JOINT LEARNING OF FULL-STRUCTURE NOISE IN HIERARCHICAL BAYESIAN REGRESSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider hierarchical Bayesian (type-II maximum likelihood) regression models for observations with latent variables for source and noise, where parameters of priors for source and noise terms need to be estimated jointly from data. This problem has application in many domains in imaging including biomagnetic inverse problems. Crucial factors influencing accuracy of source estimation are not only the noise level but also its correlation structure, but existing approaches have not addressed estimation of a full-structure noise covariance matrix. Here, we focus on sparse Bayesian learning (SBL) in regression models specifically for the application of reconstruction of brain activity from electroencephalography (EEG), which can be formulated as a linear regression with independent Gaussian scale mixture priors for both the source and noise components. As a departure from the classical SBL models where across sensor observations are assumed to be independent and identically distributed, we consider Gaussian noise with full covariance structure. Using ideas from Riemannian geometry, we derive an efficient algorithm for updating both source and the noise covariance along the manifold of positive definite matrices. We verify the theoretical advantages of our model and algorithm in extensive simulations in the context of the EEG inverse problem. Our results demonstrate that the novel framework significantly improves upon state-of-the-art techniques in the real-world scenario where the noise is indeed non-diagonal and fully-structured.

1 INTRODUCTION

Having precise knowledge of the noise distribution is a fundamental requirement for obtaining accurate solutions in many regression problems (Bungert et al., 2020). In many applications however, it is impossible to separately estimate this noise distribution, as distinct baseline "noise-only" measurements are not feasible. An alternative therefore is to design estimators that jointly optimize over the regression coefficients as well as over parameters of the noise distribution. This has been pursued both in a (penalized) maximum-likelihood settings (here referred to as *Type-I* approaches) (Petersen & Jung, 2020; Bertrand et al., 2019; Massias et al., 2018) as well as in hierarchical Bayesian settings (referred to as *Type-II*) (Wipf & Rao, 2007; Zhang & Rao, 2011; Hashemi et al., 2020; Cai et al., 2020). Most contributions in the literature are, however, limited to the estimation of only a diagonal noise covariance (i.e., independent between different measurements) (Daye et al., 2012; Van de Geer et al., 2013; Dalalyan et al., 2013; Lederer & Muller, 2015). Considering a diagonal noise covariance is a limiting assumption in practice as the noise interference in many realistic scenarios are highly correlated across measurements; and thus, have non-trivial off-diagonal elements.

In this paper, we consider the problem of electromagnetic brain source imaging (BSI) as our main application. The goal of BSI is to reconstruct brain activity from electroencephalography (EEG), which can be formulated as a linear Bayesian regression model with independent Gaussian priors on the parameters (known as sparse Bayesian learning, SBL) and noise. As a departure from the classical SBL approaches, here we specifically consider Gaussian noise with full covariance structure. Prominent source of correlated noise in this context are, for example, eye blinks, heart beats, muscular artifacts and line noise. Other realistic examples for the need for such full-structure noise can be found in the areas of array processing (Li & Nehorai, 2010) or direction of arrival (DOA) estimation (Chen et al., 2008). Designing an algorithm that can accurately estimate noise with full covariance structure is hypothesized to benefit accurate regression source coefficient estimation.

Here we argue that explicitly modeling the noise dependency structure is a key factor to improve the accuracy of a large class of regression problems including those introduced above. The main contribution of this paper is to develop an efficient optimization algorithm for jointly estimating the posterior of regression parameters as well as the noise distribution. More specifically, we consider linear regression with univariate Gaussian priors on the parameters and a full-structure multivariate Gaussian noise. We cast the problem as a hierarchical Bayesian (type-II maximum-likelihood) regression problem, in which the prior parameter variances and the noise covariance matrix are treated as hyperparameters that are optimized by maximizing the Bayesian evidence of the model. Building on theoretical results from Riemannian geometry, we derive an efficient algorithm for jointly estimating the source and noise covariances along the manifold of positive definite (P.D.) matrices.

2 Type-II Bayesian Regression

We consider the linear model $\mathbf{Y} = \mathbf{L}\mathbf{X} + \mathbf{E}$, in which a forward or design matrix, $\mathbf{L} \in \mathbb{R}^{M \times N}$, is mapped to the measurements, $\mathbf{Y}$, by a set of coefficients or source components, $\mathbf{X}$. Depending on the setting, the problem of estimating $\mathbf{X}$ given $\mathbf{L}$ and $\mathbf{Y}$ is called an inverse problem in physics, a multi-task regression problem in machine learning, or a multiple measurement vector (MMV) recovery problem in signal processing (Cotter et al., 2005). Adopting a signal processing terminology, the *measurement matrix* $\mathbf{Y} \in \mathbb{R}^{M \times T}$ captures the activity of M sensors at T time instants, $\mathbf{y}(t) \in \mathbb{R}^{M \times 1}, t = 1, \ldots, T$, while the *source matrix*, $\mathbf{X} \in \mathbb{R}^{N \times T}$, consists of the unknown activity of N sources at the same time instants, $\mathbf{x}(t) \in \mathbb{R}^{N \times 1}, t = 1, \ldots, T$. The matrix $\mathbf{E} = [\mathbf{e}(1), \ldots, \mathbf{e}(T)] \in \mathbb{R}^{M \times T}$ represents T time instances of zero-mean Gaussian noise with full covariance $\mathbf{\Lambda}$, $\mathbf{e}(t) \in \mathbb{R}^{M \times 1} \sim \mathcal{N}(0, \mathbf{\Lambda}), t = 1, \ldots, T$, which is assumed to be independent of the source activations.

In this paper, we focus on electro- or magnetoencephalography (EEG/MEG) based brain source imaging (BSI) but the proposed algorithm can be used in general regression settings, in particular for sparse signal recovery (Candès et al., 2006; Donoho, 2006) with a wide range of applications (Malioutov et al., 2005). The goal of BSI is to infer the underlying brain activity $\mathbf{X}$ from the EEG/MEG measurement $\mathbf{Y}$ given a known forward operator, called lead field matrix $\mathbf{L}$. In practice, $\mathbf{L}$ can be computed using discretization methods such as the finite element method (FEM) for a given head geometry and known electrical conductivities using the quasi-static approximation of Maxwell's equations (Hämäläinen et al., 1993; Baillet et al., 2001). As the number of sensors is typically much smaller than the number of locations of potential brain sources, this inverse problem is highly ill-posed. This problem is addressed by regularization techniques (Benning & Burger, 2018), where the main idea is to incorporate prior knowledge or assumptions about the nature of the true sources into the inverse problem. Such assumptions can be indeed encoded in various ways resulting in two main categories of algorithms including Maximum-a-Posteriori (MAP) estimation or *Type-I Bayesian learning* (Pascual-Marqui et al., 1994; Gorodnitsky et al., 1995; Haufe et al., 2008; Gramfort et al., 2012; Castaño-Candamil et al., 2015) on one hand and Type-II Maximum-Likelihood (Type-II ML) estimation or simply *Type-II Bayesian learning* (Tipping, 2001; Wipf & Nagarajan, 2009; Seeger & Wipf, 2010; Wu et al., 2016) on the other hand. Note that many regularization techniques can be also viewed as implicit ways of learning the noise level since they involve tuning a regularization parameter as a function of noise level (either explicitly embedded in the optimization of Type-I learning or implicitly for Type-II learning).

Most existing BSI approaches assume the availability of separate baseline data, from which the noise distribution can be estimated. Engemann & Gramfort (2015) proposed a method for selecting a single regularization parameter based on cross-validation and maximum-likelihood estimation, while Huizenga et al. (2002); De Munck et al. (2002); Bijma et al. (2003); De Munck et al. (2004) assume more complex spatiotemporal noise covariance structures. A common feature of these works is, however, that the noise level is not estimated as part of the source reconstruction problem on task-related data but from separate noise recordings. The approach proposed here substantially differs in this respect, as it learns the noise covariance jointly with the brain source distribution. Note that the idea of joint estimation of brain source activity and noise covariance has been previously proposed for Type-I learning methods in Massias et al. (2018); Bertrand et al. (2019).

In contrast to Type-I methods, in this paper, we focus on Type-II Bayesian learning, which assumes a family of prior distributions $p(\mathbf{X}|\mathbf{\Theta})$ parameterized by a set of hyper-parameters $\mathbf{\Theta}$. These hyper-parameters can be learned from the data along with the model parameters using a hierarchical

Bayesian approach (Tipping, 2001; Wipf & Rao, 2004) through the maximum-likelihood principle:

$$\mathbf{\Theta}^{\text{II}} := \arg\max_{\mathbf{\Theta}} p(\mathbf{Y}|\mathbf{\Theta}) = \arg\max_{\mathbf{\Theta}} \int p(\mathbf{Y}|\mathbf{X},\mathbf{\Theta})p(\mathbf{X}|\mathbf{\Theta})\mathrm{d}\mathbf{X} \ . \tag{1}$$

Here we assume a zero-mean Gaussian prior with full covariance $\mathbf{\Gamma}$ for the underlying source distribution, $\mathbf{x}(t) \in \mathbb{R}^{N\times 1} \sim \mathcal{N}(0,\mathbf{\Gamma}), t = 1,\ldots,T$. Just as most other approaches, Type-II Bayesian learning makes the simplifying assumption of statistical independence between time samples. This leads to the following expression for the distribution of the sources and measurements:

$$p(\mathbf{X}|\mathbf{\Gamma}) = \prod_{t=1}^{T} p(\mathbf{x}(t)|\mathbf{\Gamma}) = \prod_{t=1}^{T} \mathcal{N}(0,\mathbf{\Gamma}) \tag{2}$$

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{t=1}^{T} p(\mathbf{y}(t)|\mathbf{x}(t)) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{L}\mathbf{x}(t),\mathbf{\Lambda}) \ . \tag{3}$$

