bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Unification of Sparse Bayesian Learning Algorithms for Electromagnetic Brain Imaging with the Majorization Minimization Framework

Ali Hashemi[1,3,4], Chang Cai[5], Gitta Kutyniok[2,3,6], Klaus-Robert Müller[1,7,8,9,*],
Srikantan S Nagarajan[5], and Stefan Haufe[4,10,*]

[1]Machine Learning Group, Technische Universität Berlin, Germany.

[2]Institut für Softwaretechnik und Theoretische Informatik, Technische Universität Berlin, Germany

[3]Institut für Mathematik, Technische Universität Berlin, Germany

[4]Berlin Center for Advanced Neuroimaging (BCAN), Charité – Universitätsmedizin Berlin, Germany

[5]Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA, USA

[6]Department of Physics and Technology, University of Tromsø, Norway

[7]Berliner Zentrum für Maschinelles Lernen, Berlin, Germany

[8]Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea

[9]Max Planck Institute for Informatics, Saarbrücken, Germany

[10]Bernstein Center for Computational Neuroscience, Berlin, Germany

[*]Corresponding authors

Abstract

Methods for electro- or magnetoencephalography (EEG/MEG) based brain source imaging (BSI) using sparse Bayesian learning (SBL) have been demonstrated to achieve excellent performance in situations with low numbers of distinct active sources, such as event-related designs. This paper extends the theory and practice of SBL in three important ways. First, we reformulate three existing SBL algorithms under the *majorization-minimization* (MM) framework. This unification perspective not only provides a useful theoretical framework for comparing different algorithms in terms of their convergence behavior, but also provides a principled recipe for constructing novel algorithms with specific properties by designing appropriate bounds of the Bayesian marginal likelihood function. Second, building on the MM principle, we propose a novel method called *LowSNR-BSI* that achieves favorable source reconstruction performance in low signal-to-noise-ratio settings. Third, precise knowledge of the noise level is a crucial requirement for accurate source reconstruction. Here we present a novel principled technique to accurately learn the noise variance from the data either jointly within the source reconstruction procedure or using one of two proposed cross-validation strategies. Empirically, we could show that the monotonous convergence behavior predicted from MM theory is confirmed in numerical experiments. Using simulations, we further demonstrate the advantage of LowSNR-BSI over conventional SBL in low-SNR regimes, and the advantage of learned noise levels over estimates

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

derived from baseline data. To demonstrate the usefulness of our novel approach we show neurophysiologically plausible source reconstructions on averaged auditory evoked potential data.

Index Terms

Electro-/Magnetoencephalography, Brain Source Imaging, Type I/II Bayesian Learning, Non-convex, Majorization-Minimization, Noise Learning, Hyperparameter Learning.

I. Introduction

Electro- and Magnetoencephalography (EEG/MEG) are non-invasive techniques for measuring brain electrical activity with high temporal resolution. As such, both have become indispensable tools in basic neuroscience and clinical neurology. The downside of both techniques, however, is that their sensors are located far away from the neural generators of the measured brain electrical activity. EEG/MEG measurements are therefore characterized by low spatial resolution and highly overlapping contributions of multiple brain sources in each sensor. The mathematical model of the EEG/MEG sensing procedure can be described by the linear *forward model*

$$\mathbf{Y} = \mathbf{L}\mathbf{X} + \mathbf{E} \, , \tag{1}$$

which maps the electrical activity of the brain, $\mathbf{X}$, to the sensor measurements, $\mathbf{Y}$. The *measurement matrix* $\mathbf{Y} \in \mathbb{R}^{M \times T}$ captures the activity of M sensors attached at different parts of the scalp 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 brain sources located in the cortical gray matter 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 identical and independent distributed (i.i.d) zero mean white Gaussian noise with variance σ^2, $\mathbf{e}(t) \in \mathbb{R}^{M \times 1} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_M), t = 1, \ldots, T$, which is assumed to be independent of the source activations. The linear forward mapping from $\mathbf{Y}$ to $\mathbf{X}$ is given by the *lead field* matrix $\mathbf{L} \in \mathbb{R}^{M \times N}$, which is here assumed to be known. 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 [1]–[5].

The goal of brain source imaging (BSI) is to infer the underlying brain activity $\mathbf{X}$ from the EEG/MEG measurement $\mathbf{Y}$ given the lead field matrix $\mathbf{L}$. Unfortunately, this inverse problem is highly ill-posed as the number of sensors is typically much larger than the number of locations of potential brain sources. In addition, the leadfield matrix is typically highly ill-conditioned even for small numbers of sensors, introducing numerical instabilities in the inverse estimates. Regularization techniques are widely used to address both challenges by incorporating prior knowledge or assumptions about the nature of the true sources into the estimation. A common assumption is that the number of active brain sources during the execution of a specific mental task is small, i.e., that the spatial distribution of the brain activity is sparse. This assumption can be encoded in various ways. Classical approaches [6] employ super-Gaussian prior distributions to identify solutions in which most of brain regions are inactive. This approach is called Maximum-a-Posteriori (MAP) estimation or *Type-I learning*. Later work [7] has shown that hierarchical Bayesian models achieve better reconstructions of sparse brain signals by employing a separate

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Gaussian prior for each brain location. The variances at each location are treated as unknown (hyper-) parameters, which are estimated jointly with the source activity. This approach is called Sparse Bayesian Learning (SBL), Type-II Maximum-Likelihood (Type-II ML) estimation or simply *Type-II learning* [8].

Type-II learning generally leads to non-convex objective functions, which are non-trivial to optimize. A number of iterative algorithms have been proposed [7]–[11], which, due to employing distinct parameter update rules, differ in their convergence rates and overall computational complexity. Interestingly though, the published algorithms also share certain favourable properties such as the guaranteed decrease of the objective function in each iteration, and consequently the guaranteed convergence to a local minimum, which sets them apart from general purpose solvers such as (quasi-) Newton methods. Being derived using vastly different mathematical concepts such as fixed point theory and expectation-maximization (EM), it has, however, so far been difficult to explain the observed commonalities and differences, advantages and disadvantages of Type-II methods in absence of a common theoretical framework, even if the properties of individual algorithms have been extensively studied [9].

The primary contribution of this paper is to introduce *Majorization-Minimization (MM)* ([12], [13], and references therein) as a flexible algorithmic framework within which different SBL approaches can be theoretically analyzed. Briefly, MM is a family of iterative algorithms to optimize general non-linear objective functions. In a minimization setting, MM replaces the original cost function in each iteration by an upper bound, or majorization function, whose minimum is usually easy to find. The objective value at the minimum is then used to construct the bound for the following iteration, and the procedure is repeated until a local minimum of the objective is reached. Notably, MM algorithms are popular in many disciplines in which Type-II learning problems arise, such as, e.g., telecommunications [14]–[19] and finance [20], [21]. The concept of MM is, however, rarely explicitly referenced in EEG/MEG brain source imaging, even though it has been used implicitly. We demonstrate here that three popular SBL variants, denoted as *EM*, *MacKay*, and *convex-bounding based* SBL, can be cast as majorization-minimization methods employing different types of upper bounds on the marginal likelihood. This view as variants of MM help explains, among other things, the guaranteed convergence of these algorithms to a local minimum. The characteristics of the chosen bounds determine the reconstruction performance and convergence rates of the resulting algorithms. The MM framework additionally offers a principled way of constructing new SBL algorithms for specific purposes by designing appropriate bounds.

Therefore, a second contribution of this paper is the development of a new SBL algorithm, called LowSNR-BSI, that is especially suitable for low signal-to-noise ratio (SNR) regimes. Real-world applications of EEG/MEG brain source imaging are often characterized by low signal-to-noise ratio (SNR), where the power of unwanted noise sources can be comparable to the power of the signal of interest. This holds in particular for the reconstruction of ongoing as well as induced (non-phase-locked) oscillatory activity, where no averaging can be performed prior to source reconstruction. Current SBL algorithms may suffer from reduced performance in such low-SNR regimes [22]–[24]. To overcome this limitation, we propose a novel MM algorithm for EEG/MEG source imaging, which employs a bound on the SBL cost function that is particularly suitable for low-SNR regimes.

As a third contribution, this paper discusses principled ways to estimate the sensor noise variance σ^2, which is assumed to be known in the first part of the paper. Determining the goodness-of-fit of the optimal model, the

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

value of this variable exerts a strong impact on the overall reconstruction [25]. Technically being another model hyperparameter, the noise variance is, however, rarely estimated as part of the model fitting. Instead, it is often determined prior to the model fitting from a baseline recording. This approach can, however, lead to suboptimal results in practice or be even inapplicable, e.g., when resting state data are analyzed. Here we present a number of alternatives to estimate the noise variance in Type-I and Type-II brain source imaging approaches. Building on work by [26], we derive an analytic update rule, which enables the adaptive estimation of the noise variance within various SBL schemes. Moreover, we propose two novel cross-validation (CV) schemes from the machine learning field to determine the noise variance parameter.

We conduct extensive ground-truth simulations in which we compare LowSNR-BSI with popular source reconstruction schemes including existing SBL variants, and in which we systematically study the impact of different strategies to estimate the noise level σ^2 from the data.

The outline of the paper is as follows: In Section II, a comprehensive review of Type-II BSI methods is presented. In Section III, we unify the Type-II methods described in Section II within the MM framework, and in Section IV, we derive LowSNR-BSI algorithm within the same framework. Section V introduces numerous principled ways for estimating the sensor noise variance. Simulation studies, real data analysis, and discussions are presented in Sections VI, VII, and VIII, respectively. Finally, Section IX concludes the paper.

II. Bayesian Learning

The ill-posed nature of the EEG/MEG inverse problem can be overcome by assuming a prior distribution $p(\mathbf{X})$ for the source activity. The posterior distribution of the sources after observing the data $\mathbf{Y}$, $p(\mathbf{X}|\mathbf{Y})$, is given by Bayes' rule:

$$p(\mathbf{X}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})}{\int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})\mathrm{d}\mathbf{X}} \; , \tag{2}$$

where the conditional probability $p(\mathbf{Y}|\mathbf{X})$ in the numerator denotes the *likelihood*, while the term in the denominator, $\int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})\mathrm{d}\mathbf{X} = p(\mathbf{Y})$, is refered to as *model evidence* or *marginal likelihood*. Note that the posterior is often, however, difficult to evaluate, as solving the integral in the model evidence is intractable for many choices of prior distributions and likelihoods.

A. Type-I Bayesian Learning

As the model evidence in Eq. (2) only acts as a scalar normalization for the posterior, its evaluation can be avoided if one is only interested in the most probable source configuration $\mathbf{X}$ rather than the full posterior distribution. This point estimate is known as the maximum-a-posteriori (MAP) estimate:

$$\mathbf{X}^{\mathrm{MAP}} := \arg\max_{\mathbf{X}} \underbrace{p(\mathbf{Y}|\mathbf{X})}_{\text{likelihood}}\underbrace{p(\mathbf{X})}_{\text{prior}} \; . \tag{3}$$

Assuming i.i.d. Gaussian sensor noise, the likelihood reads:

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{t=1}^{T} p(\mathbf{y}(t)|\mathbf{x}(t)) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{L}\mathbf{x}(t), \sigma^2\mathbf{I}) \; , \tag{4}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and the resulting MAP estimate (3) is given by

$$\begin{aligned}\mathbf{X}^{\mathrm{MAP}} &:= \arg\max_{\mathbf{X}} \left[\prod_{t=1}^{T} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}(t)-\mathbf{L}\mathbf{x}(t)\|_2^2\right)\right] p(\mathbf{X}) \\ &= \arg\min_{\mathbf{X}} \left[\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{y}(t)-\mathbf{L}\mathbf{x}(t)\|_2^2\right] + \sigma^2\mathcal{R}^{\mathrm{I}}(\mathbf{X}) = \arg\min_{\mathbf{X}} \mathcal{L}^{\mathrm{I}}(\mathbf{X}) \ , \end{aligned} \tag{5}$$

where $\mathcal{R}^{\mathrm{I}}(\mathbf{X}) = \log\left(p(\mathbf{X})\right)$ and $\mathcal{L}^{\mathrm{I}}(\mathbf{X})$ denotes the Bayesian Type-I learning (MAP) objective function.

Note that this expression can be interpreted as a trade-off between two optimization goals, where the first (log-likelihood) term in (5) penalizes model errors using a quadratic loss function and the second (log-prior) term penalizes deviations of the solution from the assumed spatial or temporal properties of the brain sources encoded in $\mathcal{R}^{\mathrm{I}}(\mathbf{X})$. The trade-off between these two optimization goals is defined by the ratio of the noise variance σ^2 and the variance of the prior distribution. As the latter is hardly known in practice, a *regularization parameter* $\lambda \propto \sigma^2$ subsuming both variables is introduced, which can be tuned to adjust the relative importance of both penalties in the optimization.

Several existing algorithms are characterized by different choices of a prior. For instance, choosing a Gaussian prior distribution lead to the classical minimum-norm estimate [27], which also goes by the names ℓ_2-norm (or Tikhonov) regularization and "ridge regression" in the statistics and machine learning literature. The choice of a Laplace prior leads to the minimum-current estimate [6], which is also known as ℓ_1-norm regularization or "LASSO" regression. More complex priors have been used in to encode assumptions on the spatial, temporal and/or spectral structure of the sources [28]–[34].

B. *Type-II Bayesian Learning*

While in the MAP approach the prior distribution is fixed, it is sometimes desirable to consider entire families of distributions $p(\mathbf{X}|\gamma)$ parameterized by a set of hyper-parameters γ. These hyper-parameters can be learned from the data along with the model parameters using a hierarchical empirical Bayesian approach [8], [35]. In this maximum-likelihood Type-II (ML-II, or simply Type-II) approach, γ is estimated through the maximum-likelihood principle:

$$\boldsymbol{\gamma}^{\mathrm{II}} := \arg\max_{\boldsymbol{\gamma}} p(\mathbf{Y}|\boldsymbol{\gamma}) = \arg\max_{\boldsymbol{\gamma}} \int p(\mathbf{Y}|\mathbf{X},\boldsymbol{\gamma})p(\mathbf{X}|\boldsymbol{\gamma})\mathrm{d}\mathbf{X} \ . \tag{6}$$

Computation of the conditional density $p(\mathbf{Y}|\boldsymbol{\gamma})$ is formally achieved by integrating over all possible source distributions $\mathbf{X}$ for any given choice of $\boldsymbol{\gamma}$. The maximizer of Eq. (6) then determines a data-driven prior distribution $p(\mathbf{X}|\boldsymbol{\gamma}^{\mathrm{II}})$. Plugged into the MAP estimation framework Eq. (3), this gives rises to the Type-II source estimate $\mathbf{X}^{\mathrm{II}}$.

As the conditional density $p(\mathbf{Y}|\boldsymbol{\gamma})$ for a given $\boldsymbol{\gamma}$ is identical to the model evidence in Eq. (2), this approach also goes by the name evidence maximization [11], [26]. Concrete instantiations of this approach have further been introduced under the names *sparse Bayesian learning* (SBL) or *automatic relevance determination* (ARD) [8], *variational Bayes* (VB) [9], [36] and iteratively-reweighted MAP estimation [10], [37]. Interested readers are referred to [38] for a comprehensive survey on Bayesian machine learning techniques for EEG/MEG signals. To distinguish all these Type-II variants from classical ML and MAP approaches not involving hyperparameter learning, the latter are also referred to as Type-I approaches.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

C. Sparse Bayesian Learning and Champagne

A Type-II estimation framework with particular relevance for EEG/MEG source imaging is SBL. In this 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. 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. This consideration has led [35] to propose the *Champagne* algorithm for brain source imaging, which is rooted in the concept of SBL. Compared to Type-I approaches achieving sparsity through ℓ_1-norm minimization, Champagne has shown significant performance improvement with respect to EEG/MEG source localization [39, Chapter 4].

Just as most existing approaches, Champagne makes the simplifying assumption of statistical independence between time samples. This leads to the following expression for the distribution of the sources:

$$p(\mathbf{X}|\boldsymbol{\gamma}) = \prod_{t=1}^{T} p(\mathbf{x}(t)|\boldsymbol{\gamma}) = \prod_{t=1}^{T} \mathcal{N}(0, \boldsymbol{\Gamma}) \; , \tag{7}$$

where $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_N]^\top$ and $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$. Note that, in task-based analyses, the noise variance σ^2 can be estimated from a baseline (resting state) recording. In the first part of this paper, it is, therefore, assumed to be known.

