Caveats with Stochastic Gradient and Maximum Likelihood Based ICA for EEG

Jair Montoya-Martínez(✉), Jean-François Cardoso, and Alexandre Gramfort

LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, 75013 Paris, France
jmontoya@telecom-paristech.fr

Abstract. Stochastic gradient (SG) is the most commonly used optimization technique for maximum likelihood based approaches to independent component analysis (ICA). It is in particular the default solver in public implementations of Infomax and variants. Motivated by experimental findings on electroencephalography (EEG) data, we report some caveats which can impact the results and interpretation of neuroscience findings. We investigate issues raised by controlling the step size in gradient updates combined with early stopping conditions, as well as initialization choices which can artificially generate biologically plausible brain sources, so called *dipolar* sources. We provide experimental evidence that pushing the convergence of Infomax using non stochastic solvers can *reduce* the number of highly dipolar components and provide a mathematical explanation of this fact. Results are presented on public EEG data.

Keywords: Independent component analysis (ICA) · Maximum likelihood · Stochastic gradient method · Infomax · Electroencephalography (EEG) · Neuroscience

1 Introduction

Independent Component Analysis (ICA) is a multidimensional statistical method that seeks to uncover hidden latent variables in multivariate and potentially high-dimensional data. In the ICA model we consider here, the observations $\mathbf{x}$ satisfy $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{s}$ are referred to as the sources or independent components, and $\mathbf{A}$ is the mixing matrix considered unknown [12]. In the following, we assume as many sources as sensors: $\mathbf{A}$ is a square matrix. This model is usually described as a latent linear stochastic model, where $\mathbf{x}$ and $\mathbf{s}$ are random variables (r.v.) in $\mathbb{R}^N$, and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is a nonsingular matrix. The goal of ICA is, given a set of observations of the r.v. $\mathbf{x}$, to estimate the hidden sources $\mathbf{s}$ and the unknown mixing matrix $\mathbf{A}$. In order to accomplish this task, the key assumption in ICA is that the components $s_1, s_2, \ldots, s_N$ are mutually statistically independent [6], a plausible assumption if each individual source signal is thought to be generated by a process unrelated to any other source signal.

In neuroscience, and in particular when working with Electroencephalography (EEG) data, ICA is extremely popular. It is used for artifact removal as well as

P. Tichavský et al. (Eds.): LVA/ICA 2017, LNCS 10169, pp. 279–289, 2017.
DOI: 10.1007/978-3-319-53547-0_27

estimation of brain sources. Linear ICA is justified by the fact that EEG data are linear mixtures of volume-conducted neural activities [13]. Each brain source is thought to represent near-synchronous local field activity across a small cortical patch [14], which can be modeled as an electrical current dipole (ECD) located within the brain [18].

To solve the ICA estimation problem, we need to estimate a linear operator $\widehat{\mathbf{W}} \in \mathbb{R}^{N \times N}$, such that $\hat{\mathbf{s}} = \widehat{\mathbf{W}}\mathbf{x} = \widehat{\mathbf{W}}\mathbf{A}\mathbf{s} \approx \mathbf{s}$, where $\hat{\mathbf{s}}$ is an estimation of the sources. In the context of EEG, the estimated mixing matrix $\widehat{\mathbf{A}} = \widehat{\mathbf{W}}^{-1}$ gives us information about how the estimated sources are seen at the sensor level. Indeed each column of $\widehat{\mathbf{A}}$ can be visualized as a scalp map (topography). This helps the EEG users to identify plausible brain sources which correspond to ECDs. For such sources, topographies are spatially smooth and exhibit a dipolar pattern.

A common approach to tackle the ICA problem is to cast it as a maximum likelihood estimation problem [16]: given a probability density function (p.d.f.) $p_{\mathbf{s}}(\mathbf{s})$, associated with the sources, and a set $\mathcal{X} = \{\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(T)\} = \{\mathbf{x}_j\}_{j=1}^{j=T}$, containing independent and identically distributed (i.i.d) samples of the r.v. $\mathbf{x}$, one wants to find the unmixing matrix $\mathbf{W}$ that maximizes the log-likelihood function $\ell(\mathbf{W}, \mathcal{X})$ (for the sake of simplicity, we make the dependence on $\mathcal{X}$ implicit in ℓ):

$$\ell(\mathbf{W}) = \sum_{j=1}^{T} \left[\sum_{i=1}^{N} \log p_{s_i}(\mathbf{w}_i^\top \mathbf{x}_j) + \log|\det \mathbf{W}| \right] \tag{1}$$

where $\mathbf{w}_i^\top$ denotes the i-th row of $\mathbf{W}$. One of the most popular algorithm in the EEG community is Infomax [2] and it can be shown to follow this likelihood approach [5].

