A Sparse Dictionary Learning Framework to Discover Discriminative Source Activations in EEG Brain Mapping

Feng Liu,1 Shouyi Wang,1 Jay Rosenberger,1 Jianzhong Su,2 Hanli Liu3
Department of Industrial, Manufacturing & Systems Engineering1, Department of Mathematics2, Department of Bioengineering3
The University of Texas at Arlington, Arlington, TX 76019, USA
feng.liu@mavs.uta.edu, {shouyiw,jrosenb,su,hanli}@uta.edu

Abstract
Electroencephalography (EEG) source analysis is one of the most important noninvasive human brain imaging tools that provides millisecond temporal accuracy. However, discovering essential activated brain sources associated with different brain status is still a challenging problem. In this study, we propose for the first time that the ill-posed EEG inverse problem can be formulated and solved as a sparse over-complete dictionary learning problem. In particular, a novel supervised sparse dictionary learning framework was developed for EEG source reconstruction. A revised version of discriminative K-SVD (DK-SVD) algorithm is exploited to solve the formulated supervised dictionary learning problem. As the proposed learning framework incorporated the EEG label information of different brain status, it is capable of learning a sparse representation that reveal the most discriminative brain activity sources among different brain states. Compared to the state-of-the-art EEG source analysis methods, proposed sparse dictionary learning framework achieved significant superior performance in both computing speed and accuracy for the challenging EEG source reconstruction problem through extensive numerical experiments. More importantly, the experimental results also validated that the proposed sparse learning framework is effective to discover the discriminative task-related brain activation sources, which shows the potential to advance the high resolution EEG source analysis for real-time non-invasive brain imaging research.

Introduction
In the past few decades, numerous noninvasive measurements of brain activity have been proposed, implemented and applied in clinical treatment and scientific research communities. One of the most popular techniques is electroencephalography (EEG) for its advantages of low cost, portable and high temporal resolution. EEG signals measure the electrical voltage on a variety of locations on the scalp. The electrical potentials on the scalp are superimposition of electric activity of neurons inside the brain and several types of physiological and non-physiological artifacts (Haufe 2011; Haufe et al. 2013). To be specific, the EEG signal arises as a result of synchronous intracellular current flows from the depolarized membranes of apical dendrite and non-excited cell soma and basal dendrites (Baillet, Mosher, and Leahy 2001). Those above currents (sources) propagate through the conductive medium, which is approximated with the conductivity properties of different tissues overlaying each other, to the scalp which is measured by EEG electrodes.

To infer the brain sources from the scalp recorded EEG signals belongs to the class of inverse problem. Precise localization of neuronal activity inside the brain can offer an insightful understanding of how brain is functioning given certain cognitive and motion tasks. Recent years have witnessed a shift in neuroscience community from traditional “segregation” perspective to “integration” perspective in which the functional and effective connectivity between different regions of brains (Haufe 2011; Hipp et al. 2012; Liu et al. 2016) are investigated using complex network characteristics measurement (Watts and Strogatz 1998; Barabási and Albert 1999; Guan et al. 2012; Newman 2003). Connectivity between different parcellations of brain is established by measuring the similarity of reconstructed sources. As the source reconstruction or solving the inverse problem is tenably the first and primary step for connectivity analysis of the brain (Sockeel et al. 2016; Ma et al. 2016), precise localization of sources is required in order to gain solids result using complex network measurement in the latter steps.

In this paper, we aim to calculate the discriminative sources to facilitate the understanding of brain mechanism under different cognitive tasks or different neurological disorders by incorporating a simple linear classifier which can be interpreted as discriminative filters for different brain patterns. The label information is leveraged to get a consistent and robust solution to the inverse problem.

Related Work
On contrary to forward problem, which consists of modeling the contribution of each voxel to the EEG sensors by solving Maxwell’s equation, the inverse problem is ill-posed since the number of interior brain voxels taken into account is far greater than the number of sensors outside the scalp. To precisely estimate the responsible sources of EEG activity from at least several thousands of potential contributing locations which is evenly distributed across the brain requires prior knowledge. Given different neurophysiological assumptions or prior beliefs on the structure of possible source configu-
Employed the most recently developed high accurate head model, rather than approximated head model compared to previous studies.

