Basis Selection for Maximally Independent EEG Sources

Ozgur Balkan, Nima Bigdely-Shamlo, Kenneth Kreutz-Delgado, *Fellow, IEEE,* Scott Makeig

***Abstract*— We suggest a solution to the following problem: "Given multichannel linear source mixture data $\mathbf{Y}$, and an overcomplete dictionary, $\mathbf{A}$, of source projections, $\mathbf{a_i}$, how can we construct a complete basis, $\mathbf{A_0}$, by selecting columns from $\mathbf{A}$ such that the sources $\mathbf{X} = \mathbf{A_0^{-1}Y}$ are as statistically independent as possible from each other?". While conventional independent component analysis (ICA) methods learn the mixing matrix $\mathbf{A_0}$ from scratch given $\mathbf{Y}$, we restrict ourselves to selecting basis vectors from a known overcomplete dictionary. We develop two methods based on modifications of the maximum likelihood equivalent of the Infomax approach and the reconstruction-ICA (RICA) algorithm. We show that on realistic synthetic electroencephalographic (EEG) data our algorithms can find the true sources in the case of a highly coherent dictionary while requiring relatively fewer data points compared to other algorithms. On real EEG data, our algorithms obtain higher mutual information reduction.**

I. Introduction

Independent Component Analysis (ICA) has been used in many different contexts and has found a vast number of applications in diverse fields of engineering, including but not limited to blind source separation, neural networks, and biomedical source localization [1]. The common underlying model in these applications is given by

$$\mathbf{Y} = \mathbf{AX}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{\mathbf{M\times n}}$ is a data matrix in which each column $\mathbf{y_t}$ is a data vector, $\mathbf{A} \in \mathbb{R}^{\mathbf{M\times N}}$ is the mixing matrix, and $\mathbf{X} \in \mathbb{R}^{\mathbf{N\times n}}$ is composed of N component (source) activations. ICA aims to find the unknown mixing matrix $\mathbf{A}$ such that the associated source activations (rows of $\mathbf{X}^* = \mathbf{A^{-1}Y} = \mathbf{WY}$) are (maximally) statistically independent of each other. For simplicity, it is usually assumed that the mixing matrix is square, although extentions to other cases are also possible [1], [2]. Learning is usually performed without restrictions on $\mathbf{A}$. Although different constrained ICA algorithms have been considered in the past [3], the constraint we use in this paper differs from previous efforts.

Unlike common ICA algorithms, we restrict ourselves to selecting basis vectors from a pre-designated set with the goal of finding maximally independent sources. To the best of our knowledge, previous efforts for "source separation on a fixed dictionary" focused on finding a sparse solution for source activations $\mathbf{X}$ [4], [5], instead of aiming for the maximally independent solution. The dictionary of basis vectors can be known *a priori*, or can be designed (constructed) in different ways depending on the application, e.g., lead-field matrix for EEG. Although our constraint gives less freedom in learning the mixing matrix, it has some advantages over standard ICA approaches. In particular, since the solution space is finite (although combinatorially large), our approach requires less data than standard ICA to find the true mixing matrix, as we demonstrate in our experiments.

One of the biggest problems with current simultaneous sparse approximation approaches (joint sparsity or MMV) is that the number of active sources can be assumed to be sparse but the sparsity level k (the number of active sources) is usually unknown, as in EEG/ECoG (electrocorticography) problems. Given an EEG segment and a dictionary of possible sources, solving for a simultaneous sparse solution is therefore problematic. Our approach handles this problem by choosing M columns from the dictionary, and evaluating how much mutual information is reduced by projecting data onto the achieved complete basis set. Such a measure of independence is needed to assess how well we recover the unknown sources when we have no knowledge of the sparsity of the sources other than the assumption of maximal statistical independence among them. For most dictionaries, any randomly selected M columns from the dictionary would span the space that data lives in, but resulting sources would not necessarily be independent.

The outline of the paper is as follows: Section II formulates the conventional maximum likelihood (ML) ICA algorithm and modifies it to derive BASICA selection framework. Section II-B derives BASRICA algorithm using the reconstruction ICA (RICA) formulation. In Section III, we point out the connections of BASICA to M-SBL [6]. Section IV performs tests on synthetic and real data.