In order to maximize the log-likelihood function (1), we have at our disposal two different families of optimization methods: batch methods, such as gradient descent, which use at each iteration the entire set of observations $\mathcal{X}$, and stochastic methods which access at each iteration only one observation $\mathbf{x}_j$, or a small group of observations $\mathcal{B} = \{\mathbf{x}_j\} \subset \mathcal{X}$ (also known as mini-batch). When using stochastic gradient (SG) methods, the gradients used as update directions, with one sample or a mini-batch, are affected by 'noise' [4]. The consequence is that unless so-called step size annealing strategies are employed, SG will not reach a minimum of the minimized function [17]. On the contrary, gradient descent (GD), which is a non-stochastic batch method, does guarantee a decay of the minimized function at every iteration and does reach points with zero gradients (cf. Proposition 1.2.1 in [3]). Convergence rates can be up to linear for strongly convex functions with Lipschitz gradients. However, one update of the parameters by GD requires a full pass on the whole dataset while SG does already reduce the cost function after accessing a fraction of it. That is why when working with many samples, which is the case for EEG, SG exhibits a rapid convergence during the early stages of the optimization procedure, yet this convergence then slows down and the cost function reaches a plateau well before a point with zero gradient is reached. In other words, plain SG will stop too early if a high numerical precision solution is needed.

Infomax uses SG to maximize the log-likelihood function (1). In particular, it uses a mini-batch SG method in combination with a step size annealing policy, which is applied after one pass on the full data (Infomax considers one iteration as one pass on the full data). As in any stochastic method, the Infomax solver needs an initial step size.

In the first part of the paper, we explore the impact of algorithm initialization, the initial value of the step size, jointly with the annealing policy and the stopping criterion used by the standard Infomax implementation. We explain theoretically why the commonly used initialization of Infomax produces highly dipolar sources. We then explain the observation that Infomax can eventually waste a lot of computation time without converging, or worse can report convergence while the norm of the gradient is still high. Finally, using public EEG data and an alternative optimization strategy we investigate the impact of convergence on source dipolarity, highlighting specificities of EEG.

2 Infomax: Description of the Optimization Algorithm

The maximum likelihood problem tackled by Infomax can be written as the following minimization problem:

$$\widehat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmin}}\ L(\mathbf{W}) \tag{2}$$

where $L(\mathbf{W}) = -\ell(\mathbf{W})/T$ denotes the normalized negative log-likelihood function. In order to solve the problem (2), Infomax uses the relative gradient [1,7]

$$\widetilde{L}'(\mathbf{W}) = \frac{1}{T}\sum_{j=1}^{T}\left[\boldsymbol{\phi}(\mathbf{y}_j)\mathbf{y}_j^\top - \mathbf{I}_N\right] \tag{3}$$

where $\mathbf{y}_j = \mathbf{W}\mathbf{x}_j$, where $\boldsymbol{\phi}(\mathbf{y}_j) = [\phi_1(y_{j_1}), \ldots, \phi_N(y_{j_N})]^\top$, and where

$$\phi_k(y_{j_k}) = -p'_{s_k}(y_{j_k})/p_{s_k}(y_{j_k}) = \tanh\left(y_{j_k}/2\right)$$

In order to solve (2), the reference Infomax implementation, included for example in the EEGLAB software [8], uses a mini-batch stochastic gradient method, whose iterative expression can be written as follows:

$$\mathbf{W}_{k+1} = \mathbf{W}_k - \alpha \sum_{j\in\mathcal{B}_k}\left[\boldsymbol{\phi}(\mathbf{y}_j)\mathbf{y}_j^\top - \mathbf{I}_N\right]\mathbf{W}_k \tag{4}$$

where, before each pass on the full data, the set of samples $\{\mathbf{x}_j\}_{j=1}^{j=T}$ is randomly permuted, and then, during the full pass, each mini-batch $\mathcal{B}_k$ is created by taking, sequentially, a subset of samples $\{\mathbf{x}_j\}$ of size $|\mathcal{B}_k|$. Once the pass on the full data is completed, the stopping criterion is checked and the annealing policy is applied to determine whether or not the step size α should be decreased. In this policy, the step size is never increased. Finally, this process is repeated until the stopping criterion is fulfilled or until the maximum number of iterations is reached.