The parameters of the Type-II model, Θ, are the unknown source and noise covariances, i.e., $\Theta = \{\mathbf{\Gamma},\mathbf{\Lambda}\}$. The unknown parameters $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$ are optimized based on the current estimates of the source and noise covariances in an alternating iterative process. Given initial estimates of $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$, the posterior distribution of the sources is a Gaussian of the form (Sekihara & Nagarajan, 2015)

$$p(\mathbf{X}|\mathbf{Y},\mathbf{\Gamma}) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}(t),\mathbf{\Sigma}_{\mathbf{x}}) \text{ , where} \tag{4}$$

$$\boldsymbol{\mu}_{\mathbf{x}}(t) = \mathbf{\Gamma}\mathbf{L}^{\top}(\mathbf{\Sigma}_{\mathbf{y}})^{-1}\mathbf{y}(t) \tag{5}$$

$$\mathbf{\Sigma}_{\mathbf{x}} = \mathbf{\Gamma} - \mathbf{\Gamma}\mathbf{L}^{\top}(\mathbf{\Sigma}_{\mathbf{y}})^{-1}\mathbf{L}\mathbf{\Gamma} \tag{6}$$

$$\mathbf{\Sigma}_{\mathbf{y}} = \mathbf{\Lambda} + \mathbf{L}\mathbf{\Gamma}\mathbf{L}^{\top} \ . \tag{7}$$

The estimated posterior parameters $\boldsymbol{\mu}_{\mathbf{x}}(t)$ and $\mathbf{\Sigma}_{\mathbf{x}}$ are then in turn used to update $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$ as the minimizers of the negative log of the marginal likelihood $p(\mathbf{Y}|\mathbf{\Gamma},\mathbf{\Lambda})$, which is given by (Wipf et al., 2010):

$$\begin{aligned}\mathcal{L}^{\text{II}}(\mathbf{\Gamma},\mathbf{\Lambda}) &= -\log p(\mathbf{Y}|\mathbf{\Gamma},\mathbf{\Lambda}) = \log|\mathbf{\Sigma}_{\mathbf{y}}| + \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}(t)^{\top}\mathbf{\Sigma}_{\mathbf{y}}^{-1}\mathbf{y}(t) \\ &= \log|\mathbf{\Lambda} + \mathbf{L}\mathbf{\Gamma}\mathbf{L}^{\top}| + \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}(t)^{\top}\left(\mathbf{\Lambda} + \mathbf{L}\mathbf{\Gamma}\mathbf{L}^{\top}\right)^{-1}\mathbf{y}(t) \ ,\end{aligned} \tag{8}$$

where $|\cdot|$ denotes the determinant of a matrix. This process is repeated until convergence. Given the final solution of the hyperparameters $\Theta^{\text{II}} = \{\mathbf{\Gamma}^{\text{II}},\mathbf{\Lambda}^{\text{II}}\}$, the posterior source distribution is obtained by plugging these estimates into equations 3 to 6.

3 Proposed Method: Full-structure Noise (FUN) Learning

Here we propose a novel and efficient algorithm, full-structure noise *(FUN)* learning, which is able to learn the full covariance structure of the noise jointly within the Bayesian Type-II regression framework. We first formulate the algorithm in its most general form, in which both the noise distribution and the prior have full covariance structure. Later, we make the simplifying assumption of independent source priors, leading to the pruning of the majority of sources. This effect, which has also been referred to as automatic relevance determination (ARD) or sparse Bayesian learning (SBL) is beneficial in our application of interest, namely the reconstruction of parsimonious sets of brain sources underlying experimental EEG measurements.

Note that the Type-II cost function in equation 8 is non-convex, which is non-trivial to optimize. A number of iterative algorithms such as *majorization-minimization* (MM) (Sun et al., 2017) have been proposed to address this challenge. Building on the MM class of algorithms, we first construct convex surrogate functions that *majorizes* $\mathcal{L}^{\text{II}}(\mathbf{\Gamma},\mathbf{\Lambda})$ in each iteration of the optimization algorithm. Then, we show the minimization equivalence between the constructed majoring functions and equation 8. This result is presented in the following theorem:

Theorem 1. *Let $\mathbf{\Lambda}^k$ and $\mathbf{\Sigma}_{\mathbf{y}}^k$ be fixed values obtained in the (k)-th iteration of the optimization algorithm minimizing $\mathcal{L}^{II}(\mathbf{\Gamma}, \mathbf{\Lambda})$. Then, optimizing the non-convex type-II ML cost function in equation 8, $\mathcal{L}^{II}(\mathbf{\Gamma}, \mathbf{\Lambda})$, with respect to $\mathbf{\Gamma}$ is equivalent to optimizing the following convex function, which* majorizes *equation 8:*

$$\mathcal{L}_{\text{source}}^{\text{conv}}(\mathbf{\Gamma}, \mathbf{\Lambda}^k) = \text{tr}(\left(\mathbf{C}_{\text{S}}^k\right)^{-1} \mathbf{\Gamma}) + \text{tr}(\mathbf{M}_{\text{S}}^k \mathbf{\Gamma}^{-1}) \ , \tag{9}$$

where $\mathbf{C}_{\text{S}}^k$ and $\mathbf{M}_{\text{S}}^k$ are defined as:

$$\mathbf{C}_{\text{S}}^k := \left(\mathbf{L}^\top \left(\mathbf{\Sigma}_{\mathbf{y}}^k\right)^{-1} \mathbf{L}\right)^{-1} \ , \quad \mathbf{M}_{\text{S}}^k := \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}^k(t)^\top \mathbf{x}^k(t) \ . \tag{10}$$

Similarly, optimizing $\mathcal{L}^{II}(\mathbf{\Gamma}, \mathbf{\Lambda})$ with respect to $\mathbf{\Lambda}$ is equivalent to optimizing the following convex majorizing function:

$$\mathcal{L}_{\text{noise}}^{\text{conv}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}) = \text{tr}(\left(\mathbf{C}_{\text{N}}^k\right)^{-1} \mathbf{\Lambda}) + \text{tr}(\mathbf{M}_{\text{N}}^k \mathbf{\Lambda}^{-1}) \ , \tag{11}$$

where $\mathbf{C}_{\text{N}}^k$ and $\mathbf{M}_{\text{N}}^k$ are defined as:

$$\mathbf{C}_{\text{N}}^k := \left(\mathbf{\Sigma}_{\mathbf{y}}^k\right) \ , \quad \mathbf{M}_{\text{N}}^k := \frac{1}{T} \sum_{t=1}^{T} (\mathbf{y}(t) - \mathbf{L}\mathbf{x}^k(t))^\top (\mathbf{y}(t) - \mathbf{L}\mathbf{x}^k(t)). \tag{12}$$

Proof. The proof is presented in Appendix A. □

We continue by focusing on the optimization of the cost functions $\mathcal{L}_{\text{source}}^{\text{conv}}(\mathbf{\Gamma}, \mathbf{\Lambda}^k)$ and $\mathcal{L}_{\text{noise}}^{\text{conv}}(\mathbf{\Gamma}^k, \mathbf{\Lambda})$ with respect to $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$, respectively. Note that by considering full-structure source covariances, the solution of $\mathcal{L}_{\text{source}}^{\text{conv}}(\mathbf{\Gamma}, \mathbf{\Lambda}^k)$ with respect to $\mathbf{\Gamma}$ lies in an $(N^2 - N)/2$ Riemannian manifold of positive definite (P.D.) matrices. This consideration enables us to invoke efficient methods from Riemannian geometry (see Petersen et al., 2006; Berger, 2012; Jost & Jost, 2008), which ensures that the solution at each step of the optimization is contained within the low-dimensional solution space. Specifically, in order to optimize for the source covariance, at each iteration the algorithm calculates the geometric mean between the previously obtained statistical model source covariance, $\mathbf{C}_{\text{S}}^k$, and the source-space sample covariance matrix, $\mathbf{M}_{\text{S}}^k$. The same arguments analogously hold for optimization of the noise covariance matrix. Therefore, to update the noise covariance estimate, the algorithm calculates the geometric mean between the model noise covariance, $\mathbf{C}_{\text{N}}^k$, and the empirical sensor-space residuals, $\mathbf{M}_{\text{N}}^k$. Optimization of these cost-functions and the resulting update rules are summarized in the following theorem:

Theorem 2. *Considering geodesic paths that lie on the P.D. manifold, the cost functions $\mathcal{L}_{\text{source}}^{\text{conv}}(\mathbf{\Gamma}, \mathbf{\Lambda}_k)$ and $\mathcal{L}_{\text{noise}}^{\text{conv}}(\mathbf{\Gamma}_k, \mathbf{\Lambda})$ both are strictly geodesically convex on this specific domain, and their optimal solution with respect to $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$, respectively, can be attained according to the two following update rules:*

$$\mathbf{\Gamma}^{k+1} \leftarrow (\mathbf{C}_{\text{S}}^k)^{\frac{1}{2}} \left((\mathbf{C}_{\text{S}}^k)^{-1/2} \mathbf{M}_{\text{S}}^k (\mathbf{C}_{\text{S}}^k)^{-1/2}\right)^{\frac{1}{2}} (\mathbf{C}_{\text{S}}^k)^{\frac{1}{2}} \ , \tag{13}$$

$$\mathbf{\Lambda}^{k+1} \leftarrow (\mathbf{C}_{\text{N}}^k)^{\frac{1}{2}} \left((\mathbf{C}_{\text{N}}^k)^{-1/2} \mathbf{M}_{\text{N}}^k (\mathbf{C}_{\text{N}}^k)^{-1/2}\right)^{\frac{1}{2}} (\mathbf{C}_{\text{N}}^k)^{\frac{1}{2}} \ . \tag{14}$$

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

As a result of Theorem 2, the source and noise covariance matrices that optimize cost function $\mathcal{L}^{\text{II}}(\mathbf{\Gamma}, \mathbf{\Lambda})$ can be obtained by alternating between equation 13 and equation 14 until convergence.

3.1 Sparse Bayesian Learning with Full Noise Modeling

Note that, in brain source imaging, the assumption of full source covariance is often relaxed. Even if, technically, most parts of the brain are active at all times, and the concurrent activations of different brain regions can never be assumed to be fully uncorrelated, there are many experimental settings in which it is reasonable to assume only a small set of independent brain sources. Such sparse solutions

are physiologically plausible in task-based analyses, where only a fraction of the brain's macroscopic structures is expected to be consistently engaged. A common modeling strategy in this case is to model independent sources through a diagonal covariance matrix. In the Type-II Bayesian learning framework, this simplification interestingly leads to sparsity of the resulting source distributions, as, at the optimum, many of the estimated source variances are zero. This mechanism that is known as sparse Bayesian learning (SBL) and is closely related to the more general concept of automatic relevance determination. Here, we adopt the SBL assumption for the sources, leading to $\mathbf{\Gamma}$-updates previously described in the BSI literature under the name Champagne (Wipf & Nagarajan, 2009). As a novelty and main focus of this paper, we here equip the SBL framework with the capability to jointly learn full noise covariances through the geometric mean based update rule in equation 14.

