of an average human head into six different tissue types, taking into account the realistic anatomy and electrical tissue conductivities. In this model, 2004 dipolar current sources were placed evenly on the cortical

surface and 58 sensors were placed on the scalp according to the extended 10-20 system [107]. Finally, the lead field matrix was computed using the finite element method (FEM) for a given head geometry and exploiting the quasi-static approximation of Maxwell's equations [14, 19, 36, 108].

Note that in accordance with the predominant orientation of pyramidal neuron assemblies, the orientation of all simulated source currents was fixed to be perpendicular to the cortical surface, so that only the scalar deflection of each source along the fixed orientation needs to be estimated. In real data analyses in Section 6 and Appendix L, however, surface normals are hard to estimate or even undefined in case of volumetric reconstructions. Consequently, we model each source in real data analyses as a full 3-dimensional current vector. This is achieved by introducing three variance parameters for each source within the source covariance matrix, $\mathbf{\Gamma}^{\text{3D}} = \text{diag}(\boldsymbol{\gamma}^{\text{3D}}) = [\gamma_1^x, \gamma_1^y, \gamma_1^z, \ldots, \gamma_N^x, \gamma_N^y, \gamma_N^z]^\top$. As all algorithms considered here model the source covariance matrix $\mathbf{\Gamma}$ to be diagonal, this extension can be readily implemented. Correspondingly, a full 3D leadfield matrix, $\mathbf{L}^{\text{3D}} \in \mathbb{R}^{M \times 3N}$, is used.

K.2 Pseudo-EEG signal generation

We simulated a sparse set of $N_0 = 3$ active sources, which were placed at random positions on the cortex. To simulate the electrical neural activity of these sources, $T = \{10, 20, 50, 100\}$ time points were sampled from a univariate linear autoregressive (AR) process, which models the activity at time t as a linear combination of the P past values:

$$x_i(t) = \sum_{p=1}^{P} a_i(p) x_i(t-p) + \xi_i(t), \text{ for } i = 1, 2, 3 \,. \tag{73}$$

Here, $a_i(p)$ for $i = 1, 2, 3$ are linear AR coefficients, and P is the order of the AR model. The model residuals $\xi_i(\cdot)$ for $i = 1, 2, 3$ are also referred to as the innovation process; their variance determines the stability of the overall AR process. We here assume uncorrelated standard normal distributed innovations, which are independent for all sources. In this experiment, we use stable AR systems of order $P = \{1, 2, 5, 7\}$. The resulting source distribution, represented as $\mathbf{X} = [\mathbf{x}(1), \ldots, \mathbf{x}(T)]$, was projected to the EEG sensors through application of lead field matrix: $\mathbf{Y}^{\text{signal}} = \mathbf{L}\mathbf{X}$. Next we added Gaussian white noise to the sensor space signal. To this end, the same number of data points as the sources were sampled from a zero-mean normal distribution, where the time points assumed to be independent and identically distributed. The resulting noise distribution, represented as $\mathbf{E} = [\mathbf{e}(1), \ldots, \mathbf{e}(T)]$, is then normalized by its Frobenius norm and added to the signal matrix $\mathbf{Y}^{\text{signal}}$ as follows: $\mathbf{Y} = \mathbf{Y}^{\text{signal}} + \frac{(1-\alpha)\left\|\mathbf{Y}^{\text{signal}}\right\|_F}{\alpha\|\mathbf{E}\|_F}\mathbf{E}$, where α determines the signal-to-noise ratio (SNR) in sensor space. Precisely, SNR is defined as follows: $\text{SNR} = 20\log_{10}\left(\alpha/_{1-\alpha}\right)$. In this experiment the following values of α were used: α={0.55, 0.65, 0.7, 0.8}, which correspond to the following SNRs: SNR={1.7, 5.4, 7.4, 12} (dB). Interested reader can refer to Haufe and Ewald [37] for a more details on this simulation framework.