Let us denote $\Delta_k = \mathbf{W}_{k+1} - \mathbf{W}_k$. The stopping criterion used by the standard Infomax implementation is $\|\Delta_k\|_F^2 < \text{tol}$, where $\|\cdot\|_F$ is the Frobenius norm, tol is by default 10^{-6} if $N \leq 32$, and 10^{-7} otherwise. This implementation uses the following heuristic for its annealing policy: if the angle between matrices Δ_k and Δ_{k-1} is larger than 60°, i.e., $\arccos\left(\text{Trace}(\Delta_k^\top \Delta_{k-1})/(\|\Delta_k\|_F \|\Delta_{k-1}\|_F)\right) > \pi/3$, it decreases the step size by 10% ($\alpha \leftarrow 0.9\alpha$), otherwise the step size remains the same.

Regarding the stopping criterion $\|\Delta_k\|_F^2 < \text{tol}$, it is important to notice that by Eq. (4), Δ_k is proportional to the step size α so that the algorithm will stop if the gradient *or* the step size is small. Even if the stopping criterion is not met, the step size may have become small enough (even using the standard default values) to prevent any significant update. Example of such behaviors on EEG data are given in Sect. 4.

3 Assessing the Performance of ICA Using Dipolarity

Dipolarity metric. ICA can be seen an unsupervised learning method, therefore, it is in general difficult to assess its performance in real life scenarios where the ground truth is unknown. In order to help to mitigate this issue, when using EEG data, Delorme et al. [9] proposed to use the physics underlying the propagation of the electromagnetic field throughout the head. Physics states that the signals measured on the scalp can be modeled as linear mixtures of the electrical activities generated by ECDs located inside the brain. To assess the biological plausibility of ICA sources, Delorme et al. [9] proposed to take each column of the estimated mixing matrix $\widehat{\mathbf{A}}$, which can be represented as a topography, and compute how well it can be modeled by a single ECD. The following metric is defined by:

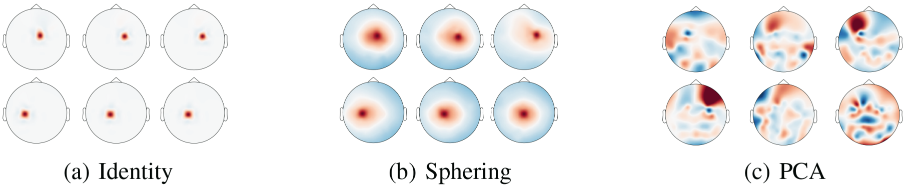
$$\text{dipolarity}(\widehat{\mathbf{A}}_j) = (1 - \|\widehat{\mathbf{A}}_j - \bar{\mathbf{A}}_j\|_2^2 / \|\widehat{\mathbf{A}}_j\|_2^2) \times 100 \tag{5}$$

where $\widehat{\mathbf{A}}_j$, $\bar{\mathbf{A}}_j$ denote respectively the j-th column of the estimated mixing matrix and the corresponding topography obtained by fitting a single dipole. Taking into account (5), the "Near-Dipolar percentage (ND%)" of an ICA decomposition is defined in [9] as: $\text{ND\%}(\widehat{\mathbf{A}}) = \{\#j : \text{dipolarity}(\widehat{\mathbf{A}}_j) > \tau\}/N$, that is, the percentage of returned components whose topographies can be modeled by a single ECD with more than a specified dipolarity threshold τ (specified as percentage of explained/residual variance). Following [9], we will consider an ICA source to be biologically plausible when its dipolarity is larger than $\tau = 90$.

Relationship between initialization and dipolarity. ICA leads to nonconvex optimization problems: solutions found by algorithms necessarily depend on their initialization. In this section we discuss connections between initialization and dipolarity.

