Title

Covariance-domain Dictionary Learning for Overcomplete EEG Source Identification

Permalink

https://escholarship.org/uc/item/7d46m44r

Authors

Balkan, Ozgur
Kreutz-Delgado, Kenneth
Makeig, Scott

Publication Date

2015-11-30

Covariance-Domain Dictionary Learning for Overcomplete EEG Source Identification

Ozgur Balkan*, *Student Member, IEEE,* Kenneth Kreutz-Delgado, *Fellow, IEEE,* and Scott Makeig

Abstract—**We propose an algorithm targeting the identification of more sources than channels for electroencephalography (EEG). Our overcomplete source identification algorithm, Cov-DL, leverages dictionary learning methods applied in the covariance-domain. Assuming that EEG sources are uncorrelated within moving time-windows and the scalp mixing is linear, the forward problem can be transferred to the covariance domain which has higher dimensionality than the original EEG channel domain. This allows for learning the overcomplete mixing matrix that generates the scalp EEG even when there may be more sources than sensors active at any time segment, i.e. when there are non-sparse sources. This is contrary to straight-forward dictionary learning methods that are based on the assumption of sparsity, which is not a satisfied condition in the case of low-density EEG systems. We present two different learning strategies for Cov-DL, determined by the size of the target mixing matrix. We demonstrate that Cov-DL outperforms existing overcomplete ICA algorithms under various scenarios of EEG simulations and real EEG experiments.**

Index Terms—**Dictionary Learning, Independent Component Analysis**

I. Introduction

As a non-invasive brain imaging modality, electroencephalography (EEG) provides high temporal resolution, applicability in mobile settings, and direct measurement of electrical brain activity as opposed to other brain imaging modalities such as BOLD activity measured in fMRI. However, a major issue in EEG signal processing is that signals measured on the scalp surface do not each index a single localized cortical source of brain activity. Because of the broad point spread function of generated potentials in the brain, EEG data collected on scalp channels is a mixture of simultaneously active brain sources distributed over many different brain areas. In addition, non-brain sources such as eye and muscle movements contribute to the mixing process as well, which makes direct channel-level EEG analysis problematic. For accurate brain activity monitoring, individual sources involved in the mixture have to be identified and extracted from scalp channel data.

Because of the fact that volume conduction and mixing at the sensors is linear [1], EEG mixing can be formulated as follows

$$\mathbf{Y} = \mathbf{A}\mathbf{X} \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{M \times N_d}$ is the matrix containing collected EEG data at M sensors for N_d data points. The matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ is the unknown mixing matrix, and $\mathbf{X} \in \mathbb{R}^{N \times N_d}$ contains the activations of N sources. The i-th column of $\mathbf{A}$, denoted as $\mathbf{a_i}$, represents the relative projection weights of the i-th source to each channel. The so-called EEG inverse problem is to identify both $\mathbf{A}$ and $\mathbf{X}$, given sensor data $\mathbf{Y}$ [2]. Learning the columns of $\mathbf{A}$, namely the scalp maps, can further enable source localization in the cortex through methods such as DIPFIT [3] or sLORETA [4]. Identifying the rows of $\mathbf{X}$ can enable the computing of time-series measures such as event-related potentials (ERPs), event-related spectral perturbation (ERSPs), and spectral components [5].

A commonly applied method to solve the EEG inverse problem has been to use independent component analysis (ICA) [2], [6]. Assuming statistical independence between source activities, ICA can separate the scalp mixture into underlying source time-series $\mathbf{X}$ and identify the mixing matrix $\mathbf{A}$. It was shown in [7] that ICA methods are well suited for solving the EEG inverse problem since independence among sources was found to be positively correlated with the number of brain sources that can be extracted from data. ICA has been extensively applied on EEG for artifact rejection and source separation [8], [9] and has been shown to increase accuracy in brain-computer-interface paradigms [10]. However, one major drawback of ICA is that the number of mixed sources is assumed to be less than or equal to the number of sensors ($N \leq M$). This assumption undermines the reliability and utility of ICA, especially in low-density EEG systems ($<$ 32 number of channels).

There are multiple reasons why an EEG source identification algorithm should be able to handle more sources than sensors. A main motivation is to increase the capabilities of EEG systems to handle large number of artifacts. Depending on the experiment settings and the length of recording, the number of distinct artifact sources could possibly outnumber the brain sources or even exceed the number of channels. In those cases, ICA solution matrix is occupied by artifact sources and only a few brain sources can be extracted from data, which limits further analysis of brain activity. Even in ideal conditions, i.e, when there are no artifacts, higher resolution is desired to better capture true brain dynamics, taking into account the possibility of more than M sources being simultaneously active and/or changing brain source locations throughout the experiment.

It is also desirable to enhance the capabilities of low-density EEG devices that are becoming increasingly popular due to their relative low-cost and ease of use. Low-density EEG

This work was supported by The Swartz Foundation (Old Field, NY). O. Balkan and K. Kreutz-Delgado are with the Department of Electrical and Computer Engineering, University of California San Diego, La Jolla CA, 92093 USA e-mail: (obalkan@ucsd.edu, kreutz@eng.ucsd.edu). S. Makeig is with the UCSD Swartz Center for Computational Neuroscience email: (smakeig@ucsd.edu).

allows for a wide range of applications by facilitating EEG recording of mobile and possibly long duration experiments. However, because they are targeted for low-cost research and consumer markets, these systems usually contain about 8-19 channels for which the results of traditional ICA results would be insufficient for reliable brain source monitoring. Extracting more sources than channels may benefit low-cost clinical research and improve consumer-oriented BCI applications.

Here, we propose a covariance-domain dictionary learning algorithm, Cov-DL, that can identify more sources than number of channels for the EEG inverse problem. We note that our algorithm does not learn the explicit source time-series activity $\mathbf{X}$ but learns the overcomplete mixing matrix $\mathbf{A}$ (projection of sources to scalp sensors) and the power of individual sources in a given data segment. In this sense, our algorithm is categorically placed between blind source identification and source separation methods.