The structure of the rest paper is as follows: In Section 2, the problem formulation is given. In Section 3, the optimization method is proposed. In Section 4, the numerical experiments and the effectiveness of our proposed framework, conclusions and future work are given in Section 5.

**Discriminative Source Reconstruction**

In this section, we first briefly review the inverse problem, and then the proposed model in form of discriminative dictionary learning is described, which comprises the source reconstruction term and label guided discriminative term. The motivation of such a discriminative inverse model will be discussed in details.

**The Inverse Problem**

The electromagnetic field measured by EEG can be described as the following linear model:

\[ X = LS + \epsilon \]  

(1)

where \( X \in \mathbb{R}^{N_e \times N_t} \) is the EEG data measured at a set of \( N_e \) electrodes for \( N_t \) time points, \( L \in \mathbb{R}^{N_c \times N_d} \) is the lead field matrix which maps the source signal to sensors on the scalp, each column of \( L \) represents the activation pattern of a particular source to the EEG electrodes, \( S \in \mathbb{R}^{N_d \times N_t} \) represents the corresponding driving potential in \( N_d \) sources locations for all the \( N_t \) time points. \( \epsilon \) is the noise. Generally, an estimate of \( S \) can be found by minimizing the following cost function, which is composed of a quadratic error and a regularization term:

\[ \arg\min_S \| X - LS \|^2_F + \lambda \Theta(S) \]  

(2)

The penalty function \( \Theta(S) \) is to discourage unnecessary complicated source configurations and enforces neurophysiologically plausible solutions, and \( \| \cdot \|^2_F \) is the Frobenius Norm. The regularization term take the form of \( \ell_2, \ell_1 \) or mixed norm, spatially smooth formulation as in LORETA estimation or spatially sparse formulation with least absolute shrinkage and selection operator estimate. For example, to restrict the total number of activated sources less than \( T \), the following \( \ell_0 \)-Norm formulation can be used:

\[ \arg\min_S \| X - LS \|^2_F \quad \text{s.t.} \quad \| s_i \|_0 \leq T, \]  

(3)

As is well known that \( \ell_0 \)-norm is the best intuitive formulation to restrict number of activated sources, almost of neuro-science researchers, if not all, for solving EEG inverse problem use approximated norm such as \( \ell_1 \) to avoid the solution being NP-hard. For the \( t \)th time point, the \( \ell_1 \) regularized formulation is given below:

\[ s_i = s^*(x_i, L) = \arg\min_x \| x_i - Ls_i \|^2_2 + \gamma \| s_i \|_1 \]  

(4)

The ill-posed problem of Eqn.1 originates from the fact that \( L \) is a matrix with column number far much greater than the row number. We view the \( L \) matrix as a dictionary, and each column in \( L \) is an atom of the dictionary. Given the EEG recordings at a time point, which is denoted as \( x_i \) column \( X \) matrix, we want to represent the signal with minimum error by trying to find the best linear representation from activation patterns (atoms) in the over-complete dictionary \( L \). The solution \( s_i \) is the sparse coding for the \( x_i \) in the dictionary \( L \), the non-zero entries in \( s_i \) corresponding to a column in the dictionary matrix \( L \) represent the activated regions inside the brain.
Dictionary Learning Fused with Label Information