Learning of a separating matrix $\mathbf{W}$ starts with some initial value $\mathbf{W}_0$. While it is generally possible to start with the identity matrix $\mathbf{W}_0 = \mathbf{I}_N$, it is a sound and common practice to start with some whitening matrix, that is, with a matrix $\mathbf{W}_0$ such that $\mathbf{W}_0\mathbf{\Sigma}_\mathbf{x}\mathbf{W}_0^\top = \mathbf{I}_N$ where $\mathbf{\Sigma}_\mathbf{x} = \mathrm{Cov}(\mathbf{x})$ denotes the covariance matrix of $\mathbf{x}$. There are infinitely many such matrices; two popular choices are 'PCA' and 'sphering', which can be defined in terms of the eigen-value decomposition $\mathbf{\Sigma}_\mathbf{x} = \mathbf{UDU}^\top$:

$$\mathbf{W}^{\mathrm{pca}} = \mathbf{D}^{-1/2}\mathbf{U}^\top, \ \mathbf{W}^{\mathrm{sph}} = \mathbf{UD}^{-1/2}\mathbf{U}^\top.$$


(a) Identity

(b) Sphering

(c) PCA

Fig. 1. Topographies associated with different initializations (actually a subset of 6 of them).

The topographies (the columns of matrix $\mathbf{A}_0 = \mathbf{W}_0^{-1}$) associated with the three aforementioned initializations $\mathbf{W}_0 = \mathbf{I}_N, \mathbf{W}^{\mathrm{sph}}, \mathbf{W}^{\mathrm{pca}}$, are displayed on Fig. 1. Of course, the topographies associated with $\mathbf{W}_0 = \mathbf{I}_N$ (Fig. 1(a)) are 'quasi-dipolar' in the sense that activating only one channel could be interpreted as the effect of a single source located just beneath the scalp. Much more striking is the fact that the sphering $\mathbf{W}_0 = \mathbf{W}^{\mathrm{sph}}$ produces topographies which all look dipolar. Figure 1(b) shows 6 of them, randomly selected. Nothing similar is observed in Fig. 1(c) after PCA $\mathbf{W}_0 = \mathbf{W}^{\mathrm{pca}}$. See also Fig. 4, which shows the dipolarity index for all components after sphering or PCA, sorted in decreasing order.

If the dipolarity criterion is to be used for assessing the biological plausibility of a source, one has to understand why a simple sphering would produce dipolar topographies. An explanation can be provided by the observation that somehow *sphering is the 'smallest' whitening transform.* Indeed, if a whitening matrix is close to the identity, then it should not modify much the 'quasi-dipolar' patterns of Fig. 1(a) and this is what seems to happen upon observation of Fig. 1(b). So, in which sense would sphering be the 'smallest whitening transform'? An answer is provided by Theorem 1 of Eldar *et al.* [10], which implies that, among all whitening matrices $\mathbf{W}$, the sphering matrix is the one with the minimal mean-squared difference $\mathrm{E}\|\mathbf{x}-\mathbf{Wx}\|^2$. In other words, among all white random vectors $\mathbf{Wx}$, $\mathbf{W}^{\mathrm{sph}}\mathbf{x}$ is the closest to $\mathbf{x}$ with closeness measured in the mean-squared sense. In other words, sphering is the whitening transform which moves the data the least. In terms of matrix norms, we can write the mean-squared difference $\mathrm{E}\|\mathbf{x} - \mathbf{Wx}\|^2$ as $\mathrm{E}\|\left(\mathbf{I}_N - \mathbf{W}\right)\mathbf{x}\|^2 = \mathrm{Trace}\left[\left(\mathbf{I}_N - \mathbf{W}\right)\mathbf{\Sigma}_\mathbf{x}\left(\mathbf{I}_N - \mathbf{W}\right)^\top\right]$, so that, the sphering matrix is the closest to the identity in the matrix norm $\|\mathbf{M}\|_{\mathbf{\Sigma}}^2 = \mathrm{Trace}\left[\mathbf{M}\mathbf{\Sigma}_\mathbf{x}\mathbf{M}^\top\right]$. For later reference, we note that sphering is the default initialization used by the current Infomax implementation in the EEGLAB package [8].

4 Numerical Experiments

Comparison of EEGLAB and MNE implementations. Our numerical experiments were conducted with the Infomax implementation of MNE-Python [11]. We checked that this implementation matches the reference Infomax implementation in EEGLAB [8] by reproducing the Infomax results published in [9] based on 13 anonymized EEG datasets (publicly available at http://sccn.ucsd.edu/wiki/BSSComparison [9,15]). This comparison is presented in Table 1. It shows the average across the 13 EEG datasets used in [9]. In this table, "MIR" stands for Mutual Information Reduction [9], whereas "ND 90%" denotes the percentage of ICA components with dipolarity larger than $\tau = 90$. In order to fit a single ECD to a topography, we used the same four-sphere model used in [9] for forward computation. The radius of each sphere was equal to 71, 72, 79 and 85 mm, and their corresponding conductivities relative to the cerebrospinal fluid were equal to 0.33, 1.0, 0.0042 and 0.33, respectively.