The parameters of the SBL model are the unknown sources as well as their variances. As computation of the integral in Eq. (6) is infeasible, Champagne considers an approximation, where the variances $\gamma_n, n = 1, \ldots, N$, are optimized based on the current estimates of the sources in an alternating iterative process. Given an initial estimate of the variances, the posterior distribution of the sources is a Gaussian of the form [7], [39, Chapter 4]

$$p(\mathbf{X}|\mathbf{Y}, \boldsymbol{\gamma}) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}(t), \boldsymbol{\Sigma}_{\mathbf{x}}) \text{ , where} \tag{8}$$

$$\boldsymbol{\mu}_{\mathbf{x}}(t) = \boldsymbol{\Gamma}\mathbf{L}^\top(\boldsymbol{\Sigma}_{\mathbf{y}})^{-1}\mathbf{y}(t) \tag{9}$$

$$\boldsymbol{\Sigma}_{\mathbf{x}} = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}\mathbf{L}^\top(\boldsymbol{\Sigma}_{\mathbf{y}})^{-1}\mathbf{L}\boldsymbol{\Gamma} \tag{10}$$

$$\boldsymbol{\Sigma}_{\mathbf{y}} = \sigma^2\mathbf{I} + \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top \; . \tag{11}$$

The estimated posterior parameters $\boldsymbol{\mu}_{\mathbf{x}}(t)$ and $\boldsymbol{\Sigma}_{\mathbf{x}}$ are then in turn used to update the estimate of the variances $\gamma_n, n = 1, \ldots, N$ as the minimizer of the negative log of the marginal likelihood $p(\mathbf{Y}|\boldsymbol{\gamma})$, which is given by [7]:

$$\mathcal{L}^{\mathrm{II}}(\boldsymbol{\gamma}) = -\log p(\mathbf{Y}|\boldsymbol{\gamma}) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{y}(t)^\top \boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{y}(t) + \log|\boldsymbol{\Sigma}_{\mathbf{y}}| \; , \tag{12}$$

where $|\cdot|$ denotes the determinant of a matrix. This process is repeated until convergence. Given the final solution of the hyperparameter $\boldsymbol{\gamma}^{\mathrm{II}}$, the point estimate $\mathbf{x}^{\mathrm{II}}$ of the source activity is obtained from the posterior mean of the estimated source distribution: $\mathbf{x}^{\mathrm{II}}(t) = \boldsymbol{\mu}_{\mathbf{x}}(t)$. Note that given the definition of the empirical sample covariance

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

matrix as $\mathbf{C}_\mathbf{y} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}(t)\mathbf{y}(t)^\top$, the term $\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}(t)^\top\boldsymbol{\Sigma}_\mathbf{y}^{-1}\mathbf{y}(t)$ in Eq. (12) can be rewritten as $\mathrm{tr}(\mathbf{C}_\mathbf{y}\boldsymbol{\Sigma}_\mathbf{y}^{-1})$, so that Eq. (12) becomes [7, Section II]

$$\mathcal{L}^{\mathrm{II}}(\boldsymbol{\gamma}) = \mathrm{tr}(\mathbf{C}_\mathbf{y}\boldsymbol{\Sigma}_\mathbf{y}^{-1}) + \log|\boldsymbol{\Sigma}_\mathbf{y}| \; . \tag{13}$$

Note that, in this form, the loss function Eq. (13) bears an interesting similarity to the *log-determinant (log-det) Bregman divergence* in information geometry [40]. This perspective on Type-II loss function enables a common viewpoint for Type-I and Type-II methods.

Given Legendre-Fenchel duality theory, the cost function Eq. (12) can be formulated equivalently as a joint minimization over $\mathbf{X}$ and $\boldsymbol{\gamma}$ [11, see also Section II-B] [41], [42]:

$$\begin{aligned}\mathcal{L}^{\mathrm{II}-x}(\mathbf{X},\boldsymbol{\gamma}) &= \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{y}(t)-\mathbf{L}\mathbf{x}(t)\|_2^2 + \sigma^2\mathcal{R}^{\mathrm{II}-x}(\mathbf{X},\boldsymbol{\gamma})\\ \mathcal{R}^{\mathrm{II}-x}(\mathbf{X},\boldsymbol{\gamma}) &:= \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\frac{x_n(t)^2}{\gamma_n} + \log|\boldsymbol{\Sigma}_\mathbf{y}| \; ,\end{aligned} \tag{14}$$

where $\mathcal{R}^{\mathrm{II}-x}(\mathbf{X},\boldsymbol{\gamma})$ denotes a regularizer that depends on the data, $\mathbf{x}(t)$, and where $x_n(t)$ denotes the activity of source n at time instant t. Then, as each source $x_n(t)$ is also a function of γ_n according to Eq. (9), the term $\frac{x_n(t)^2}{\gamma_n}$ goes to zero when $\gamma_n \to 0$.

We will use the above formulation to derive alternative optimization schemes for Champagne in Sections II-C2 and II-C3.

1) EM Champagne: As the cost function Eq. (12) is non-convex in $\boldsymbol{\gamma}$, the quality of the obtained solution depends substantially on the properties of the employed numerical optimization algorithm. Crucially, algorithms might not only differ w.r.t. their convergence properties but may also lead to different solutions representing distinct local minima of Eq. (12). The first algorithm for mimimizing Eq. (12) has been introduced by [9] and is an application of the expectation-maximization (EM) formalism [43]. As can be shown, Eqs. (9)–(11) correspond to the expectation (E) step of the EM algorithms w.r.t. the posterior distribution $p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma})$. The maximization (M) step of the EM formalism with respect to $\boldsymbol{\gamma}$ then leads to the update rule

$$\gamma_n := [\boldsymbol{\Sigma}_\mathbf{x}]_{n,n} + \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{\mu}_\mathbf{x}(t))_n^2 \text{ for } n = 1,\ldots,N \; . \tag{15}$$

Final estimates of both parameters are obtained by iterating the updates (9)–(11) and (15) until convergence. The resulting algorithm is known as the EM variant of the Champagne algorithm [39, Chapter 4] [9] in the field of brain source imaging.

2) Convex-bounding based Champagne: As the EM algorithm outlined above has high computational complexity, alternative minimization strategies have been proposed. Two such variants, a convex-approximation based approach and the so-called MacKay update, have been proposed in [9] and further practically investigated in [23]. Considering that the log-determinant in Eq. (14) is concave, the convex-bounding based variant of Champagne constructs a linear upper bound based on the concave conjugate of $\log\left|\sigma^2\mathbf{I}+\mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top\right|$, defined as $w^*(\mathbf{z})$,

$$\log\left|\sigma^2\mathbf{I}+\mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top\right| = \min_{\mathbf{z}>0} \mathbf{z}^T\boldsymbol{\gamma} - w^*(\mathbf{z}) \; . \tag{16}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

With this upper bound, and for a fixed value of $\boldsymbol{\gamma}$, the auxiliary variable $\mathbf{z}$ can be derived as the tangent hyperplane of the $\log|\boldsymbol{\Sigma}_{\mathbf{y}}|$:

$$\mathbf{z} = \nabla_{\boldsymbol{\gamma}} \log \left|\sigma^2\mathbf{I} + \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top\right| \ .$$

Note that the definition of the concave conjugate requires some background from convex analysis, in particular Legendre-Fenchel duality theory, which is out of the scope of this paper (see, e.g., [41], [42]).

By inserting Eq. (16) instead of $\log|\boldsymbol{\Sigma}_{\mathbf{y}}|$ into Eq. (14), the non-convex penalty function Eq. (14) is replaced by the convex function

$$\mathcal{R}^{\text{II}-x}_{\text{conv}}(\mathbf{X}, \boldsymbol{\gamma}) = \min_{\boldsymbol{\gamma}\geq 0, \mathbf{z}>0} \left[\frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\frac{x_n(t)^2}{\gamma_n}\right] + \mathbf{z}^T\boldsymbol{\gamma} - w^*(\mathbf{z})$$

in each step of the optimization. The final estimates of $\mathbf{X}$, $\boldsymbol{\gamma}$ and $\mathbf{z}$ are obtained by iterating between following update rules until convergence:

$$z_n = \mathbf{L}_n^\top(\boldsymbol{\Sigma}_{\mathbf{y}})^{-1}\mathbf{L}_n, n = 1, \ldots, N \tag{17}$$

$$\boldsymbol{\mu}_{\mathbf{x}}(t) = \boldsymbol{\Gamma}\mathbf{L}^\top(\boldsymbol{\Sigma}_{\mathbf{y}})^{-1}\mathbf{y}(t) \tag{18}$$

$$\gamma_n = \sqrt{\frac{\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{\mu}_{\mathbf{x}}(t))_n^2}{z_n}}, n = 1, \ldots, N \ . \tag{19}$$

Here, $\mathbf{L}_n$ in (17) denotes the n-th column of the lead field matrix.

3) MacKay Update for Champagne: The MacKay update proposed in [9, Section III.A-2] can be derived in a similar fashion as the convex-bounding based update using different auxiliary functions and variables. Defining new variables $\kappa_n := \log(\gamma_n)$ for $n = 1, \ldots, N$, one can introduce another surrogate function [9, Appendix-B]

$$\log\left|\sigma^2\mathbf{I} + \sum_{n=1}^{N}\exp\left(\kappa_n\right)\mathbf{L}_n^\top\mathbf{L}_n\right| = \max_{\mathbf{z}>0} \mathbf{z}^T\log(\boldsymbol{\gamma}) - h^*(\mathbf{z}) \tag{20}$$

for the non-convex term $\log\left|\sigma^2\mathbf{I} + \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top\right|$, where $h^*(\mathbf{z})$ denotes the convex conjugate of $\log\left|\sigma^2\mathbf{I} + \sum_{n=1}^{N}\exp\left(\kappa_n\right)\mathbf{L}_n^\top\mathbf{L}_n\right|$ in contrast to the concave conjugate counterpart, $w^*(\mathbf{z})$ used in Eq. (16). Substituting (20) into Eq. (14) leads to a so-called *min-max optimization program* for optimizing the non-convex penalty function $\mathcal{R}^{\text{II}-x}(\mathbf{X})$, which alternates between minimizations over $\boldsymbol{\gamma}$ and maximizations of the bound in (20):

$$\mathcal{R}^{\text{MacKay}}_{\text{conv}}(\mathbf{X}, \boldsymbol{\gamma}) = \min_{\boldsymbol{\gamma}\geq 0}\max_{\mathbf{z}>0} \left[\frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\frac{x_n(t)^2}{\gamma_n}\right] + \mathbf{z}^T\log(\boldsymbol{\gamma}) - h^*(\mathbf{z}) \ . \tag{21}$$

Let γ_n^k denote the value of γ_n in the k-th iteration. Inserting $\boldsymbol{\gamma}^k$ into Eq. (21) and minimizing with respect to $\boldsymbol{\gamma}_n^{k+1}$ requires that the derivates

$$\frac{\partial}{\partial\gamma_n^k}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{\mathbf{x}(t)^\top\mathbf{x}(t)}{\gamma_n^k} + \mathbf{z}^T\log(\gamma_n^k) - h^*(\mathbf{z})\right] = 0 \ ,$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

TABLE I

THIS TABLE SUMMARIZES THE UPDATE RULES PRESENTED IN SECTION II-C AND THEIR CORRESPONDING MM UPPER-BOUNDS THAT WILL BE UTILIZED IN SECTIONS III AND IV.

	Update Rule	Mathematical Formalism Used	Inequality Used
EM	$\gamma_n^{k+1} := [\mathbf{\Sigma}_{\mathbf{x}}^k]_{n,n} + \frac{1}{T}\sum_{t=1}^T(\boldsymbol{\mu}_{\mathbf{x}}(t))_n^2$	Expectation-Maximization Formalism	Jensen's Inequality
Convex Bounding	$\gamma_n^{k+1} := \sqrt{\left[\frac{1}{T}\sum_{t=1}^T\left(\mathbf{x}_n(t)\right)^2\right]}\left(\mathbf{L}_n^\top\left(\mathbf{\Sigma}_{\mathbf{y}}^k\right)^{-1}\mathbf{L}_n\right)^{-1/2}$	Concave Conjugate	Taylor Expansion
MacKay	$\gamma_n^{k+1} := \gamma_n^k\left[\frac{1}{T}\sum_{t=1}^T\left(\mathbf{x}_n(t)\right)^2\right]\left(\mathbf{L}_n^\top\left(\mathbf{\Sigma}_{\mathbf{y}}^k\right)^{-1}\mathbf{L}_n\right)^{-1}$	Change of Variable + Convex Conjugate	Taylor Expansion
LowSNR-BSI	$\gamma_n^{k+1} := \sqrt{\left[\frac{1}{T}\sum_{t=1}^T\left(\mathbf{x}_n(t)\right)^2\right]}\left(\mathbf{L}_n^\top\mathbf{L}_n\right)^{-1/2}$	MM Principle in Low-SNR Setting	Taylor Expansion

for $n = 1, \ldots, N$, vanish. The resulting function is then maximized with respect to $\mathbf{z}$ [9, Appendix-B], which leads to the so-called MacKay update for optimizing Eq. (14) [9, Section A-2]:

$$\begin{aligned}\gamma_n^{k+1} &:= \left[\frac{1}{T}\sum_{t=1}^T\left((\gamma_n^k)\mathbf{x}_n(t)\right)^2\right]\left(\gamma_n^k\mathbf{L}_n^\top\left(\mathbf{\Sigma}_{\mathbf{y}}^k\right)^{-1}\mathbf{L}_n\right)^{-1} \\ &= \gamma_n^k\left[\frac{1}{T}\sum_{t=1}^T\left(\mathbf{x}_n(t)\right)^2\right]\left(\mathbf{L}_n^\top\left(\mathbf{\Sigma}_{\mathbf{y}}^k\right)^{-1}\mathbf{L}_n\right)^{-1} \ ,\end{aligned} \tag{22}$$

for $n = 1, \ldots, N$.

III. UNIFICATION OF SPARSE BAYESIAN LEARNING ALGORITHMS WITH THE MAJORIZATION-MINIMIZATION (MM) FRAMEWORK

In this section, we first briefly review theoretical concepts behind the MM algorithmic framework [12], [44]–[46]. Then, we formally characterize Champagne variants as MM algorithms by suggesting upper bounds on the cost function Eq. (14) that, when employed within the MM framework, yield the same update rules as the original algorithms. The first three rows of Table I list the update rules and mathematical formalism used in this section.

A. *Majorization-Minimization*

Majorization-minimization is a promising strategy for solving general non-linear optimization programs. Compared to other popular optimization paradigms such as (quasi)-Newton methods, MM algorithms enjoy guaranteed convergence to a stationary point [13]. The MM class covers a broad range of common optimization algorithms such as *proximal methods* and *convex-concave procedures (CCCP)* [13, Section IV], [47]–[49]. While such algorithms have been applied in various contexts, such as non-negative matrix factorization [50] and massive MIMO systems for wireless communication [16], [19], their advantages have so far rarely been made explicit in the context of brain source imaging [51], [52].

Given the objective of minimizing a continuous function $f(\mathbf{u})$ within a closed convex set $\mathcal{U} \subset \mathbb{R}^n$:

$$\min_{\mathbf{u}} f(\mathbf{u}) \quad \text{subject to } \mathbf{u} \in \mathcal{U} \ , \tag{23}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the idea of MM can be summarized as follows. First, construct a continuous *surrogate function* $g(\mathbf{u}|\mathbf{u}^k)$ that upper-bounds, or *majorizes*, 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

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

Theorem 1. *Any MM algorithm satisfying conditions [A1]–[A3] 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. The proof is included in Appendix B. □

Theorem 1 implies that if a surrogate function is constructed to fulfill conditions [A1] and [A2], and if the next feasible point of the algorithm is always assigned as the minimizer of the surrogate function based on [A3], the resulting algorithm decreases $f(\mathbf{u})$ in each step. Algorithms possessing this property are often referred to as MM [12]. However, theoretical guarantees regarding the convergence of such algorithms [44, Theorem 1] require additional assumptions on particular properties of f and g [44], [45]. For the smooth functions considered in this paper, we require that the derivatives of the original and surrogate functions coincide at $\mathbf{u}^k$:

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

Then, the following, stronger, theorem holds.

Theorem 2. *Assume the conditions [A1]–[A4] are satisfied for a MM algorithm. Then, every limit point of the sequence of minimizers generated in [A3], is a stationary point of the original optimization problem Eq.* (23).

Proof. A detailed proof can be found in [44, 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 2 (see [12], [44], [46] and [43], [53]) for proofs of the convergence behaviour of other MM algorithms such as expectation maximization.

Note that the performance of MM algorithms heavily depends on the choice of a suitable surrogate function, which should, on one hand, faithfully reflect the behavior of the original non-convex function Eq. (23) while, on the other hand, be easy to minimize.

We now show that three algorithms that have been proposed for solving the SBL cost function Eq. (12) can all be cast as instances of the MM framework invoking different majorization functions on $\mathcal{R}^{\mathrm{II}-x}(\mathbf{X})$. For the convex-bounding based approach as well the algorithm using MacKay updates, the full set of conditions [A1]–[A4] in Theorem 2 are proven. Due to the considerations made above, we, however, only prove Theorem 1 for the EM-based Champagne algorithm.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1) EM Update as MM: It is known that the EM algorithm is a special case of MM framework using Jensens inequality to construct the surrogate function [13], [46]. Here, we work out the specific surrogate function for the SBL cost function Eq. (12) (i.e., the negative log marginal likelihood).

As Wipf and Nagarajan have shown [9, Section III.A-1], the EM algorithm for Type-II problems consists of the following two parts: For the E-step, the posterior $p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k)$ is obtained given the value of $\boldsymbol{\gamma}$ at k-th iteration, $\boldsymbol{\gamma}^k$. The M-step then solves:

$$\begin{aligned}\boldsymbol{\gamma}^{k+1} &:= \arg\min_{\boldsymbol{\gamma}} \mathrm{E}_{p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k)}\left[-\log p(\mathbf{Y},\mathbf{X}|\boldsymbol{\gamma})\right] \text{ , where} \\ -\log p(\mathbf{Y},\mathbf{X}|\boldsymbol{\gamma}) &= \frac{T}{2}\log|\boldsymbol{\Gamma}| + \frac{1}{2}\sum_{t=1}^{T}\mathbf{x}(t)^\top\boldsymbol{\Gamma}^{-1}\mathbf{x}(t) \\ &+ \frac{T}{2}\log\left|\sigma^2\mathbf{I}\right| + \sum_{t=1}^{T}\frac{1}{2\sigma^2}||\mathbf{y}(t)-\mathbf{L}\mathbf{x}(t)||_2^2 \,, \end{aligned} \tag{24}$$

which leads to the update rule in Eq. (15).