II. Related Work

An important family of blind source identification methods is comprised of cumulant-based algorithms that incorporate second order (SOBI) [11] or fourth order statistics (FOOBI) [12]. In non-EEG settings, it was shown that FOOBI can identify a number of sources that are roughly quadratic in the number of sensors [12]. However, multiple studies [7], [13] showed that cumulant-based methods perform relatively poorly in EEG source separation tasks compared to maximum likelihood based methods such as Infomax [14]. Among all methods, AMICA, an EM-based maximum likelihood ICA framework with flexible source densities, [15], performed best in terms of extracting the most number of plausible brain sources while providing the highest independence among sources [7].

An extension of traditional ICA for the overcomplete case is provided by the ICA mixture model [16], [15]. This approach learns N_{model} mixing matrices, $\mathbf{A}_i \in \mathbb{R}^{M\times M}$, instead of learning one overcomplete mixing matrix $\mathbf{A}$, in order to provide tractable computation. An adaptation of this method with AMICA, Multiple Model AMICA, was shown to be successful in identifying more sources than electrodes in some non-stationary EEG paradigms [15]. However, the mixture model has some drawbacks; because it assumes that at most M sources are active at any given time and there are only a few disjoint sets of simultaneously active sources (N_{model}). This is problematic especially when M is low. An ideal algorithm should be able to handle cases where any of $\binom{N}{k}$ sources, $1 \leq k \leq N$, can be jointly active. Our algorithm targets this case.

Another set of overcomplete ICA algorithms [17], [18] model the source estimates as $\hat{\mathbf{X}} = \mathbf{W}\mathbf{Y}$, where $\mathbf{W} \in \mathbb{R}^{N\times M}$ is a tall unmixing matrix with full column rank. These algorithms optimize $\mathbf{W}$ and return the mixing matrix as $\mathbf{A} = \mathbf{W^T}$. One of the recent algorithms of this type is RICA [18], an efficient method used for unsupervised feature learning in neural networks. We have found that in the complete mixing matrix case ($M = N$), RICA gives almost identical results with Infomax on EEG data. In this paper, we are considering the overcomplete setting for RICA and multiple model AMICA for comparison with our overcomplete method Cov-DL.

Dictionary learning-based sparse coding algorithms are closely related to overcomplete ICA methods. In the dictionary learning framework, the inverse problem is formulated as the following optimization problem,

$$\min_{A,X} \frac{1}{2}\sum_{t=1}^{N_d}\|\mathbf{Y}-\mathbf{A}\mathbf{X}\|_F^2+\lambda\sum_{t=1}^{N_d}g(\mathbf{x_t}) \tag{2}$$

where $g(\cdot)$ is a function that promotes sparsity of the source vector $\mathbf{x_t}$ at time index t and λ is the regularization parameter controlling the sparsity of the sources. Optimization is generally performed on $\mathbf{A}$ and $\mathbf{X}$ iteratively, namely learning $\mathbf{X}$ while keeping $\mathbf{A}$ fixed, and vice versa [19], [20]. Given a fixed dictionary $\hat{\mathbf{A}}$, the sources $\hat{\mathbf{X}}$ are learned by solving the following optimization

$$\min_{X} \frac{1}{2}\sum_{t=1}^{N_d}\|\mathbf{Y}-\hat{\mathbf{A}}\mathbf{X}\|_F^2+\lambda\sum_{t=1}^{N_d}g(\mathbf{x_t}) \tag{3}$$

The true dictionary can be recovered if the sources $\mathbf{x_t}$ are sparse ($k_t < M$), where k_t is the number of active sources at time t. The accuracy of recovery is strongly dependent on the level of sparsity as higher accuracy is achieved if $k \ll M$. It was shown for various dictionary learning algorithms that the performance significantly drops as k approaches M [21]. Indeed, when $k \geq M$, any full-row rank dictionary can provide a source decomposition with sparsity k and zero representation error $\|\mathbf{Y}-\mathbf{AX}\|_\mathbf{F}^\mathbf{2}$ for (2), thus the true mixing matrix becomes unrecoverable. In the case of EEG, this allows at most k = $\mathcal{O}(M)$ EEG sources to be simultaneously active which limits direct applicability of dictionary learning to low-density EEG systems.

Recently it was shown that given the true dictionary $\mathbf{A}$, and a data segment $\mathbf{Y_s} \in \mathbb{R}^{M\times L_s}$, where L_s is the length of the segment in data frames, M-SBL (multiple measurement Sparse Bayesian Learning) applied directly on $\mathbf{Y_s}$ can identify active sources under the assumption that sources are uncorrelated in the time segment [22]. The number of sources identified in this case is not limited by the number of channels M, $1 \leq k \leq M(M+1)/2$. This finding is supported by [23], where LASSO is applied on the covariance matrix off the data segment $\mathbf{Y_s}$ to obtain probability bounds on the identification of active sources. Under the assumption of uncorrelated sources $\mathbf{X_s}$, the sample-covariance matrix $\frac{1}{L_s}\mathbf{X_s}\mathbf{X_s^T}$ is assumed to be nearly diagonal ("pseudo-diagonal") and expressible as $\mathbf{\Sigma_{X_s}} = \frac{1}{L_s}\mathbf{X_s}\mathbf{X_s^T} = \mathbf{\Delta}+\mathbf{E}$, where $\mathbf{\Delta}$ is a diagonal matrix composed of diagonal entries of Σ_{X_s}. Hence in [23], $\mathbf{Y_s} = \mathbf{A}\mathbf{X_s}$ is modeled as

$$\begin{aligned}
\mathbf{Y_s}\mathbf{Y_s^T} &= \mathbf{A}\mathbf{X_s}\mathbf{X_s^T}\mathbf{A^T}\\
\mathbf{\Sigma_{Y_s}} &= \mathbf{A}\mathbf{\Sigma_{X_s}}\mathbf{A^T}\\
\mathbf{\Sigma_{Y_s}} &= \mathbf{A}\mathbf{\Delta}\mathbf{A^T}+\mathbf{E} = \sum_{\mathbf{i=1}}^{\mathbf{N}}\mathbf{\Delta_{ii}}\mathbf{a_i}\mathbf{a_i^T}+\mathbf{E}.
\end{aligned} \tag{4}$$