In the course of this comparison, we found that EEGLAB does not constrain the fitted dipoles to be located *inside* the brain, whereas MNE-Python does. For this reason, EEGLAB tends to report an artificially high number of dipolar components as compared to MNE-Python (second row of Table 1). However, we checked that EEGLAB and MNE-Python agree in the number of high dipolar components for dipoles located inside the brain (third row of Table 1). Even better, we checked that they agree on the *locations* of those dipoles.

Table 1. Comparison of EEGLAB and MNE-Python Infomax implementations.

Metric	EEGLAB	MNE-Python
MIR	43.092901	43.092938
ND 90%	43.445287	31.744312
EEGLAB loc. in ND 90%	22.751896	22.751896

Evaluation of the stochastic gradient approach. We proceed to evaluate the performance of the SG method used by Infomax. We use subject kb77 from the same study [9,15]. The EEG dataset is composed of 306600 samples of 71 channels sampled at 250 Hz. The algorithm is evaluated by monitoring the step size, as well as the Frobenius norm of the relative gradient after each pass on the full data. We consider the following scenarios, differing by the initial step size α_0 and tolerance tol:

- **Infomax a:** EEGLAB defaults: $\alpha_0 = 6.5 \times 10^{-4}/\log(N) \approx 1.5 \times 10^{-4}$ and $\mathrm{tol} = 10^{-7}$.

- **Infomax b:** $\alpha_0 = 10^{-3}$ as in [9], and EEGLAB default value for tol $= 10^{-7}$.
- **Infomax c:** same α_0 as in scenario "Infomax b" and tol $= 10^{-14}$.

The left panel of Fig. 2 displays the evolution of the relative gradient norm $\|\widetilde{L}'(\mathbf{W})\|_F$ across passes on the full data. As we can see in this figure, in none of the scenarios the relative gradient ever go to zero. Scenarios **a** and **b**, which use the default tolerance, either reach a plateau or stop early. Yet, reducing the tolerance in scenario **c**, reveals that it is not sufficient to push convergence compared to **b**: the trajectory plateaus just after the stopping point for **b**. However, the two plateaus of cases **a** and **c** are of different natures, as revealed by the right panel of Fig. 2, which displays the step size trajectories. For **a**, the step size remains constant after pass ~83, hence we observe the known plateau of SG

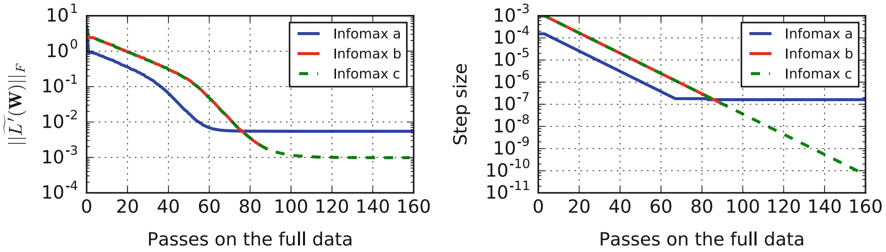
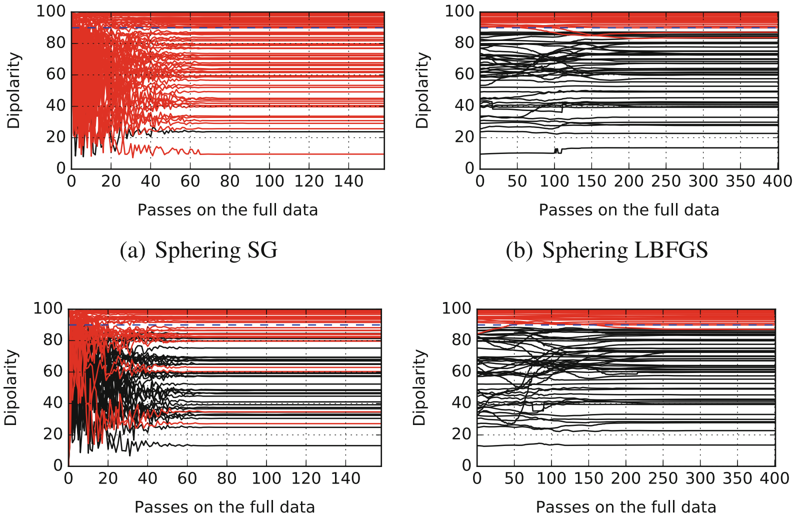


Fig. 2. Left: Frobenius norm of the relative gradient vs. iterations. Right: Step size α vs. iterations.