Proposition 3. *The EM based Champagne algorithm is an MM algorithm fulfilling Theorem 1, where the negative log-likelihood loss,* $-\log p(\mathbf{Y}|\boldsymbol{\gamma})$*, is majorized by the following surrogate function*

$$\begin{aligned}\mathcal{L}^k_{\mathrm{EM}}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k) &= \frac{T}{2}\log|\boldsymbol{\Gamma}| + \mathrm{E}_{p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k)}\left[\frac{1}{2}\sum_{t=1}^{T}\mathbf{x}(t)^\top\boldsymbol{\Gamma}^{-1}\mathbf{x}(t)\right] \\ &+ \frac{T}{2}\log\left|\sigma^2\mathbf{I}\right| + \mathrm{E}_{p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k)}\left[\sum_{t=1}^{T}\frac{1}{2\sigma^2}||\mathbf{y}(t)-\mathbf{L}\mathbf{x}(t)||_2^2\right] + \mathrm{E}_{p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k)}p\left(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k\right) \,. \end{aligned} \tag{25}$$

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

Note that the *EM* algorithm is also equivalent to the restricted maximum likelihood (ReML) [54] and dynamic statistical parametric mapping (dSPM) approaches [55] for solving the sparse EEG/MEG inverse problem, which, thereby, can also be interpreted as instances of minimization-majorization.

2) Convex-bounding Based Approach as MM: We start by recalling the non-convex penalty $\mathcal{R}^{\mathrm{II}-x}(\mathbf{X},\boldsymbol{\gamma})$ as defined in Eq. (14). By setting $\mathbf{x} = \mathbf{x}^k$ to the value obtained by the convex-bounding based method in the k-th iteration, the following holds:

Proposition 4. *The convex-bounding based Champagne algorithm is an MM algorithm fulfilling Theorem 2, where* $\mathcal{R}^{\mathrm{II}-x}(\mathbf{X},\boldsymbol{\gamma})$ *is majorized by the following surrogate function:*

$$\mathcal{R}^k_{\mathrm{conv}}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k) = \left[\frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\frac{x_n(t)^2}{\gamma_n}\right] + \log\left|\boldsymbol{\Sigma}^k_{\mathbf{y}}\right| + \mathrm{tr}\left[\left(\boldsymbol{\Sigma}^k_{\mathbf{y}}\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{y}}\right] - \mathrm{tr}\left[\left(\boldsymbol{\Sigma}^k_{\mathbf{y}}\right)^{-1}\boldsymbol{\Sigma}^k_{\mathbf{y}}\right] \,. \tag{26}$$

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

3) MacKay Update as MM: Similar to convex-bounding, we can show that the Mackay updates for Champagne can be viewed as an MM algorithm.

Proposition 5. *The Champagne variant employing MacKay updates is an MM algorithm fulfilling Theorem 2, where* $\mathcal{R}^{\mathrm{II}-x}(\mathbf{X},\boldsymbol{\gamma})$ *is majorized by* $\mathcal{R}^k_{\mathrm{conv}}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k)$*.*

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Proof. The proof is similar to that of Proposition 4 and provided in Appendix E. □

To summarize this section, we have shown that three popular strategies for solving the SBL problem in Eq. (12), namely the EM, the MacKay, and the convex bounding based approaches, can be characterized as MM algorithms. Importantly, this perspective provides a common framework for comparing different Champagne algorithms. For example, we can derive and compare certain characteristics of Champagne algorithms directly based on the properties of the majorization functions they employ. Conversely, it is also possible to design specific majorization functions that are optimal in a specific sense, leading to new source reconstruction algorithms.

IV. LowSNR-Brain Source Imaging (LowSNR-BSI)

Here, we assume a low signal-to-noise ratio (SNR) regime, as it is common in BSI applications. SNR is defined as signal power, $\mathbb{E}\{||x(t)||^2\}$, divided by noise power, σ^2: SNR $= \frac{\mathbb{E}\{||x(t)||^2\}}{\sigma^2}$, and can be expressed in dB scale as $\text{SNR}_{\text{dB}} = 10\log_{10}(\text{SNR})$. In many practical applications, we are interested in solving the BSI problem for $\text{SNR}_{\text{dB}} \leq 0$; that is, when the noise power is comparable to the power of the signal or even larger. Although the algorithms presented in Sections III-A1–III-A3 achieve satisfactory performance in terms of computational complexity, their reconstruction performance degrades significantly in low-SNR regimes. This behavior has been theoretically shown in [22, Section VI-E] and has also been confirmed in several simulation studies [23], [24].

In order to improve the performance of SBL in low-SNR settings, we propose a novel MM algorithm by constructing a surrogate function for Eq. (12) specifically for this setting. Based on [16], we propose the following convex surrogate function:

$$\mathcal{L}_{\text{conv}}^{\text{Low-SNR}}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k) = \text{tr}(\mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top) + \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}(t)^\top\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{y}(t) \ . \tag{27}$$

The following proposition is based on results in [16].

Proposition 6. *The surrogate function Eq.* (27) *majorizes the Type-II loss function Eq.* (12) *and results in an MM algorithm that fulfills Theorem 2. For SNR* $\to 0$*, Eq.* (12) *converges to Eq.* (27)*:*

$$\mathcal{L}^{\text{II}}(\boldsymbol{\gamma}) = \mathcal{L}_{\text{conv}}^{\text{Low-SNR}}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k) + \mathcal{O}(\text{SNR}) \ . \tag{28}$$

Proof. The proof of this result is presented in Appendix F. □

Note that, as a result of Proposition 6, the behaviour of the non-convex SBL cost function Eq. (12) is more and more well approximated in the vicinity of the current estimate by the proposed surrogate function Eq. (27) as the noise level increases, which sets it apart from existing surrogate functions. Therefore, the proposed bound is particularly suitable in low-SNR regimes.

In contrast to the original SBL cost function Eq. (12), the surrogate function Eq. (27) is convex and has unique minimum that can be found analytically in each iteration of the optimization. To find the optimal value of $\boldsymbol{\gamma} =$

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$[\gamma_1, \dots, \gamma_N]^\top$, we first take the derivative of (27) with respect to each γ_n for $n = 1, \dots, N$, and then set it to zero, which yields the following closed-form solution for $\boldsymbol{\gamma} = [\gamma_1, \dots, \gamma_N]^\top$:

$$\gamma_n^{k+1} := \sqrt{\frac{\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{\mu}_{\mathbf{x}}(t))_n^2}{\mathbf{L}_n^\top \mathbf{L}_n}} \text{ for } n = 1, \dots, N \, . \tag{29}$$

A detailed derivation of Eq. (29) can be found in Appendix G. We call the algorithm obtained by iterating between (9)–(11) and (29) *LowSNR-Brain Source Imaging (LowSNR-BSI)*. In practice, values exactly equal to zero may not be obtained for the γ_n. Therefore, an *active-set* strategy is employed. Given a threshold γ_{thresh}, those variances γ_n for which $\gamma_n < \gamma_{\text{thresh}}$ holds are set to zero in each iteration of the algorithm. Algorithm 1 summarizes the steps of LowSNR-BSI. Table I allows for a direct comparison of the LowSNR-BSI update rule (last column) and the corresponding update rules of other Champagne variants derived within the MM framework.

Algorithm 1: LowSNR-BSI algorithm.

```
Input: The lead field matrix $\mathbf{L} \in \mathbb{R}^{M \times N}$, the measurement vectors $\mathbf{y}(t) \in \mathbb{R}^{M \times 1}, t = 1, \dots, T$, and the noise
       variance $\sigma^2$.
Result: The estimated prior source variances $[\gamma_1, \dots, \gamma_N]^\top$, the posterior mean $\boldsymbol{\mu}_{\mathbf{x}}(t)$ and covariance $\boldsymbol{\Sigma}_{\mathbf{x}}$ of
        the sources.
Set a random initial value for $\boldsymbol{\gamma} = [\gamma_1, \dots, \gamma_N]^\top$ and construct $\boldsymbol{\Gamma} = \mathrm{diag}(\boldsymbol{\gamma})$.
Calculate the statistical covariance $\boldsymbol{\Sigma}_{\mathbf{y}} = \sigma^2\mathbf{I} + \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top$.
Repeat
    Calculate the posterior mean as $\boldsymbol{\mu}_{\mathbf{x}}(t) = \boldsymbol{\Gamma}\mathbf{L}^\top(\boldsymbol{\Sigma}_{\mathbf{y}})^{-1}\mathbf{y}(t)$.
    Update $\gamma_n$ for $n = 1, \dots, N$ based on Eq. (29).
    Recalculate the active set of brain sources by selecting the values of $\gamma_n$ that are greater than a pre-defined
      threshold: $\gamma_n > \gamma_{\text{thresh}}, \; n = 1, \dots, N$.
Until stopping condition is satisfied;
Calculate the posterior covariance as $\boldsymbol{\Sigma}_{\mathbf{x}} = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}\mathbf{L}^\top(\boldsymbol{\Sigma}_{\mathbf{y}})^{-1}\mathbf{L}\boldsymbol{\Gamma}$.
```

V. Automatic Estimation of the Noise Level

A. Adaptive Noise Learning

It is common practice to estimate the noise variance σ^2 from baseline data prior to solving the EEG/MEG inverse problem [23], [24], [56]–[62]. However, a baseline estimate may not always be available or may not be accurate enough, say, due to inherent non-stationarities in the data/experimental setup. Here, we argue that estimating the noise parameter from the to-be-reconstructed data can significantly improve the reconstruction performance even compared to a baseline estimate. To this end, we here derive data-driven update rules that allow us to tune estimate the noise variance, σ^2 within the source reconstruction procedure using the Champagne and LowSNR-BSI algorithms, where we build on prior work by [8], [26], [63], [64]. Practically we introduce the shortcut $\lambda = \sigma^2$ to underscore that λ is a tunable parameter whose estimate can substantially deviate from the baseline estimate in practice. We

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

then treat λ as another model hyperparameter, similar to the noise variances γ_n. Thus, in each step of learning cycles of the Champagne and LowSNR-BSI algorithms, we also minimize the loss function $\mathcal{L}^{\text{II}}$ with respect to λ, where the remaining parameters $\mathbf{\Gamma}$ and $\mathbf{\Sigma}_{\mathbf{x}}$ are fixed to the values obtained in the preceding iteration. This leads to the following theorem:

Theorem 7. *The minimization of $\mathcal{L}^{\text{II}}(\lambda)$ with respect to λ,*

$$\lambda^* := \arg\min_{\lambda} \mathcal{L}^{\text{II}}(\lambda) = \arg\min_{\lambda} \left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{y}(t)^\top \mathbf{\Sigma}_{\mathbf{y}}^{-1} \mathbf{y}(t) + \log|\mathbf{\Sigma}_{\mathbf{y}}| \right) ,$$

yields the following update rule for λ at the $(k+1)$-th iteration, assuming $\mathbf{\Gamma}^k$ and $\mathbf{\Sigma}_{\mathbf{x}}^k$ be fixed values obtained in the (k)-th iteration:

$$\lambda^{k+1} := \frac{\frac{1}{T}\sum_{t=1}^{T} ||\mathbf{y}(t) - \mathbf{L}\mathbf{x}^k(t)||^2}{M - \text{tr}\left[\mathbf{I}_N\right] + \text{tr}\left[(\mathbf{\Sigma}_{\mathbf{x}}^k)^{-1}(\mathbf{\Gamma}^k)^{-1}\right]} \; . \tag{30}$$

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

B. Cross-validation Strategies

In the previous section, we proposed to estimate the noise variance $\lambda = \sigma^2$ *in-sample* such that the SBL likelihood according to Eq. (12) was maximized, which led to an analytic update rule. As, under our assumption of homoscedastic sensor noise, λ is only a single scalar parameter, it moreover becomes feasible make use of robust model selection techniques employing the concept of cross-validation (CV), whose aim it is to maximize the *out-of-sample* likelihood [65]–[67]. To this end, the data are split into two parts. On the so-called *training set*, the model parameters is fitted for a wide range of possible values of γ, which are fixed within each individual optimization. The likelihoods of the fitted models ares then evaluated on the held-out data parts, called the *test sets*. The choice of λ that maximizes the empirical likelihood on the test data is then used as an unbiased estimate of the noise variance. It is well-known from the field of machine learning that cross-validation effectively overcomes the problem of model overfitting in small samples. Here, we introduce two CV strategies employing different ways of splitting the data.

1) Temporal Cross-validation: In temporal CV, the temporal sequence of the data samples is split into k different contiguous blocks (folds) [68], [69]. Here, we use $k = 4$. Three folds form the training set, $\mathbf{Y}^{\text{train\_temp}} \in \mathbb{R}^{M \times T^{\text{train\_temp}}}$, on which we fit the Champagne and LowSNR-BSI models for a range of λs. On the remaining fold, $\mathbf{Y}^{\text{test\_temp}} \in \mathbb{R}^{M \times T^{\text{test\_temp}}}$, the Type-II log-likelihood (c.f. Eqs. (12) and (13))

$$\begin{aligned} \mathcal{L}^{\text{II}}(\mathbf{Y}^{\text{train\_temp}}, \mathbf{Y}^{\text{test\_temp}}) &= \frac{1}{T}\sum_{t=1}^{T} \mathbf{y}^{\text{test\_temp}}(t)^\top \mathbf{\Sigma}_{\mathbf{y}^{\text{train\_temp}}}^{-1} \mathbf{y}^{\text{test\_temp}}(t) + \log|\mathbf{\Sigma}_{\mathbf{y}^{\text{train\_temp}}}| \\ &= \text{tr}(\mathbf{C}_{\mathbf{y}^{\text{test\_temp}}} \mathbf{\Sigma}_{\mathbf{y}^{\text{train\_temp}}}^{-1}) + \log|\mathbf{\Sigma}_{\mathbf{y}^{\text{train\_temp}}}| \end{aligned} \tag{31}$$

is then evaluated. Note that in Eq. (31) the model covariance $\mathbf{\Sigma}_{\mathbf{y}^{\text{train\_temp}}}$ that has been determined on the training data $\mathbf{Y}^{\text{train\_temp}}$ is combined with the empirical covariance of the hold-out data $\mathbf{Y}^{\text{test\_temp}}$, which were not used during model fitting. Thus, Eq. (31) is the *out-of-sample* Type-II log-likelihood. It has been theoretically shown [22], [70] that the Type-II log-likelihood function is a metric on the second-order information of the sensors closely related to

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the log-det Bregman divergence (discrepancy) the between statistical (model) and empirical covariance [40], [71]. The choice of λ that minimizes that discrepancy on hold-out data is, therefore, a sensible estimate for the true noise variance. We provide further details on the relation between the SBL likelihood and the log-det Bregman divergence in Appendix A.

2) Spatial Cross-validation: In spatial CV, the data are not split into temporal segments but by dividing the available EEG/MEG sensors into the training and test sets. This variant has been proposed by [25], [31]. Here, we again use $k = 4$ folds, where we randomly assign 75% of the sensors to the training set, $\mathbf{Y}^{\text{train\_spat}} \in \mathbb{R}^{M^{\text{train\_spat}}} \times T$, and the remaining 25% to the test set, $\mathbf{Y}^{\text{test\_spat}} \in \mathbb{R}^{M^{\text{test\_spat}}} \times T$. On the training sensors, Champagne and LowSNR-BSI are fitted using the corresponding portion of the leadfield matrix, $\mathbf{L}^{\text{train\_spat}}$, for the same range of λs as used in temporal CV. The sources, $\mathbf{X}^{\text{train\_spat}} \in \mathbb{R}^{N \times T}$, estimated from the fitted models are then mapped back to the sensor space, and the out-of-sample Type-I log-likelihood (c.f. Eq. (5)) is evaluated on the hold-out (test) sensors:

$$\begin{aligned}\mathcal{L}^{\text{I}}(\mathbf{Y}^{\text{train\_spat}}, \mathbf{Y}^{\text{test\_spat}}) &= \frac{1}{T}\sum_{t=1}^{T} \left\| \mathbf{y}^{\text{test\_spat}}(t) - \mathbf{L}^{\text{test\_spat}}\mathbf{x}^{\text{train\_spat}}(t) \right\|_2^2 \\ &:= \left\| \mathbf{Y}^{\text{test\_spat}} - \mathbf{L}^{\text{test\_spat}}\mathbf{X}^{\text{train\_spat}} \right\|_F^2 \ . \end{aligned} \tag{32}$$

Note that, while the Type-II log-likelihood has an interpretation as a Bregman divergence beween model and empirical covariance matrices, the Type-I log-likelihood is the Frobenius norm or mean-squared error (MSE) $\|\cdot\|_F^2$ of the model residuals, i.e., the average squared Euclidean distance between empirical and modeled observation vectors. Thus, while the Type-II likelihood compares model and observations in terms of their second-order statistics, the Type-I likelihood uses only first-order information. As in temporal CV, the value of λ that minimizes the MSE on the test sensors is selected as the final noise estimate.