II. Methods

It was shown by [7] that the Infomax approach to the ICA problem is equivalent to the maximum likelihood formulation of data. The likelihood of the data, which is to be maximized, can be expressed as

$$p(\mathbf{Y}) = \prod_{t=1}^{n} |\det \mathbf{A}^{-1}| p_s(\mathbf{A^{-1}y_t}) = \prod_{t=1}^{n} \frac{1}{|\det \mathbf{A}|} p_s(\mathbf{A^{-1}y_t}) \tag{2}$$

where $p_s(.)$ is the vector source density function. When $\mathbf{A}$ or $\mathbf{W}$ is not invertible, the nonexistence of $|\det \mathbf{W}|$ is handled by the substitution of $|\det \mathbf{WW^T}|^{\frac{1}{2}}$ in the undercomplete case [2], which becomes equal to $|\det \mathbf{W}|$ when $\mathbf{W}$ is complete. Similar term for $|\det \mathbf{A}|$ would be $|\det \mathbf{AA^T}|^{\frac{1}{2}}$ when $\mathbf{A}$ is overcomplete.

O. Balkan and K. Kreutz-Delgado are with the Department of Electrical and Computer Engineering, University of California, San Diego. O. Balkan, N. Bigdely-Shamlo, S. Makeig are with the Swartz Center for Computational Neuroscience, UCSD

Email for correspondence:obalkan@ucsd.edu

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

A. BASICA

Our goal is to select among columns of $\mathbf{A}$ the basis set which provides independence among resulting source activations. In order to be able to use the general maximum likelihood ICA framework, we model the generating matrix as $\mathbf{A_{gen}} = \mathbf{A\Gamma^{\frac{1}{2}}}$, a weighted selection of columns of $\mathbf{A}$[1]. The weight matrix $\mathbf{\Gamma} = \text{diag}(\gamma)$ is a diagonal matrix of size $N \times N$ with nonnegative unknown values on the diagonal which we will learn, with the idea that at the end of the learning phase we are going to achieve M nonzero elements in γ, therefore effectively selecting M columns of the dictionary $\mathbf{A}$. We denote the matrix of nonzero columns of $\mathbf{A_{gen}}$ as $\mathbf{A}_0$.

Adopting the methodology proposed in [2], we substitute $|\det \mathbf{A}|$ in (2) with $|\det \mathbf{A\Gamma^{\frac{1}{2}}(A\Gamma^{\frac{1}{2}})^T}|^{\frac{1}{2}} = |\det \mathbf{A\Gamma A^T}|^{\frac{1}{2}}$. One should note that when γ has M nonzero values, $|\det \mathbf{A\Gamma A^T}|^{\frac{1}{2}}$ is equal to the determinant of the underlying forward model $\mathbf{A}_0$, and if γ has more than M nonzero values, the determinant still exists. Here, we assume that $\text{rank}(\mathbf{Y}) = M$, so at least M columns of $\mathbf{A}$ will be needed to explain the data, which ensures at least M elements of γ will be nonzero.

The backward transformation to the source domain is given by $\mathbf{x_t} = \mathbf{A_{gen}^{\dagger}y_t} = (\mathbf{A\Gamma^{\frac{1}{2}}})^{\dagger}\mathbf{y_t}$, where $A^{\dagger}$ denotes the pseudoinverse of the noninvertible matrix A. Here, we should note that although the pseudoinverse mapping to the source domain for overcomplete dictionaries is not always used to perform the ML overcomplete ICA framework [8], it suits well for our problem, which is rather different than the conventional overcomplete ICA. Since our goal is to find a basis set of size $M \times M$ instead of an overcomplete one, $\mathbf{x_t} = (\mathbf{A\Gamma^{\frac{1}{2}}})^{\dagger}\mathbf{y_t}$ serves to compute the regular inverse on the selected M columns of $\mathbf{A}$ in a weighted manner and place it at the corresponding locations in N-dimensional $\mathbf{x_t}$ vector, such that $\mathbf{x_t}$ has $N-M$ zero values. With this source mapping, we can write the data likelihood for our problem as,

$$p(\mathbf{Y}) = \prod_{t=1}^{n} \frac{1}{|\det \mathbf{A\Gamma A^T}|^{\frac{1}{2}}} p_s((\mathbf{A\Gamma^{\frac{1}{2}}})^{\dagger}\mathbf{y_t}) \tag{3}$$