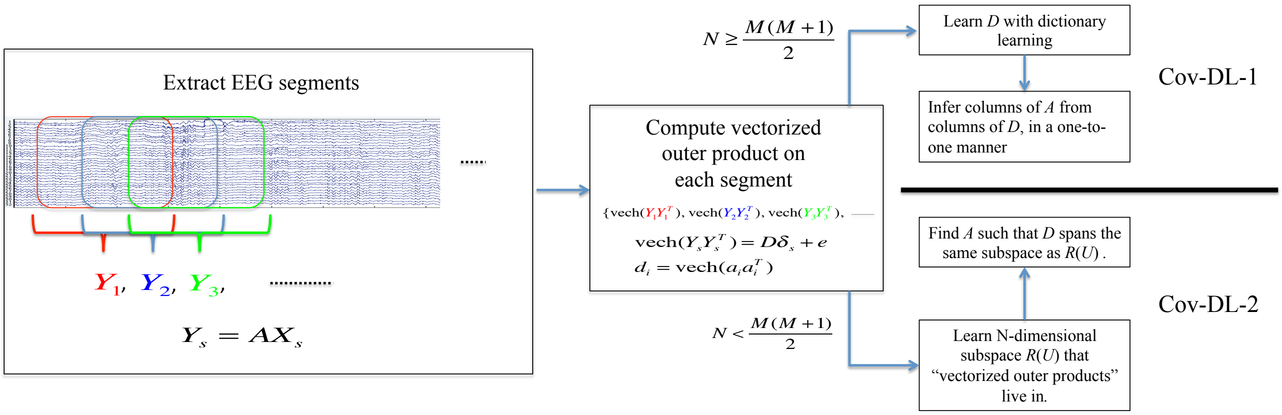

Extract EEG segments
......
Y_1, Y_2, Y_3, ...
$Y_s = AX_s$
Compute vectorized outer product on each segment
$\{\text{vech}(Y_1Y_1^T), \text{vech}(Y_2Y_2^T), \text{vech}(Y_3Y_3^T),$ ....
$\text{vech}(Y_sY_s^T) = D\delta_s + e$
$d_i = \text{vech}(a_ia_i^T)$
$N \geq \frac{M(M+1)}{2}$
Learn D with dictionary learning
Infer columns of A from columns of D, in a one-to-one manner
Cov-DL-1
$N < \frac{M(M+1)}{2}$
Find A such that D spans the same subspace as $R(U)$.
Learn N-dimensional subspace $R(U)$ that "vectorized outer products" live in.
Cov-DL-2

Fig. 1: The summary of two different strategies of Cov-DL for overcomplete EEG source identification. Cov-DL-1 involves a dictionary learning stage requiring the assumption that $k < M(M+1)/2$ sources are active at any given segment. Cov-DL-2 does not require sparsity of sources.

Since the covariance matrix is symmetric, we can vectorize the lower triangular part of both sides and obtain,

$$\begin{aligned}\text{vech}\left(\Sigma_{Y_s}\right) &= \sum_{\mathbf{i=1}}^{\mathbf{N}} \text{vech}\left(\mathbf{a_i a_i^T}\right) \mathbf{\Delta_{ii}} + \text{vech}\left(\mathbf{E}\right) \\ \text{vech}\left(\Sigma_{Y_s}\right) &= \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{d_i \Delta_{ii}} + \text{vech}\left(\mathbf{E}\right) \\ \text{vech}\left(\Sigma_{Y_s}\right) &= \mathbf{D}\delta + \text{vech}\left(\mathbf{E}\right) \qquad (5)\end{aligned}$$

where $\mathbf{D} = [\mathbf{d_1}, \mathbf{d_2}, \dots, \mathbf{d_N}]$, $\mathbf{d_i} = \text{vech}\left(\mathbf{a_i a_i^T}\right)$ and vech$(\cdot)$ is a function that maps a symmetric matrix $S \in \mathbb{R}^{M \times M}$ to its vectorized lower triangular matrix, of size $\frac{M(M+1)}{2}$. Here, we also define the inverse function vech$^{-1}(\cdot)$, which takes as an input an $\frac{M(M+1)}{2}$ dimesional vector v and outputs a symmetric matrix of size $M \times M$ whose lower triangular matrix consists of entries in v. Thus, for any vector v, we have $v = \text{vech}\left(\text{vech}^{-1}(v)\right)$.

It was shown in [23] that this formulation, together with the correlation constraint (4) can identify $\mathcal{O}(M^2)$ sources given the true dictionary. We leverage this idea to also learn the dictionary $\mathbf{A}$ from EEG data considering multiple segments from the overall recording. We also note that assumption of uncorrelated sources, albeit being a weaker constraint, is implied by the independence of sources, an assumption which was shown to be successful for EEG source separation [7].