In the SBL framework, the N modeled brain sources are assumed to follow independent univariate Gaussian distributions with zero mean and distinct unknown variances γ_n: $x_n(t) \sim \mathcal{N}(0, \gamma_n), n = 1, \ldots, N$. In the SBL solution, the majority of variances is zero, thus effectively inducing spatial sparsity of the corresponding source activities. For FUN learning, we also impose a diagonal structure on the source covariance matrix, $\mathbf{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$, where $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_N]^\top$. By constraining $\mathbf{\Gamma}$ in equation 9 to the set of diagonal matrices, $\mathcal{W}$, we can show that the update rule equation 13 for the source variances simplifies to the following form:

$$\gamma_n^{k+1} \leftarrow \sqrt{\frac{\left[\left(\mathbf{M}_{\mathrm{S}}^k\right)^{-1}\right]_{n,n}}{\left[\left(\mathbf{C}_{\mathrm{S}}^k\right)^{-1}\right]_{n,n}}} = \sqrt{\frac{\frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_n^k(t))^2}{\mathbf{L}_n^\top \left(\mathbf{\Sigma}_{\mathbf{y}}^k\right)^{-1} \mathbf{L}_n}} \quad \text{for } n = 1, \ldots, N \ , \tag{15}$$

where $\mathbf{L}_n$ denotes the n-th column of the lead field matrix. Interestingly, equation 15 is identical to the update rule of the *Champagne* algorithm. A detailed derivation of equation 15 can be found in Appendix C.

Summarizing, the FUN learning approach, just like Champagne and other SBL algorithms, assumes independent Gaussian sources with individual variances (thus, diagonal source covariances), which are updated through equation equation 15. Departing from the classical SBL setting, which assumes the noise distribution to be known, FUN models noise with full covariance structure, which is updated using equation 14. Algorithm 1 summarizes the used update rules.

Algorithm 1: Full-structure noise (FUN) learning

Input: The lead field matrix $\mathbf{L} \in \mathbb{R}^{M \times N}$ and the measurement vectors $\mathbf{y}(t) \in \mathbb{R}^{M \times 1}, t = 1, \ldots, T$.

Result: The estimated prior source variances $[\gamma_1, \ldots, \gamma_N]^\top$, noise covariance $\mathbf{\Lambda}$, the posterior mean $\boldsymbol{\mu}_{\mathbf{x}}(t)$ and covariance $\mathbf{\Sigma}_{\mathbf{x}}$ of the sources.

1 Set a random initial value for $\mathbf{\Lambda}$ as well as $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_N]^\top$, and construct $\mathbf{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$.

2 Calculate the statistical covariance $\mathbf{\Sigma}_{\mathbf{y}} = \mathbf{\Lambda} + \mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top$.

Repeat

3 Calculate the posterior mean as $\boldsymbol{\mu}_{\mathbf{x}}(t) = \mathbf{\Gamma}\mathbf{L}^\top(\mathbf{\Sigma}_{\mathbf{y}})^{-1}\mathbf{y}(t)$.

4 Calculate $\mathbf{C}_{\mathrm{S}}^k$ and $\mathbf{M}_{\mathrm{S}}^k$ based on equation 10, and update γ_n for $n = 1, \ldots, N$ based on equation 15.

5 Calculate $\mathbf{C}_{\mathrm{N}}^k$ and $\mathbf{M}_{\mathrm{N}}^k$ based on equation 12, and update $\mathbf{\Lambda}$ based on equation 14.

Until *stopping condition is satisfied*;

6 Calculate the posterior covariance as $\mathbf{\Sigma}_{\mathbf{x}} = \mathbf{\Gamma} - \mathbf{\Gamma}\mathbf{L}^\top(\mathbf{\Sigma}_{\mathbf{y}})^{-1}\mathbf{L}\mathbf{\Gamma}$.

Note that various recent Type-II noise learning schemes for diagonal noise covariance matrices (Hashemi et al., 2020; Cai et al., 2020) that are rooted in the concept of SBL can be also derived as special cases of FUN learning assuming diagonal source and noise covariances, i.e., $\mathbf{\Gamma}, \mathbf{\Lambda} \in \mathcal{W}$. Specifically imposing diagonal structure on the noise covariance matrix for the FUN algorithm, $\mathbf{\Lambda}$, results in identical noise variance update rules as derived in Cai et al. (2020) for *heteroscedastic*, and in Hashemi et al. (2020) for *homoscedastic* noise. We explicitly demonstrate this connection in Appendix D. Here, we note that *heteroscedasticity* refers to the common phenomenon that measurements are contaminated with non-uniform noise levels across channels, while *homoscedasticity* only accounts for uniform noise levels.

4 NUMERICAL SIMULATIONS

Source, Noise and Forward Model: We simulated a sparse set of $N_0 = 5$ active brain sources that were placed at random positions on the cortex. To simulate the electrical neural activity of these sources, $T = 200$ identically and independently distributed (i.i.d) points were sampled from a Gaussian distribution, yielding sparse source activation vectors $\mathbf{x}(t)$. 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 as the forward operator: $\mathbf{Y}^{\text{signal}} = \mathbf{L}\mathbf{X}$. The lead field matrix, $\mathbf{L} \in \mathbb{R}^{58 \times 2004}$, was generated using the New York Head model (Huang et al., 2016) taking into account the realistic anatomy and electrical tissue conductivities of an average human head. Further details regarding forward modeling is provided in Appendix E. Gaussian additive noise was randomly sampled from a zero-mean normal distribution with *full* covariance matrix $\mathbf{\Lambda}$: $\mathbf{e}(t) \in \mathbb{R}^{M \times 1} \sim \mathcal{N}(0, \mathbf{\Lambda}), t = 1, \ldots, T$. This setting is further referred to as *full-structure noise*. Note that we also generated noise with diagonal covariance matrix, referred to as *heteroscedastic noise*, in order to investigate the effect of model violation on reconstruction performance. The noise matrix $\mathbf{E} = [\mathbf{e}(1), \ldots, \mathbf{e}(T)] \in \mathbb{R}^{M \times T}$ is normalized by it Frobenius norm and added to the signal matrix $\mathbf{Y}^{\text{signal}}$ as follows: $\mathbf{Y} = \mathbf{Y}^{\text{signal}} + \left({}^{(1-\alpha)} \|\mathbf{Y}^{\text{signal}}\|_F / {}_{\alpha} \|\mathbf{E}\|_F \right) \mathbf{E}$, 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 the subsequently described experiments the following values of α were used: α={0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.65, 0.7, 0.8}, which correspond to the following SNRs: SNR={-12, -7.4, -5.4, -3.5, -1.7, 0, 1.7, 3.5, 5.4, 7.4, 12} (dB).

Evaluation Metrics and Simulation Set-up: We applied the full-structure noise learning approach on the synthetic datasets described above to recover the locations and time courses of the active brain sources. In addition to our proposed approach, two further Type-II Bayesian learning schemes, namely homoscedastic and heteroscedastic Champagne (Hashemi et al., 2020; Cai et al., 2020), were also included as benchmarks with respect to source reconstruction performance and noise covariance estimation accuracy. Source reconstruction performance was evaluated according to the *earth mover's distance* (EMD) (Rubner et al., 2000)), the error in the reconstruction of the source time courses, the average Euclidean distance (EUCL) (in mm) between each simulated source and the best (in terms of absolute correlations) matching reconstructed source, and finally F1-measure score (Chinchor & Sundheim, 1993). A detailed definition of evaluation metrics is provided in Appendix E. To evaluate the accuracy of the noise covariance matrix estimation, the following two metrics were calculated: the Pearson correlation between the original and reconstructed noise covariance matrices, $\mathbf{\Lambda}$ and $\hat{\mathbf{\Lambda}}$, denoted by $\mathbf{\Lambda}^{\text{sim}}$, and the normalized mean squared error (NMSE) between $\mathbf{\Lambda}$ and $\hat{\mathbf{\Lambda}}$, defined as: $\text{NMSE} = ||\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}||_F^2 / ||\mathbf{\Lambda}||_F^2$. Note that NMSE measures the reconstruction of the true scale of the noise covariance matrix, while $\mathbf{\Lambda}^{\text{sim}}$ is scale-invariant and hence only quantifies the overall structural similarity between simulated and estimated noise covariance matrices. 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. Convergence of the optimization programs for each run 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.

Figure 1 shows two simulated datasets with five active sources in presence of full-structure noise (upper panel) as well as heteroscedastic noise (lower panel) at 0 (dB) SNR. Topographic maps depict the locations of the ground-truth active brain sources (first column) along with the source reconstruction result of three noise learning schemes assuming noise with homoscedastic (second column), heteroscedastic (third column), and full (fourth column) structure. For each algorithm, the estimated noise covariance matrix is also plotted above the topographic map. Source reconstruction performance was measured in terms of EMD and time course correlation (Corr), and is summarized in the table next to each panel. Besides, the accuracy of the noise covariance matrix reconstruction was measured on terms of $\mathbf{\Lambda}^{\text{sim}}$ and NMSE. Results are included in the same table.