K.3 Source reconstruction and evaluation metrics

We applied thin Dugh to the synthetic datasets described above. In addition to thin Dugh, one further Type-II Bayesian learning scheme, namely Champagne [33], and two Type-I source reconstruction schemes, namely S-FLEX [35] and eLORETA [34], were also included as benchmarks with respect to source reconstruction performance. S-FLEX is used as an example of a sparse Type-I Bayesian learning method based on ℓ_1-norm minimization. As spatial basis functions, unit impulses were used, so that the resulting estimate was identical to the so-called minimum-current estimate (MCE) [109]. eLORETA estimate, as an example of a smooth inverse solution based on weighted ℓ_2^2-norm minimization, was used with 5% regularization, whereas S-FLEX was fitted so that the residual variance was consistent with the ground-truth noise level. Note that the 5% rule is chosen as it gives the best performance across a subset of regularization values ranging between 0.5% to 15%. For thin Dugh, the noise variances as well as the variances of all voxels were initialized randomly by sampling from a standard normal distribution. The optimization program was terminated after reaching convergence. Convergence was defined if the relative change of the Frobenius-norm of the reconstructed sources between subsequent iterations was less than 10^{-8}. A maximum of 1000 iterations was carried out if no convergence was reached beforehand.

Source reconstruction performance was evaluated according to two different measures, the *earth mover's distance* (EMD), used to quantify the spatial localization accuracy, and the correlation

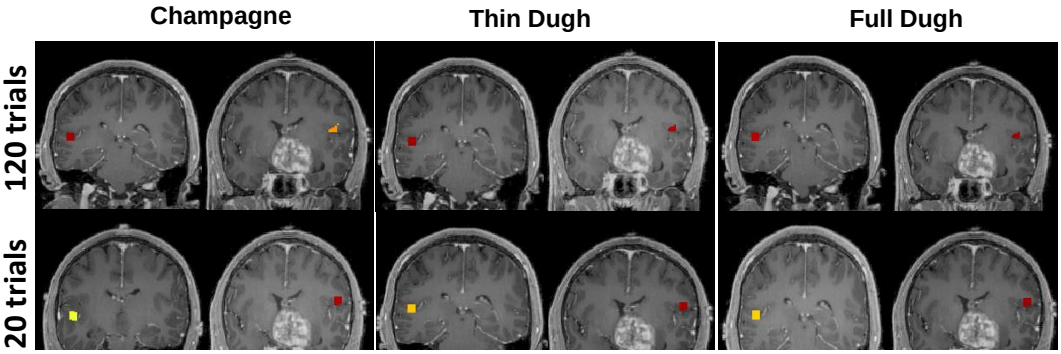


Figure 6: Robustness of Dugh and Champagne performance when the number of trials is increased to 20 and 120.

between the original and reconstructed sources, $\hat{\mathbf{X}}$ and $\mathbf{X}$. The EMD metric measures the cost needed to map two probability distributions, defined on the same metric domain, into each other, see [21, 38]. It was applied here to the power of the true and estimated source activations defined on the cortical surface of the brain, which were obtained by taking the voxel-wise ℓ_2-norm along the time domain. EMD was normalized to $[0, 1]$. The correlation between simulated and reconstructed source time courses was assessed as the mean of the absolute correlations obtained for each source, after optimally matching simulated and reconstructed sources. To this end, Pearson correlation between all pairs of simulated and reconstructed (i.e., those with non-zero activations) sources was measured. Each simulated source was matched to a reconstructed source based on maximum absolute correlation. Time-course correlation error (TCE) was then defined as one minus the average of these absolute correlations across sources. Each simulation was carried out 100 times using different instances of $\mathbf{X}$ and $\mathbf{E}$, and the mean and standard error of the mean (SEM) of each performance measure across repetitions was calculated.