III. Covariance-Domain Dictionary Learning (Cov-DL)

Here, we describe our covariance based dictionary learning algorithm that leverages the assumed uncorrelated nature of EEG sources. We start by segmenting the overall EEG data matrix $\mathbf{Y} \in \mathbb{R}^{M \times N_d}$, sampled with frequency S_f, into possibly overlapping segments $\mathbf{Y_s} \in \mathbb{R}^{M \times t_s S_f}$ of t_s seconds, where s denotes the index for the corresponding segment. For each segment, the following equation holds under the linear mixture model of EEG,

$$\mathbf{Y_s} = \mathbf{A X_s}, \forall s \qquad (6)$$

and thus, $\mathbf{Y_s Y_s^T} = \mathbf{A X_s X_s^T A^T}$. Then, we calculate the sample data covariance $\mathbf{\Sigma_{Y_s}} = \frac{1}{L_s}\mathbf{Y_s Y_s^T}$, for each segment s. We have,

$$\begin{aligned}\mathbf{\Sigma_{Y_s}} &= \mathbf{A \Delta_s A^T} + \mathbf{E_s} \\ \text{vech}\left(\mathbf{\Sigma_{Y_s}}\right) &= \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{\Delta_{s_{ii}}} \text{vech}\left(\mathbf{a_i a_i^T}\right) + \text{vech}\left(\mathbf{E_s}\right), \\ \text{vech}\left(\mathbf{\Sigma_{Y_s}}\right) &= \mathbf{D}\delta_\mathbf{s} + \text{vech}\left(\mathbf{E_s}\right), \forall s. \qquad (7)\end{aligned}$$

where the vector δ_s contains the diagonal entries of the source sample-covariance matrix $\mathbf{\Sigma_{X_s}} = \frac{1}{L_s}\mathbf{X_s X_s^T}$, and the matrix $\mathbf{D} \in \mathbb{R}^{M(M+1)/2 \times N}$ consists of columns $\mathbf{d_i} = \text{vech}\left(\mathbf{a_i a_i^T}\right)$. Note that, for each segment, the left hand side of the equations are obtained from data while $\mathbf{D}$ and $\delta_\mathbf{s}$ are not known. Our goal is to first learn $\mathbf{D}$ and then find the associated matrix $\mathbf{A}$. We propose two different approaches to recover $\mathbf{D}$ and $\mathbf{A}$ which depend on the relation between the target number of total sources N and the number of channels M. See Fig. 1.

A. *Overcomplete* D *(Cov-DL-1)*

When N, the number of total sources to be identified for the whole EEG session, is larger than or equal to $M(M+1)/2$, $\mathbf{D}$ in (7) is overcomplete. If we assume that at any given segment s, there are less than $M(M+1)/2$ active sources, namely $\delta_\mathbf{s}$ is sparse, then we can learn $\mathbf{D}$ by applying traditional dictionary learning methods on the set of data points $\{\text{vech}\left(\mathbf{\Sigma_{Y_s}}\right), \forall s\}$. Note that, the sparsity constraint imposed here, that is $k < M(M+1)/2$ is much weaker than the traditional sparsity constraint $k < M$ and is not necessarily violated when $k > M$.

After learning dictionary $\mathbf{D}$, we can find the mixing matrix $\mathbf{A}$ that generated $\mathbf{D}$ through the relation $\mathbf{d_i} = \text{vech}\left(\mathbf{a_i a_i^T}\right)$. For each column of the dictionary we optimize,

$$\min_{\mathbf{a_i}} \|\mathbf{d_i} - \text{vech}\left(\mathbf{a_i a_i^T}\right)\|_2^2 \qquad (8)$$

or equivalently,

$$\min_{\mathbf{a_i}} \|\text{vech}^{-1}(\mathbf{d_i}) - \mathbf{a_i}\mathbf{a_i^T}\|_F^2 \tag{9}$$

The global minimum for this optimization problem is $\hat{\mathbf{a_i}} = \sqrt{\lambda_1} b_1$, where λ_1 is the largest eigenvalue of $\text{vech}^{-1}(\mathbf{d_i})$, and b_1 is the associated eigenvector. For a visualization of the algorithm, see Fig. 2a.

B. Undercomplete $\mathbf{D}$ *(Cov-DL-2)*

When, $N < M(M+1)/2$, the data points $\{\text{vech}(\mathbf{\Sigma_{Y_s}}), \forall s\}$ live on or near a subspace of dimension N, which is spanned by the columns of $\mathbf{D}$. We denote this subspace as $\mathcal{R}(\mathbf{D})$. We can learn $\mathcal{R}(\mathbf{D})$ with methods such as Principal Component Analysis (PCA) without imposing any sparsity constraints on δ_s. However, the set of basis vectors $\mathbf{U}$ that a subspace learning algorithm, such as PCA, returns only guarantee $\mathcal{R}(\mathbf{D}) = \mathcal{R}(\mathbf{U})$, not $\mathbf{U} = \mathbf{D}$. Therefore, we can extract $\mathcal{R}(\mathbf{D})$ but there is an ambiguity about the basis vectors $\mathbf{D}$. Note, however, that we can enforce the conditions that the columns of $\mathbf{D}$ satisfy $\mathbf{d_i} = \text{vech}(\mathbf{a_i}\mathbf{a_i^T})$ and also span $\mathcal{R}(\mathbf{U})$ as closely as possible. Furthermore, since the projection operator for a given subspace is unique, namely $\mathcal{R}(\mathbf{D}) = \mathcal{R}(\mathbf{U})$ if and only if $\mathbf{D}(\mathbf{D^T}\mathbf{D})^{-1}\mathbf{D^T} = \mathbf{U}(\mathbf{U^T}\mathbf{U})^{-1}\mathbf{U^T}$, we can obtain $\mathbf{A}$ by solving the following optimization problem.

$$\begin{aligned} &\min_{\mathbf{a_i}} \|\mathbf{D}(\mathbf{D^T}\mathbf{D})^{-1}\mathbf{D^T} - \mathbf{U}(\mathbf{U^T}\mathbf{U})^{-1}\mathbf{U^T}\|_F^2 \\ &\quad \text{s.t} \quad \mathbf{d_i} = \text{vech}(\mathbf{a_i}\mathbf{a_i^T}) \end{aligned} \tag{10}$$

where U is learned through a subspace learning algorithm on data points $\{\text{vech}(\mathbf{\Sigma_{Y_s}}), \forall s\}$. We compute the above cost function's gradient w.r.t $\mathbf{A}$ using the chain rule and can minimize the cost function using quasi-Newton optimization methods. We emphasize that although $\mathbf{D}$ is not overcomplete in this case, the mixing matrix $\mathbf{A}$, which relates the cortical sources to the scalp EEG sensors, can still be complete or overcomplete. For a visualization of the algorithm, see Fig. 2b.