VI. Simulations

We conducted an extensive set of simulations, in which we compared the reconstruction performance of the proposed LowSNR-BSI algorithm to that of Champagne and two additional widely-used source reconstruction schemes for a range of different SNRs. We also tested impact of the proposed noise learning schemes (adaptive, temporal CV and spatial CV) on the source reconstruction performance compared to estimating the noise level from baseline data.

A. Pseudo-EEG Signal Generation

Forward Modeling: Populations of pyramidal neurons in the cortical gray matter are known to be the main drivers of the EEG signal [72]–[74]. Here, we use a realistic volume conductor model of the human head to model the linear relationship between primary electrical source currents in these populations and the scalp surface potentials captured by EEG electrodes. The New York Head model [4] provides a segmentation of an average human head into six different tissue types. 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 [75]. In accordance with the predominant orientation of pyramidal neuron assemblies, the orientation of all source currents was fixed to be

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

perpendicular to the cortical surface, so that only scalar source amplitudes needed to be estimated. Finite-element modeling was used to compute the lead field matrix, $\mathbf{L} \in \mathbb{R}^{58\times 2004}$, which serves as the forward model in our simulations.

Source Generation: We simulated a sparse set of $N_0 = 3$ active sources, which were placed at random positions on the cortex. The temporal activity of each source was generated by a univariate linear autoregressive (AR) process, which models the activity at time t as a linear combination of the P past values:

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

Here, $a_i(p)$ for $i = 1, 2,$ and 3 are linear AR coefficients, and P is the order of the AR model. The model residuals $\xi_i(\cdot)$ for $i = 1, 2$ and 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 the following, we use stable AR systems of order $P = 5$.

Noise Model: To simulate the electrical neural activity of the underlying brain sources, $T = 20$ data points were sampled from the AR process described above. Corresponding dipolar current sources were then placed at random locations, yielding sparse source activation vectors $\mathbf{x}(t)$. Source activations $\mathbf{X} = [\mathbf{x}(1), \ldots, \mathbf{x}(T)]$ were mapped to the 58 EEG sensors through application of the lead field matrix $\mathbf{L}$:

$$\mathbf{Y}^{\text{signal}} = \mathbf{L}\mathbf{X} \tag{33}$$

Next, we added Gaussian white noise to the sensor-space signal. To this end, noise was randomly sampled from a standard normal distribution and normalized with respect to its Frobenius norm. A weighted sum of signal and noise contributions then yielded the pseudo-EEG signal

$$\mathbf{Y} = \mathbf{Y}^{\text{signal}} + \alpha \frac{\mathbf{Y}^{\text{noise}}}{\|\mathbf{Y}^{\text{noise}}\|_F}, \tag{34}$$

where α determines the signal-to-noise ratio in sensor space. For a given α, the noise variance is obtained as $\sigma^2 = {}^1\!/\!_M\, \text{tr}[\mathbf{\Sigma_e}]$, for $\mathbf{\Sigma_e} = \text{Cov}[\alpha \frac{\mathbf{Y}^{\text{noise}}}{\|\mathbf{Y}^{\text{noise}}\|_F}]$, and the SNR (in dB) is calculated as $\text{SNR} = 20\log_{10}\left(\|\mathbf{Y}^{\text{signal}}\|_F/\alpha\right)$. Note that since our goal is to investigate the effect of noise variance estimation on the performance of the proposed algorithms, we fixed the noise variance in each set of simulations so as to obtain distributions of performance metrics for a number of discrete SNR values. We conducted four sets of simulations using $\alpha = \{2, 1.5, 1, 0.5\}$, corresponding to average noise variances of $\sigma^2 = \{37.4 \times 10^{-3}, 21.0 \times 10^{-3}, 9.4 \times 10^{-3}, 2.3 \times 10^{-3}\}$ and average SNRs of $\text{SNR} = \{0.33, 2.17, 4.87, 11.40\}$ (dB). Each set of simulations consists of 100 experiments, in which source locations and time series as well as noise realizations were randomly sampled.

In addition to the pseudo-EEG signal, a pseudo baseline measurement containing only noise but no signal was generated. The sole purpose of this measurement was to provide an empirical estimate of the noise variance as a baseline for our joint source reconstruction and noise estimation approaches, which estimate the same quantity from the summed pseudo-EEG signal. To ensure sufficiently precise baseline estimation, 300 noise samples were generated, normalized, and scaled by α as in Eq. (34) for each experiment.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

B. Source reconstruction

We applied Champagne and LowSNR-BSI to the synthetic datasets described above. All parameters were initialized randomly. The optimization programs were terminated either after reaching convergence (defined by a relative change of the Frobenius-norm of the reconstructed sources between subsequent iterations of less than 10^{-8} or after reaching a maximum of 3000 iterations).

In each experiment, we evaluated the algorithms using 40 predefined choices of the noise variance ranging from $\lambda = {}^{1}\!/_{3}\sigma^2$ to $\lambda = 30\sigma^2$. In addition, λ was estimated from data using the techniques introduced in Section V. We observed that the variance estimated from baseline data, $\hat{\sigma}^2$ (averaged over all EEG channels) was typically almost identical to the ground-truth value $\lambda = \sigma^2$ used to simulate the data. The reconstruction performance obtained using this value was therefore included in the comparison as a baseline. Performance at baseline noise level was compared to the performance obtained using adaptive learning of the noise using Eq. (30) as well as using spatial or temporal cross-validation. Note that, for temporal CV, we generated $T = 80$ samples, so that we obtained 60 samples in each training set and 20 samples in each test fold. Due to the increased number of training samples, this method, therefore, has an advantage over the remaining ones.

In addition to Champagne and LowSNR-BSI, two non-SBL source reconstruction schemes were included for comparison. As an example of a sparse Type-I method based on ℓ_1-norm minimization, S-FLEX [31] was used. As spatial basis functions, unit impulses were used, so that the resulting estimate was identical to the so-called minimum-current estimate [6]. In addition, the eLORETA estimate, a smooth inverse solution based on ℓ_2-norm minimization was used. eLORETA 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%.

C. Evaluation Metrics

Source reconstruction performance was evaluated according to the following metrics. First, the *earth mover's distance* (EMD, [30], [76]) was used to quantify the spatial localization accuracy. The EMD metric measures the cost needed to transform two probability distributions, defined on the same metric domain, into each other. It was applied here to the $N \times 1$ amplitude distributions of the true and estimated sources, which were obtained by taking the voxel-wise ℓ_2-norm along the time domain. EMD scores were normalized to be in $[0, 1]$. 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) sources was measured. Each simulated source was matched to a reconstructed source based on maximum absolute correlation. Time course reconstruction error was then defined as one minus the average of these absolute correlations across sources. Finally, the runtime of the algorithms was measured in seconds (s).

D. Results

Figure 1 shows the EMD (upper row), the time course reconstruction error (middle row) and the negative log-likelihood loss value (lower low) incurred by Champagne and LowSNR-BSI for two SNR settings $(\mathrm{SNR} = 0.33\ \mathrm{dB}$

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and $\mathrm{SNR} = 11.40$ dB). Four different schemes of estimating the noise level from data (estimation from baseline data, adaptive learning, spatial CV, and temporal CV) are compared. Note that we found previously that the ground-truth noise variance $\lambda = \hat{\sigma}^2$ used in the simulation is generally accurately estimated from baseline data, which is referred to as 'baseline' in the figure. Interestingly, however, this baseline is optimal only for LowSNR-BSI, and only with respect to temporal source reconstruction. For Champagne, and with respect to the spatial source reconstruction performance of LowSNR-BSI, the choice of the baseline noise variance turns out to be suboptimal, as it is outperformed by all three proposed schemes that estimate the noise variance from the actual (task) data to be reconstructed ('Adaptive Learning', 'Spatial CV' and 'Temporal CV'). Interestingly, noise levels estimated using Spatial CV lead to near-optimal reconstruction performance in a broad variety of settings, in line with observations made in [25], [31]. It is also worth noting that all proposed noise learning schemes converge to points in the vicinity of the minimum of the loss function.

Figure 2 further compares the source reconstruction performance of the four noise estimation variants separately for Champagne and LowSNR-BSI for a range of four SNR values. As already observed in Figure 1, all three proposed approaches for noise variance estimation (adaptive learning, spatial CV, and temporal CV) lead to better source reconstruction performance than the estimation from baseline data. Overall, spatial CV for Champagne and temporal CV for LowSNR-BSI achieve the best combination of spatial and temporal reconstruction performance.

The superior performance of CV techniques, however, comes at the expense of higher computational complexity of the source reconstruction. As Figure 2 demonstrates, using CV techniques with the specified numbers of folds increases the runtime of Champagne and LowSNR-BSI by approximately two orders of magnitude ($10^3 s \sim 10^4 s$) compared to the runtimes of eLORETA, S-FLEX, and the baseline and adaptive learning variants of Champagne and LowSNR-BSI ($1s \sim 10s$).

Figure 3 provides an alternative depiction of the data presented in Figure 2, which allows for a more direct comparison of Champagne and LowSNR-BSI. As benchmark algorithms, eLORETA [29] and S-FLEX [31] are also included in the comparison. It can be seen that LowSNR-BSI in the baseline mode, using adaptive noise learning, and using temporal CV consistently outperforms Champagne in terms of spatial localization accuracy, in particular in low-SNR settings. This behavior indeed confirms the advantage of the surrogate function, $\mathcal{L}_{\text{conv}}^{\text{Low-SNR}}(\gamma|\gamma^k)$, which is designed to provide a better approximation of the non-convex SBL cost function in low-SNR regimes, as presented in Section IV. Consequently, as the SNR decreases, the gap between LowSNR-BSI and Champagne further increases. In terms of the time course reconstruction error, LowSNR-BSI shows a similar improvement over Champagne when the SNR is low. However, the magnitude of this improvement is not as pronounced as observed for the EMD metric. The only setting in which Champagne consistently outperforms LowSNR-BSI is when spatial CV is used to estimate the noise variance, and spatial reconstruction performance is evaluated.

It can further be observed that S-FLEX yields higher spatial localization accuracy (lower EMD) than eLORETA, while eLORETA yields higher temporal accuracy (lower time course error) than S-FLEX across all SNR values. With respect to spatial accuracy, both approaches, however, are consistently outperformed by Champagne and LowSNR-BSI. Note that the superior spatial reconstruction of sparsity-inducing algorithms (Champagne, LowSNR-BSI and S-FLEX) compared to eLORETA is expected here, because the simulated spatial distributions are indeed sparse.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

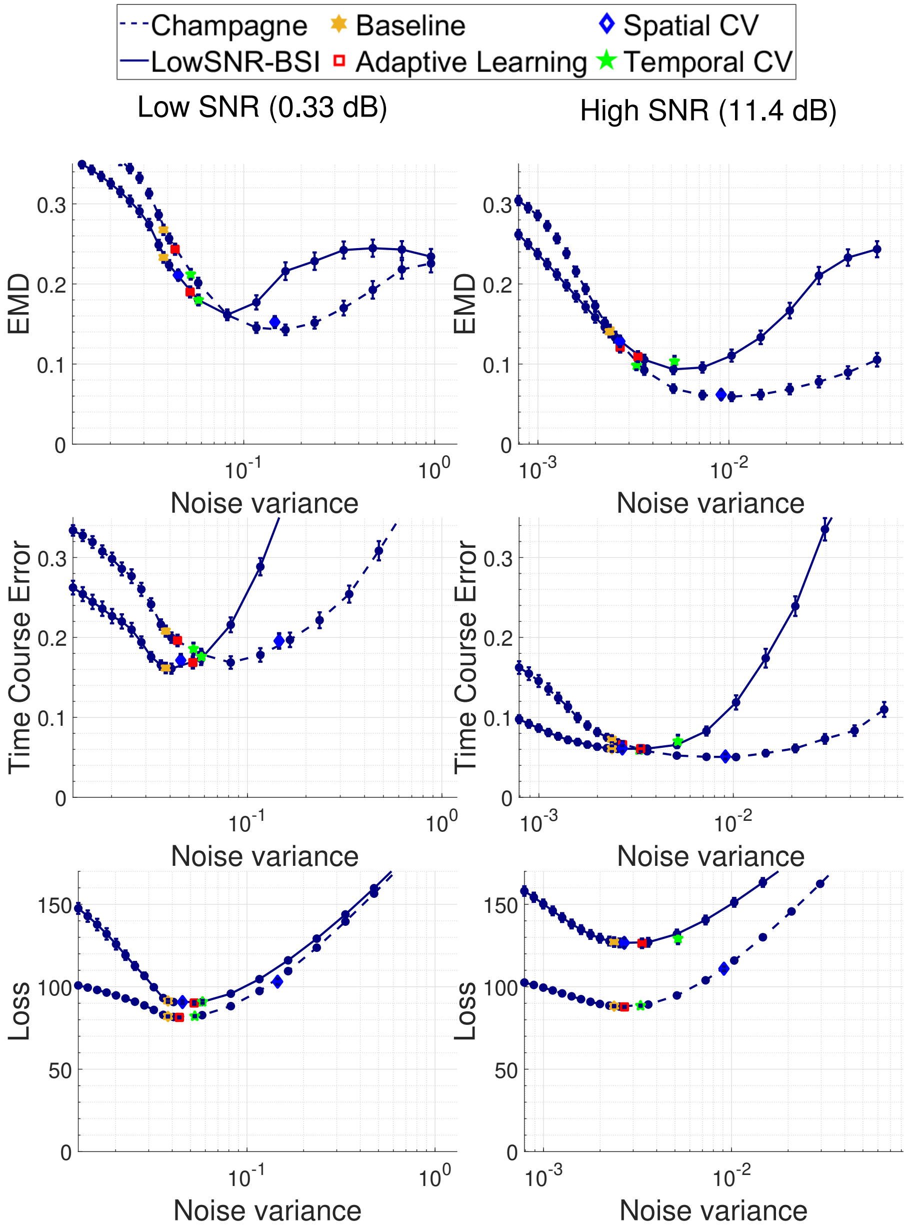

Fig. 1. Source reconstruction performance of Champagne and LowSNR-BSI in two different SNR regimes (low SNR: 0.33 dB, left column; high SNR: 11.4 dB, right column). Spatial reconstruction error is measured in terms of the earth-mover's distance, and is shown in the upper row, while time course reconstruction error is shown in the lower row.

The superiority of SBL methods (Champagne, LowSNR-BSI) over S-FLEX that is observed here confirms observations and theoretical considerations made in [7], [23], [24]. eLORETA shows comparable temporal reconstruction performance as LowSNR-BSI and Champagne, while S-FLEX is outperformed by all other methods.

The convergence behavior of the different SBL variants discussed and introduced in Sections III–V is illustrated in Figure 4. LowSNR-BSI variants have faster convergence rates at the early stage of the optimization procedure compared to standard Champagne as well as Champagne with MacKay updates. They, however, reach lower negative

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

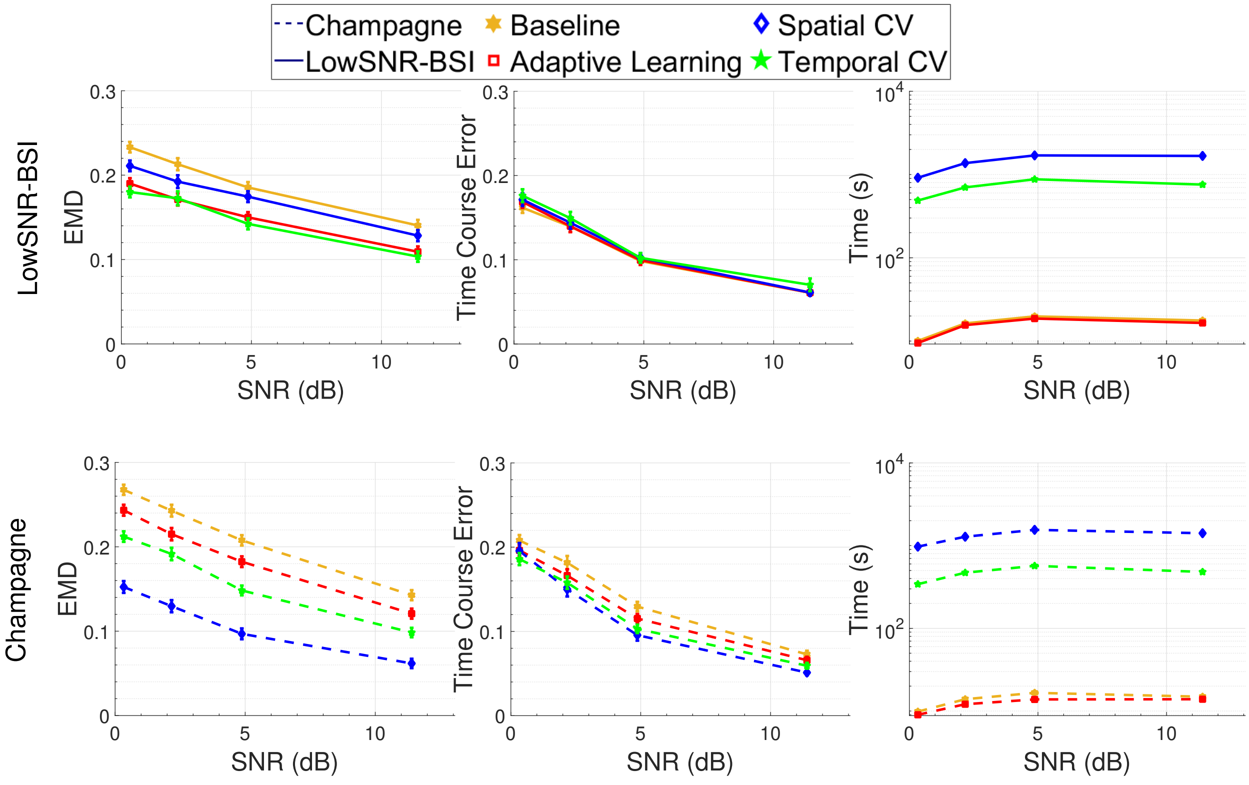

Fig. 2. Source reconstruction performance of four different variants of LowSNR-BSI (upper row) and Champagne (lower row). The noise variance was estimated from baseline data (ground truth), using adaptive learning, or using spatial or temporal cross-validation. Performance was evaluated for four SNRs ($\text{SNR} = \{0.33, 2.17, 4.87, 11.40\}$ dB) and with respect to three different metrics (spatial reconstruction according to the earth-mover's distance – left column, time course reconstruction error – middle column, and computational complexity – right column).

log-likelihood values eventually, which indicates that they find better maxima of the model evidence. Furthermore, the adaptive-learning variants of Champagne and LowSNR-BSI reach lower negative log-likelihood values than their counterparts estimating the noise variance from baseline data, suggesting that learning the noise variance, or in other words overestimating the noise variance, improves the reconstruction performance through better model evidence maximization.