L Real data analysis

L.1 Auditory and visual evoked fields (AEF and VEF)

The MEG data used in this article were acquired in the Biomagnetic Imaging Laboratory at the University of California San Francisco (UCSF) with a CTF Omega 2000 whole-head MEG system from VSM MedTech (Coquitlam, BC, Canada) with 1200 Hz sampling rate. The neural responses for one subject's auditory evoked fields (AEF) and visual evoked fields (VEF) were localized. The AEF response was elicited while subjects were passively listening to 600 ms duration tones (1 kHz) presented binaurally. Data from 120 trial epochs were analysed. The VEF response was elicited while subjects were viewing images of objects projected onto a screen and subjects were instructed to name the objects verbally. Both AEF and VEF data were first digitally filtered from 1 to 70 Hz to remove artifacts and DC offset, time-aligned to the stimulus. Different number of trials were included for algorithm analyses. The pre-stimulus window was selected to be -100 ms to -5 ms and the post-stimulus time window was selected to be 60 ms to 180 ms, where 0 ms is the onset of the tone. The lead field for each subject was calculated with NUTMEG [110] using a single-sphere head model (two spherical orientation lead fields) and an 8 mm voxel grid. Each column was normalized to have a norm of unity. Further details on these datasets can be found in [33, 111–113].

Figure 6 shows the reconstructed sources of the Auditory Evoked Fields (AEF) from a representative subject using Champagne, thin and full Dugh. In this case, we tested the reconstruction performance of all algorithms with the number of trials limited to 20 and 120. As Figure 6 demonstrates, the performance of Dugh remains robust as the number of trials is increased to 20 and 120 in Figure 6. Finally, the VEF performance of benchmark algorithm eLORETA is demonstrated in Figure 7.

L.2 EEG Data: Faces vs scrambled pictures

A publicly available EEG dataset (128-channel Biosemi ActiveTwo system) was downloaded from the SPM website (http://www.fil.ion.ucl.ac.uk/spm/data/mmfaces) and the lead field was calculated

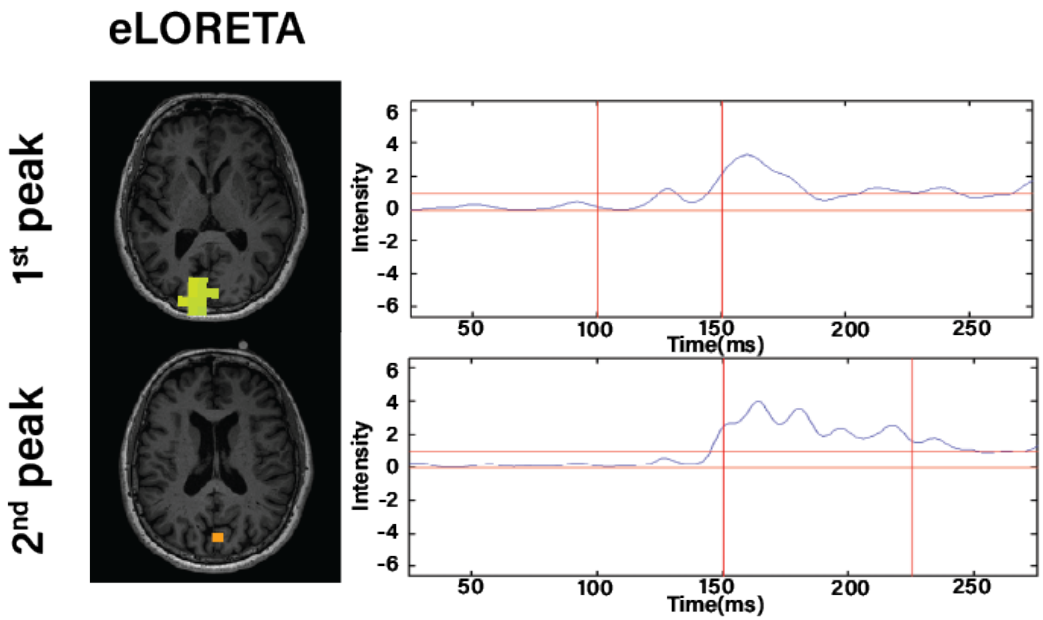


Figure 7: VEF performance of benchmark algorithm eLORETA. Both these benchmarks did not yield reliable results for 5 trial epochs. Even when the number of trials were increased to 20, benchmarks performance yielded neither good spatial localization of the two visual cortical areas nor good estimation of the time courses of these activations.