Figure 1 (upper panel) allows for a direct comparison of the estimated noise covariance matrices obtained from the three different noise learning schemes. It can be seen that FUN learning can better capture the overall structure of ground truth full-structure noise as evidenced by lower NMSE and similarity errors compared to the heteroscedastic and homoscedastic algorithm variants that are only able to recover a diagonal matrix while enforcing the off-diagonal elements to zero. This behaviour

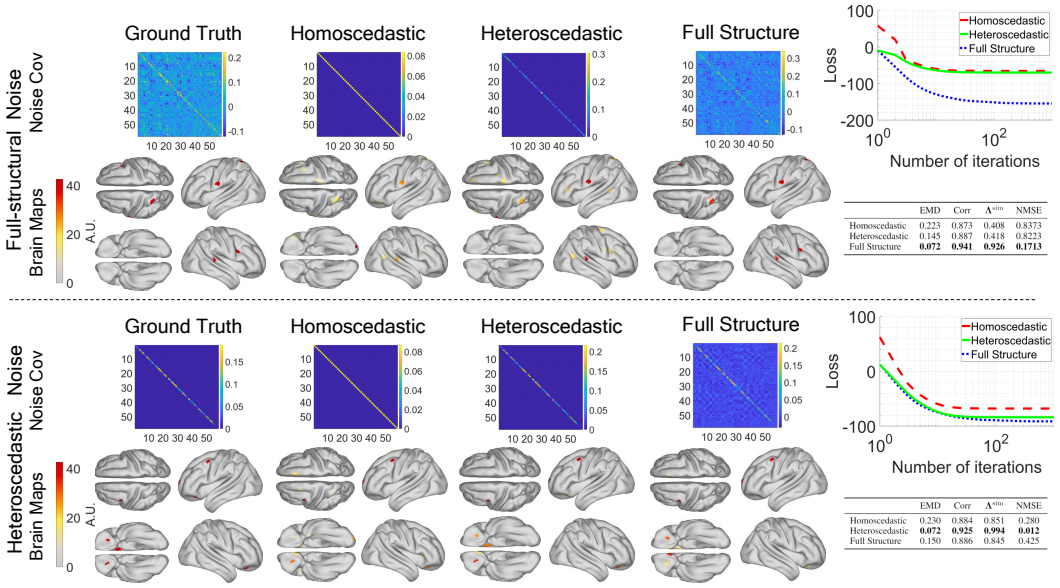

Figure 1: Two examples of the simulated data with five active sources in presence of full-structure noise (upper panel) as well as heteroscedastic noise (lower panel) at 0 (dB) SNR. Topographic maps depict the locations of the ground-truth active brain sources along with the source reconstruction results of three noise learning schemes. For each algorithm, the estimated noise covariance matrix is also plotted above the topographic maps. The source reconstruction performance of these examples in terms of EMD and time course correlation (Corr) is summarized in the associated table next to each panel. We also report the accuracy with which the ground-truth noise covariance was estimated in terms of the $\mathbf{\Lambda}^{\text{sim}}$ and NMSE. The convergence behaviour of all three noise estimation approaches is also shown.

results in higher spatial and temporal accuracy (lower EMD and time course error) for FUN learning compared to competing algorithms assuming diagonal noise covariance. This advantage is also visible in the topographic maps. The lower-panel of Figure 1 presents analogous results for the setting where the noise covariance is generated according to a heteroscedastic model. Note that the superior spatial and temporal reconstruction performance of the heteroscedastic noise learning algorithm compared to the full-structure scheme is expected here because the simulated ground truth noise is indeed heteroscedastic. The full-structure noise learning approach, however, provides fairly reasonable performance in terms of EMD, time course correlation (corr), and $\mathbf{\Lambda}^{\text{sim}}$, although it is designed to estimate a full-structure noise covariance matrix. The convergence behaviour of all three noise learning variants is also illustrated in Figure 1. Note that the full-structure noise learning approach reaches lower negative log-likelihood values eventually in both scenarios, namely full-structure and heteroscedastic noise.

Figure 2 shows the EMD, the time course reconstruction error, the EUCL and the F1 measure score incurred by three different noise learning approaches assuming homoscedastic (red), heteroscedastic (green) and full-structure (blue) noise covariances for a range of 10 SNR values. The upper panel represents the evaluation metrics for the setting where the noise covariance is full-structure model, while the lower-panel depicts the same metric for simulated noise with heteroscedastic diagonal covariance. Concerning the first setting, FUN learning consistently outperforms its homoscedastic and heteroscedastic counterparts according to all evaluation metrics in particular in low-SNR settings. Consequently, as the SNR decreases, the gap between FUN learning and the two other variants increases. Conversely, heteroscedastic noise learning shows an improvement over FUN learning according to all evaluation metrics when the simulated noise is indeed heteroscedastic. However, note that the magnitude of this improvement is not as large as observed for the setting where the noise covariance is generated according to a full-structure model and then is estimated using the FUN approach.

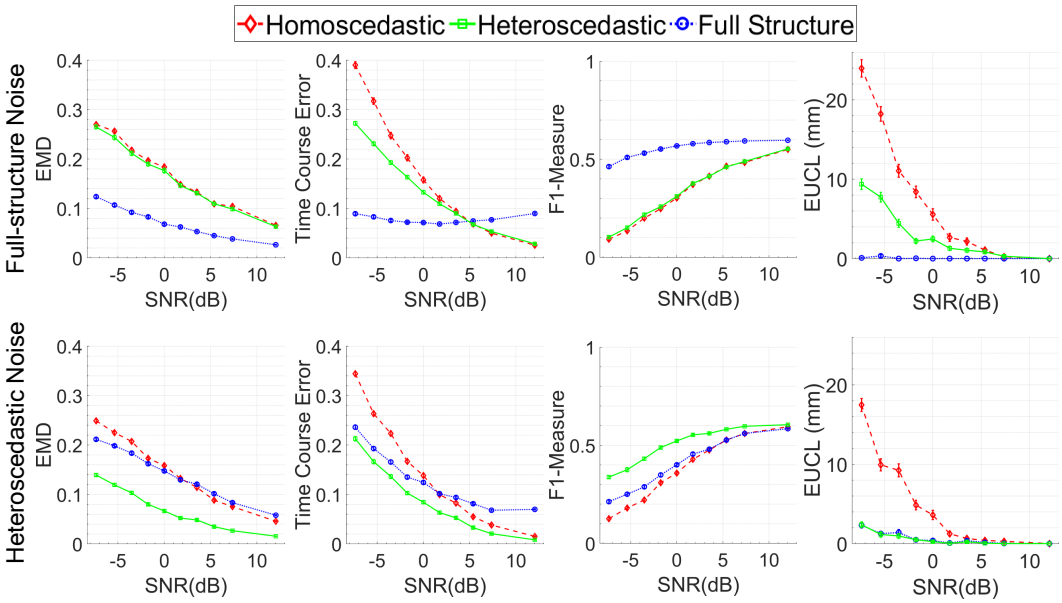

Figure 2: Source reconstruction performance (mean $\pm$ SEM) of the three different noise learning schemes for data generated by a realistic lead field matrix. Generated sensor signals were superimposed by either full-structure or heteroscedastic noise covering a wide range of SNRs. Performance was measured in terms of the earth mover's distance (EMD), time-course correlation error, F1-measure and Euclidean distance (EUCL) in (mm) between each simulated source and the reconstructed source with highest maximum absolute correlation.

5 Discussion

5.1 FUN Learning as Majorization-Minimization

Several recent Type-I noise learning algorithms benefit from guaranteed convergence (Bertrand et al., 2019). Here, we show that convergence to a local optimum is guaranteed also for FUN. Precisely, we proof that FUN is an instance of the wider class of majorization-minimization (MM) algorithms, for which convergence is guaranteed by construction (see Appendix F).

5.2 Limitations and Future Work

In this article, we assumed Gaussian noise, which made the joint Bayesian learning of source and full-structure noise distributions tractable. Although this assumption is common, it would be desirable to also include more robust (e.g., heavy-tailed) non-Gaussian noise distributions in our framework. Another limitation is that the superior performance of the full-structure noise learning technique comes at the expense of higher computational complexity compared to the variants assuming homoscedastic or heteroscedastic noise. Imposing physiologically plausible constraints on the noise model, e.g., low-rank Toeplitz structure, could reduce the computational cost in two ways: a) by reducing the number of parameters and b) by taking advantage of efficient implementations using circular embeddings and the fast Fourier transform (Babu, 2016).

6 Conclusion

This paper proposes an efficient optimization algorithm for jointly estimating Gaussian regression parameter distributions as well as Gaussian noise distributions with full covariance structure within a hierarchical Bayesian framework. Using the Riemannian geometry of positive definite matrices, we derived an efficient algorithm for jointly estimating source and noise covariances. The benefits of our proposed framework were evaluated within an extensive set of experiments in the context of electromagnetic brain source imaging inverse problem and showed significant improvement upon state-of-the-art techniques in the realistic scenario where the noise has full covariance structure.

REFERENCES

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

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

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.

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

Martin Benning and Martin Burger. Modern regularization methods for inverse problems. *Acta Numerica*, 27:1, 2018.

Marcel Berger. *A panoramic view of Riemannian geometry*. Springer Science & Business Media, 2012.

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*, pp. 3959–3970, 2019.

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

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.

Ingo Bojak. Neural population models and cortical field theory: overview. 2014.

Ingo Bojak and Michael Breakspear. Neuroimaging, neural population models for. 2014.

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.

Leon Bungert, Martin Burger, Yury Korolev, and Carola-Bibiane Schoenlieb. Variational regularisation for inverse problems with imperfect forward operators and general noise models. *arXiv preprint arXiv:2005.14131*, 2020.

Chang Cai, Mithun Diwakar, Ali Hashemi, Stefan Haufe, Kensuke Sekihara, and Srikantan S Nagarajan. Robust estimation of noise for electromagnetic brain imaging with the Champagne algorithm. *bioRxiv*, 2020.

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.

Sebastián Castaño-Candamil, Johannes Höhne, Juan-David Martínez-Vargas, Xing-Wei An, German Castellanos-Domínguez, and Stefan Haufe. Solving the EEG inverse problem based on space–time–frequency structured sparsity constraints. *NeuroImage*, 118:598–612, 2015.

Chiao En Chen, Flavio Lorenzelli, Ralph E Hudson, and Kung Yao. Stochastic maximum-likelihood DOA estimation in the presence of unknown nonuniform noise. *IEEE Transactions on Signal Processing*, 56(7):3038–3044, 2008.

Nancy Chinchor and Beth M Sundheim. Muc-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5)*, 1993.

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.

Arnak Dalalyan, Mohamed Hebiri, Katia Meziani, and Joseph Salmon. Learning heteroscedastic models by convex programming under group sparsity. In *International Conference on Machine Learning*, pp. 379–387, 2013.

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*, pp. 209–216, 2007.

Z John Daye, Jinbo Chen, and Hongzhe Li. High-dimensional heteroscedastic regression with an application to eQTL data analysis. *Biometrics*, 68(1):316–326, 2012.