As the brain has different emotion/task related states, classification of different status is important in Brain Computer Interface (BCI) application, also it helps us understand the mechanism how brain is functioning. Mapping the EEG to the source give us a direct sense of how the sources are evoked and evolved in different states. The motivation of the supervised inverse problem formulation can be explained using a simple demonstrative example as it’s illustrates in Fig.1. The electrical potential mentioned at $x_1$ can be formulated as $x_1 = a_1s_1 + a_2s_2 + a_3s_3 + \epsilon$ and the same case for $x_2$ channels, where $a_i (i = 1, 2, 3)$ describe the conductivity from the for electricity traveling from site $s_i$ to channel $x_1$. According previous studies (Raichle 2006), only a small portion of electrical energy are task related and it’s reasonable to assume that $s_1$ represents the non-task resting state source and contribute most of the potential measured in sensors. Assume $s_2$ is activated when performing task A and $s_3$ is related to task B. Under the condition of low signal noise ratio (SNR), the reconstructed source tends to be only $s_1$ without explicitly using the supervising label. Here we leverage the label information explicitly in the hope of successful reconstruction of the discriminative source $s_2$ and $s_3$. Here we present a new framework that can infer the source signal guided by the label information. A classification of different brain status based on the sparse coder $s_i$ is obtained by determining its model parameters $W$, where

$$W = \arg \min_W \sum_i \ell \{ h_i, f(s_i, W) \} + \lambda ||W||^2_F$$

where $\ell \{ \cdot \}$ is the loss function for classification accuracy based on the ground truth and classification model $f(\cdot)$, and $h_i$ is the label vector where non-zero entry denotes the corresponding class. Traditional procedure is first solve the pure inverse problem ignoring the supervising label and then train the sparse coding $s_i$ with classification model. Separating the inverse problem and classification problem can be misleading, we argue that since we have the brain status information, it’s better to use it as a label to make the inverse solution exhibiting discriminative capability. With this thought and inspired by literatures in computer vision community (Jiang, Lin, and Davis 2013; Yang et al. 2014; Zhang and Li 2010; Pham and Venkatesh 2008), the following sparse discriminant inverse model is given:

$$\begin{align*}
(W, S) &= \arg \min_{W,S} ||X - LS||^2_F + \beta \sum_i \ell \{ h_i, f(s_i, W) \} \\
&+ \lambda ||W||^2_F \text{ s.t. } \forall i, ||s_i||_0 \leq T
\end{align*}$$

(6)

The first term is the reconstruction error, the second term represents the classification loss, the third term is the regularization of $W$ to avoid over-fitting. This formulation aims to simultaneously learn the sparse coding and the classification model. Using the multi-class classifier $f(\cdot)$ instead of one-against-all classifiers is efficient for classification, by suppressing features sharing among classes and trying to explicitly extract different sparse representation among different classes. In this paper, We focus on an inverse solution with more balanced reconstructive and discriminative power by adding the classification regularization term $\lambda$. A summary of our proposed framework is illustrated in Fig.2.

Source Reconstruction Based on Linear Classifier

From Eqn.6, we reduce to the following optimization problem by using a simple linear classifier.

$$\begin{align*}
(W, S) &= \arg \min_{W,S} ||X - LS||^2_F + \beta ||H - WS||^2_F \\
&+ \lambda ||W||^2_F \text{ s.t. } \forall i, ||s_i||_0 \leq T
\end{align*}$$

(7)

Here $H = [h_1, h_2, \ldots, h_N] \in \mathbb{R}^{m \times N_t}$, with each row $h_i, i = 1, \ldots, N_t$ being the label vector corresponding to an EEG signal $x_i$. In order to solve the optimization problem (7), the K-SVD algorithm and its derivatives can be used. However, our proposed method is different from these previous methods in several aspects, owing to it being tailored to solve the EEG inverse problem.

Optimization with K-SVD Algorithm

For Equation 7, it can be rewritten as

$$\begin{align*}
(W, S) &= \arg \min_{W,S} \left\| \left( \frac{X}{\sqrt{\beta}H} \right) - \left( \frac{L}{\sqrt{\beta}W} \right) S \right\|^2_F \\
&+ \lambda ||W||^2_F \text{ s.t. } \forall i, ||s_i||_0 \leq T
\end{align*}$$

(8)

Let $X_{new} = (X^t, \sqrt{\beta}W^t)^t, L_{new} = (L^t, \sqrt{\beta}W^t)^t$, the optimization of Equation 8 is equivalent to solving the following problem:

$$\begin{align*}
(L_{new}, S) &= \arg \min_{L_{new}, S} \left\| X_{new} - L_{new}S \right\|^2_F \\
&+ \lambda ||W||^2_F \text{ s.t. } \forall i, ||s_i||_0 \leq T
\end{align*}$$

(9)

In neuroscience community, the lead field matrix is rarely normalized (Grech et al. 2008). We use normalized lead field matrix $L$ to meet the requirement of K-SVD algorithm. It’s more important to find an explanatory activation pattern compared to magnitude of the signal as a common practice.