VII. Analysis of Auditory Evoked Fields (AEF)

The MEG data used here 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 of one subject to an Auditory Evoked Fields (AEF) stimulus were localized. The AEF response was elicited with single 600 ms duration tones (1 kHz) presented binaurally. The data were averaged across 120 trials (after the trials were time-aligned to the stimulus). The pre-stimulus window was selected to be 100 ms to 5 ms and the post-stimulus time window was selected to be 5 ms to 250 ms, where 0 ms is the onset of the tone. Further details on this dataset can be found in [7], [23], [24], [77].

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

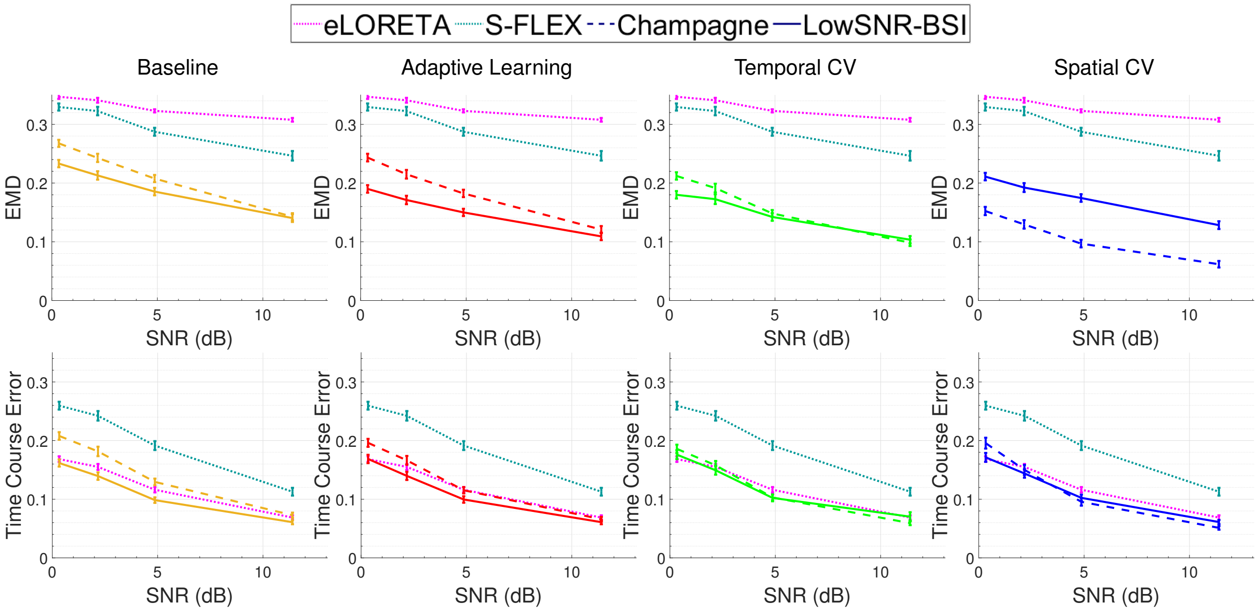

Fig. 3. Source reconstruction performance of Champagne (dashed line) and LowSNR-BSI (solid line) for four SNR values ($\mathrm{SNR} = \{0.33, 2.17, 4.87, 11.40\}$ dB). The noise variance was estimated from baseline data baseline as well as using adaptive learning, spatial and temporal CV. Spatial reconstruction error was measured in terms of the earth-mover's distance and is shown in the upper row, while time course reconstruction error is shown in the lower row.

The lead field for each subject was calculated with NUTMEG (http://bil.ucsf.edu) using a single-sphere head model (two spherical orientation lead fields) and an 8 mm voxel grid.

Figure 5 shows the reconstructed sources of the AEF of one subject using conventional Champagne with pre-estimated $\lambda = \sigma^2$, adaptive noise learning, and spatial CV. LowSNR-BSI with pre-estimated $\lambda = \sigma^2$ was also included in the comparison. Shown in the top panel are the reconstructions at the time of the maximal deflection of the auditory N100 component (shown in bottom panel).

All reconstructions are able to correctly localize bilateral auditory activity to Heschel's gyrus, which is the location of the primary auditory cortex. Note that an additional source in the midbrain, which is indicated by all three Champagne variants, is absent for lowSNR-BSI.

VIII. Discussion

We have provided a unifying theoretical platform for deriving different sparse Bayesian learning algorithms for electromagnetic brain imaging using the Majorization-Minimization (MM) framework. First, we demonstrated that the choice of upper bounds of the Type-II non-convex loss function within the MM framework influences the reconstruction performance and convergence rates of the resulting algorithms. Second, focusing on commonly occurring low-SNR settings, we derived a novel Type-II Bayesian algorithm, LowSNR-BSI, using a novel convex bounding MM function that converges to the original loss function as the SNR goes to zero. We demonstrated the advantage of LowSNR-BSI over existing benchmark algorithms including Champagne, eLORETA and S-FLEX.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

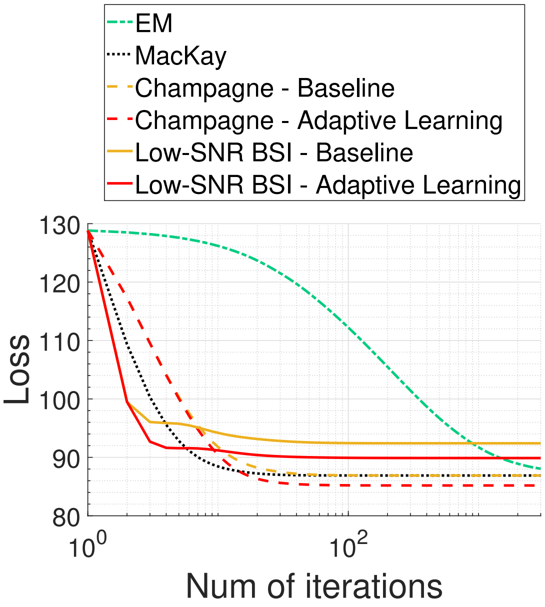

Fig. 4. Convergence behavior of LowSNR-BSI as well as Champagne using the standard (convex-bounding based) updates (Champagne) as well as EM and MacKay updates. For standard Champagne and LowSNR-BSI, the use of a fixed noise variance estimated from baseline data is compared with adaptive noise learning. LowSNR-BSI variants have faster convergence rate at early stages of the optimization procedure, but later converge to less optimal log-likelihood values. Adaptive learning variants of Champagne and LowSNR-BSI reach better log-likelihood values than their counterparts using a fixed noise variance estimated from baseline data.

Consistent with the theoretical considerations, the advantage of LowSNR-BSI over Champagne decreases with increasing SNR. Third, we have derived an analytic solution that allows us to estimate of the noise variance jointly within the source estimation procedure on the same (task-related) data that are used for the reconstruction. We have also adopted cross-validation schemes to empirically estimate the noise variance from hold-out data through a line search. We have proposed spatial and temporal CV schemes, where either subsets of EEG/MEG channels or recorded samples are left out of the source reconstruction, and where the noise variance is selected as the minimizer of a divergence between model and hold-out data. We also demonstrate that precise knowledge of the noise variance is required in order to determine the optimal algorithm performance. Finally, according to our empirical results, all three proposed techniques for estimating the noise variance lead to superior source reconstruction performance compared to the setting in which the noise variance is estimated from baseline data.

A. *Cross-validation vs. adaptive noise learning*

Spatial CV for Champagne and Temporal CV for LowSNR-BSI achieved the best performances and are generally applicable to any distributed inverse solution. Their long computation time can, however, be challenging as their computational complexity is drastically higher (around two orders of magnitude) than using baseline data or adaptive learning schemes. The high complexity of CV techniques is a potential limitation in settings where the efficiency of the algorithm or immediate access to the outcome is crucial. What is more, this approach quickly becomes infeasible

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

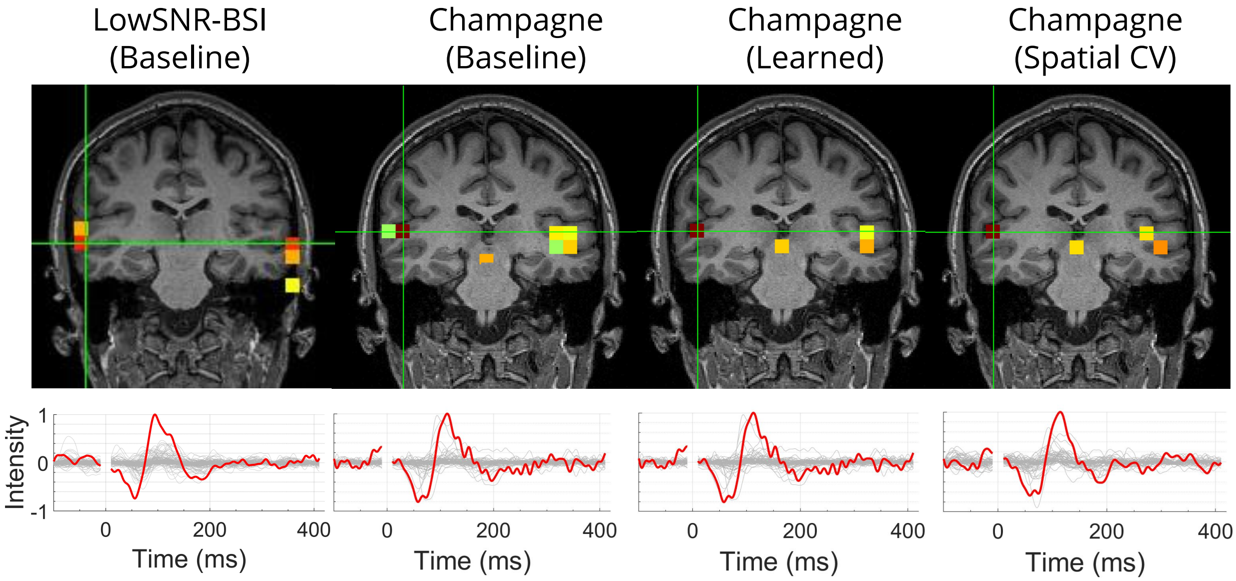

Fig. 5. Analysis of auditory evoked fields (AEF) of one subject using conventional Champagne with pre-estimated $\lambda = \sigma^2$, adaptive noise learning, and spatial CV as well as LowSNR-BSI. Shown in the top panel are the reconstructions at the time of the maximal deflection of the auditory N100 component (shown in bottom panel). All reconstructions show sources at the expected locations in the left and right auditory cortex.

if more than one parameter needs to be estimated through a grid search. In contrast, the computational complexity of the proposed noise level estimation scheme using adaptive learning is of the same order as the complexity of the baseline approach. Moreover, we have successfully extended this approach to the estimation of heteroscedastic noise, where a distinct variance is estimated for each M/EEG sensor [78]. Hence, the adaptive-learning approach can be seen as an advancement of the baseline algorithm that combines performance improvement and computational efficiency. It is also worth noting that the computational complexity of CV techniques heavily relies on tunable parameters such as the number of folds/splits of the data and the total number of candidate points in the grid search.

B. *Interpretation of Type-I and Type-II loss functions as divergences*

We have pointed out (see Section V-B and Appendix A) that Type-I and Type-II Bayesian approaches implicitly use different metrics to compare the empirical sensor-space observations to the signal proportion explained by the reconstructed brain sources. Type-I approaches measure first-order differences between modeled and reconstructed time series using variants of the MSE, while Type-II approaches amount to using the log-det Bregman divergence to measure differences in the second-order statistics of the empirically observed and modeled data as summarized in the respective covariance matrices. While the connection between the Type-II loss function and the log-det Bregman divergence has been investigated and exploited in numerous forms such as *Stein's loss* [40] or the *graphical Lasso* [70], [79], [80], and has found applications in disciplines such as information theory and metric learning [81],

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[82], wireless communication [19], and signal processing [22], [83], [84], it has not received much attention in the BSI literature to the best of authors' knowledge. Here, we have used this insight to devise a novel cross-validation scheme, temporal CV, in which model fit is measured in terms of the log-det Bregman divergence (or, Type-II likelihood) on held-out samples. In contrast, the previously introduced spatial CV uses the mean-squared error to measure out-of-sample model fit. Importantly, however, this difference does not imply that the application of spatial CV is restricted to Type-I approaches or that the use of temporal CV is restricted to Type-II approaches. Rather, both approaches are universally applicable. In fact, it is straightforward to evaluate the Type-I likelihood based on the source times series reconstructed with Type-II methods. Conversely, it is also possible to estimate the Type-II likelihood for Type-I approaches such as S-FLEX. Here, the model source and noise covariances are first estimated from the reconstructed sources as $\hat{\mathbf{\Gamma}} = \mathrm{Cov}[\boldsymbol{x}(t)]$ and $\hat{\sigma}^2 = {}^{1}\!/{}_{M} \sum_m [\mathbf{C}_{\mathbf{y}} - \mathbf{L}\hat{\mathbf{\Gamma}}\mathbf{L}^\top]_{[m,m]}$, after which $\mathbf{\Sigma}_{\boldsymbol{y}}$ can be calculated. The optimal Type-I regularization parameter is then selected as the minimizer of $\mathcal{L}^{\mathrm{II}}(\mathbf{Y}^{\mathrm{train\_temp}}, \mathbf{Y}^{\mathrm{test\_temp}})$ in Eq. (31).

C. Limitations and future work

The current results have been obtained for the scalar setting, where the orientation of the brain sources are assumed to be perpendicular to the surface of cortex and, hence, only the scalar deflection of each source along the fixed orientation needs to be estimated. However, as all algorithms considered here model the source covariance matrix $\mathbf{\Gamma}$ to be diagonal, they can be readily extended to the case where each source is modeled as a full 3-dimensional current vector. This can be achieved by introducing three variance parameters for each source within the source covariance matrix. In parallel, a full 3D leadfield matrix mapping the three components of each source to the sensors can be used.

Another assumption of the current work is that the activity of the sources is modeled to be independent across voxels, spatial orientations, and time samples. Analogously, the noise is assumed to be independent across times samples, and homoscedastic (independent with equal variance across sensors). Note that these assumptions merely act as prior information whose purpose is to bias the inverse reconstruction towards solutions with lower complexity. Thus, they do not prevent the reconstruction of brain and noise sources with more complex structure if the observed data are inconsistent with these priors. On the other hand, modeling dependency structures that are in fact present in real data has the potential to substantially improve the source reconstruction. We have recently proposed adaptive noise learning algorithms that relax the rather unrealistic assumption of homoscedastic noise [78]. Going further, it would be possible to also model spatial covariances of the sources between voxels and/or between source orientation within voxels, which would encode the realistic assumption that individual brain regions do not work in isolation. Similary, the spatial covariance structure of the noise could be modeled in order to accommodate spatially distributed artifacts due to, for example, heart beat or line noise interference. Finally, electrophysiological data are known to possess a complex intrinsic autocorrelation structure, which is not modeled by the majority of existing BSI algorithms. We have recently proposed ways to also learn temporal correlations within the Type-II framework and have obtained promising results with respect to time course reconstruction [51], [85].

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

IX. Conclusion

We have provided a unifying theoretical platform for deriving different sparse Bayesian learning algorithms for electromagnetic brain imaging using the Majorization-Minimization (MM) framework. This unification perspective not only provides a useful theoretical framework for comparing different algorithms in terms of their convergence behavior, but also provides a principled recipe for constructing novel algorithms with specific properties by designing appropriate bounds of the Bayesian marginal likelihood function. Building on MM principles, we then proposed a novel method called *LowSNR-BSI* that achieves favorable source reconstruction performance in low signal-to-noise-ratio settings. Recognizing the importance of noise estimation for algorithm performance, we present both analytical and cross-validation approaches for noise estimation. Empirically, we show that the monotonous convergence behavior predicted from MM theory is confirmed in numerical experiments. Using simulations, we further demonstrate the advantage of LowSNR-BSI over conventional Champagne in low-SNR regimes, and the advantage of learned noise levels over estimates derived from baseline data. To demonstrate the usefulness of our novel approach, we show neurophysiologically plausible source reconstructions on averaged auditory evoked potential data.