(a) Sphering SG

(b) Sphering LBFGS

(c) PCA SG

(d) PCA LBFGS

Fig. 3. Evolution for scenario **c** of dipolarity during Infomax (left) followed by LBFGS (right). Initialization with Sphering (top row) and PCA (bottow row). A line is colored red if it exceeds the value of 90% during the iterations. (Color figure online)

methods with fixed step size, whereas for **b** and **c** the annealing policy drives the step size to zero exponentially fast, therefore preventing the algorithm to make any further progress.

Figure 3 shows the evolution of dipolarity of the components across passes on the full data. One can see in Fig. 3(a) that when starting with sphering most of the lines are red which means that they exceed at some point the value of 90. This is due to initialization as explained previously. When initializing Infomax with PCA, see Fig. 3(c), much less components reach this high dipolarity threshold. In both cases, we observe that the dipolarity stops evolving after approximately 70 passes on the dataset, which is consistent with the plateaus of Fig. 2. The two plots on the right column show the same dipolarity metrics, but this time using the quasi-Newton method known as LBFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno), taking as initialization the unmixing matrices estimated by Infomax. In Fig. 3(b), we can see that two highly dipolar sources according to Infomax leave the region of high dipolarity. In other words, pushing the convergence with LBFGS reduces here the number of highly dipolar components as quantified in [9]. To evaluate the convergence of LBFGS towards a stationary point, we computed the Frobenius norm of the relative gradient at the end of the iterations.

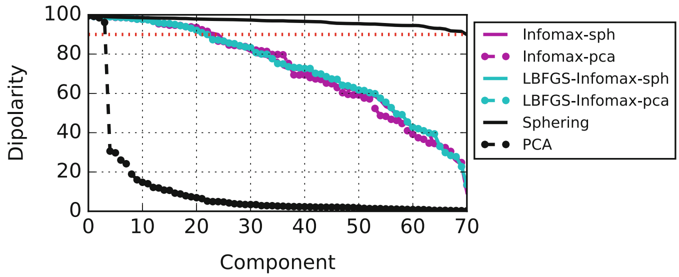


Fig. 4. Dipolarity of the components sorted in decreasing order. Plain lines correspond to sphering initialization while dashed lines correspond to PCA initialization.

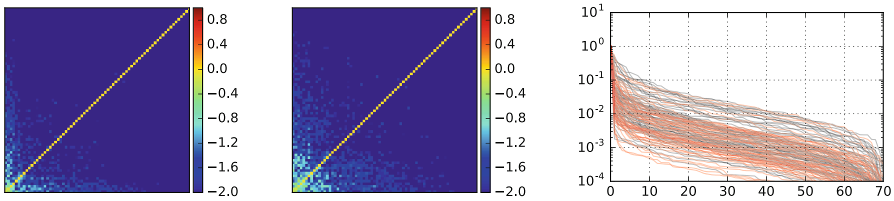


Fig. 5. LBFGS further transforms the Infomax components by a matrix T. Left shows $\log_{10}(|T|)$ using sphering initialization (for display, the sources are sorted to have the largest T_{ij} in the lower left corner). Middle shows the update following PCA initialization. Right plots the rows of the same matrices after sorting each row (red is sphering and black is PCA). (Color figure online)