in SPM8 using a three-shell spherical model at the coarse resolution of 5124 voxels at approximately 8 mm spacing. These EEG data were also obtained during a visual response paradigm that involved randomized presentation of at least 86 faces and 86 scrambled faces. To examine the differential responses to faces across all trials, the averaged responses to scrambled-faces were subtracted from the averaged responses to faces. The result is demonstrated in Figure 8.

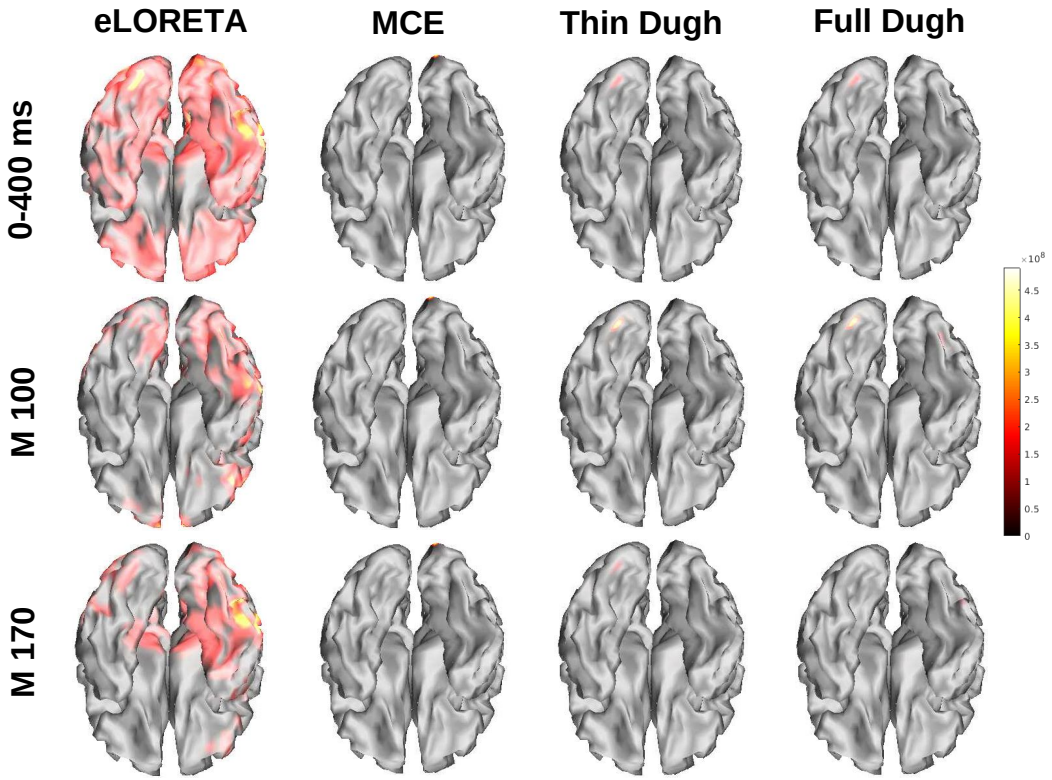


Figure 8: Performance of Dugh and benchmarks on EEG data acquired during a face recognition task. Dugh was able to provide more focal and distinct activations for the M100 and M170 responses that were not clearly identified using the benchmarks eLORETA and MCE.