Likelihood-based population independent component analysis

ANI ELOYAN*, CIPRIAN M. CRAINICEANU, BRIAN S. CAFFO
Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA
aeloyan@jhsph.edu

SUMMARY
Independent component analysis (ICA) is a widely used technique for blind source separation, used heavily in several scientific research areas including acoustics, electrophysiology, and functional neuroimaging. We propose a scalable two-stage iterative true group ICA methodology for analyzing population level functional magnetic resonance imaging (fMRI) data where the number of subjects is very large. The method is based on likelihood estimators of the underlying source densities and the mixing matrix. As opposed to many commonly used group ICA algorithms, the proposed method does not require significant data reduction by a 2-fold singular value decomposition. In addition, the method can be applied to a large group of subjects since the memory requirements are not restrictive. The performance of our approach is compared with a commonly used group ICA algorithm via simulation studies. Furthermore, the proposed method is applied to a large collection of resting state fMRI datasets. The results show that established brain networks are well recovered by the proposed algorithm.

Keywords: Functional MRI; Signal processing; Source separation.

1. INTRODUCTION
Independent component analysis (ICA, Jutten and Herault, 1991) is a technique for separating mixed random signals (blind source separation) that assumes linear mixing and independence of the sources. ICA is commonly used in a variety of fields, including acoustics, electrophysiology, and neuroimaging. Particularly, in resting state (rs) functional connectivity (fc) functional magnetic resonance imaging (fMRI), it has become the standard tool for discovery, exploration, and modeling of brain networks. An active scientific discussion is ongoing as to the exact (patho-) physiological interpretation and importance of rs-fc-fMRI brain networks discovered via ICA. It is clear that large-scale implementations of ICA will be an important component of resolving these issues. Large-scale databases of rs-fMRI scans are becoming increasingly common, ideally allowing rs-fMRI to become a part of population-based research. As an example, the 1000 Connectome Project combines scans from several sites resulting in a database of over 1400 scans of healthy adults. In addition, the ADHD 200 dataset has rs scans of roughly 200 attention deficit hyperactive children and 500 typically developing control children where some of the children have several scans resulting in over 1000 fMRI scans. Moreover, the US National Institutes of Health has spearheaded the Human Connectome Project, a 30 million dollar venture to compile a comprehensive database of connectivity data,
including rs-fMRI. In addition to addressing the importance of such large-scale implementations of ICA to rs-fc-fMRI, our work generalizes many high-dimensional implementations of ICA.

ICA is an umbrella term that includes several different algorithmic implementations and theoretical foundations. At their core, the primary commonality of ICA algorithms is a linear factor analytic model (Harman, 1976) with the assumption of the independence of underlying factors. We focus on so-called noise-free ICA, a version of ICA that simply results in an “unmixing” (non-singular linear multiplication) of an input data matrix. This results in including the measurement error or other noise in the data as a part of the independent components (ICs). The estimation of the linear unmixing matrix involves iterative algorithms. A common starting point for all algorithms is a first-stage singular value decomposition (SVD) where dimension reduction is performed, after de-meaning, to avoid an overdetermined system. Thus, the data input to ICA have mean zero and are uncorrelated. Hence, Gaussian distributional assumptions provide little further insight into linear reorganizations. This motivates the search for solutions that are as non-Gaussian as possible. Hyvärinen and others (2001) and Comon and Jutten (2010) present an extensive overview of such algorithms. Most notably, Hyvärinen and Oja (1997) introduced a fast fixed point algorithm (called fastICA) for finding the independent signals by maximizing an approximation to the kurtosis. Cardoso (1990) introduced the JADE algorithm which is based on cumulant tensors. The Bell–Sejnowski algorithm (Bell and Sejnowski, 1995) finds maximum likelihood (ML) estimates of the underlying independent signals. Karvanen and Koivunen (2002) showed a Pearson system-based approach to ICA. Chen and Bickel (2006) presented a semiparametric approach to ICA by modeling the score functions of the underlying sources.

As in Bell and Sejnowski (1995), we focus on ML implementations of ICA, which require a fully specified likelihood. Eloyan and Ghosh (2011a) discuss parameter identifiability in ICA and present a set of sufficient conditions that ensure model identifiability. This manuscript builds on their work, extending it to high-dimensional applications, focusing on fMRI.