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.

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.

David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

Denis A Engemann and Alexandre Gramfort. Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals. *NeuroImage*, 108:328–342, 2015.

D. Fagot, H. Wendt, C. Févotte, and P. Smaragdis. Majorization-minimization algorithms for convolutive NMF with the beta-divergence. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

Irina F Gorodnitsky, John S George, and Bhaskar D Rao. Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm. *Electroencephalography and Clinical Neurophysiology*, 95(4):231–251, 1995.

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.

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

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.

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

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. *bioRxiv*, 2020.

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

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.

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.

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.

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

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.

Jürgen Jost and Jèurgen Jost. *Riemannian geometry and geometric analysis*, volume 42005. Springer, 2008.

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)*, pp. 1–7. IEEE, 2020.

Johannes Lederer and Christian L Muller. Don't fall for tuning parameters: tuning-free variable selection in high dimensions with the TREX. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2729–2735, 2015.

Tao Li and Arye Nehorai. Maximum likelihood direction finding in spatially colored noise fields using sparse sensor arrays. *IEEE Transactions on Signal Processing*, 59(3):1048–1062, 2010.

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

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

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.

Mudassir Masood, Ali Ghrayeb, Prabhu Babu, Issa Khalil, and Mazen Hasna. A minorization-maximization algorithm for an-based MIMOE secrecy rate maximization. In *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 975–980. IEEE, 2016.

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*, pp. 998–1007, 2018.

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

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.

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

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

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.

Hendrik Bernd Petersen and Peter Jung. Robust instance-optimal recovery of sparse signals at unknown noise levels. *arXiv preprint arXiv:2008.08385*, 2020.

Peter Petersen, S Axler, and KA Ribet. *Riemannian geometry*, volume 171. Springer, 2006.

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

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.

Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121, 2000.

Matthias W Seeger and David P Wipf. Variational Bayesian inference techniques. *IEEE Signal Processing Magazine*, 27(6):81–91, 2010.

Kensuke Sekihara and Srikantan S Nagarajan. *Electromagnetic brain imaging: a Bayesian perspective*. Springer, 2015.

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

Suvrit Sra and Reshad Hosseini. Geometric optimization in machine learning. In *Algorithmic Advances in Riemannian Geometry and Applications*, pp. 73–91. Springer, 2016.

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.

Michael E Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1(Jun):211–244, 2001.

Sara Van de Geer, Johannes Lederer, et al. The Lasso, correlated design, and improved oracle inequalities. In *From Probability to Statistics and Back: High-Dimensional Models and Processes– A Festschrift in Honor of Jon A. Wellner*, pp. 303–316. Institute of Mathematical Statistics, 2013.

CJ van Rijsbergen. Information retrieval. 1979.

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

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

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

David P Wipf and Bhaskar D Rao. Sparse Bayesian learning for basis selection. *IEEE Transactions on Signal Processing*, 52(8):2153–2164, 2004.

David P Wipf and Bhaskar D Rao. An empirical Bayesian strategy for solving the simultaneous sparse approximation problem. *IEEE Transactions on Signal Processing*, 55(7):3704–3716, 2007.

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.

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

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

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

Pourya Zadeh, Reshad Hosseini, and Suvrit Sra. Geometric mean metric learning. In *International Conference on Machine Learning*, pp. 2464–2471, 2016.

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.

A PROOF OF THEOREM 1

Proof. We start the proof by recalling equation 8:

$$\mathcal{L}^{\mathrm{II}}(\mathbf{\Gamma}, \mathbf{\Lambda}) = -\log p(\mathbf{Y}|\mathbf{\Gamma}, \mathbf{\Lambda}) = \log|\mathbf{\Sigma_y}| + \frac{1}{T}\sum_{t=1}^{T} \mathbf{y}(t)^\top \mathbf{\Sigma_y}^{-1}\mathbf{y}(t) \ . \tag{16}$$

The upper bound on the $\log|\mathbf{\Sigma_y}|$ can then be directly inferred from the concavity of the log-determinant function and its first-order Taylor expansion around the value from the previous iteration, $\mathbf{\Sigma_y}^k$, which provides the following inequality (Sun et al., 2017, Example 2):

$$\begin{aligned}\log|\mathbf{\Sigma_y}| &\leq \log\left|\mathbf{\Sigma_y}^k\right| + \mathrm{tr}\left[\left(\mathbf{\Sigma_y}^k\right)^{-1}\left(\mathbf{\Sigma_y} - \mathbf{\Sigma_y}^k\right)\right] \\ &= \log\left|\mathbf{\Sigma_y}^k\right| + \mathrm{tr}\left[\left(\mathbf{\Sigma_y}^k\right)^{-1}\mathbf{\Sigma_y}\right] - \mathrm{tr}\left[\left(\mathbf{\Sigma_y}^k\right)^{-1}\mathbf{\Sigma_y}^k\right] \ .\end{aligned} \tag{17}$$

Note that the first and last terms in equation 17 do not depend on $\mathbf{\Gamma}$; hence, they can be ignored in the optimization procedure. Now, we decompose $\mathbf{\Sigma_y}$ into two terms, each of which only contains either the noise or source covariances:

$$\mathrm{tr}\left[\left(\mathbf{\Sigma_y}^k\right)^{-1}\mathbf{\Sigma_y}\right] = \mathrm{tr}\left[\left(\mathbf{\Sigma_y}^k\right)^{-1}\left(\mathbf{\Lambda} + \mathbf{L\Gamma L}^\top\right)\right] = \mathrm{tr}\left[\left(\mathbf{\Sigma_y}^k\right)^{-1}\mathbf{\Lambda}\right] + \mathrm{tr}\left[\left(\mathbf{\Sigma_y}^k\right)^{-1}\mathbf{L\Gamma L}^\top\right] \ . \tag{18}$$

In next step, we decompose the second term in equation 8, $\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}(t)^\top\mathbf{\Sigma_y}^{-1}\mathbf{y}(t)$, into two terms, each of which is a function of either only the noise or only the source covariances. To this end, we exploit the following key relationship between the sensor and source space covariances:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}(t)^\top\mathbf{\Sigma_y}^{-1}\mathbf{y}(t) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}^k(t)^\top\mathbf{\Gamma}^{-1}\mathbf{x}^k(t) + \frac{1}{T}\sum_{t=1}^{T}(\mathbf{y}(t) - \mathbf{L}\mathbf{x}^k(t))^\top\mathbf{\Lambda}^{-1}(\mathbf{y}(t) - \mathbf{L}\mathbf{x}^k(t)) \ . \tag{19}$$

By combining equation 18 and equation 19, rearranging the terms, and ignoring all terms that do not depend on $\mathbf{\Gamma}$, we have:

$$\begin{aligned}\mathcal{L}^{\mathrm{II}}(\mathbf{\Gamma}) &\leq \mathrm{tr}\left[\left(\mathbf{\Sigma_y}^k\right)^{-1}\mathbf{L\Gamma L}^\top\right] + \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}^k(t)^\top\mathbf{\Gamma}^{-1}\mathbf{x}^k(t) + \mathrm{const} \\ &= \mathrm{tr}(\left(\mathbf{C}_{\mathrm{S}}^k\right)^{-1}\mathbf{\Gamma}) + \mathrm{tr}(\mathbf{M}_{\mathrm{S}}^k\mathbf{\Gamma}^{-1}) + \mathrm{const} = \mathcal{L}_{\mathrm{source}}^{\mathrm{conv}}(\mathbf{\Gamma}, \mathbf{\Lambda}^k) + \mathrm{const} \ ,\end{aligned} \tag{20}$$

where $\mathbf{C}_{\mathrm{S}}^k = \left(\mathbf{L}^\top\left(\mathbf{\Sigma_y}^k\right)^{-1}\mathbf{L}\right)^{-1}$ and $\mathbf{M}_{\mathrm{S}}^k = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}^k(t)^\top\mathbf{x}^k(t)$. Note that constant values in equation 20 do not depend on $\mathbf{\Gamma}$; hence, they can be ignored in the optimization procedure. This proves the equivalence of equation 8 and equation 9 when the optimization is performed with respect to $\mathbf{\Gamma}$.

The equivalence of equation 8 and equation 11 can be shown analogously, with the difference that we only focus on noise-related terms in equation 18 and equation 19:

$$\begin{aligned}\mathcal{L}^{\mathrm{II}}(\mathbf{\Lambda}) &\leq \mathrm{tr}\left[\left(\mathbf{\Sigma_y}^k\right)^{-1}\mathbf{\Lambda}\right] + \frac{1}{T}\sum_{t=1}^{T}(\mathbf{y}(t) - \mathbf{L}\mathbf{x}^k(t))^\top\mathbf{\Lambda}^{-1}(\mathbf{y}(t) - \mathbf{L}\mathbf{x}^k(t)) + \mathrm{const} \\ &= \mathrm{tr}(\left(\mathbf{C}_{\mathrm{N}}^k\right)^{-1}\mathbf{\Lambda}) + \mathrm{tr}(\mathbf{M}_{\mathrm{N}}^k\mathbf{\Lambda}^{-1}) + \mathrm{const} = \mathcal{L}_{\mathrm{noise}}^{\mathrm{conv}}(\mathbf{\Gamma}^k, \mathbf{\Lambda}) + \mathrm{const} \ ,\end{aligned} \tag{21}$$

where $\mathbf{C}_{\mathrm{N}}^k = \mathbf{\Sigma_y}^k$, and $\mathbf{M}_{\mathrm{N}}^k = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{y}(t) - \mathbf{L}\mathbf{x}^k(t))^\top(\mathbf{y}(t) - \mathbf{L}\mathbf{x}^k(t))$. Constant values in equation 21 do not depend on $\mathbf{\Lambda}$; hence, they can again be ignored in the optimization procedure. Summarizing, we have shown that optimizing equation 8 is equivalent to optimizing $\mathcal{L}_{\mathrm{noise}}^{\mathrm{conv}}(\mathbf{\Gamma}^k, \mathbf{\Lambda})$ and $\mathcal{L}_{\mathrm{source}}^{\mathrm{conv}}(\mathbf{\Gamma}, \mathbf{\Lambda}^k)$, which concludes the proof. □

B PROOF OF THEOREM 2

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

Definition 3 (Geodesic path)**.** *Let $\mathcal{M}$ be a Riemannian manifold, i.e., a differentiable manifold whose tangent space 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 $l = 1$ results in 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 positive definite (P.D.) manifold joining two P.D. matrices, $\mathbf{P}_0, \mathbf{P}_1 > 0$. The specific pairs of matrices we will deal with are $\{\mathbf{C}_{\mathrm{S}}^k, \mathbf{M}_{\mathrm{S}}^k\}$ and $\{\mathbf{C}_{\mathrm{N}}^k, \mathbf{M}_{\mathrm{N}}^k\}$.