Figure 1: source to electrode
Figure 2: Discriminative Source Reconstruction Framework: The left two topoplots represent the recorded EEG potentials on the scalp for two stimulus status (e.g. finger tapping and comedy video stimulus), and the lead field matrix are represented as overcomplete dictionary, the sparse coefficients are the codes for the source activation location and activation potentials; The sparse coefficients and $W$ matrix are estimated simultaneously. Each row of $W$ matrix is termed as discriminative filter because the Hadamard product of the source code coefficient and the discriminative filter can highlight the corresponding stimulus activated source signal by masking the common background or resting activation signals which is share by other different brain stimulus inputs. The rightmost pictures are exemplary reconstructed discriminative source activation patterns on the cortex (Haufe 2011). Later we show that the normalization doesn’t effect the solution in case of $\ell_0$ norm. The normalization is defined as:

$$
\langle L', S \rangle = \arg \min_{L', S} \|X_{\text{new}} - L'S\|_F^2
$$

s.t. $\forall i, \|s_i\|_0 \leq T$

(11)

For similarity, we omit the apostrophe (') notation when there is no confusion. When fixing $S$, solving $L$ matrix can be regarded as solving a simple regression problem:

$$
\hat{L} = \arg \min_L \|X - LS\|_F^2,
$$

(12)

where $\hat{L} = XS^T(SS^T)^{-1}$. The computational complexity of $XS^T(SS^T)^{-1}$ is $O(n^3)$, it is advisable to solve it using K-SVD by updating the dictionary atom-by-atom. This optimization problem of Eqn.11 is exactly what K-SVD algorithm (Aharon, Elad, and Bruckstein 2006) solves and the only difference is that the upper $L$ part of dictionary $L_{\text{new}}$...
will not be updated. We adopt the procedure in the original K-SVD algorithm.

Following K-SVD, denote $l_k^T$ as the $k$th column in the $L_{new}$ and $s_k$ as the corresponding $k$th row in $S$. The second term $L_{new}S$ can be decomposed into the following formulation:

$$L_{new} = \sum_{k=1}^{N_s} l_k * s_k$$

Let $E_k = (X - \sum_{j \neq k}(l_j * s_j))$, representing the error without using the atom $l_k$, the main idea of K-SVD is to update each atom in the dictionary sequentially to the projected direction that most reduces the error. Let $s_R^k$ and $\hat{E}_k$ denote the result of discarding the zero entries in $x_R^k$ and $E_k$, respectively. As a result, $l_k$ and $s_R^k$ can be computed using

$$\langle l_k, s_R^k \rangle = \arg \min_{l_k, s_R^k} \| \hat{E}_k - l_k s_R^k \|_F^2$$

The above optimization problem can be easily solved by employing an SVD composition of $E_k$, namely, $USV^T = SVD(E_k)$, and using the SVD result and update the $l_k$ and $s_R^k$ with $l_k = U(:, 1), s_R^k = \Sigma(1, 1)V(1,:)$.

Algorithm 1 Revised DK-SVD algorithm

**INPUT:** Lead field matrix $L$, preprocessed EEG signal matrix $X$, relative controlling scalar $\beta$, label matrix $H$

**OUTPUT:** classification matrix $W$, EEG source matrix $S$

**Initialization:** Using K-SVD initialization described in Ref. (Aharon, Elad, and Bruckstein 2006)

set $m = 1$

while not converged do

Solve the following sparse coding problem using matching pursuit algorithm for $i = 1, 2, \ldots, N$:

$$\min_{s_i} ||x_i - Ls_i||_2^2 \quad s.t. \quad ||s_i||_0 \leq T$$

while $i$ is not equal to $N_d$ do

(1) Compute the representation error without atom $l_i, E_i = (X - \sum_{j \neq i}(l_j * s_j))$

(2) Extract the nonzero entries of $s_i$ and truncate the $E_i$ to $E_i^p$ accordingly.