Calhoun and others (2001a) introduced the use of ICA for group inferences of fMRI data. The proposed algorithm is based on reducing the dimensions of the original images using an SVD and then applying fastICA to a data matrix obtained by concatenating subject-specific singular matrices. In related work, Beckmann and Smith (2005) presented tensorial extensions to ICA for group fMRI data analysis and Guo and Pagnoni (2008) provide an expectation-maximization (EM) algorithm-based ICA method for the case when there is a Gaussian noise term in the model. Another approach by Zhang and others (2010) is subject order-independent ICA. That is, the method is independent of the order in which subject-specific data matrices are concatenated after the SVD for analyses. Group ICA methods are widely used in the fMRI and EEG literature; for instance, Grin-Yatsenko and others (2010) used group ICA for analyzing EEG data at early stages of depressive disorder. An fMRI example was the motivation for this paper and is used for illustrating the proposed algorithm. However, the methods can be used for any other application where the underlying components can be assumed common across subjects and the mixing is subject specific.

We propose a two-stage ML algorithm for group ICA applicable to datasets where the number of subjects is very large. Following general assumptions in ICA models, we assume that each independent source is a vector of independent and identically distributed (i.i.d.) draws from a statistical distribution. The goal of ICA is to simultaneously estimate the parameters of these distributions as well as a non-singular linear transformation that “unmixes” the data matrix, whose columns are assumed to be “mixes” (linear combinations) of independent draws from the source distributions. The densities associated with the source distributions are modeled using finite mixtures of smooth densities.

In group ICA each subject has their own unmixing matrix while the ICs are assumed common across subjects. The subject-specific unmixing matrices are updated using an optimization algorithm. The method is based on iteratively updating the unmixing matrix of each subject separately, estimating the ICs and using the ICs to estimate the underlying source distributions. In fMRI, the result is common spatial statistical maps and subject-specific time courses. The proposed method is aimed at finding the ML estimates of the
unknown parameters. Hence, it is radically different from the commonly used fastICA, JADE, etc. where the goal is to maximize non-Gaussianity via an approximation of the kurtosis or negentropy (defined by Comon, 1994) of the underlying source densities. Our work in fMRI is most closely associated to that of Guo and Pagnoni (2008), who used Gaussian mixture models and an EM framework for estimation. We, however, formally handle identifiability constraints and use pre-specified means and variances for the sieve structure to find the number of mixture elements. In addition, our proposed method is easily generalized to handle large datasets.

Most methods developed for group ICA require high memory capacity, since the data for all subjects are loaded into the memory simultaneously. Furthermore, the observed image matrices from each subject are usually concatenated (the matrices are stacked together) to obtain a full matrix which requires additional memory. The additional analysis may be impossible for the resulting very high-dimensional full matrix. If the number of subjects is very large, then loading the concatenated matrix requires excessive memory and becomes implausible. A solution to the problem of dimensionality uses a two-stage SVD approach for dimension reduction. In this approach, a first-stage SVD is applied to the image for each subject, where a few vectors are retained before concatenating the matrices. Next, a second-stage SVD is applied to the now concatenated (over subjects) spatial singular vectors and the first few spatial singular vectors are retained to force a determined linear system for the noise-free group ICA model. The ICA algorithm is applied to the resulting (twice projected) data to find the ICs. In this standard approach to group ICA, one SVD is required by the algorithm, while another is done purely for computational convenience. If the number of subjects is very large, the concatenation of the matrices may not be computationally feasible, which limits direct applications of current group ICA methods to only a few subjects. Our methods are linear in the number of subjects and are scalable to high-dimensional data, because they require sequential access to subject-specific data instead of the entire group data matrix.

A specific important case where a large set of fMRI data were analyzed using group ICA is shown by Biswal and others (2010) in the 1000 Functional Connectome Project hosted on the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) website (http://www.nitrc.org/projects/fcon_1000). These data form our primary example application. Biswal and others (2010) describe the data collection and acquisition parameters from contributing sites and provide an analysis of rs fc. The dataset contains structural and rs-fMRI images for 1414 healthy adults collected independently from 35 centers worldwide. The scanner, time of repetition, total scan time, and other experimental parameters differ across sites. Hence, we focus on subsets of the data that share the same scanning characteristics. The largest number of subjects considered was 150, but being linear in the number of subjects the methods can easily be applied to thousands or tens of thousands of subjects. We label our method “likelihood-based population ICA” because of this scalability to populations of subjects.