C. Remarks

We provide some comments about important aspects of above described algorithms. First, notice that the number of data points that Cov-DL is trained on is substantially reduced because of segmenting and learning in the covariance-domain (there is now effectively one data point per segment). For example, if t_s is 4 seconds and sampling rate is 250Hz, the total number of data points used is $\frac{1}{1000}N_d$ if the segments are non-overlapping. The number of data points for Cov-DL will increase as the overlap ratio increases. However we have found that algorithm performance does not improve when the overlap ratio of consecutive segments increases beyond 0.5. The reduced number of data points in the Cov-DL-1 framework linearly speeds up the dictionary learning computation time and makes its application to EEG feasible.

The segment length t_s is an important parameter that affects the performance of the algorithms. If the segment length t_s is short, the sample-covariance $\frac{1}{L_s}\mathbf{X_s}\mathbf{X_s^T}$ is no longer pseudo-diagonal and thus the derivation in (7) is not accurate. On the other hand, as t_s gets longer, the number of active sources in a segment increases (becomes less sparse), thus the performance of Cov-DL-1 will decrease. We have found that the choice $t_s \in [2, 4]$sec. provides a good compromise in our experiments.

We also note that for both algorithms to succeed, the power of the individual sources in segments δ_s should not stay constant throughout the recording. This is required to ensure that $\mathbf{D}$ is identifiable for algorithm Cov-DL-1 and that the data points $\mathbf{\Sigma_{Y_s}}$ obtained by $\mathbf{D}\delta_\mathbf{s}$ fill the space spanned by $\mathbf{D}$ for Cov-DL-2. This requirement holds for most EEG sources, including event-related potentials/oscillations and eye/head movement related artifact sources. To the best of our knowledge, the only EEG source that has constant power across the whole recording is electronic noise/line noise. Yet, the characteristics of this source is available (a 50Hz/60Hz sine wave) and can be filtered in the pre-processing step of EEG analysis.

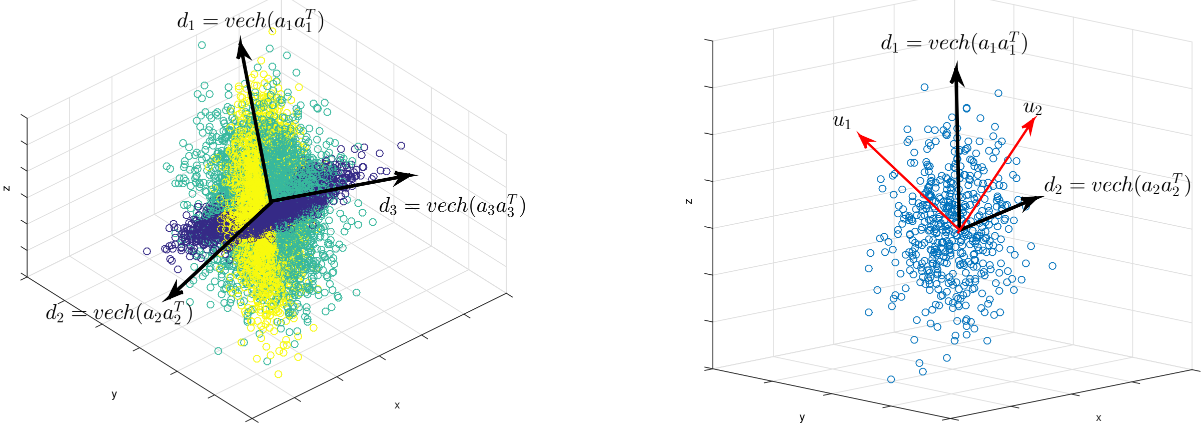
Finally we note that for algorithm Cov-DL-1, one can choose any dictionary learning algorithm for learning $\mathbf{D}$. Here, we use Bilinear Generalized Approximate Message Passing (BiGAMP-DL) [21], an EM-based bayesian dictionary learning method leveraging approximate message passing. This method has the advantage of automatically learning the sparsity level and signal-to-noise ratio (SNR). For Cov-DL-2, we have used the robust PCA method described in [24] to identify $\mathbf{U}$.

IV. Experiments

A. EEG Simulation

First we test our algorithm on three simulated data scenarios, for which we exactly know the ground truth mixing matrix $\mathbf{A}_{\text{true}}$. We simulate the placement of 32 electrodes on the scalp as shown in Fig. 3b. To generate the mixing matrix, we place dipolar sources in the brain using the Montreal Head Institute (MNI) head model. We assign random locations and random orientations for each dipole. Using the FieldTrip toolbox [3], we compute the projection weights of the i-th dipole to each channel (scalp maps) and obtain the true $\mathbf{a_i}$. See Fig. 3b. For realistic source activations $\mathbf{X}$, we generate an AR (auto-regressive) model via Source Information Flow Toolbox (SIFT) [25] under EEGLAB [26] and obtain super-Gaussian source activations of duration 66 minutes with 100Hz sampling rate. We choose a segment length $t_s = 2$sec. (200 frames) and scale the sources in each segment with a random weight uniformly assigned in the continuous interval [1,2] to model the possibly varying power dynamics of brain sources across the recording.

For the first scenario, we first test and compare algorithms for the case of a complete mixing matrix ($M = N$). We select $M = N = 32$, for an overcompleteness ratio of $N/M = 1$. We also let $k = N = 32$, so that all the sources are active in any given segment. We generate scalp EEG with $\mathbf{Y} = \mathbf{A}_{\text{true}}\mathbf{X}$ and apply Cov-DL-2 on $\mathbf{Y}$ with $t_s = 2$sec non-overlapping segments. The accuracy of the result is measured as the ratio of the number of scalp maps that are recovered (having correlation higher than 0.99 with true scalp maps) to N. We compare our algorithm with the 1-model AMICA [15] and RICA [18].