Definition 4 (Geodesic on the manifold of positive definite matrices)**.** *Now we restrict our attention to geodesics on the specific manifold P.D. matrices, which 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 (Bhatia, 2009, Chapter. 6):*

$$\xi_l(\mathbf{P}_0, \mathbf{P}_1) = (\mathbf{P}_0)^{\frac{1}{2}} \left((\mathbf{P}_0)^{-1/2} \mathbf{P}_1 (\mathbf{P}_0)^{-1/2} \right)^l (\mathbf{P}_0)^{\frac{1}{2}} \quad l \in [0, 1] \ . \tag{22}$$

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*. Besides, note that according to Definition 4 the following equality holds :

$$\begin{aligned} \xi_l(\mathbf{\Gamma}_0, \mathbf{\Gamma}_1)^{-1} &= \left((\mathbf{\Gamma}_0)^{1/2} \left((\mathbf{\Gamma}_0)^{-1/2} \mathbf{\Gamma}_1 (\mathbf{\Gamma}_0)^{-1/2} \right)^l (\mathbf{\Gamma}_0)^{1/2} \right)^{-1} \\ &= \left((\mathbf{\Gamma}_0)^{-1/2} \left((\mathbf{\Gamma}_0)^{1/2} (\mathbf{\Gamma}_1)^{-1} (\mathbf{\Gamma}_0)^{1/2} \right)^l (\mathbf{\Gamma}_0)^{-1/2} \right) = \xi_l(\mathbf{\Gamma}_0^{-1}, \mathbf{\Gamma}_1^{-1}) \ . \end{aligned} \tag{23}$$

Definition 5 (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 with domain $\mathcal{A} \subset \mathcal{M}$ with $f : \mathcal{A} \to \mathbb{R}$ is called* geodesic convex *(g-convex) if the following relation holds:*

$$f\left(\zeta_l(\mathbf{p}_0, \mathbf{p}_1)\right) \leq l f(\mathbf{p}_0) + (1 - l) f(\mathbf{p}_1) \ , \tag{24}$$

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 3. 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 6. *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:*

$$\zeta_l(\mathbf{p}_0, \mathbf{p}_1) = l\mathbf{p}_0 + (1 - l)\mathbf{p}_1 \ . \tag{25}$$

For the sake of brevity, we will omit a detailed theoretical introduction of g-convexity, and only borrow a result from Zadeh et al. (2016); Sra & Hosseini (2015). Interested readers are referred to Wiesel et al. (2015, Chapter 1) for a gentle introduction to this topic, and Papadopoulos (2005, Chapter. 2) Rapcsak (1991); Ben-Tal (1977); Liberti (2004); Pallaschke & Rolewicz (2013); Bonnabel & Sepulchre (2009); Moakher (2005); Sra & Hosseini (2016); Vishnoi (2018) for more in-depth technical details.

Now we are ready to state the proof:

Proof. The proof parallels the one provided in Zadeh et al. (2016, Theorem. 3). Here, we only show the proof for $\mathcal{L}_{\text{source}}^{\text{conv}}(\mathbf{\Gamma}, \mathbf{\Lambda}_k)$. The proof for $\mathcal{L}_{\text{noise}}^{\text{conv}}(\mathbf{\Gamma}_k, \mathbf{\Lambda})$ can be presented analogously; and therefore, is omitted here for brevity. We present the proof in two parts. In the first part, we limit our attention to P.D. manifolds and express equation 24 in terms of geodesic paths and functions that lie on this particular space. By focusing on P.D. manifolds, we then show that $\mathcal{L}_{\text{source}}^{\text{conv}}(\mathbf{\Gamma}, \mathbf{\Lambda}_k)$ is strictly g-convex on this specific domain. In the second part, we then derive the updates rules proposed in equation 13 and equation 14.

We start by focusing on geodesics along the P.D. manifold, by setting $\zeta_l(\mathbf{p}_0, \mathbf{p}_1)$ to $\xi_l(\mathbf{\Gamma}_0, \mathbf{\Gamma}_1)$ as presented in Definition 4, and defining $f(.)$ as $f(\mathbf{\Gamma}) = \mathrm{tr}(\mathbf{C}_{\mathrm{S}}^k \mathbf{\Gamma}) + \mathrm{tr}(\mathbf{M}_{\mathrm{S}}^k \mathbf{\Gamma}^{-1})$ that represents cost function $\mathcal{L}_{\text{source}}^{\text{conv}}(\mathbf{\Gamma}, \mathbf{\Lambda}_k)$.

After restricting our attention to geodesic paths and functions on the P.D. manifold, we now show that $f(\mathbf{\Gamma})$ is strictly g-convex on this specific domain. For continuous functions as considered in this paper, fulfilling equation 24 for $f(\mathbf{\Gamma})$ and $\xi_l(\mathbf{\Gamma}_0, \mathbf{\Gamma}_1)$ with $l = 1/2$ is sufficient to prove strictly g-convexity:

$$\begin{aligned} &\mathrm{tr}\left(\mathbf{C}_{\mathrm{S}}^k \xi_{1/2}(\mathbf{\Gamma}_0, \mathbf{\Gamma}_1)\right) + \mathrm{tr}\left(\mathbf{M}_{\mathrm{S}}^k \xi_{1/2}(\mathbf{\Gamma}_0, \mathbf{\Gamma}_1)^{-1}\right) \\ &< \frac{1}{2}\mathrm{tr}\left(\mathbf{C}_{\mathrm{S}}^k \mathbf{\Gamma}_0\right) + \frac{1}{2}\mathrm{tr}\left(\mathbf{M}_{\mathrm{S}}^k {\mathbf{\Gamma}_0}^{-1}\right) \\ &\quad + \frac{1}{2}\mathrm{tr}\left(\mathbf{C}_{\mathrm{S}}^k \mathbf{\Gamma}_1\right) + \frac{1}{2}\mathrm{tr}\left(\mathbf{M}_{\mathrm{S}}^k {\mathbf{\Gamma}_1}^{-1}\right) \ . \end{aligned} \tag{26}$$

Given $\mathbf{C}_{\mathrm{S}}^k \in \mathcal{S}_{++}$, i.e., $\mathbf{C}_{\mathrm{S}}^k > 0$, and considering the operator inequality property as (Bhatia, 2009, Chapter. 4)

$$\xi_{1/2}(\mathbf{\Gamma}_0, \mathbf{\Gamma}_1) \prec \frac{1}{2}\mathbf{\Gamma}_0 + \frac{1}{2}\mathbf{\Gamma}_1 \ , \tag{27}$$

we have:

$$\mathrm{tr}\left(\mathbf{C}_{\mathrm{S}}^k \xi_{1/2}(\mathbf{\Gamma}_0, \mathbf{\Gamma}_1)\right) < \frac{1}{2}\mathrm{tr}\left(\mathbf{C}_{\mathrm{S}}^k \mathbf{\Gamma}_0\right) + \frac{1}{2}\mathrm{tr}\left(\mathbf{C}_{\mathrm{S}}^k \mathbf{\Gamma}_1\right) \ , \tag{28}$$

which is derived by multiplying both sides of equation 27 by $\mathbf{C}_{\mathrm{S}}^k$ followed by taking the trace of both sides.

Similarly, we can write the operator inequality property for $\{\mathbf{\Gamma}_0^{-1}, \mathbf{\Gamma}_1^{-1}\}$ using equation 23 as:

$$\xi_{1/2}(\mathbf{\Gamma}_0, \mathbf{\Gamma}_1)^{-1} = \xi_{1/2}(\mathbf{\Gamma}_0^{-1}, \mathbf{\Gamma}_1^{-1}) \prec \frac{1}{2}\mathbf{\Gamma}_0^{-1} + \frac{1}{2}\mathbf{\Gamma}_1^{-1} \ , \tag{29}$$

Multiplying both sides of equation 29 by $\mathbf{M}_{\mathrm{S}}^k \in \mathcal{S}_{++}$, and applying the trace operator on both sides leads to:

$$\mathrm{tr}\left(\mathbf{M}_{\mathrm{S}}^k \xi_{1/2}(\mathbf{\Gamma}_0, \mathbf{\Gamma}_1)^{-1}\right) < \frac{1}{2}\mathrm{tr}\left(\mathbf{M}_{\mathrm{S}}^k {\mathbf{\Gamma}_0}^{-1}\right) + \frac{1}{2}\mathrm{tr}\left(\mathbf{M}_{\mathrm{S}}^k {\mathbf{\Gamma}_1}^{-1}\right) \ . \tag{30}$$

Adding equation 28 and equation 30 proves equation 26 and concludes the first part of the proof.