Existing ICA applications, if applied directly without modification, are not applicable to such large (or even considerably smaller) collections of subjects. Typical strategies for handling large numbers of subjects employ stratification, resampling, and multistage procedures. In Biswal and others (2010), the data were stratified by gender, age, and data collecting site and group ICA was used to recover the ICs from each stratum separately. The results were combined by using a bootstrap approach. In a second example, Anderson and others (2011) showed an application of group ICA to large sample fMRI data where bootstrapped clustering and atlas-based methods were used for dimension reduction.

2. Methods

2.1 Likelihood-based group ICA for fMRI

We present the model in generality. Suppose that, for each subject, indexed by $i = 1, \ldots, I$, a $T \times V$-dimensional matrix $X_i$ is observed. Here, the columns, indexed by $v = 1, \ldots, V$, are the
independent observed repetitions of each mixture. The rows of each $X_i$, indexed by $t = 1, \ldots, T$, are the mixed signals. Contextually, $v$ represents voxels in an fMRI series, while $t$ indexes the scans.

When necessary, we use $X_i(t, v)$ to represent row $t$, column $v$ of $X_i$, and apply the same convention to other vectors and matrices. We assume that a group ICA decomposition implies the equation

$$X_i(t, v) = \sum_{q=1}^{Q} A_i(t, q) S(q, v), \quad (2.1)$$

for all $i = 1, \ldots, I$. Model (2.1) assumes that the spatio-temporal process, $X_i(t, v)$, for each subject $i$ can be decomposed into a finite sum of products between subject-specific time series $A_i(t, q)$ and spatial maps $S(q, v)$ that are common across subjects. This equation is equivalent to $X_i = A_i S$ or $X = AS$ where $X = [X_1^T \cdots X_I^T]^T$ and $A = [A_1^T \cdots A_I^T]^T$ are the $IT \times V$ and $IT \times Q$ matrices obtained by stacking the $X_i$ and $A_i$, respectively. Here, the $V$-dimensional vectors $S(q, \cdot)$ are the underlying ICs and $A_i(t, \cdot)$ are the subject-specific linear mixing vectors. In the context of fMRI, the $S(q, \cdot)$ are spatial maps that are often interpreted as brain networks while the $A_i$ are the subject-specific temporal mixing matrices.

The number of components $Q$ is primarily chosen in an ad hoc manner. In fMRI analysis, a common approach is to choose a high $Q$ and then discard the redundant components. After selecting the $Q$, principal component analysis (PCA) is applied to reduce the dimension of the matrix $X_i$ for $i = 1, \ldots, I$ to $Q \times V$. Hence, we can assume that the $A_i$ are square $Q \times Q$ matrices. In order for the ICA model to be fully identifiable, we assume that the square mixing matrices $A_i$ are of full rank and hence we define $W_i = A_i^{-1}$.

We define the densities of the underlying sources as $f_1, \ldots, f_Q$. That is, we assume that $\{S(q, v)\}_{q=1}^{Q}$ is a collection of i.i.d. draws from $f_q$ for all $q = 1, \ldots, Q$. Intrinsically, the spatial ICA (sICA) model used in the analysis assumes that the values of the ICs are independent draws from a density. The assumption of independence is discussed in more detail in single subject ICA literature (Comon, 1994). Most ICA algorithms (such as the commonly used fastICA, JADE, etc.) are based on this assumption. Calhoun and others (2001b) discuss the possible violations of independence of draws in task-based fMRI. They show that, for some task-based fMRI studies, this assumption fails and the results are contaminated heavily by the high dependence within the sources; however, for most of the task-based studies as well as for rs data the correlation between the voxels is negligible. Hence, the sICA is applicable for our example.

Group ICA makes the parsimonious assumption that the ICs are common across subjects, while how they mix can differ among subjects. This is exactly a standard ICA model on the data concatenated across subjects to obtain an $IT \times V$ matrix. Allowing for separate ICs and mixing matrices across subjects is equivalent to simply applying ICA separately to each subject and is not discussed here. Having separate ICs across subjects and a common mixing matrix is another possible parsimonious assumption. This is analogous to an ICA model for the data having concatenated subjects to obtain a $T \times IV$ matrix.

In the context of single-subject fMRI, assuming common spatial ICs and temporal subject-specific mixing matrices results in the so-called sICA (Calhoun and others, 2001b). Alternatively, assuming temporal ICs and spatial mixing matrices is often referred to as temporal ICA (tICA). As mentioned above, the assumptions on spatial or temporal independence in different fMRI experiments leading to the choice of using sICA or tICA have been widely discussed. Concatenating to obtain an $IT \times V$ matrix and hence using the sICA has been settled upon for group ICA analysis of rs-fMRI. For rs data, only this variation of concatenation is sensible, since subjects are spatially co-registered into a common template space, whereas they are not temporally registered. In other words, time 1 for subject 1 is not the same as time 1 for subject 2. We develop our method for group sICA; however, it can easily be modified to obtain the tICA model if necessary.