Note that $p_s(\mathbf{x_t})$ is a vector source distribution and can be decomposed as $p_s(\mathbf{x_t}) = \prod_{i=1}^{N} p_{s_i}(\mathbf{x_{ti}})$ under the independence formulation. Here, we choose an analytical source density function $p_{s_i}(.)$ to be able to explicitly write and optimize (3). We use the super-Gaussian density $p_{s_i}(x) = c\,\text{sech}(x)$, which has been shown to be suitable for EEG sources in the past [9]. Moreover, the use of a super-Gaussian source density also enhances the selection property of the algorithm, namely the convergence to sparse γ and equivalently large number of zero rows of sources $\mathbf{X}$. It should also be emphasized that although the actual sources might have arbitrary variances, we fix $p_{s_i}(x) = C\text{sech}\,(x)$ for each source, with zero mean and a fixed variance. The zero mean condition is easy to satisfy by removing the mean of the data, and we are able to allow fixed variance for each source s_i due to the source equation $\mathbf{x_t} = (\mathbf{A\Gamma^{\frac{1}{2}}})^{\dagger}\mathbf{y_t}$ in (3), i.e. the actual source variances are embedded in Γ. Sources with higher variance will have higher γ_i.

Separating the vector source distribution in (3) into individual scalar source distributions and taking the $-2\log(.)$ transformation of likelihood gives the following, which is to be minimized:

$$\begin{aligned} L(\gamma) =& n \log |\det \mathbf{A\Gamma A^T}| \\ & - 2\sum_{t=1}^{n}\sum_{i=1}^{N} \log p_{s_i}(\gamma_{\mathbf{i}}^{\frac{1}{2}} \mathbf{a_i^T}(\mathbf{A\Gamma A^T})^{-1}\mathbf{y_t}). \end{aligned} \tag{4}$$

The vector $\mathbf{a_i}$ is the i-th column of $\mathbf{A}$. We minimize the above quantity over γ, which is the only unknown in the model. This approach involves computing the gradient of $L(\gamma)$ with respect to γ and rearranging the terms to achieve the following fixed point update at the k^{th} iteration,

$$\gamma_{\mathbf{i}}^{(\mathbf{k+1})} = \gamma_{\mathbf{i}}^{(\mathbf{k})} \frac{\mathbf{2}\sum_{\mathbf{t=1}}^{\mathbf{n}} \nabla_{\mathbf{i}}(\mathbf{t})}{\mathbf{n a_i^T \Sigma^{-1} a_i}}. \tag{5}$$

where $\mathbf{\Sigma} = (\mathbf{A\Gamma A^T})$ and the numerator is the derivative of the second term in (4). We initialize the algorithm with $\gamma_i^{(0)} = 1, \forall i$, and perform the updates for a fixed number of iterations, or stop once the changes in γ are below a threshold.

B. BASRICA

In [10], Le et al. proposed an ICA algorithm called Reconstruction ICA (RICA) using a soft reconstruction cost, which is also applicable to the overcomplete ICA case. Instead of finding the mixing matrix, RICA optimizes the tall unmixing matrix $\mathbf{W} \in \mathbb{R}^{N \times M}$ with the following objective function.

$$\min \frac{1}{2}\sum_{t=1}^{n} \|\mathbf{W^T W y_t} - \mathbf{y_t}\|_2^2 + \lambda \sum_{j=1}^{N}\sum_{t=1}^{n} g(\mathbf{W_j} y_t) \tag{6}$$

We modify the RICA objective function such that it allows for a basis selection from a known dictionary $\mathbf{A}$. Using the same idea as in Section II.A, we regard the mixing matrix as a weighted selection of columns from the dictionary $\mathbf{A}$, namely $\mathbf{A_{gen}} = \mathbf{A\Gamma}$, with the projection to the source domain as $\mathbf{x_t} = \mathbf{W}^{'}\mathbf{y_t} = (\mathbf{A\Gamma})^{\dagger}\mathbf{y_t}$ [2]. If the data is whitened by the sphering matrix $\mathbf{S}$, such that $\hat{\mathbf{y}}_\mathbf{t} = \mathbf{S y_t}$, the source equation can be rewritten as $\mathbf{x_t} = \mathbf{W}\hat{\mathbf{y}}_\mathbf{t} = (\mathbf{A\Gamma})^{\dagger}\mathbf{S^{-1}}\hat{\mathbf{y}}_\mathbf{t}$. Plugging $\mathbf{W} = (\mathbf{A\Gamma})^{\dagger}\mathbf{S^{-1}} = \mathbf{\Gamma A}(\mathbf{A\Gamma^2 A^T})^{-1}\mathbf{S^{-1}}$ into (6) gives the following function to be optimized for BASRICA.