Our characterization of the Type-II likelihood as a divergence measure provides a novel perspective on the construction of BSI algorithms and might open new avenues of research in this field. It is conceivable that alternative divergence metrics can be used for solving the M/EEG source reconstruction problem in the future by modeling specific neurophysiologically valid aspects of similarity between data and model output. Promising metrics in that respect are information divergences such as Kullback-Leibler (KL) [86], Rényi [87], Itakura-Saito (IS) [88], [89] and β divergences [90]–[94] as well as transportation metrics such as the Wasserstein distance between empirical and statistical covariances (e.g., [95]–[98]).

Although this paper focuses on electromagnetic brain source imaging, Type-II methods have also been successfully developed in other fields such as direction of arrival (DoA) and channel estimation in massive Multiple Input Multiple Output (MIMO) systems [15], [16], [19], [99], robust portfolio optimization in finance [20], covariance matching and estimation [83], [100]–[105], graph learning [106], and brain functional imaging [86]. The methods introduced in this work may also prove useful in these domains.

Appendix

A. Bregman Divergence Formulation of the Loss Function

We start by recalling the definition of log-det Bregman matrix divergence - also known as Stein's loss [40] - between any two $M \times M$ positive semidefinite (PSD) matrices $\mathbf{Q}$ and $\mathbf{W}$:

$$\mathcal{D}_{\text{log-det}}(\mathbf{Q}, \mathbf{W}) = \text{tr}(\mathbf{Q}\mathbf{W}^{-1}) - \log\left|\mathbf{Q}\mathbf{W}^{-1}\right| - M \;, \tag{35}$$

where the "log-det" Bregman matrix divergence in (35) is an special case of Bregman matrix divergence [71], where $-\log|\cdot|$ is selected as a strictly convex function. By substituding $\mathbf{C_y}$ and $\mathbf{\Sigma_y}$ in (35) instead of $\mathbf{Q}$ and $\mathbf{W}$,

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the *log-det* Bregman matrix divergence can be written as follows [19], [22], [70], [79]–[83]:

$$\begin{aligned}\mathcal{D}_{\text{log-det}}(\mathbf{C}_{\mathbf{y}},\mathbf{\Sigma}_{\mathbf{y}}) &= \operatorname{tr}(\mathbf{C}_{\mathbf{y}}\mathbf{\Sigma}_{\mathbf{y}}^{-1}) - \log\left|\mathbf{C}_{\mathbf{y}}\mathbf{\Sigma}_{\mathbf{y}}^{-1}\right| - M \\ &= \operatorname{tr}(\mathbf{C}_{\mathbf{y}}\mathbf{\Sigma}_{\mathbf{y}}^{-1}) + \log|\mathbf{\Sigma}_{\mathbf{y}}| - \log|\mathbf{C}_{\mathbf{y}}| - M \\ &= \log|\mathbf{\Sigma}_{\mathbf{y}}| + \operatorname{tr}(\mathbf{C}_{\mathbf{y}}\mathbf{\Sigma}_{\mathbf{y}}^{-1}) + \text{const}\ , \qquad (36)\end{aligned}$$

where (36) is the same as (13) up to a constant.

B. Proof of Theorem 1

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. □

C. Proof of Proposition 3

Proof. We first show that the objective function of the M-step is derived by upper-bounding the negative log-likelihood, $-\log p(\mathbf{Y}|\boldsymbol{\gamma})$, using Jensen's inequality (J):

$$\begin{aligned}-\log p(\mathbf{Y}|\boldsymbol{\gamma}) &= -\log \mathrm{E}_{p(\mathbf{X}|\boldsymbol{\gamma})} p(\mathbf{Y}|\mathbf{X},\boldsymbol{\gamma}) = -\log \mathrm{E}_{p(\mathbf{X}|\boldsymbol{\gamma})}\left(\frac{p\left(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k\right)p(\mathbf{Y}|\mathbf{X},\boldsymbol{\gamma})}{p\left(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k\right)}\right) \\ &\overset{\text{(I)}}{=} -\log \mathrm{E}_{p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k)}\left(\frac{p(\mathbf{Y}|\mathbf{X},\boldsymbol{\gamma})}{p\left(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k\right)}p(\mathbf{X}|\boldsymbol{\gamma})\right) \\ &\overset{\text{(J)}}{\leq} -\mathrm{E}_{p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k)}\log\left(\frac{p(\mathbf{Y}|\mathbf{X},\boldsymbol{\gamma})}{p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k)}p(\mathbf{X}|\boldsymbol{\gamma})\right) \\ &\overset{\text{(II)}}{=} -\mathrm{E}_{p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k)}\log p(\mathbf{Y},\mathbf{X}|\boldsymbol{\gamma}) + \underbrace{\mathrm{E}_{p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k)}\log p\left(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k\right)}_{\text{const}} \\ &:= \mathcal{L}_{\text{EM}}^k(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k)\ . \qquad (37)\end{aligned}$$

The resulting bound is a majorizing function for $-\log p(\mathbf{Y}|\boldsymbol{\gamma})$, so that condition [A2] holds. Note that the term $\mathrm{E}_{p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k)}p\left(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k\right)$ does not depend on $\boldsymbol{\gamma}$ and, therefore, does not influence the optimization. According to the definition of Jensen's inequality, the equality constraint – condition [A1] – holds if and only if the argument of the convex function is a constant. Therefore, to establish the equivalence of both sides of (J) when $\gamma = \gamma^k$, it is sufficient to show that the argument of the log function, $\frac{p(\mathbf{Y}|\mathbf{X},\boldsymbol{\gamma})}{p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k)}p(\mathbf{X}|\boldsymbol{\gamma})$, is constant when $\gamma = \gamma^k$. This can be verified by invoking Bayes rule:

$$\frac{p(\mathbf{Y}|\mathbf{X},\boldsymbol{\gamma}^k)}{p(\mathbf{X}|\mathbf{Y},\boldsymbol{\gamma}^k)}p(\mathbf{X}|\boldsymbol{\gamma}^k) = p(\mathbf{Y}|\boldsymbol{\gamma}^k)\ .$$

Since $p(\mathbf{Y}|\boldsymbol{\gamma}^k)$ is a constant, equality condition [A1] holds.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

After inserting the analytic form of $-\log p(\mathbf{Y},\mathbf{X}|\gamma)$ in Eq. (24):

$$-\log p(\mathbf{Y},\mathbf{X}|\gamma) = \frac{T}{2}\log|\mathbf{\Gamma}| + \frac{1}{2}\sum_{t=1}^{T}\mathbf{x}(t)^{\top}\mathbf{\Gamma}^{-1}\mathbf{x}(t) + \frac{T}{2}\log\left|2\sigma^2\mathbf{I}\right| + \sum_{t=1}^{T}\frac{1}{\sigma^2}||\mathbf{y}(t)-\mathbf{L}\mathbf{x}(t)||_2^2 \ ,$$

we are ready to prove that $\mathcal{L}_{\mathrm{EM}}^{k}(\gamma|\gamma^k)$ fulfills condition [A3]. We have:

$$\mathcal{L}_{\mathrm{EM}}^{k}(\gamma|\gamma^k) \propto \log|\mathbf{\Gamma}| + \mathrm{E}_{p(\mathbf{X}|\mathbf{Y},\gamma^k)}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}(t)^{\top}\mathbf{\Gamma}^{-1}\mathbf{x}(t)\right] + \mathrm{const} \ , \tag{38}$$

where const comprises all terms of Eq. (24) that are not a function of γ. To prove that $\mathcal{L}_{\mathrm{EM}}^{k}(\gamma|\gamma^k)$ satisfies condition [A3], we need to show that $\mathcal{L}_{\mathrm{EM}}^{k}(\gamma|\gamma^k)$ reaches to its global minimum in each MM iteration. This can be easily guaranteed if Eq. (38) is convex. While the second term in (38) is convex, the first term, $\log|\mathbf{\Gamma}|$, is in fact concave, which hampers conclusions concerning the convexity of their sum. However, we can use the concept of *geodesic convexity* or *g-convexity* from non-Euclidean and geometric optimization, which enables us to prove that any local minimum of Eq. (38) is actually a global minimum. For the sake of brevity, we will omit a detailed theoretical introduction of g-convexity, and only borrow the following required propositions, Propositions 8 and 10, from the literature (an interested reader can refer to [84, Chapter 1] for a gentle introduction to this topic, and to [107, Chapter 2] [108]–[114] for more in-depth technical details). Now, we state the following preliminary results:

Proposition 8. *The function* $\log|\mathbf{\Gamma}|$ *is g-convex in* $\mathbf{\Gamma}$*, where* $\mathbf{\Gamma}$ *belongs to the manifold of positive definite (PD) matrices.*

Proof. A detailed proof can be found in [84, Lemma. 1.13]. The main idea is to leverage the geodesic $\mathbf{Q}_q = \mathbf{V}\mathbf{D}^q\mathbf{V}^{\top}$, $q \in [0,1]$ between two matrices, $\mathbf{Q}_0 = \mathbf{V}\mathbf{V}^{\top}$ and $\mathbf{Q}_1 = \mathbf{V}\mathbf{D}\mathbf{V}^{\top}$, in order to transfer the problem into the following form:

$$f(\mathbf{Q}_q) = \log|\mathbf{V}\mathbf{D}^q\mathbf{V}^{\top}| = 2\log|\mathbf{V}| + q\log|\mathbf{D}| \ ,$$

where $f(\mathbf{Q}_q)$ is a linear function and, therefore, convex in q. □

Remark 9. *The log-determinant function is concave in classical Euclidean analysis. However, Proposition 8 demonstrates that it is g-convex with respect to the PD manifold.*

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 [108, Theorem 2.1]. □

Given that g-convexity is an extension of classical convexity to non-Euclidean geometry, it is straightforward to show that all convex functions are also g-convex, where the geodesics between pairs of matrices are simply line segments. Therefore, given Proposition 8, we can conclude that Eq. (38) is g-convex; hence, any local minimum of $\mathcal{L}_{\mathrm{EM}}^{k}(\gamma|\gamma^k)$ is a global minimum according to Proposition 10. This proves that condition [A3] is fulfilled and completes the proof of Proposition 3. □

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

D. *Proof of Proposition 4*

Proof. We start by recalling $\mathcal{R}^{\text{II}-x}(\mathbf{X}, \boldsymbol{\gamma})$ in Eq. (14):

$$\mathcal{R}^{\text{II}-x}(\mathbf{X}, \boldsymbol{\gamma}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\frac{x_n(t)^2}{\gamma_n} + \log|\boldsymbol{\Sigma}_{\mathbf{y}}| \ .$$

Based on [13, Example 2], [A2] can be directly inferred from the concavity of the log-determinant function and its first-order Taylor expansion around the value from the previous iteration, $\boldsymbol{\Sigma}_{\mathbf{y}}^k$, which leads to the following inequality:

$$\begin{aligned}\log|\boldsymbol{\Sigma}_{\mathbf{y}}| &\leq \log\left|\boldsymbol{\Sigma}_{\mathbf{y}}^k\right| + \operatorname{tr}\left[\left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1}\left(\boldsymbol{\Sigma}_{\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{y}}^k\right)\right] \\ &= \log\left|\boldsymbol{\Sigma}_{\mathbf{y}}^k\right| + \operatorname{tr}\left[\left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{y}}\right] - \operatorname{tr}\left[\left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{y}}^k\right] \ .\end{aligned} \tag{39}$$

Note that the first and last term in (39) do not depend on $\boldsymbol{\gamma}$; hence, they can be ignored in the optimization procedure. Conditions [A1] and [A4] are automatically satisfied by construction because the majorizing function is obtained through a Taylor expansion around $\boldsymbol{\Sigma}_{\mathbf{y}}^k$. Concretely, [A1] is satisfied because the equality in Eq. (39) holds for $\boldsymbol{\Sigma}_{\mathbf{y}} = \boldsymbol{\Sigma}_{\mathbf{y}}^k$. Similarly, [A4] is satisfied because the gradient of $\log|\boldsymbol{\Sigma}_{\mathbf{y}}|$ at point $\boldsymbol{\Sigma}_{\mathbf{y}}^k$, $\left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1}$, defines the linear Taylor approximation $\log\left|\boldsymbol{\Sigma}_{\mathbf{y}}^k\right| + \operatorname{tr}\left[\left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1}\left(\boldsymbol{\Sigma}_{\mathbf{y}} - \boldsymbol{\Sigma}_{\mathbf{y}}^k\right)\right]$. Thus, both gradients coincide in $\boldsymbol{\Sigma}_{\mathbf{y}}^k$ by construction. Now, we show that [A3] can be satisfied easily using standard optimization algorithms by proving that $\mathcal{R}_{\text{conv}}^k(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k)$ is a convex function with respect to $\boldsymbol{\gamma}$. To this end, we rewrite Eq. (26):

$$\mathcal{R}_{\text{conv}}^k(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k) = \left[\frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N}\frac{x_n^k(t)^2}{\gamma_n}\right] + \log\left|\boldsymbol{\Sigma}_{\mathbf{y}}^k\right| + \operatorname{tr}\left[\left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{y}}\right] - \operatorname{tr}\left[\left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{y}}^k\right] \ ,$$

as follows:

$$\mathcal{R}_{\text{conv}}^k(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k) = \operatorname{diag}\left[\mathbf{U}\right]\boldsymbol{\gamma}^{-1} + \operatorname{diag}\left[\mathbf{V}\right]\boldsymbol{\gamma} + \text{const} \ , \tag{40}$$

where $\mathbf{U} := \frac{1}{T}\sum_{t=1}^{T}\left[\mathbf{x}^k(t)^\top\mathbf{x}^k(t)\right]$ and $\mathbf{V} := \mathbf{L}^\top\left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1}\mathbf{L}$ are defined as parameters that do not depend on $\boldsymbol{\gamma}$. The term const also collects constant terms in (39), i.e. $\text{const} := \log\left|\boldsymbol{\Sigma}_{\mathbf{y}}^k\right| + \sigma^2\operatorname{tr}\left[(\boldsymbol{\Sigma}_{\mathbf{y}}^k)^{-1}\right] - M$. Besides, $\boldsymbol{\gamma}^{-1} = [\gamma_1^{-1}, \ldots, \gamma_N^{-1}]^\top$ is defined as the element-wise inversion of $\boldsymbol{\gamma}$. The convexity of $\mathcal{R}_{\text{conv}}^k(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k)$ can be directly inferred from the convexity of $\operatorname{diag}\left[\mathbf{U}\right]\boldsymbol{\gamma}^{-1}$ and $\operatorname{diag}\left[\mathbf{V}\right]\boldsymbol{\gamma}$ with respect to $\boldsymbol{\gamma}$ [115, Chapter. 3]. The convexity of $\mathcal{R}_{\text{conv}}^k(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k)$, which ensures that condition [A3] can be satisfied using standard optimization, along with fulfillment of conditions [A1], [A2] and [A4], ensure that Theorem 2 holds.