(a)

(b)

Fig. 2: A geometrical explanation of Cov-DL for $M = 2, k = 2$. (a) If $N = 3$, then $\mathbf{A} \in \mathbb{R}^{2\times3}$, and $\mathbf{D} \in \mathbb{R}^{3\times3}$. In this case $\mathbf{d_1}, \mathbf{d_2}, \mathbf{d_3}$ are identifiable with a dictionary learning algorithm applied on the data of vectorized outer products of segments. Associated $\mathbf{a_1}, \mathbf{a_2}, \mathbf{a_3}$ can then be found via solving Eg. (9) (Cov-DL-1). (b) If N = 2, then $\mathbf{D} \in \mathbb{R}^{3\times2}$, and data is not sparse since $k = N = 2$. $\mathbf{D}$ is not identifiable through learning the 2-dimensional subspace (PCA results in $\mathbf{u_1}, \mathbf{u_2}$). In this case, we solve Eq. (10) to directly find $\mathbf{A}$ such that $\mathbf{D}$ will span $\mathcal{R}(\mathbf{U})$ (Cov-DL-2).

For the second scenario, we have $M = 32, N = 64$, and overcompleteness ratio $N/M = 2$. We also let $k = N = 64$, and again all the sources are active in any given segment. We generate scalp EEG with $\mathbf{Y} = \mathbf{A}_{\text{true}}\mathbf{X}$ and apply Cov-DL-2 on $\mathbf{Y}$ with $t_s = 2$sec non-overlapping segments. We compare our algorithm with the overcomplete ICA method RICA [18] and concatenated dictionary obtained from multi-model AMICA ($N/M = 2$ models in this case) [15].

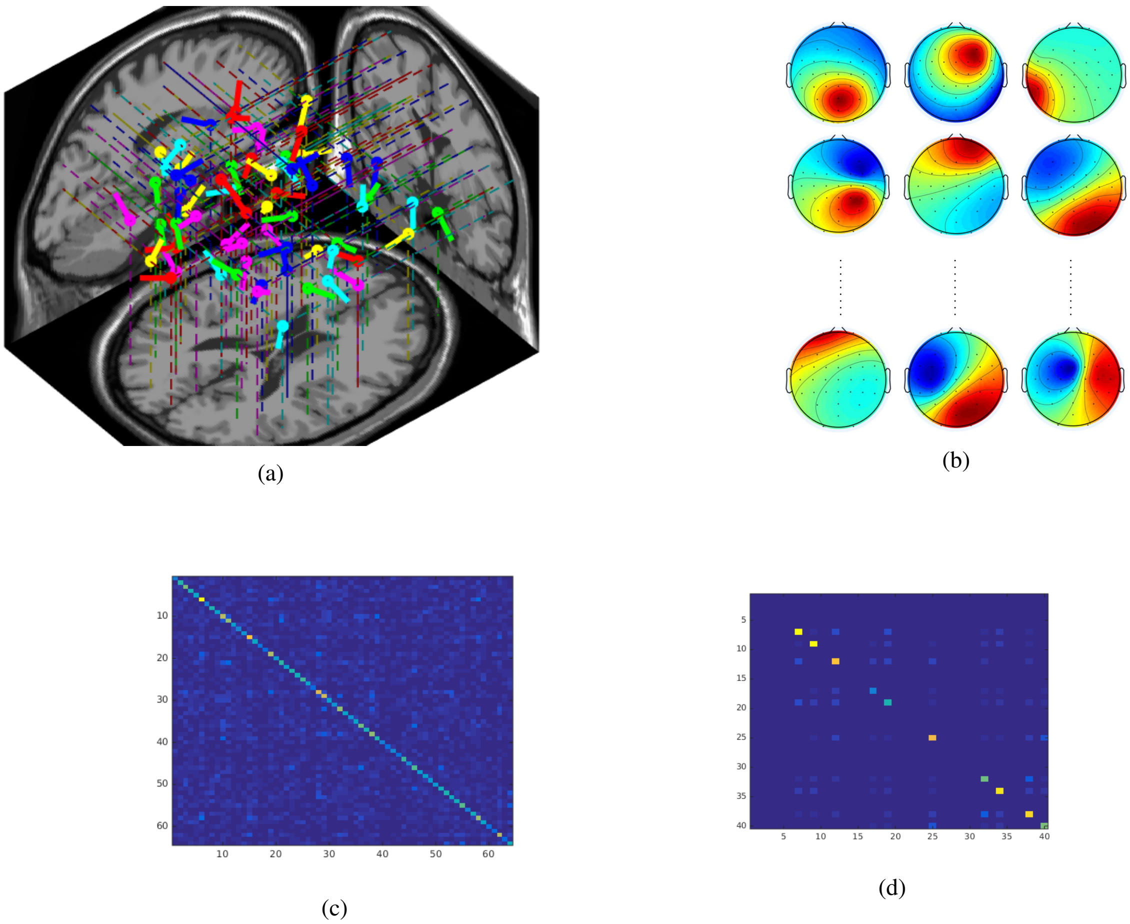
For the third scenario, we have $M = 8, N = 40, k = 10$, and overcompleteness ratio $N/M = 5$. In each segment a randomly selected k out of N sources are retained and $N - k$ sources are assigned no activation. At any given segment there are more active sources than channels ($k > M$) and the set of active sources are changing throughout the recording. We select $M = 8$ channels out of the 32 channels shown in Fig. 3b such that we uniformly cover the whole head. In this case, since $N \geq M(M+1)/2$ and $k < M(M+1)/2$, we use Cov-DL-1. We compare with the results obtained from RICA and 5-model AMICA ($N/M = 5$).