(3) SVD decomposition for $E_i^p$ as $E_i^p = UAV$

(4) Update $l_i$ and $s_R^i$:

$$l_i(N_c + 1 : end) \leftarrow U(:, 1)(N_c + 1 : end),
\hat{s}_R^i \leftarrow \Sigma(1, 1)V(1,:).$$

(5) Update index $i \leftarrow i + 1$

end while

$m \leftarrow m + 1$

end while

Numerical Experiments

Numerical simulations were conducted given different SNR. We compared our proposed framework with two different baseline methods. Computation time (in s) combined with three different accuracy criteria and two solution quality measurements were used as gauges. The first baseline method is formulated as $\ell_1$ sparse representation ($\ell_1SR$) and solved with Efficient Projections onto the $\ell_1$-Ball (EP-$\ell_1$B) (Liu and Ye 2009), and the second baseline method are the recently developed MxNE (Gramfort, Kowalski, and Hämäläinen 2012).

Head and Source Model

Head model is a volume conductor model which is used to describe the flow of electric current in the head. Usually, the brain model was built in 3 steps, (1) collect the MRI images; (2) tissue segmentations (3) Mesh generation and assignment of conductivities for different tissues. We used a newly developed lead field model called ICBM-NY or “New York Head” (Huang, Parra, and Haufe 2016) which is based on highly detailed standardized finite element model (FEM) of the non-linear averaged anatomical template-ICBM152. The brain tissue segmentation is divided into 6 tissue type (scalp, skull, cerebro-spinal fluid(CSF), gray matter, white matter and air cavities) with native MRI resolution of 0.5mm$^3$. We imposed biological pink noise and EEG sensor measurement noise to test accuracy of different algorithm with SNR from 1.5 to 0.5. The pink noise is generated from 100 sources located randomly inside the brain. The sensor measurement noise is directly added to the measured EEG signal. Based on the criteria given by Baillet and Garnero (Baillet and Garnero 1997), the spatial and temporal accuracy should be at least better than 5 mm and 5 ms respectively. Also, we divided the brain into 8 region of interests (ROI) called Right Anterior Inferior (RAI), Right Anterior Superior (RAS), Right Posterior inferior (RPI), Right Posterior Superior (RPS), Left Anterior Inferior (LAI), Left Anterior Superior (LAS), Left Posterior inferior (LPI), Left Posterior Superior (LPS). The simulated source dynamics is generated linear autoregressive (AR) model with order of 6.

Reconstruction Accuracy

We used three different accuracy criteria to measure the reconstructed source accuracy. The first one is perfect reconstruction accuracy (PRA), which compare the calculated source location and the exact ground true. The second measurement is to use Baillet-Garnerno’s reconstruction accuracy (BRA) criteria (Baillet and Garnero 1997). The third measurement is to use the criteria proposed in (Haufe and Ewald 2016) denoted as Haufe reconstruction accuracy (HRA), which is to measure whether the reconstructed source is located in the ROI. To make the solution be more informative, a sparse solution is always preferred for its interpretability. The averaged number non-zero entries (NZE) in the solution is also included to measure the sparsity. $\|X - LS\|_F$ is the reconstruction error (RE).

Detailed numerical experiments performance are listed in Table 1-3. From the tables 1-3, our proposed framework outperforms the other two baseline methods for less computa-
tion time, more accurate based on 3 different criteria, more sparse solution, better reconstruction error and also with the capability of extracting discriminative sources. For the other two baseline methods, the results are based on optimized regularization parameters with a balanced trade-off between the sparsity and reconstruction error.