$$\begin{aligned} \min_{\gamma} \frac{1}{2}\sum_{t=1}^{n} & \|\mathbf{S^{-T}}(\mathbf{A\Gamma^2 A^T})^{-1}\mathbf{S^{-1}}\hat{\mathbf{y}}_\mathbf{t} - \hat{\mathbf{y}}_\mathbf{t}\|_2^2 + \dots \\ & \dots + \lambda \sum_{j=1}^{N}\sum_{t=1}^{n} g(\gamma_\mathbf{j} \mathbf{a_j^T}(\mathbf{A\Gamma^2 A^T})^{-1}\mathbf{S^{-1}}\hat{\mathbf{y}}_\mathbf{t}) \end{aligned} \tag{7}$$

[1]The square root weighting $\Gamma^{\frac{1}{2}}$ is used instead of Γ to make the connection to M-SBL more obvious

[2]Without loss of generality, we choose the weighting Γ contrary to $\Gamma^{\frac{1}{2}}$ in BASICA because it will enable us to optimize without the constraint $\gamma \geq 0$.

To be consistent with the BASICA formulation we choose $g(.) = \log(\text{sech}(.))$. One of the benefits of this optimization problem is that it allows us to use unconstrained solvers, e.g. L-FBGS. Since RICA results in a degenerate solution $\mathbf{W}$ (only M nonzero rows) without row normalization, we expect $\mathbf{\Gamma A}(\mathbf{A\Gamma^2 A^T})^{-1}$ to converge to $N - M$ zero rows as well, equivalently to a sparse γ. Moreover, the sparsity of the solutions can be altered with the trade-off parameter λ.

III. BASICA and M-SBL

In this section, we explore the connection of BASICA to M-SBL which is the modification of sparse bayesian learning (SBL) algorithm to the simultaneous sparse approximation [6]. In [6], the negative log-likelihood for M-SBL is given as,

$$L_{M-SBL}(\gamma, \lambda) = n \log |\det \left(\lambda \mathbf{I} + \mathbf{A\Gamma A^T}\right)| + \ldots \\ \ldots + \sum_{t=1}^{n} \mathbf{y_t^T} \left(\lambda \mathbf{I} + \mathbf{A\Gamma A^T}\right)^{-1} \mathbf{y_t}. \quad (8)$$

where λ is the noise variance parameter. Plugging $p_{s_i}(x) = \mathcal{N}(0, 1)$ into (4) for BASICA gives,

$$\begin{aligned} L(\gamma) &\propto n \log |\det \mathbf{A\Gamma A^T}| + \ldots \\ &\ldots + \sum_{t=1}^{n} \mathbf{y_t^T} (\mathbf{A\Gamma A^T})^{-1} \mathbf{A\Gamma^{\frac{1}{2}} \Gamma^{\frac{1}{2}} A^T} (\mathbf{A\Gamma A^T})^{-1} \mathbf{y_t} \\ &= n \log |\det \mathbf{A\Gamma A^T}| + \sum_{t=1}^{n} \mathbf{y_t^T} (\mathbf{A\Gamma A^T})^{-1} \mathbf{y_t} \\ &\equiv L_{M-SBL}(\gamma, \lambda) \qquad , \lambda \to 0 \end{aligned}$$

We can see that M-SBL in the noiseless limit is a special case of BASICA.

IV. Experiments

We test our algorithms on synthetic data and real EEG data to compare results with those of M-SBL and reweighted $l_{1,1}$ which induces a Laplacian prior on the sources.

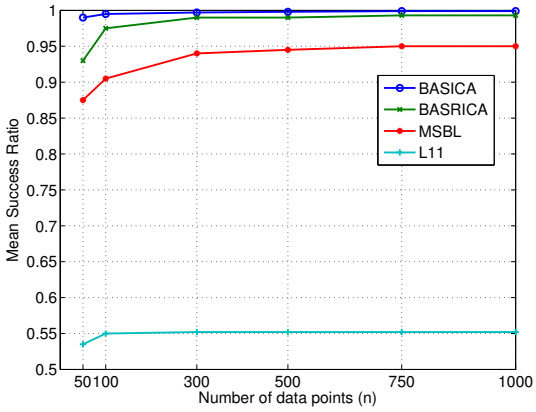
A. Simulated Data

Here, we compare the performance of our algorithms in a realistic EEG scenario. We construct a coherent dictionary $\mathbf{A}$ of EEG scalp maps of size $M = 32, N = 100$. We obtain this dictionary by selecting a subset of columns that are coherent with each other from the lead-field matrix. The subset we choose creates a highly coherent dictionary with mutual coherence of $\mu = 0.998$, and average spatial map correlation of 0.85.