We now present the second part of the proof by deriving the update rules in equations 13 and 14. Since the cost function $\mathcal{L}_{\mathrm{source}}^{\mathrm{conv}}(\mathbf{\Gamma}, \mathbf{\Lambda}_k)$ 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 equation 9 and setting the result to zero as follows:

$$\nabla \mathcal{L}_{\mathrm{source}}^{\mathrm{conv}}(\mathbf{\Gamma}, \mathbf{\Lambda}_k) = \left(\mathbf{C}_{\mathrm{S}}^k\right)^{-1} - \mathbf{\Gamma}^{-1}\mathbf{M}_{\mathrm{S}}^k\mathbf{\Gamma}^{-1} = 0 \ , \tag{31}$$

which results in

$$\mathbf{\Gamma}\left(\mathbf{C}_{\mathrm{S}}^k\right)^{-1}\mathbf{\Gamma} = \mathbf{M}_{\mathrm{S}}^k \ . \tag{32}$$

This solution is known as the *Riccati equation*, and is the geometric mean between $\mathbf{C}_{\mathrm{S}}^k$ and $\mathbf{M}_{\mathrm{S}}^k$ (Davis et al., 2007; Bonnabel & Sepulchre, 2009):

$$\mathbf{\Gamma}^{k+1} = (\mathbf{C}_{\mathrm{S}}^k)^{\frac{1}{2}}\left((\mathbf{C}_{\mathrm{S}}^k)^{-1/2}\mathbf{M}_{\mathrm{S}}^k(\mathbf{C}_{\mathrm{S}}^k)^{-1/2}\right)^{\frac{1}{2}}(\mathbf{C}_{\mathrm{S}}^k)^{\frac{1}{2}} \ .$$

The update rule for the full noise covariance matrix can be derived analogously:

$$\mathbf{\Lambda}^{k+1} = (\mathbf{C}_{\mathrm{N}}^k)^{\frac{1}{2}}\left((\mathbf{C}_{\mathrm{N}}^k)^{-1/2}\mathbf{M}_{\mathrm{N}}^k(\mathbf{C}_{\mathrm{N}}^k)^{-1/2}\right)^{\frac{1}{2}}(\mathbf{C}_{\mathrm{N}}^k)^{\frac{1}{2}} \ .$$

Remark 7. *Note that the obtained update rules are closed-form solutions for the surrogate cost functions, equations 9 and 11, which stands in contrast to conventional majorization minimization algorithms (see section F in the appendix), which require iterative procedures in each step of the optimization.*

Deriving the update rules in equation 13 and equation 14 concludes the second part of the proof of Theorem 2. □

C Derivation of Champagne as a Special Case of FUN Learning

We start the derivation of update rule in equation 15 by constraining $\boldsymbol{\Gamma}$ to the set of diagonal matrices $\mathcal{W}$: $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$, where $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_N]^\top$. We continue by rewriting the constrained optimization with respect to the source covariance matrix,

$$\boldsymbol{\Gamma}^{k+1} = \underset{\boldsymbol{\Gamma} \in \mathcal{W},\ \boldsymbol{\Lambda} = \boldsymbol{\Lambda}^k}{\arg\min} \ \mathrm{tr}(\mathbf{C}_{\mathrm{S}}^k \boldsymbol{\Gamma}) + \mathrm{tr}(\mathbf{M}_{\mathrm{S}}^k \boldsymbol{\Gamma}^{-1}) \ , \tag{33}$$

as follows:

$$\boldsymbol{\gamma}^{k+1} = \underset{\boldsymbol{\gamma},\ \boldsymbol{\Lambda} = \boldsymbol{\Lambda}^k}{\arg\min} \ \underbrace{\mathrm{diag}\left[\left(\mathbf{C}_{\mathrm{S}}^k\right)^{-1}\right] \boldsymbol{\gamma} + \mathrm{diag}\left[\mathbf{M}_{\mathrm{S}}^k\right] \boldsymbol{\gamma}^{-1}}_{\mathcal{L}_{\mathrm{source}}^{\mathrm{diag}}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k)} \ , \tag{34}$$

where $\boldsymbol{\gamma}^{-1} = [\gamma_1^{-1}, \ldots, \gamma_N^{-1}]^\top$ is defined as the element-wise inversion of $\boldsymbol{\gamma}$. The optimization with respect to the scalar source variances is then carried out by taking the derivative of equation 34 with respect to $\gamma_n,\ \text{for } n = 1, \ldots, N$, and setting it to zero,

$$\begin{aligned}
&\frac{\partial}{\partial \gamma_n} \left(\left[\left(\mathbf{C}_{\mathrm{S}}^k\right)^{-1}\right] \gamma_n + \left[\left(\mathbf{M}_{\mathrm{S}}^k\right)^{-1}\right] \gamma_n^{-1} \right) \\
&\quad = \left[\left(\mathbf{C}_{\mathrm{S}}^k\right)^{-1}\right]_{n,n} - \frac{1}{(\gamma_n)^2} \left[\left(\mathbf{M}_{\mathrm{S}}^k\right)^{-1}\right]_{n,n} \\
&\quad = 0 \quad \text{for } n = 1, \ldots, N \ ,
\end{aligned}$$

where $\mathbf{L}_n$ denotes the n-th column of the lead field matrix. This yields the following update rule

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

which is identical to the update rule of Champagne (Wipf & Nagarajan, 2009).

D Derivation of Champagne with Heteroscedastic Noise Learning as a Special Case of FUN Learning

Similar to Appendix C, we start by constraining $\boldsymbol{\Lambda}$ to the set of diagonal matrices $\mathcal{W}$: $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_M]^\top$. We continue by reformulating the constrained optimization with respect to the noise covariance matrix,

$$\boldsymbol{\Lambda}^{k+1} = \underset{\boldsymbol{\Lambda} \in \mathcal{W},\ \boldsymbol{\Gamma} = \boldsymbol{\Gamma}^k}{\arg\min} \ \mathrm{tr}(\mathbf{C}_{\mathrm{N}}^k \boldsymbol{\Lambda}) + \mathrm{tr}(\mathbf{M}_{\mathrm{N}}^k \boldsymbol{\Lambda}^{-1}) \ , \tag{35}$$

as follows:

$$\boldsymbol{\lambda}^{k+1} = \underset{\boldsymbol{\lambda},\ \boldsymbol{\Gamma} = \boldsymbol{\Gamma}^k}{\arg\min} \ \underbrace{\mathrm{diag}\left[\left(\mathbf{C}_{\mathrm{N}}^k\right)^{-1}\right] \boldsymbol{\lambda} + \mathrm{diag}\left[\mathbf{M}_{\mathrm{N}}^k\right] \boldsymbol{\lambda}^{-1}}_{\mathcal{L}_{\mathrm{noise}}^{\mathrm{diag}}(\boldsymbol{\lambda}|\boldsymbol{\lambda}^k)} \ , \tag{36}$$

where $\boldsymbol{\lambda}^{-1} = [\lambda_1^{-1}, \ldots, \lambda_M^{-1}]^\top$ is defined as the element-wise inversion of $\boldsymbol{\lambda}$. The optimization with respect to the scalar noise variances then proceeds by taking the derivative of equation 36 with respect to $\lambda_m,\ \text{for } m = 1, \ldots, M$, and setting it to zero,

$$\begin{aligned}
&\frac{\partial}{\partial \lambda_m} \left(\left[\left(\mathbf{C}_{\mathrm{N}}^k\right)^{-1}\right] \lambda_m + \left[\mathbf{M}_{\mathrm{N}}^k\right] \lambda_m^{-1} \right) \\
&\quad = \left[\left(\mathbf{C}_{\mathrm{N}}^k\right)^{-1}\right]_{m,m} - \frac{1}{(\lambda_m)^2} \left[\mathbf{M}_{\mathrm{N}}^k\right]_{m,m} \\
&\quad = 0 \quad \text{for } m = 1, \ldots, M \ .
\end{aligned}$$

This yields the following update rule:

$$\lambda_m^{k+1} \leftarrow \sqrt{\frac{\left[\left(\mathbf{M}_{\mathrm{N}}^k\right)^{-1}\right]_{m,m}}{\left[\left(\mathbf{C}_{\mathrm{N}}^k\right)^{-1}\right]_{m,m}}} = \sqrt{\frac{\left[\frac{1}{T}\sum_{t=1}^{T}(\mathbf{y}(t)-\mathbf{L}\mathbf{x}^k(t))^\top(\mathbf{y}(t)-\mathbf{L}\mathbf{x}^k(t))\right]_{m,m}}{\left[\left(\mathbf{\Sigma}_{\mathbf{y}}^k\right)^{-1}\right]_{m,m}}} \quad \text{for } m = 1,\ldots,M\ , \tag{37}$$

which is identical to the update rule of the Champagne with heteroscedastic noise learning as presented in Cai et al. (2020).

E PSEUDO-EEG SIGNAL GENERATION

Our simulation setting is an adoption of the EEG inverse problem, where brain activity is to be reconstructed from simulated pseudo-EEG data (Haufe & Ewald, 2016).

Forward Modeling: Populations of pyramidal neurons in the cortical gray matter are known to be the main drivers of the EEG signal (Bojak, 2014; Bojak & Breakspear, 2014). 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, $\mathbf{L} \in \mathbb{R}^{58\times 2004}$, was generated using the New York Head model (Huang et al., 2016) taking into account the realistic anatomy and electrical tissue conductivities of an average human head. In this model, 2004 dipolar current sources were placed evenly on the cortical surface and 58 sensors were considered. The lead field matrix, $\mathbf{L} \in \mathbb{R}^{58\times 2004}$ was computed using the finite element method. Note that the orientation of all source currents was fixed to be perpendicular to the cortical surface, so that only scalar source amplitudes needed to be estimated.

Evaluation Metrics: Source reconstruction performance was evaluated according to the following metrics. First, the *earth mover's distance* (EMD) (Rubner et al., 2000; Haufe et al., 2008)) was used to quantify the spatial localization accuracy. The EMD measures the cost needed to transform two probability distributions defined on the same metric domain (in this case, distributions of the true and estimated sources defined in 3D Euclidean brain space) into each other. EMD scores were normalized to $[0, 1]$. Second, the error in the reconstruction of the source time courses was measured. To this end, Pearson correlation between all pairs of simulated and reconstructed (i.e., those with non-zero activations) source time courses was assessed as the mean of the absolute correlations obtained for each source, after optimally matching simulated and reconstructed sources based on maximal absolute correlation. We also report another metric for evaluating the localization error as the average Euclidean distance (EUCL) (in mm) between each simulated source and the best (in terms of absolute correlations) matching reconstructed source. For assessing the recovery of the true support, we also compute F_1-*measure* scores (Chinchor & Sundheim, 1993; van Rijsbergen, 1979): $\mathrm{F}_1 = {2\times TP}/{P+TP+FP}$, where P denotes the number of true active sources, while TP and FP are the numbers of true and false positive predictions. Note that perfect support recovery, i.e., $\mathrm{F}_1 = 1$, is only achieved when there is a perfect correspondence between ground-truth and estimated support.