We found that even when the SNR is high, the traditional algorithm can only find the resting state or background noise which is set to be much larger, missing unanimously the discriminative source, meaning that the task related source can’t be estimated correctly. We are not surprised that the baseline method achieved bad performance since some of the atoms are highly coherent each other and the sparsity property is hard to achieve, the \(\ell_1\) or \(\ell_2\) norm tend to assign non-zero value to those high coherent atoms simultaneously.

We demonstrated an exemplary results in Fig.3 and Fig.4. Fig.3 demonstrates the reconstructed source and calculated discriminative filters \(W\), the discriminative filters suppress common source shared by different status while extracting and magnifying distinguished ones. Fig.4 illustrates the EEG potentials topoplots on the scalp before and after the application of our method, distinctive source activation patterns can be clearly revealed.

### Table 1: Performance comparison at SNR=1.2

<table>
<thead>
<tr>
<th>Method</th>
<th>Time</th>
<th>PRA</th>
<th>BRA</th>
<th>HRA</th>
<th>NZE</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DKSVD</td>
<td>2.04</td>
<td>0.77</td>
<td>0.81</td>
<td>0.93</td>
<td>4.00</td>
<td>0.91</td>
</tr>
<tr>
<td>(\ell_1)-SR</td>
<td>10.9</td>
<td>0.33</td>
<td>0.37</td>
<td>0.52</td>
<td>308</td>
<td>118</td>
</tr>
<tr>
<td>MxNE</td>
<td>10.6</td>
<td>0.22</td>
<td>0.25</td>
<td>0.50</td>
<td>449</td>
<td>85.6</td>
</tr>
</tbody>
</table>

### Table 2: Performance comparison at SNR=0.8

<table>
<thead>
<tr>
<th>Method</th>
<th>Time</th>
<th>PRA</th>
<th>BRA</th>
<th>HRA</th>
<th>NZE</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DKSVD</td>
<td>2.26</td>
<td>0.77</td>
<td>0.79</td>
<td>0.91</td>
<td>4.00</td>
<td>1.86</td>
</tr>
<tr>
<td>(\ell_1)-SR</td>
<td>10.9</td>
<td>0.31</td>
<td>0.35</td>
<td>0.50</td>
<td>316</td>
<td>117</td>
</tr>
<tr>
<td>MxNE</td>
<td>12.9</td>
<td>0.32</td>
<td>0.32</td>
<td>0.53</td>
<td>418</td>
<td>95.8</td>
</tr>
</tbody>
</table>

### Table 3: Performance comparison at SNR=0.5

<table>
<thead>
<tr>
<th>Method</th>
<th>Time</th>
<th>PRA</th>
<th>BRA</th>
<th>HRA</th>
<th>NZE</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DKSVD</td>
<td>2.38</td>
<td>0.63</td>
<td>0.64</td>
<td>0.68</td>
<td>4.00</td>
<td>12.4</td>
</tr>
<tr>
<td>(\ell_1)-SR</td>
<td>11.4</td>
<td>0.30</td>
<td>0.32</td>
<td>0.50</td>
<td>410</td>
<td>111</td>
</tr>
<tr>
<td>MxNE</td>
<td>12.0</td>
<td>0.25</td>
<td>0.28</td>
<td>0.50</td>
<td>507</td>
<td>89.6</td>
</tr>
</tbody>
</table>

**Conclusion and Future Work**

We aim to reconstruct discriminative sources given different brain status. A label guided dictionary learning formulation is given for the first time with \(\ell_0\)-norm and is solved using our revised version of DK-SVD algorithm. Through numerical simulations, we showed that in terms of accuracy and speed, our method is better than the \(\ell_1\) or \(\ell_2\) related ones. The reason is high coherence of lead field matrix and sparsity constraints is easy to fail. The classification component trained a \(W\) matrix with each row corresponding certain type of brain status, which is physically meaningful, we termed as discriminative filter. Our proposed framework can achieve satisfactory result compared to traditional methods and can be extended to more specific priors such as spatially smoothness requirement or depth compensation requirement.

**References**


Baillet, S., and Garnero, L. 1997. A bayesian approach...