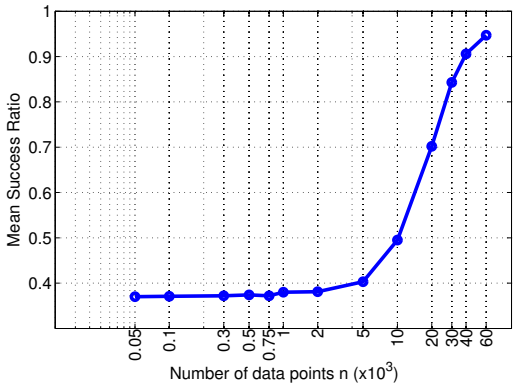
For each trial, we randomly choose the support set of size M, and obtain M realistic EEG sources from ICA decompositions of earlier EEG studies. We generate data as $\mathbf{Y} = \mathbf{AX}$, and applying different algorithms we try to recover the true support set of sources. After convergence, we extract the support set of size M by choosing M rows of the resulting source matrix that has the highest power. We calculate the success ratio for each algorithm with the below formula after 100 trials

$$r = \frac{1}{100} \sum_{k=1}^{100} |s_k \cap \hat{s}_k| / M. \quad (9)$$

where s_k is the true support set for trial k and $\hat{s}_k$ is the support set returned by the algorithm. Figure 1(a) shows the comparison of 4 algorithms. It is seen that BASICA and BASRICA outperform M-SBL and reweighted $l_{1,1}$ in terms of converging to the true support set in the highly coherent dictionary. Our algorithms require fewer data points to successfully identify the true sources.



(a)



(b)

Fig. 1. (a) Comparison of 4 algorithms on synthetically generated data with a coherent EEG scalp maps dictionary (b) Perfomance of ICA + column matching on the same type of data. ICA + column matching requires 500x data points compared to BASICA and BASRICA (compare with (a)) to recover 95% of the true sources

Another approach we investigated is performing regular ICA, e.g. Infomax, on data $\mathbf{Y}$ and finding a mixing matrix $\mathbf{A}'$, followed by matching the columns of $\mathbf{A}'$ to the closest columns in dictionary $\mathbf{A}$, in a one-to-one manner. Comparing Figure 1(a) and 1(b), it can be seen that unconstrained ICA requires many more data points to converge to the true mixing matrix. This example shows an important benefit of our proposed approaches for direct basis selection.

TABLE I

PERCENTAGE OF EPOCHS FOR WHICH ALGORITHM i (ROW) PRODUCES MORE MIR THAN ALGORITHM j (COLUMN).

Algorithms	Reweighted $l_{1,1}$	M-SBL	BASICA	BASRICA	$\text{MIR}_{\text{AMICA}}$
Reweighted $l_{1,1}$	·	97.14	0	17.14	2.86
M-SBL	2.86	·	2.86	5.71	2.86
BASICA	100	97.14	·	17.14	14.29
BASRICA	82.86	94.29	82.86	·	82.86
$\text{MIR}_{\text{AMICA}}$	97.14	97.14	85.71	17.14	·

B. Experiments on real EEG data

Given a real EEG data segment and a dictionary of possible sources, it is a challenging task to assess how the algorithms perform, due to the unknown nature of true sources. Yet, for EEG source separation tasks it is widely accepted that the sources are instantaneously statistically independent of each other. Therefore, a measure of independence among the sources, e.g. mutual information reduction (MIR) [11], can serve well to compare the results of different algorithms.

$$\begin{aligned}\text{MIR}_Y(W) &= I(x) - I(y) \\ &= \sum_{i=1}^{M} h(x_i) - \sum_{i=1}^{M} h(y_i) - h(x) + \dots \\ &\quad \dots + \log|\det W| + h(x) \\ &= \log|\det W| + \sum_{i=1}^{M} h(x_i) - \sum_{i=1}^{M} h(y_i). \quad (10)\end{aligned}$$

The calculation of MIR requires a square unmixing matrix $\mathbf{W}$ that relates the sources and data as $\mathbf{X} = \mathbf{WY}$. This fits well with our methods, since the support set s our algorithms select from the dictionary is of size M and we can assign $\mathbf{W} = \mathbf{A}_s^{-1}$, where $\mathbf{A}_s$ is the matrix of selected columns.