To evaluate the accuracy of the noise covariance matrix estimation, the following two metrics were calculated: the Pearson correlation between the original and reconstructed noise covariance matrices, $\mathbf{\Lambda}$ and $\hat{\mathbf{\Lambda}}$, denoted by $\mathbf{\Lambda}^{\mathrm{sim}}$, and the normalized mean squared error (NMSE) between $\mathbf{\Lambda}$ and $\hat{\mathbf{\Lambda}}$, defined as: NMSE $= ||\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}||_F^2 / ||\mathbf{\Lambda}||_F^2$. Similarity error was then defined as one minus the Pearson correlation: $1 - \mathbf{\Lambda}^{\mathrm{sim}}$. Note that NMSE measures the reconstruction of the true scale of the noise covariance matrix, while $\mathbf{\Lambda}^{\mathrm{sim}}$ is scale-invariant and hence only quantifies the overall structural similarity between simulated and estimated noise covariance matrices.

Evaluating the accuracy of the noise covariance matrix estimation: Figure 3 depicts the accuracy with which the covariance matrix is reconstructed by three different noise learning approaches assuming noise with homoscedastic (red), heteroscedastic (green) and full (blue) structure. The ground truth noise covariance matrix either had full (upper row) or heteroscedastic (lower row) structure. Performance was measured in terms of similarity error and NMSE. Similar to the trend observed in Figure 2, full-structure noise learning leads to better noise covariance estimation accuracy (lower NMSE and similariy error) for the full-structure noise model, while superior reconstruction perfor-

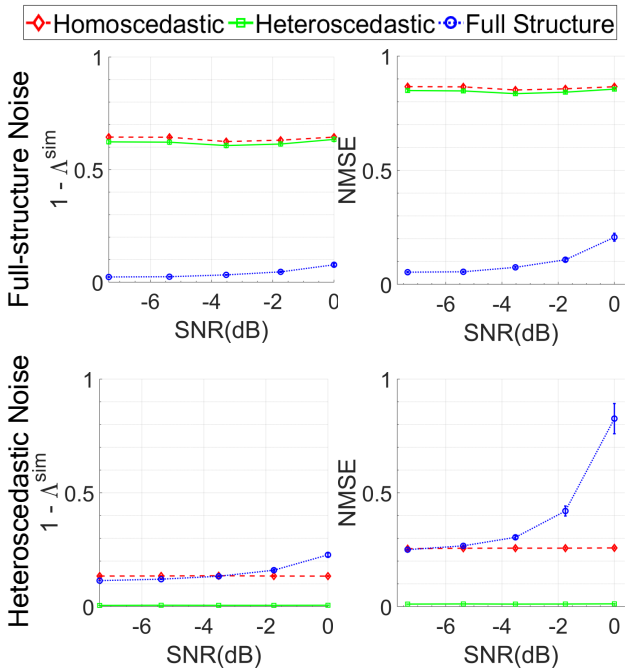

Figure 3: Accuracy of the noise covariance matrix reconstruction incurred by three different noise learning approaches assuming homoscedastic (red), heteroscedastic (green) and full-structure (blue) noise covariances. The ground-truth noise covariance matrix is either full-structure (upper row) or heteroscedastic diagonal (lower row). Performance is assessed in terms of the Pearson correlation between the entries of the original and reconstructed noise covariance matrices, $\mathbf{\Lambda}$ and $\hat{\mathbf{\Lambda}}$, denoted by $\mathbf{\Lambda}^{\text{sim}}$ (left column). Shown is the similarity error $1 - \mathbf{\Lambda}^{\text{sim}}$. Further, the normalized mean squared error (NMSE) between $\mathbf{\Lambda}$ and $\hat{\mathbf{\Lambda}}$, defined as $\text{NMSE} = ||\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}||_F^2 / ||\mathbf{\Lambda}||_F^2$ is reported (right column).

mance is achieved for heteroscedastic noise learning when true noise covariance is heteroscedastic.

F Convergence Guarantee for FUN Learning

In the following, we prove that FUN learning is guaranteed to converge to a local minimum of the Bayesian Type-II likelihood equation 8. To this end, we will prove that FUN learning is an instance of the general class of majorization-minimization (MM) algorithms, for which this property follows by construction. We first briefly review theoretical concepts behind the majorization-minimization (MM) algorithmic framework (Hunter & Lange, 2004; Razaviyayn et al., 2013; Jacobson & Fessler, 2007; Wu et al., 2010).

F.1 Majorization-Minimization

MM encompasses a family of iterative algorithms for optimizing general non-linear cost functions. 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. The MM class covers a broad range of common optimization algorithms such as *convex-concave procedures (CCCP)* and *proximal methods* (Sun et al., 2017, Section IV), (Mjolsness & Garrett, 1990; Yuille & Rangarajan, 2003; Lipp & Boyd, 2016). The MM class of algorithms have been applied in various domains such as brain source imaging (Hashemi & Haufe, 2018; Bekhti et al., 2018; Cai et al., 2020; Hashemi et al., 2020), wireless communication systems with massive MIMO technology (Masood et al., 2016; Haghighatshoar & Caire, 2017; Khalilsarai et al., 2020), and non-negative matrix factorization (Fagot et al., 2019). (Interested readers are referred to Sun et al. (2017) for an extensive list of applications on MM).

The idea of MM for minimizing a continuous function $f(\mathbf{u})$ within a closed convex set $\mathcal{U} \subset \mathbb{R}^n$:

$$\min_{\mathbf{u}} f(\mathbf{u}) \quad \text{subject to } \mathbf{u} \in \mathcal{U} \ , \tag{38}$$

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 $\mathbf{u}^k$:

$$\text{[A1]} \qquad g(\mathbf{u}^k|\mathbf{u}^k) = f(\mathbf{u}^k) \qquad \forall\, \mathbf{u}^k \in \mathcal{U}$$

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

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

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

If a surrogate function fulfills all [A1]–[A3] conditions, then the resulting algorithm has a *descending trend* property, whereby the value of the cost function f decreases in each iteration: $f(\mathbf{u}^{k+1}) \leq f(\mathbf{u}^k)$. For the smooth functions considered in this paper, if the derivatives of the original and surrogate functions coincide at $\mathbf{u}^k$:

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

then, the following theorem holds:

Theorem 8. *Assume that an MM algorithm fulfills conditions [A1]–[A4]. Then, every limit point of the sequence of minimizers generated in [A3], is a stationary point of the original optimization problem in equation 38.*

Proof. A detailed proof is provided in Razaviyayn et al. (2013, Theorem 1). □

F.2 FUN Learning as Majorization-Minimization

We now show that FUN learning is an instance of majorization-minimization as defined above.

Theorem 9. *Optimizing the non-convex type-II ML cost function in equation 8, $\mathcal{L}^{II}(\mathbf{\Gamma}, \mathbf{\Lambda})$, with respect to $\mathbf{\Gamma}$ using the update rule 13 leads to an MM algorithm fulfilling Theorem 8, where $\mathcal{L}^{II}(\mathbf{\Gamma}, \mathbf{\Lambda})$ is majorized by $\mathcal{L}^{\text{conv}}_{\text{source}}(\mathbf{\Gamma}, \mathbf{\Lambda}^k)$. Similarly, optimizing $\mathcal{L}^{II}(\mathbf{\Gamma}, \mathbf{\Lambda})$ with respect to $\mathbf{\Lambda}$ using update rule 14 leads to an MM algorithm fulfilling Theorem 8, where $\mathcal{L}^{II}(\mathbf{\Gamma}, \mathbf{\Lambda})$ is majorized by $\mathcal{L}^{\text{conv}}_{\text{noise}}(\mathbf{\Gamma}^k, \mathbf{\Lambda})$ in equation 11.*

Proof. We need to prove that conditions [A1]–[A4] are fulfilled for FUN learning. To this end, we recall the upper bound on $\log|\mathbf{\Sigma_y}|$ in equation 17, which fulfills condition [A2] since it majorizes $\log|\mathbf{\Sigma_y}|$ as a result of leveraging the concavity of the log-determinant function and its first-order Taylor expansion around $\mathbf{\Sigma}^k_{\mathbf{y}}$. Besides, it automatically satisfies conditions [A1] and [A4] by construction, because the majorizing function in equation 17 is obtained through a Taylor expansion around $\mathbf{\Sigma}^k_{\mathbf{y}}$. Concretely, [A1] is satisfied because the equality in equation 17 holds for $\mathbf{\Sigma_y} = \mathbf{\Sigma}^k_{\mathbf{y}}$. Similarly, [A4] is satisfied because the gradient of $\log|\mathbf{\Sigma_y}|$ at point $\mathbf{\Sigma}^k_{\mathbf{y}}$, $\left(\mathbf{\Sigma}^k_{\mathbf{y}}\right)^{-1}$, defines the linear Taylor approximation $\log\left|\mathbf{\Sigma}^k_{\mathbf{y}}\right| + \operatorname{tr}\left[\left(\mathbf{\Sigma}^k_{\mathbf{y}}\right)^{-1}\left(\mathbf{\Sigma_y} - \mathbf{\Sigma}^k_{\mathbf{y}}\right)\right]$. Thus, both gradients coincide in $\mathbf{\Sigma}^k_{\mathbf{y}}$ by construction. Now, we prove that [A3] can be satisfied by showing that $\mathcal{L}^{\text{conv}}_{\text{source}}(\mathbf{\Gamma}, \mathbf{\Lambda}^k)$ reaches its global minimum in each MM iteration. This is guaranteed if $\mathcal{L}^{\text{conv}}_{\text{source}}(\mathbf{\Gamma}, \mathbf{\Lambda}^k)$ can be shown to be convex or g-convex with respect to $\mathbf{\Gamma}$. To this end, we first require the subsequent proposition:

Proposition 10. *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 (1991, Theorem 2.1). □

Given the presented proof in appendix B, we can conclude that equation 20 is g-convex; hence, any local minimum of $\mathcal{L}^{\text{conv}}_{\text{source}}(\mathbf{\Gamma}, \mathbf{\Lambda}^k)$ is a global minimum according to Proposition 10. This proves that condition [A3] is fulfilled and completes the proof that the optimization of equation 8 with respect to $\mathbf{\Gamma}$ using the convex surrogate cost function equation 9 leads to an MM algorithm. The same holds for the optimization with respect to $\mathbf{\Lambda}$ based on the convex surrogate function in equation 11, $\mathcal{L}^{\text{conv}}_{\text{noise}}(\mathbf{\Gamma}^k, \mathbf{\Lambda})$. The proof is omitted here for the sake of brevity. □