The results of three scenarios are shown in Fig. 4. It can be seen that when $M = N$, single model AMICA shows perfect source identification whereas Cov-DL performs slightly worse but still has accuracy of 0.9687 (recovers 31/32 components). This might be because fewer number of data points (number of segments) are fed to Cov-DL compared to AMICA and AMICA has the ability to model arbitrary source probability densities in an adaptive way. RICA performs the worst with an identification ratio of 0.9375 even under ideal complete conditions. This is likely due to the high coherence of the realistic mixing matrix, since other experiments showed that RICA demonstrates perfect recovery with random mixing matrices (a low coherence situation). In the overcomplete scenarios, we see that there is a drop in the performance of all algorithms. AMICA and RICA perform poorly due to their differences in modeling the overcompleteness. Multi-model AMICA considers a mixture ICA model which has only few distinct states and can handle at most M sources active at a given time. RICA fits a super-Gaussian distribution to sources obtained as $\mathbf{WY}$ where $\mathbf{W}$ is a tall unmixing matrix. However, sources derived in this form cannot be truly independent simply due to the necessary linear dependence of the rows of a tall matrix. Cov-DL is free of these drawbacks of existing ICA algorithms, and can handle more sources than sensors without requiring sparsity of any form as opposed to traditional dictionary learning algorithms which prohibit their use when $k > M$.

B. *Experiments on Real EEG*

Unlike simulated EEG, the true mixing matrix for real EEG is not known beforehand. In order to test our algorithm's performance on real EEG data, we follow the strategy proposed below.

Suppose we have an actual dataset that has M_{orig} channel recordings. After rejection of the artifact windows and contaminated channels, suppose that N channels remain. Then, we apply Extended Infomax ICA and extract N sources and their associated scalp maps. We regard these scalp maps as ground truth mixing matrix and measure how well the proposed algorithms recover these scalp maps from using only a subset of M channels out of N ($M < N$). We choose M channels in a spatially uniform manner as in the previous section. We compare algorithms on 3 different types of datasets; 1) EEGLAB sample data, 2) a Motor Imagery task, 3) a Arrow Flanker task. The results are shown in Fig. 5. The segment length for Cov-DL is $t_s = 2$sec, with an overlap ratio of



(a)

(b)

(c)

(d)

Fig. 3: (a) Randomly located and oriented $N = 64$ dipoles/sources in the MNI head model that generate the simulated EEG. (b) Some of the scalp maps (with 32 channel locations) associated with the dipoles in (a). These constitute columns of true mixing matrix $\mathbf{A}_{\text{true}} \in \mathbb{R}^{32\times 64}$. Dictionary $\mathbf{A}_{\text{true}}$ has maximum spatial coherence of 0.9888. (c) Outer product of the source matrix in a 2sec. segment (sample-covariance) from Scenario 1; $M = 32, N = 64$, $k = 64$, all the sources are active at any give time. (d) Outer product of the source matrix in a 2sec. segment (sample-covariance) from Scenario 2; $M = 8, N = 40$. $k = 10$ sources are active at any given segment.

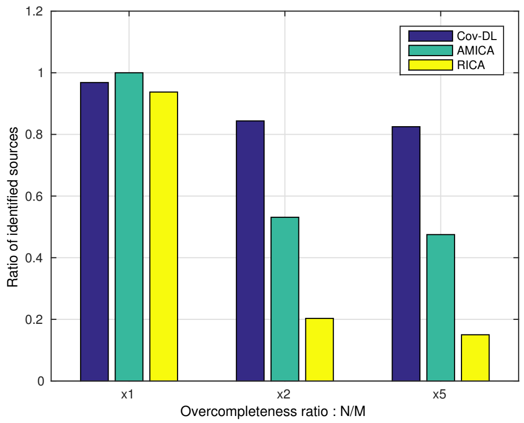
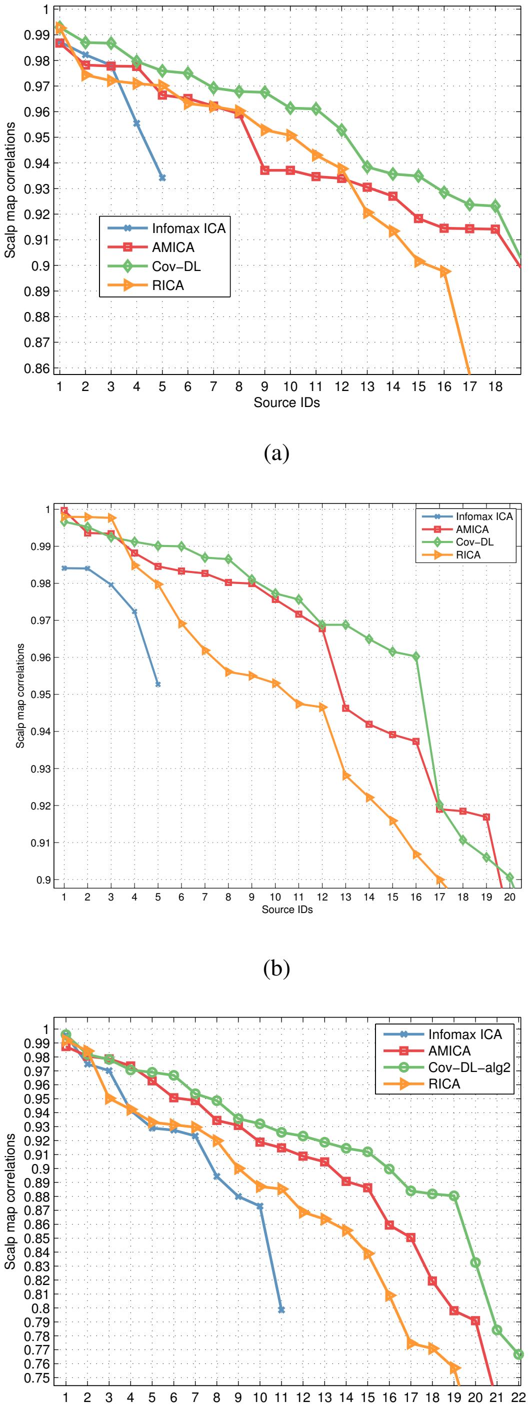


Fig. 4: Simulation results for three cases: complete, two times overcomplete, five times overcomplete. In the complete case. $M = 32, N = 32, k = 32$, Cov-DL-2 is used. In the second scenario $M = 32, N = 64, k = 64$, Cov-DL-2 is used. Third case: $M = 8, N = 40, k = 10$, Cov-DL-1 is used.