We use 32-channel 256-Hz EEG data collected during a rapid serial visual presentation task (RSVP). We first perform an ICA mixture model on the entire dataset using multi-model AMICA [12] with 10 mixture models, returning 10 square mixing matrices $\{\mathbf{A_1}, \mathbf{A_2}, \dots, \mathbf{A_{10}}\}$. Dictionary $\mathbf{A}$ is then formed by concatenating those individual ICA models and removing the identical scalp maps. Multi-model AMICA was previously shown to capture the possible non-stationarities inherent in EEG data [12], thus forming a tractable way of obtaining more sources than sensors from EEG. The overcomplete scalp maps dictionary we obtain with the above described method is of size $M = 32, N = 63$.

We extract 70 EEG epochs (data segments around events of interest) of length 6 seconds. Using the dictionary $\mathbf{A}$, we run our algorithms separately on each epoch and compare the resulting MIR values. In Table I, we perform a pairwise comparison of the algorithms and measure the percentage of epochs for which one algorithm results in a larger MIR than the other. In addition to the algorithms examined before, we also compare results with the maximum MIR for the individual ICA models returned by AMICA, namely $\text{MIR}_{\text{AMICA}} = \max_i \text{MIR}(\mathbf{A_i^{-1}})$. It can be seen that BASRICA has the highest likelihood of returning a larger mutual information reduction over all pairwise comparisons. BASRICA obtained a higher MIR than individual AMICA models on $\sim 82\%$ of the epochs ($p < 0.025$ on Wilcoxon signed-rank test). On real EEG, BASRICA performs better than BASICA possibly due to the data representation (error/noise) term in (7).

V. CONCLUSIONS

We modify Infomax and RICA to construct two algorithms aimed at finding jointly active sources in the case of a known overcomplete set of possible sources. While previous attempts at underdetermined source recovery problems focus on finding the sparsest solution, our algorithms aim at finding the maximally independent sources. We show that on simulated realistic EEG data our algorithms can recover the true sources in the case of a highly coherent dictionary while requiring relatively fewer data points compared to other algorithms. In real EEG experiments, our algorithms obtain higher mutual information reduction.

REFERENCES

[1] Pierre Comon and Christian Jutten, "Handbook of blind source separation: Independent component analysis and applications," *Academic Press*, 2010.

[2] James V. Stone and John Porrill, "Undercomplete independent component analysis for signal separation and dimension reduction," Tech. Rep., 1997.

[3] Wei Lu and Jagath C. Rajapakse, "Approach and applications of constrained ica," *IEEE Trans. Neural Netw*, vol. 16, no. 1, pp. 203–212, 2005.

[4] Yuanqing Li, Andrzej Cichocki, and Shun-ichi Amari, "Analysis of sparse representation and blind source separation," *Neural Comput.*, vol. 16, no. 6, pp. 1193–1234, June 2004.

[5] Michael Zibulevsky and Barak A Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural computation*, vol. 13, no. 4, pp. 863–882, 2001.

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

[7] Jean-Fran cois Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal processing letters*, 1997.

[8] Michael S. Lewicki, Terrence J. Sejnowski, and Howard Hughes, "Learning overcomplete representations," *Neural Computation*, vol. 12, pp. 337–365, 1998.

[9] S. Haufe, R. Tomioka, G. Nolte, K.-R. M andller, and M. Kawanabe, "Modeling sparse connectivity between underlying brain sources for eeg/meg," *Biomedical Engineering, IEEE Transactions on*, vol. 57, no. 8, pp. 1954 –1963, aug. 2010.

[10] Quoc V Le, Alexandre Karpenko, Jiquan Ngiam, and Andrew Y Ng, "Ica with reconstruction cost for efficient overcomplete feature learning," *Advances in Neural Information Processing Systems*, vol. 24, pp. 1017–1025, 2011.

[11] Arnaud Delorme, Jason Palmer, Julie Onton, Robert Oostenveld, and Scott Makeig, "Independent eeg sources are dipolar," *PLoS ONE*, vol. 7, 02 2012.

[12] Jason A. Palmer, Scott Makeig, Kenneth Kreutz-Delgado, and Bhaskar D. Rao, "Newton method for the ica mixture model.," in *ICASSP*. 2008, pp. 1805–1808, IEEE.