In order to establish the equivalence of the MM algorithm using the majorization function Eq. (26) and the convex-bounding based Champagne variant presented in Section II-C2, we here decompose $\boldsymbol{\Sigma}_{\mathbf{y}}$ into rank-one matrices as introduced in [116]. The first term of Eq. (26) can be reformulated as follows:

$$\begin{aligned}\operatorname{tr}\left[\left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{y}}\right] &= \operatorname{tr}\left[\left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1}\left(\sigma^2\mathbf{I} + \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^\top\right)\right] \\ &= \operatorname{tr}\left[\left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1}\tilde{\mathbf{L}}\tilde{\boldsymbol{\Gamma}}\tilde{\mathbf{L}}^\top\right] = \operatorname{diag}\left[\tilde{\mathbf{L}}^\top\left(\boldsymbol{\Sigma}_{\mathbf{y}}^k\right)^{-1}\tilde{\mathbf{L}}\right]^\top\tilde{\boldsymbol{\gamma}} \ ,\end{aligned} \tag{41}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where $\tilde{\mathbf{\Gamma}} = \text{diag}(\gamma_1, \ldots, \gamma_N, \sigma^2, \ldots, \sigma^2)$, and $\tilde{\mathbf{L}} = [\mathbf{L}, \mathbf{I}]$. Since we are optimizing Eq. (26) with respect to γ_n, for $n = 1, \ldots, N$, the elements of $\tilde{\mathbf{\Gamma}}$ and $\tilde{\mathbf{L}}$ related to the sensor noise σ^2 vanish. Thus, by inserting Eq. (41) into Eq. (26), taking the derivative with respect to γ_n, for $n = 1, \ldots, N$, and setting it to zero,

$$\begin{aligned}\frac{\partial}{\partial \gamma_n}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{x}_n^k(t)\right)^2 \gamma_n^{-1} + \left[\mathbf{L}_n^\top \left(\mathbf{\Sigma}_\mathbf{y}^k\right)^{-1}\mathbf{L}_n\right]\gamma_n\right)\\ = -\frac{1}{(\gamma_n)^2}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{x}_n^k(t)\right)^2\right) + \left[\mathbf{L}_n^\top \left(\mathbf{\Sigma}_\mathbf{y}^k\right)^{-1}\mathbf{L}_n\right]\\ = 0 \quad \text{for } n = 1, \ldots, N \;,\end{aligned}$$

where $\mathbf{L}_n$ denotes the n-th column of the lead field matrix, we obtain an update rule in terms of the original variables $\mathbf{\Gamma}$ and $\mathbf{L}$:

$$\gamma_n^{k+1} := \sqrt{\frac{\frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_n^k(t))^2}{\mathbf{L}_n^\top \left(\mathbf{\Sigma}_\mathbf{y}^k\right)^{-1}\mathbf{L}_n}}\;, \tag{42}$$

which is identical to the update rule of the convex-bounding based approach discussed in Section II-C2, Eqs. (17)–(19) with $[x_1^k(t), \ldots, x_N^k(t)]^\top = \boldsymbol{\mu}_\mathbf{x}(t)$. □

E. Proof of Proposition 5

Proof. The proof that conditions [A1]–[A4] are satisfied is directly analogous to that of Proposition 4; therefore, it is omitted here. The equivalence of the Champagne variant based on MacKay updates [9, Section III.A-2] presented in Section II-C3 and the solution derived within the MM framework can be derived by transforming the update rule Eq. (42) into a fixed-point iteration of the form $\boldsymbol{\gamma}^{k+1} = f(\boldsymbol{\gamma}^k)$, which is an alternative way of minimizing the same surrogate function (Eq. (26)). By squaring the left and right hand sides of Eq. (42), one can divide both sides by γ_n^{k+1} and re-interpret the term on the right hand side as the estimate from the previous (k-th) iteration:

$$\gamma_n^{k+1} := \gamma_n^k\left[\frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_n(t))^2\right]\left(\mathbf{L}_n^\top \left(\mathbf{\Sigma}_\mathbf{y}^k\right)^{-1}\mathbf{L}_n\right)^{-1} \tag{43}$$

for $n = 1, \ldots, N$. This is indeed identical to the MacKay update in Eq. (22), which concludes the proof. □

F. Proof of Proposition 6

Proof. (following [16, Appendix C-A]) Without loss of generality, we here consider the case $\sigma^2 = 1$, which can be obtained by normalizing the sensor and source covariance matrices by σ^2: $\mathbf{\Gamma} \leftarrow {}^{\mathbf{\Gamma}}/_{\sigma^2}$, $\mathbf{\Sigma}_\mathbf{y} \leftarrow {}^{\mathbf{\Sigma}_\mathbf{y}}/_{\sigma^2} = \mathbf{I} + \mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top$. Also, due to the concavity of the $\log(\cdot)$ function and by using a Taylor expansion around point a, we have:

$$\log(x) = \log a + \frac{x}{a} - 1 + \mathcal{O}(x), \forall a > 0\;. \tag{44}$$

Assuming that $\mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top$ has an eigenvalue decomposition $\mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top = \mathbf{U}\mathbf{P}\mathbf{U}^\top$ with $\mathbf{P} = \text{diag}(p_1, \ldots, p_M)$, the majorizing function $\mathcal{L}_{\text{conv}}^{\text{Low-SNR}}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k)$ as well as Eq. (28) are derived as follows:

$$\log|\mathbf{\Sigma}_\mathbf{y}| = \log|\mathbf{I} + \mathbf{U}\mathbf{P}\mathbf{U}^\top| \overset{\text{(I)}}{=} \sum_{i=1}^{M}\log(1 + p_i) \overset{\text{(II)}}{=} \sum_{i=1}^{M} p_i + \mathcal{O}(p_i) = \text{tr}(\mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top) + \mathcal{O}(\text{SNR})\;, \tag{45}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where the p_i denote the diagonal elements of $\mathbf{P}$, which are equivalent to the eigenvalues of $\mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top$. The term $\mathcal{O}(p_i)$ represents the second and higher-order residuals of the Taylor expansion. Note that (45)-(I) is obtained by expanding $\mathbf{P}$ over its diagonal elements, while (45)-(II) is derived by exploiting the concavity of the $\log(.)$ function and its first-order Taylor expansion around $a = 1$ based on Eq. (44). Given the eigenvalue decomposition of $\mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top = \mathbf{U}\mathbf{P}\mathbf{U}^\top$ and the normalization with respect to the noise variance, the sum over all sensor-space variance components represents the ratio between the power of the signal and the power of the noise; hence, one can replace $\sum_{i=1}^{M} \mathcal{O}(p_i)$ in Eq. (45) with $\mathcal{O}(\text{SNR})$. As we have shown that $\log|\mathbf{\Sigma}_\mathbf{y}| = \text{tr}(\mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top) + \mathcal{O}(\text{SNR})$, condition [A2] holds and $\mathcal{L}^{\text{II}}(\boldsymbol{\gamma})$ converges to $\mathcal{L}^{\text{Low-SNR}}_{\text{conv}}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k)$ when SNR $\to 0$. Moreover, as Eq. (28) is constructed using a linear Taylor approximation, [A1] and [A4] hold due to the same arguments made in the proof of Proposition 4. It remains to be shown that condition [A3] can be easily fulfilled due to the convexity of $\mathcal{L}^{\text{Low-SNR}}_{\text{conv}}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k)$. To this end, we exploit the following key relationship between the sensor and source space covariances:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{y}(t)^\top \mathbf{\Sigma}_\mathbf{y}^{-1} \mathbf{y}(t) = \frac{1}{T}\sum_{t=1}^{T} \left[\frac{1}{\lambda}||\mathbf{y}(t) - \mathbf{L}\mathbf{x}^k(t)||_2^2 + \mathbf{x}^k(t)^\top \mathbf{\Gamma}^{-1}\mathbf{x}^k(t)\right] . \tag{46}$$

By replacing $\frac{1}{T}\sum_{t=1}^{T} \mathbf{y}(t)^\top \mathbf{\Sigma}_\mathbf{y}^{-1} \mathbf{y}(t)$ in Eq. (27) with its source space equivalence in (46), we have:

$$\mathcal{L}^{\text{Low-SNR}}_{\text{conv}}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k) = \text{tr}(\mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top) + \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}^k(t)^\top \mathbf{\Gamma}^{-1}\mathbf{x}^k(t) + \text{const} , \tag{47}$$

where const denotes the terms that do not depend on $\boldsymbol{\gamma}$. Reformulating (47) as

$$\mathcal{L}^{\text{Low-SNR}}_{\text{conv}}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k) = \text{diag}\left[\mathbf{W}\right]\boldsymbol{\gamma} + \text{diag}\left[\mathbf{Q}\right]\boldsymbol{\gamma}^{-1} + \text{const} ,$$

with $\mathbf{W} := \mathbf{L}^\top\mathbf{L}$, $\mathbf{Q} := \frac{1}{T}\sum_{t=1}^{T}\left[\mathbf{x}^k(t)^\top\mathbf{x}^k(t)\right]$ and $\boldsymbol{\gamma}^{-1} = [\gamma_1^{-1}, \dots, \gamma_N^{-1}]^\top$ proves the convexity of $\mathcal{L}^{\text{Low-SNR}}_{\text{conv}}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k)$ using the same arguments made for proving convexity in Proposition 4. Thus, we have shown that conditions [A1]–[A4] hold, which concludes the proof. □

G. Detailed Derivation of the LowSNR-BSI Algorithm

To find the optimal value of $\boldsymbol{\gamma} = [\gamma_1, \dots, \gamma_N]^\top$, we take the derivative of $\mathcal{L}^{\text{Low-SNR}}_{\text{conv}}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k)$ in (27) with respect to each γ_n for $n = 1, \dots, N$:

$$\begin{aligned}
\frac{\partial}{\partial\gamma_n}\mathcal{L}^{\text{Low-SNR}}_{\text{conv}}(\boldsymbol{\gamma}|\boldsymbol{\gamma}^k) &= \frac{\partial}{\partial\gamma_n}\left[\text{tr}(\mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top) + \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}(t)^\top\mathbf{\Sigma}_\mathbf{y}^{-1}\mathbf{y}(t)\right] \\
&\overset{\text{(I)}}{=} \frac{\partial}{\partial\gamma_n}\left[\sum_{n=1}^{N}\gamma_n\mathbf{L}_n^\top\mathbf{L}_n\right] + \frac{\partial}{\partial\gamma_n}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}(t)^\top\mathbf{\Sigma}_\mathbf{y}^{-1}\mathbf{y}(t)\right] \\
&\overset{\text{(II)}}{=} \mathbf{L}_n^\top\mathbf{L}_n + \frac{\partial}{\partial\gamma_n}\sum_{t=1}^{T}\frac{1}{T}\left[\frac{1}{\sigma^2}||\mathbf{y}(t) - \mathbf{L}\mathbf{x}^k(t)||_2^2 + \mathbf{x}^k(t)^\top\mathbf{\Gamma}^{-1}\mathbf{x}^k(t)\right] \\
&\overset{\text{(III)}}{=} \mathbf{L}_n^\top\mathbf{L}_n + \left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}^k(t)^\top\left(\frac{\partial}{\partial\gamma_n}\mathbf{\Gamma}^{-1}\right)\mathbf{x}^k(t)\right] \\
&= \mathbf{L}_n^\top\mathbf{L}_n + \left(-\frac{1}{\gamma_n^2}\right)\left[\frac{1}{T}\sum_{t=1}^{T}((\mathbf{x}_n^k(t))^2)\right] ,
\end{aligned} \tag{48}$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where Eq. (48)-I is derived based on a *sum-of-rank-one matrices* reformulation of the term tr$(\mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top)$ by exploiting the diagonal structure of $\mathbf{\Gamma}$. Equality (48)-II is the direct implication of the duality between $\boldsymbol{\gamma}$-space and $\mathbf{X}$-space that has been pointed out in (14). Finally, $\frac{1}{\sigma^2}||\mathbf{y}(t) - \mathbf{L}\mathbf{x}^k(t)||_2^2$ does not appear in (48)-III and is ignored since it does not depend on $\boldsymbol{\gamma}$. Setting the derivative in Eq. (48) to zero yields the following closed-form update for $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_N]^\top$:

$$\gamma_n^{k+1} := \sqrt{\frac{\frac{1}{T}\sum_{t=1}^{T}((\mathbf{x}_n^k(t))^2}{\mathbf{L}_n^\top \mathbf{L}_n}} \text{ for } n = 1, \ldots, N \; ,$$

which is identical to the update rule in Eq. (29) with $[x_1^k(t), \ldots, x_N^k(t)]^\top = \boldsymbol{\mu}_{\mathbf{x}}(t)$. This completes the derivation of the LowSNR-BSI algorithm.

H. Proof of Theorem 7

Proof. We start by taking the derivative of $\mathcal{L}^{\text{II}}(\lambda)$ with respect to λ:

$$\frac{\partial}{\partial\lambda}\mathcal{L}^{\text{II}}(\lambda) = \frac{\partial}{\partial\lambda}\left(\log|\mathbf{\Sigma}_\mathbf{y}|\right) + \frac{\partial}{\partial\lambda}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}(t)^\top\mathbf{\Sigma}_\mathbf{y}^{-1}\mathbf{y}(t)\right] \; . \tag{49}$$

We first calculate the first term, $\frac{\partial}{\partial\lambda}\left(\log|\mathbf{\Sigma}_\mathbf{y}|\right)$. Using the matrix inversion equality

$$\log|\mathbf{\Sigma}_\mathbf{y}| = \log|\lambda\mathbf{I} + \mathbf{L}\mathbf{\Gamma}\mathbf{L}^\top| = \log|\frac{1}{\lambda}\mathbf{L}^\top\mathbf{L} + \mathbf{\Gamma}^{-1}| + \log|\mathbf{\Gamma}| + \log|\lambda\mathbf{I}| \; ,$$

we have

$$\frac{\partial}{\partial\lambda}\left(\log|\mathbf{\Sigma}_\mathbf{y}|\right) = \frac{\partial}{\partial\lambda}M\log\lambda + \log|\mathbf{\Sigma}_\mathbf{x}^{-1}| \; ,$$

where the term $\log|\mathbf{\Gamma}|$ is omitted since it is does not depend on λ. Then, the derivative of $\log|\mathbf{\Sigma}_\mathbf{y}|$ with respect to λ can be obtained as follows:

$$\frac{\partial}{\partial\lambda}\left(\log|\mathbf{\Sigma}_\mathbf{y}|\right) = \frac{M}{\lambda} - \left(\frac{1}{\lambda^2}\right)\text{tr}\left[\mathbf{\Sigma}_\mathbf{x}\mathbf{L}^\top\mathbf{L}\right] \; , \tag{50}$$

where the second term in (50) is derived according to the equality $\mathbf{\Sigma}_\mathbf{x}^{-1} = (\mathbf{\Gamma}^{-1} + \frac{1}{\lambda}\mathbf{L}^\top\mathbf{L})$, which holds for the inverse of the posterior covariance in Eq. (10) [39, Chapter 4]:

$$\begin{aligned}\frac{\partial}{\partial\lambda}\left(\log|\mathbf{\Sigma}_\mathbf{x}^{-1}|\right) &= \text{tr}\left[\mathbf{\Sigma}_\mathbf{x}\frac{\partial}{\partial\lambda}\mathbf{\Sigma}_\mathbf{x}^{-1}\right] = \text{tr}\left[\mathbf{\Sigma}_\mathbf{x}\frac{\partial}{\partial\lambda}\left(\mathbf{\Gamma}^{-1} + \frac{1}{\lambda}\mathbf{L}^\top\mathbf{L}\right)\right] \\ &= \text{tr}\left[\mathbf{\Sigma}_\mathbf{x}\frac{\partial}{\partial\lambda}\left(\frac{1}{\lambda}\mathbf{L}^\top\mathbf{L}\right)\right] = -\left(\frac{1}{\lambda^2}\right)\text{tr}\left[\mathbf{\Sigma}_\mathbf{x}\mathbf{L}^\top\mathbf{L}\right] \; .\end{aligned}$$

In the next step, we calculate the derivative of the second term in Eq. (49) using the following key relation between the sensor and source space covariances presented in Appendix F. Given (46), we have

$$\frac{\partial}{\partial\lambda}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}(t)^\top\mathbf{\Sigma}_\mathbf{y}^{-1}\mathbf{y}(t)\right] = \left(-\frac{1}{\lambda^2}\right)\frac{1}{T}\sum_{t=1}^{T}||\mathbf{y}(t) - \mathbf{L}\mathbf{x}^k(t)||^2 \; , \tag{51}$$

where the term $\mathbf{x}^k(t)^\top\mathbf{\Gamma}^{-1}\mathbf{x}^k(t)$ is neglected since it does not depend on λ. Let $\mathbf{\Gamma}^k$ and $\mathbf{\Sigma}_\mathbf{x}^k$ be fixed values obtained in the (k)-th iteration. Then, by substituting Eqs. (50) and (51) into Eq. (49) and setting the derivative to zero, the update rule for λ at the $(k+1)$-th iteration is obtained as

$$\lambda^{k+1} := \frac{\frac{1}{T}\sum_{t=1}^{T}||\mathbf{y}(t) - \mathbf{L}\mathbf{x}^k(t)||^2}{M - \text{tr}\left[\mathbf{I}_N\right] + \text{tr}\left[(\mathbf{\Sigma}_\mathbf{x}^k)^{-1}(\mathbf{\Gamma}^k)^{-1}\right]} \; ,$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where the denominator is expressed in terms of the values at the k-th iteration according to the following matrix equality [63]:

$$\mathrm{tr}\left[\boldsymbol{\Sigma}_{\mathbf{x}}^{k}\mathbf{L}^{\top}\mathbf{L}\right] = \mathrm{tr}[\boldsymbol{\Sigma}_{\mathbf{x}}^{k}\lambda^{k}\left((\boldsymbol{\Sigma}_{\mathbf{x}}^{k})^{-1} - (\boldsymbol{\Gamma}^{k})^{-1}\right)] = \mathrm{tr}[\lambda^{k}\mathbf{I}_{N}] - \mathrm{tr}\left[\lambda^{k}(\boldsymbol{\Sigma}_{\mathbf{x}}^{k})^{-1}(\boldsymbol{\Gamma}^{k})^{-1}\right] \ .$$

This completes the proof. □

Acknowledgment

This result is part of a project that has received funding from the European Research Council (ERC) under the European Unions 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 Berlin International Graduate School in Model and Simulation based Research (BIMoS), the Berlin Mathematical School (BMS), and the Berlin Mathematics Research Center MATH+ for partial support. GK acknowledges partial support by the Bundesministerium fur Bildung und Forschung (BMBF) through the Berliner Zentrum for Machine Learning (BZML), Project AP4, RTG DAEDALUS (RTG 2433), Projects P1 and P3, RTG BIOQIC (RTG 2260), Projects P4 and P9, and by the Berlin Mathematics Research Center MATH+, Projects EF1-1 and EF1-4. KRM was supported in part by the Institute for Information & Communications Technology Promotion and funded by the Korea government (MSIT) (No. 2017-0-01779), and was partly supported by the German Ministry for Education and Research (BMBF) under Grants 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18025A and 01IS18037A; the German Research Foundation (DFG) under Grant Math+, EXC 2046/1, Project ID 390685689. SSN was funded in part by National Institutes of Health grants (R01DC004855, R01EB022717, R01DC176960, R01DC010145, R01NS100440, R01AG062196, and R01DC013979), University of California MRPI MRP-17454755, the US Department of Defense grant (W81XWH-13-1-0494).