0.5 between consecutive segments. We plot the the sorted correlation values of resulting scalp maps with the best column match in the ground truth mixing matrix. In all 3 datasets, Cov-DL shows consistently higher correlations than multi-model AMICA and RICA. We also plot the correlation results of complete extended Infomax applied on M channels to show the importance of overcomplete approaches for accurate source identification in low-density EEG systems.

V. Conclusion

We proposed a dictionary learning framework, Cov-DL, that incorporates the presumed uncorrelated nature of EEG sources, which is a related but a weaker assumption than EEG source independence [2], [7]. Identification of the mixing matrix is carried to a higher dimensional covariance-domain, which enables source identification even if the number of sources active at any time is larger than the number of sensors - sparsity is not required. We proposed two different algorithms which depend on the relation between the number of sources targeted and number of sensors available. We have shown



(a)

(b)

(c)

Fig. 5: (a) EEGLAB sample data. $M = 5, N = 30$. AMICA is trained with 6 models. Cov-DL-1 is performed. (b) Motor Imagery Task, $M = 5, N = 30$. Cov-DL-1 is performed. (c) Arrow Flanker task. $M = 11, N = 30$. AMICA is trained with 3 models. Cov-DL-2 is used.

that the proposed algorithm Cov-DL is more successful than existing overcomplete ICA algorithms for finding the true generating matrix in EEG simulations. We have also demonstrated the power of Cov-DL on real data. The proposed algorithm, because of its ability to provide higher resolution than the number of sensors, can potentially increase the applicability of low-cost, low-density EEG systems in biomedical research.

References

[1] Paul L Nunez et al., "EEG coherency: I: statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales," *Electroencephalography and clinical neurophysiology*, vol. 103, no. 5, pp. 499–515, 1997.

[2] Scott Makeig et al., "Independent component analysis of electroencephalographic data," *Advances in neural information processing systems*, pp. 145–151, 1996.

[3] Robert Oostenveld et al., "FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data," *Computational intelligence and neuroscience*, vol. 2011, 2010.

[4] Roberto Domingo Pascual-Marqui et al., "Standardized low-resolution brain electromagnetic tomography (sloreta): technical details," *Methods Find Exp Clin Pharmacol*, vol. 24, no. Suppl D, pp. 5–12, 2002.

[5] Scott Makeig et al., "Mining event-related brain dynamics," *Trends in cognitive sciences*, vol. 8, no. 5, pp. 204–210, 2004.

[6] Scott Makeig et al., "Blind separation of auditory event-related brain responses into independent components," *Proceedings of the National Academy of Sciences*, vol. 94, no. 20, pp. 10979–10984, 1997.

[7] Arnaud Delorme et al., "Independent EEG Sources Are Dipolar," *PLoS ONE*, vol. 7, 02 2012.

[8] Tzyy-Ping Jung et al., "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology*, vol. 37, no. 02, pp. 163–178, 2000.

[9] Tzyy-Ping Jung et al., "Imaging brain dynamics using independent component analysis," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1107–1122, 2001.

[10] Yijun Wang and Tzyy-Ping Jung, "Improving brain–computer interfaces using independent component analysis," in *Towards Practical Brain-Computer Interfaces*, pp. 67–83. Springer, 2013.

[11] Adel Belouchrani et al., "A blind source separation technique using second-order statistics," *Signal Processing, IEEE Transactions on*, vol. 45, no. 2, pp. 434–444, 1997.

[12] Lieven De Lathauwer et al., "Fourth-order cumulant-based blind identification of underdetermined mixtures," *Signal Processing, IEEE Transactions on*, vol. 55, no. 6, pp. 2965–2973, 2007.

[13] Laurent Albera et al., "Ica-based eeg denoising: a comparative analysis of fifteen methods," *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol. 60, no. 3, pp. 407–418, 2012.

[14] Anthony J. Bell and Terrence J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *NEURAL COMPUTATION*, vol. 7, pp. 1129–1159, 1995.

[15] Jason A. Palmer et al., "Newton method for the ica mixture model.," in *ICASSP*. 2008, pp. 1805–1808, IEEE.

[16] Te-Won Lee et al., "ICA mixture models for unsupervised classification of non-Gaussian classes and automatic context switching in blind signal separation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 10, pp. 1078–1089, 2000.

[17] Shun-Ichi Amari, "Natural gradient learning for over-and under-complete bases in ica," *Neural Computation*, vol. 11, no. 8, pp. 1875–1883, 1999.

[18] Quoc V Le et al., "ICA with reconstruction cost for efficient overcomplete feature learning," in *Advances in Neural Information Processing Systems*, 2011, pp. 1017–1025.

[19] Michal Aharon et al., "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[20] Kenneth Kreutz-Delgado et al., "Dictionary learning algorithms for sparse representation," *Neural computation*, vol. 15, no. 2, pp. 349–396, 2003.

[21] Jason T Parker et al., "Bilinear generalized approximate message passing," *arXiv preprint arXiv:1310.2632*, 2013.

[22] Ozgur Balkan et al., "Localization of more sources than sensors via jointly-sparse bayesian learning," *Signal Processing Letters, IEEE*, vol. 21, no. 2, pp. 131–134, 2014.

[23] Piya Pal and PP Vaidyanathan, "Pushing the limits of sparse support recovery using correlation information," *Signal Processing, IEEE Transactions on*, 2015.

[24] Soren Hauberg et al., "Grassmann averages for scalable robust PCA," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 3810–3817.

[25] Arnaud Delorme et al., "EEGLAB, SIFT, NFT, BCILAB, and ERICA: new tools for advanced EEG processing," *Computational intelligence and neuroscience*, vol. 2011, pp. 10, 2011.

[26] Arnaud Delorme and Scott Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, 2004.