There is a technical consideration in that (2.1) is overdetermined. Hence, we first preprocess the data at the subject level by centering and whitening the observed matrices via an SVD and retaining only the
functions were used to model the model, noise in the data is absorbed into the estimated ICs and the mixing matrix. A similar approach by using mixture density estimates (MDE) introduced by Eloyan and Ghosh (2011b).

2.2 Density estimation in high dimensions

In the early literature on ML-based estimation of the independent sources, well-known distribution functions were used to model the \( f_q \) (Hyvärinen and others, 2001, p. 204). Boscolo and others (2004) suggested using kernel density estimation to model the underlying densities non-parametrically. We take a similar approach by using mixture density estimates (MDE) introduced by Eloyan and Ghosh (2011b).

To elaborate, we parameterize the density as follows:

\[
f_q(s) = \sum_{j=1}^{J_q} \theta_{qj} \frac{1}{\sigma_q} \phi\left( \frac{s - \mu_{qj}}{\sigma_q} \right),
\]

where \( \phi(\cdot) \) is the standard normal density function. Our treatment fixes the means \( \mu_{qj} \) and variance \( \sigma_q \) for each mixture density. We define the \( \mu_{qj} \) to lie on a grid of not necessarily equidistant points. To illustrate, consider a strict density estimation problem where \( S(q, \cdot) \) are observed with empirical mean 0 and variance 1.

As a starting set of fixed parameters for the mixture densities define the number of densities in the mixture as \( J_q = 1 + \frac{2}{3} \text{Range}_v(S(q, v)) \). The set of fixed means is given by \( \mu_{qj} = \min_v S(q, v) + ((j - 1)/(J_q - 1)) \text{Range}_v(S(q, v)) \) for \( j = 1, \ldots, J_q \), and the variance component \( \sigma_q^2 = 2(\mu_q V - \mu_{q1})/3(J_q - 1) \). The rationale behind these choices is to set the \( \sigma_q^2 \) such that \( \sigma_q = o(1) \) as \( J_q \to \infty \). Suppose \( \mathcal{M}_q = \{\mu_q 1 < \cdots < \mu_q J_q\} \) is the set of fixed means of the mixture densities. As the number of mixture densities \( J_q \) increases, the set \( \mathcal{M}_{q+1} \) is constructed by adding the median of one of the intervals \([\mu_q j, \mu_q j-1]\). Hence, the sequence \( \mathcal{M}_1, \mathcal{M}_2, \ldots \) maintains the sieve structure, i.e. each consecutive set contains the previous sets as subsets.

The weights of the mixture densities in (2.2) given by \( (\theta_{q1}, \ldots, \theta_{qJ_q}) \) are estimated using a constrained EM algorithm. The resulting density estimates satisfy the moment constraints required for full identifiability of the model given by

\[
E[S(q, \cdot)] = 0, \quad E[S(q, \cdot)^2] = 1, \quad \text{and} \quad 0 < E[S(1, \cdot)^3] < \cdots < E[S(q, \cdot)^3], \quad q = 1, \ldots, Q.
\]

The non-parametric estimation of the density of a vector \( S(q, \cdot) \), which has a large sample size (\( \sim 70,000 \) voxels in this case), can be computationally problematic. To address this issue, we propose a binning algorithm for the density estimation, essentially looking at the approximation to the histogram of the data. Choose \( p \ll V \) and suppose, for each IC, \( q \), the set \( \{r_{q1}, r_{q2}, \ldots, r_{qp}, r_{q,p+1}\} \) consists of the \( p + 1 \) quantiles of \( S(q, \cdot) \). Next, bins, \( B_{q1}, B_{q2}, \ldots, B_{qp} \), are constructed using the above quantiles as the endpoints for the bins. In addition, let \( c_q = (c_{q1}, c_{q2}, \ldots, c_{qp}) \) denote the counts of the observations in each of the bins. The underlying density of the components can be found by using the midpoints, say \( M_q = (M_{q1}, M_{q2}, \ldots, M_{qp}) \), of the bins and the counts by slightly modifying the proposed constrained EM algorithm. In other words, the updates of the mixture weights are computed as