References

[1] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, "Magnetoencephalographytheory, instrumentation, and applications to noninvasive studies of the working human brain," *Reviews of modern Physics*, vol. 65, no. 2, p. 413, 1993.

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

[3] A. Gramfort, "Mapping, timing and tracking cortical activations with MEG and EEG: Methods and application to human vision," Ph.D. dissertation, Ecole nationale supérieure des telecommunications-ENST, 2009.

[4] Y. Huang, L. C. Parra, and S. Haufe, "The New York head a precise standardized volume conductor model for EEG source localization and tES targeting," *NeuroImage*, vol. 140, pp. 150–162, 2016.

[5] K. Maksymenko, "Novel algorithmic approaches for the forward and inverse m/eeg problems," Ph.D. dissertation, Université Côte d'Azur, 2019.

[6] K. Matsuura and Y. Okabe, "Selective minimum-norm solution of the biomagnetic inverse problem," *IEEE Transactions on Biomedical Engineering*, vol. 42, no. 6, pp. 608–615, 1995.

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

[8] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, no. Jun, pp. 211–244, 2001.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[10] ——, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 317–329, 2010.

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

[12] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[13] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2017.

[14] P. Oguz-Ekim, J. P. Gomes, J. Xavier, and P. Oliveira, "Robust localization of nodes and time-recursive tracking in sensor networks using noisy range measurements," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3930–3942, 2011.

[15] R. Prasad, C. R. Murthy, and B. D. Rao, "Joint channel estimation and data detection in mimo-ofdm systems: A sparse bayesian learning approach," *IEEE Transactions on Signal Processing*, vol. 63, no. 20, pp. 5369–5382, 2015.

[16] S. Haghighatshoar and G. Caire, "Massive MIMO channel subspace estimation from low-dimensional projections," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 303–318, 2017.

[17] A. Fengler, G. Caire, P. Jung, and S. Haghighatshoar, "Massive MIMO unsourced random access," *arXiv preprint arXiv:1901.00828*, 2019.

[18] K. Shen, W. Yu, L. Zhao, and D. P. Palomar, "Optimization of mimo device-to-device networks via matrix fractional programming: A minorization–maximization approach," *IEEE/ACM Transactions on Networking*, vol. 27, no. 5, pp. 2164–2177, 2019.

[19] M. B. Khalilsarai, T. Yang, S. Haghighatshoar, and G. Caire, "Structured channel covariance estimation from limited samples in massive mimo," *arXiv preprint arXiv:1910.14467*, 2019.

[20] Y. Feng, D. P. Palomar *et al.*, "A signal processing perspective on financial engineering," *Foundations and Trends® in Signal Processing*, vol. 9, no. 1–2, pp. 1–231, 2016.

[21] K. Benidis, Y. Feng, D. P. Palomar *et al.*, "Optimization methods for financial index tracking: From theory to practice," *Foundations and Trends® in Optimization*, vol. 3, no. 3, pp. 171–279, 2018.

[22] S. Khanna and C. R. Murthy, "On the support recovery of jointly sparse Gaussian sources using sparse Bayesian learning," *arXiv preprint arXiv:1703.04930*, 2017.

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

[24] C. Cai, M. Diwakar, D. Chen, K. Sekihara, and S. S. Nagarajan, "Robust empirical Bayesian reconstruction of distributed sources for electromagnetic brain imaging," *IEEE Transactions on Medical Imaging*, vol. 39, no. 3, pp. 567–577, 2019.

[25] C. Habermehl, J. M. Steinbrink, K.-R. Müller, and S. Haufe, "Optimizing the regularization for image reconstruction of cerebral diffuse optical tomography," *Journal of Biomedical Optics*, vol. 19, no. 9, p. 096006, 2014.

[26] D. P. Wipf and B. D. Rao, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Transactions on Signal Processing*, vol. 55, no. 7, pp. 3704–3716, 2007.

[27] M. S. Hämäläinen and R. J. Ilmoniemi, "Interpreting magnetic fields of the brain: minimum norm estimates," *Medical & Biological Engineering & Computing*, vol. 32, no. 1, pp. 35–42, 1994.

[28] R. D. Pascual-Marqui, C. M. Michel, and D. Lehmann, "Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain," *International Journal of psychophysiology*, vol. 18, no. 1, pp. 49–65, 1994.

[29] R. D. Pascual-Marqui, "Discrete, 3d distributed, linear imaging methods of electric neuronal activity. part 1: exact, zero error localization," 2007.

[30] S. Haufe, V. V. Nikulin, A. Ziehe, K.-R. Müller, and G. Nolte, "Combining sparsity and rotational invariance in EEG/MEG source reconstruction," *NeuroImage*, vol. 42, no. 2, pp. 726–738, 2008.

[31] S. Haufe, R. Tomioka, T. Dickhaus, C. Sannelli, B. Blankertz, G. Nolte, and K.-R. Müller, "Large-scale EEG/MEG source localization with spatial flexibility," *NeuroImage*, vol. 54, no. 2, pp. 851–859, 2011.

[32] A. Gramfort, M. Kowalski, and M. Hämäläinen, "Mixed-norm estimates for the M/EEG inverse problem using accelerated gradient methods," *Physics in Medicine and Biology*, vol. 57, no. 7, p. 1937, 2012.

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

[34] S. Castaño-Candamil, J. Höhne, J.-D. Martínez-Vargas, X.-W. An, G. Castellanos-Domínguez, and S. Haufe, "Solving the EEG inverse problem based on space–time–frequency structured sparsity constraints," *NeuroImage*, vol. 118, pp. 598–612, 2015.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

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

[38] W. Wu, S. Nagarajan, and Z. Chen, "Bayesian machine learning: EEG\MEG signal processing measurements," *IEEE Signal Processing Magazine*, vol. 33, no. 1, pp. 14–36, 2016.

[39] K. Sekihara and S. S. Nagarajan, *Electromagnetic brain imaging: a Bayesian perspective*. Springer, 2015.

[40] W. James and C. Stein, "Estimation with quadratic loss," in *Breakthroughs in Statistics*. Springer, 1992, pp. 443–460.

[41] H. H. Bauschke and P. L. Combettes, "Fenchel–rockafellar duality," in *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2017, pp. 247–262.

[42] R. T. Rockafellar, *Convex analysis*. Princeton University Press, 1970, no. 28.

[43] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[44] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.

[45] M. W. Jacobson and J. A. Fessler, "An expanded theoretical treatment of iteration-dependent majorize-minimize algorithms," *IEEE Transactions on Image Processing*, vol. 16, no. 10, pp. 2411–2422, 2007.

[46] T. T. Wu, K. Lange *et al.*, "The MM alternative to EM," *Statistical Science*, vol. 25, no. 4, pp. 492–505, 2010.

[47] E. Mjolsness and C. Garrett, "Algebraic transformations of objective functions," *Neural Networks*, vol. 3, no. 6, pp. 651–669, 1990.

[48] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural computation*, vol. 15, no. 4, pp. 915–936, 2003.

[49] T. Lipp and S. Boyd, "Variations and extension of the convex–concave procedure," *Optimization and Engineering*, vol. 17, no. 2, pp. 263–287, 2016.

[50] D. Fagot, H. Wendt, C. Fvotte, 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. [Online]. Available: https://www.irit.fr/~Cedric.Fevotte/publications/proceedings/icassp2019a.pdf

[51] A. Hashemi and S. Haufe, "Improving EEG source localization through spatio-temporal sparse Bayesian learning," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1935–1939.

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

[53] C. J. Wu, "On the convergence properties of the EM algorithm," *The Annals of Statistics*, pp. 95–103, 1983.

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

[55] A. M. Dale, A. K. Liu, B. R. Fischl, R. L. Buckner, J. W. Belliveau, J. D. Lewine, and E. Halgren, "Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity," *Neuron*, vol. 26, no. 1, pp. 55–67, 2000.

[56] H. M. Huizenga, J. C. De Munck, L. J. Waldorp, and R. P. Grasman, "Spatiotemporal EEG/MEG source analysis based on a parametric noise covariance model," *IEEE Transactions on Biomedical Engineering*, vol. 49, no. 6, pp. 533–539, 2002.

[57] J. C. De Munck, H. M. Huizenga, L. J. Waldorp, and R. Heethaar, "Estimating stationary dipoles from MEG/EEG data contaminated with spatially and temporally correlated background noise," *IEEE Transactions on Signal Processing*, vol. 50, no. 7, pp. 1565–1572, 2002.

[58] S. C. Jun, S. M. Plis, D. M. Ranken, and D. M. Schmidt, "Spatiotemporal noise covariance estimation from limited empirical magnetoencephalographic data," *Physics in Medicine & Biology*, vol. 51, no. 21, p. 5549, 2006.

[59] S. M. Plis, D. M. Schmidt, S. C. Jun, and D. M. Ranken, "A generalized spatiotemporal covariance model for stationary background in analysis of MEG data," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2006, pp. 3680–3683.

[60] F. Bijma, J. C. De Munck, H. M. Huizenga, and R. M. Heethaar, "A mathematical approach to the temporal stationarity of background noise in MEG/EEG measurements," *NeuroImage*, vol. 20, no. 1, pp. 233–243, 2003.

[61] D. A. Engemann and A. Gramfort, "Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals," *NeuroImage*, vol. 108, pp. 328–342, 2015.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

[63] Z. Zhang and B. D. Rao, “Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 912–926, 2011.

[64] Y. Wu and D. P. Wipf, “Dual-space analysis of the sparse linear model,” in *Advances in Neural Information Processing Systems*, 2012, pp. 1745–1753.

[65] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[66] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[67] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[68] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K.-R. Müller, “Single-trial analysis and classification of ERP componentsa tutorial,” *NeuroImage*, vol. 56, no. 2, pp. 814–825, 2011.

[69] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller, “Introduction to machine learning for brain imaging,” *NeuroImage*, vol. 56, no. 2, pp. 387–399, 2011.

[70] J. Friedman, T. Hastie, and R. Tibshirani, “Sparse inverse covariance estimation with the graphical Lasso,” *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[71] L. M. Bregman, “The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming,” *USSR Computational Mathematics and Mathematical Physics*, vol. 7, no. 3, pp. 200–217, 1967.

[72] F. L. Da Silva, A. Hoeks, H. Smits, and L. Zetterberg, “Model of brain rhythmic activity,” *Kybernetik*, vol. 15, no. 1, pp. 27–37, 1974.

[73] I. Bojak, “Neural population models and cortical field theory: overview,” 2014.

[74] P. L. Nunez, R. Srinivasan *et al.*, *Electric fields of the brain: the neurophysics of EEG*. Oxford University Press, USA, 2006.

[75] 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*, vol. 112, no. 4, pp. 713–719, 2001.

[76] Y. Rubner, C. Tomasi, and L. J. Guibas, “The earth mover’s distance as a metric for image retrieval,” *International Journal of Computer Vision*, vol. 40, pp. 99–121, 2000.

[77] S. S. Dalal, J. M. Zumer, A. G. Guggisberg, M. Trumpis, D. D. Wong, K. Sekihara, and S. S. Nagarajan, “MEG/EEG source reconstruction, statistical evaluation, and visualization with NUTMEG,” *Computational Intelligence and Neuroscience*, vol. 2011, 2011.

[78] C. Chang, M. Diwakar, A. Hashemi, S. Haufe, K. Sekihara, and S. Nagarajan, “Robust estimation of noise for electromagnetic brain imaging with the Champagne algorithm,” *Preprint*, 2020.

[79] P. Ravikumar, M. J. Wainwright, G. Raskutti, B. Yu *et al.*, “High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence,” *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.

[80] R. Mazumder and T. Hastie, “The graphical lasso: New insights and alternatives,” *Electronic Journal of Statistics*, vol. 6, p. 2125, 2012.

[81] P. Zadeh, R. Hosseini, and S. Sra, “Geometric mean metric learning,” in *International Conference on Machine Learning*, 2016, pp. 2464–2471.

[82] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, “Information-theoretic metric learning,” in *Proceedings of the 24th International Conference on Machine Learning*, 2007, pp. 209–216.

[83] T. Tsiligkaridis and A. O. Hero, “Covariance estimation in high dimensions via Kronecker product expansions,” *IEEE Transactions on Signal Processing*, vol. 61, no. 21, pp. 5347–5360, 2013.

[84] A. Wiesel, T. Zhang *et al.*, “Structured robust covariance estimation,” *Foundations and Trends® in Signal Processing*, vol. 8, no. 3, pp. 127–216, 2015.

[85] A. Hashemi, C. Chang, G. Kutyniok, S. Nagarajan, K.-R. Müller, and S. Haufe, “Spatio-temporal brain source imaging using sparse Bayesian learning: Mathematical guarantees and trade-off,” *Preprint*, 2020.

[86] H. Wei, A. Jafarian, P. Zeidman, V. Litvak, A. Razi, D. Hu, and K. J. Friston, “Bayesian fusion and multimodal dcm for EEG and fMRI,” *NeuroImage*, vol. 211, p. 116595, 2020.

[87] S. Khanna and C. R. Murthy, “Rényi divergence based covariance matching pursuit of joint sparse support,” in *18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2017, pp. 1–5.

[88] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis,” *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[89] C. Févotte, "Majorization-minimization algorithm for smooth itakura-saito nonnegative matrix factorization," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 1980–1983.

[90] S. Eguchi and S. Kato, "Entropy and divergence associated with power function and the statistical application," *Entropy*, vol. 12, no. 2, pp. 262–274, 2010.

[91] A. Cichocki and S.-i. Amari, "Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, 2010.

[92] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[93] W. Samek, M. Kawanabe, and K.-R. Müller, "Divergence-based framework for common spatial patterns algorithms," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 50–72, 2013.

[94] W. Samek, D. Blythe, K.-R. Müller, and M. Kawanabe, "Robust spatial filtering with beta divergence," in *Advances in Neural Information Processing Systems*, 2013, pp. 1007–1015.

[95] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.

[96] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems*, 2013, pp. 2292–2300.

[97] A. Gramfort, G. Peyré, and M. Cuturi, "Fast optimal transport averaging of neuroimaging data," in *International Conference on Information Processing in Medical Imaging*. Springer, 2015, pp. 261–272.

[98] H. Janati, T. Bazeille, B. Thirion, M. Cuturi, and A. Gramfort, "Group level MEG/EEG source imaging via optimal transport: minimum Wasserstein estimates," in *International Conference on Information Processing in Medical Imaging*. Springer, 2019, pp. 743–754.

[99] P. Gerstoft, C. F. Mecklenbräuker, A. Xenaki, and S. Nannuru, "Multisnapshot sparse Bayesian learning for DOA," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1469–1473, 2016.

[100] B. Ottersten, P. Stoica, and R. Roy, "Covariance matching estimation techniques for array signal processing applications," *Digital Signal Processing*, vol. 8, no. 3, pp. 185–210, 1998.

[101] K. Werner, M. Jansson, and P. Stoica, "On estimation of covariance matrices with Kronecker product structure," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 478–491, 2008.

[102] K. Greenewald and A. O. Hero, "Robust kronecker product PCA for spatio-temporal covariance estimation," *IEEE Transactions on Signal Processing*, vol. 63, no. 23, pp. 6368–6378, 2015.

[103] T. Tsiligkaridis, A. O. Hero III, and S. Zhou, "On convergence of Kronecker graphical lasso algorithms," *IEEE Transactions on Signal Processing*, vol. 61, no. 7, pp. 1743–1755, 2013.

[104] A. M. Zoubir, V. Koivunen, E. Ollila, and M. Muma, *Robust statistics for signal processing*. Cambridge University Press, 2018.

[105] E. Ollila, D. P. Palomar, and F. Pascal, "Shrinking the eigenvalues of M-estimators of covariance matrix," *arXiv preprint arXiv:2006.10005*, 2020.

[106] S. Kumar, J. Ying, J. V. de Miranda Cardoso, and D. P. Palomar, "A unified framework for structured graph learning via spectral constraints." *Journal of Machine Learning Research*, vol. 21, no. 22, pp. 1–60, 2020.

[107] A. Papadopoulos, *Metric spaces, convexity and nonpositive curvature*. European Mathematical Society, 2005, vol. 6.

[108] T. Rapcsak, "Geodesic convexity in nonlinear optimization," *Journal of Optimization Theory and Applications*, vol. 69, no. 1, pp. 169–183, 1991.

[109] A. Ben-Tal, "On generalized means and generalized convex functions," *Journal of Optimization Theory and Applications*, vol. 21, no. 1, pp. 1–13, 1977.

[110] L. Liberti, "On a class of nonconvex problems where all local minima are global," *Publications de lInstitut Mathémathique*, vol. 76, no. 90, pp. 101–109, 2004.

[111] D. E. Pallaschke and S. Rolewicz, *Foundations of mathematical optimization: convex analysis without linearity*. Springer Science & Business Media, 2013, vol. 388.

[112] S. Bonnabel and R. Sepulchre, "Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1055–1070, 2009.

[113] M. Moakher, "A differential geometric approach to the geometric mean of symmetric positive-definite matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 26, no. 3, pp. 735–747, 2005.

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

bioRxiv preprint doi: https://doi.org/10.1101/2020.08.10.243774; this version posted August 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[115] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[116] Y. Sun, P. Babu, and D. P. Palomar, “Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions,” *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3576–3590.