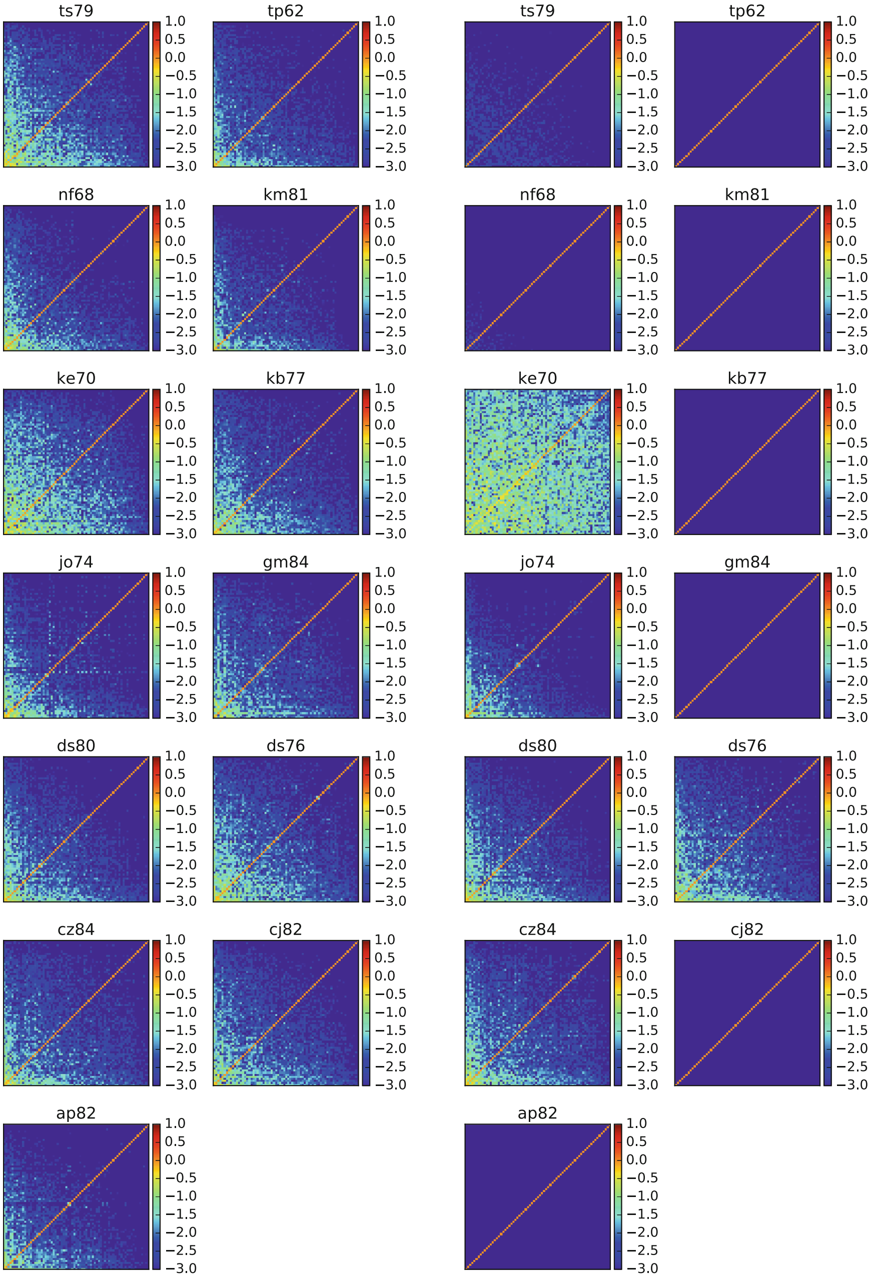


Fig. 6. Changing the initialization changes the local minimum otherwise the transform T linking the sources obtained with the two initializing whiteners (PCA and sphering) would be the identity. Each plot shows $\log_{10}|T|$. Left: Infomax, right: LBFGS. By completing the convergence process, we find that in 8 cases out of 13, the resulting sources do not depend on initialization.

While this norm was about 10^{-4} after SG, it is about 10^{-7} after LBFGS, which confirms that LBFGS does push significantly the convergence.

The dipolarity of the components after sphering or PCA, as well as following Infomax and LBFGS in the same four cases as Fig. 3 is presented in Fig. 4. This plot is an extra evidence that simple sphering already yields almost only highly dipolar sources. PCA, on the contrary, contains far less dipolar sources. This is also in line with Fig. 1. This plot also reveals that Infomax followed by LBFGS reaches almost identical dipolarities. This suggests that LBFGS manages to wash out the effect of initialization by converging to the same local minimum.

To gain further evidence, we ran a number of checks. Figure 5 quantifies how close the unmixing matrix estimated after LBFGS is to the inverse of the mixing matrix obtained by SG. The multiplication of these matrices should be close to the identity (up to permutation). Figure 5 reports that it is far from it, demonstrating that LBFGS deviates non-trivially from the output of Infomax. One can also see that the change operated by LBFGS is larger in the PCA case. In other words, Infomax following PCA brings the estimate further away from a stationary point than the sphering initialization. This conclusion only holds if the same stationary point is reached in both settings. Evidence for this is presented in Fig. 6, where one can see that for this subject (kb77) the estimated unmixing matrix obtained in the PCA condition is close to the inverse of the mixing matrix obtained following sphering (up to a permutation). To assess if there is convergence to the same local minimum when SG is followed by LBFGS, we ran the same computation on all the subjects. Figure 6 shows that for 8 out of 13 subjects the result perfectly replicates.

5 Conclusion

We explored the annealing policy, the initial step size and the stopping criterion used by the SG Infomax. We reported results where this algorithmic choices lead Infomax to stop before reaching an accurate stationary point, despite a high number of iterations and long computation times. We explained theoretically why the sphering initialization used by Infomax produces highly dipolar sources. By further pushing the convergence using a quasi-Newton method, we showed that the initialization influences the output of Infomax, hence overestimating the number of highly dipolar sources. This observation could explain why practitioners tend to avoid dimensionality reduction when Infomax is used on EEG. Indeed sphering cannot be used in this case. This paper should be seen as an instantaneous picture on current usage of ICA for EEG data. Given the massive use of such techniques, we hope that it will motivate the development and dissemination of better optimization schemes in this scientific community.

Acknowledgments. This work was funded by the Paris-Saclay Center for Data Science with partial support from the European Research Council (ERC-YStG-676943).

References

1. Amari, S.-I.: Natural gradient works efficiently in learning. Neural Comput. **10**(2), 251–276 (1998)
2. Bell, A.J., Sejnowski, T.J.: An information-maximization approach to blind separation and blind deconvolution. Neural Comput. **7**(6), 1129–1159 (1995)
3. Bertsekas, D.: Nonlinear Programming. Athena Scientific, Cambridge (1999)
4. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Lechevallier, Y., Saporta, G. (eds.) Proceedings of COMPSTAT'2010, pp. 177–186. Springer, Heidelberg (2010)
5. Cardoso, J.F.: Infomax and maximum likelihood for blind source separation. IEEE Signal Process. Lett. **4**, 112–114 (1997)
6. Cardoso, J.F.: Blind signal separation: statistical principles. Proc. IEEE **86**(10), 2009–2025 (1998)
7. Cardoso, J.F., Laheld, B.H.: Equivariant adaptive source separation. IEEE Trans. Sig. Proc. **44**(12), 3017–3030 (1996)
8. Delorme, A., Makeig, S.: EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Methods **134**(1), 9–21 (2004)
9. Delorme, A., Palmer, J., Onton, J., Oostenveld, R., Makeig, S., et al.: Independent EEG sources are dipolar. PloS One **7**(2), e30135 (2012)
10. Eldar, Y.C., Oppenheim, A.V.: MMSE whitening and subspace whitening. IEEE Trans. Inf. Theor. **49**(7), 1846–1851 (2003)
11. Gramfort, A., et al.: MNE software for MEG and EEG data. Neuroimage **86**, 446–460 (2014)
12. Jutten, C., Herault, J.: Blind separation of sources, Part I: an adaptive algorithm based on neuromimetic architecture. Signal Process. **24**(1), 1–10 (1991)
13. Makeig, S., Bell, A.J., Jung, T.P., Sejnowski, T.J.: Independent component analysis of electroencephalographic data. In: NIPS, pp. 145–151 (1996)
14. Nunez, S.: Electric Fields of the Brain: The Neurophysics of EEG. Oxford University Press, New York (2006)
15. Onton, J., Delorme, A., Makeig, S.: Frontal midline EEG dynamics during working memory. Neuroimage **27**(2), 341–356 (2005)
16. Pham, D.T., Garat, P.: Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. IEEE Trans. Signal Process. **45**(7), 1712–1725 (1997)
17. Schaul, T., Zhang, S., LeCun, Y.: No more pesky learning rates. In: Proceedings of ICML Conference (2013)
18. Scherg, M., Berg, P.: Use of prior knowledge in brain electromagnetic source analysis. Brain Topogr. **4**(2), 143–150 (1991)