\[
\hat{\theta}_{qj}^{(k+1)} = \frac{\sum_{i=1}^n w_{ij} \theta_{qj}^{(k)} c_q M_q}{\hat{\lambda}_1 + \hat{\lambda}_2 \mu_{qj} + \hat{\lambda}_3 \mu_{qj}^2},
\]
where the Lagrange multipliers are computed by the following system of equations:

\[
\begin{align*}
\sum_{j=1}^{N} \sum_{i=1}^{n} w_{ij}(\theta_q^{(k)}, c_q, M_q) &= 1,
\sum_{j=1}^{N} \sum_{i=1}^{n} \frac{w_{ij}(\theta_q^{(k)}, c_q, M_q)\mu_{qj}}{\lambda_1 + \lambda_2 \mu_{qj} + \lambda_3 \mu_{qj}^2} &= 0, \\
\sum_{j=1}^{N} \sum_{i=1}^{n} \frac{w_{ij}(\theta_q^{(k)}, c_q, M_q)\mu_{qj}^2}{\lambda_1 + \lambda_2 \mu_{qj} + \lambda_3 \mu_{qj}^2} &= 1 - \sigma_q^2,
\end{align*}
\]

(2.5)

where

\[
w_{ij}(\theta_q^{(k)}, c_q, M_q) = \frac{\theta_q^{(k)}(M_{qi} - \mu_{qj})/\sigma_q}{\sum_{j=1}^{J} \theta_q^{(k)}((M_{qi} - \mu_{qj})/\sigma_q)c_{qi}}.
\]

See Eloyan and Ghosh (2011b) for more details on the construction of the constrained EM algorithm.

The estimated densities maintain the constraints on the moments of the densities (mean and variance). By (2.3), it is stated that the ICs are ordered so that the third moments are in increasing order. This is necessary to avoid label switching issues, as the model is invariant to permutations of the ICs (provided the mixing matrix is permuted correspondingly). The ordered third moment assumption is a straightforward fix to this issue. Thus, within each iteration, the ICs are permuted to preserve the order of the estimated third moments.

### 2.3 Semiparametric iterative algorithm for group ICA

Based on the density estimation algorithm described in Section 2.2, we develop a semiparametric iterative algorithm for true group ICA model (2.1).

By the independence of the underlying sources \(S(q, \cdot)\), the likelihood of the unmixing matrix \(W = [W_1^T \cdots W_I^T]^T\) is given by

\[
L(W, f) = \prod_{i=1}^{I} \prod_{v=1}^{V} \prod_{q=1}^{Q} f_q \left( \sum_{l=1}^{Q} w_{iqlxlv} \right) |\det W_i|,
\]

(2.6)

where \(f = (f_1, \ldots, f_Q)\). Assuming that the densities of the underlying sources are estimated using the finite mixtures of Gaussian densities defined as \(\hat{f}_1, \ldots, \hat{f}_Q\), the loglikelihood function of the unmixing matrix \(W\) can be constructed analytically as

\[
I(W, \hat{f}) = \sum_{i=1}^{I} \left[ \sum_{v=1}^{V} \sum_{q=1}^{Q} \log \left\{ \hat{f}_q \left( \sum_{l=1}^{Q} w_{iqlxlv} \right) \right\} + V \log |\det W_i| \right],
\]

(2.7)

where

\[
\hat{f}_q(s) = \sum_{j=1}^{J} \hat{\theta}_{qj} \phi \left( \frac{s - \mu_{qj}}{\sigma_q} \right) \frac{1}{\sigma_q}.
\]

Note that, by construction of the densities of the underlying sources, the derivative and Hessian matrices of the loglikelihood can also be found analytically.
We need to obtain starting values for the unmixing matrices to start the algorithm. This can be done by choosing the subject-specific unmixing matrices given by $\hat{W}_i^{(0)}$. Alternatively, we can find starting values that satisfy the condition that the underlying ICs are the same for all subjects. We can use the population value decomposition (Crainiceanu and others, 2011) of the full matrix $X = [X_1 \cdots X_I]^T$ given by

$$X = U\Sigma V^T.$$  

(2.8)

The starting values of the $W_i$ then are chosen as the $r$th block of the rows of $U\Sigma$.

Even though the derivative and Hessian matrices of the loglikelihood can be computed analytically, we can also use an approximation to the Hessian given by $l''(w) \approx l'(w)l'(w)^T$ to obtain a more robust optimization algorithm.

The iterative high dimensional ICA (HDICA) algorithm for finding the ML estimate of the unmixing matrices $W_1, \ldots, W_I$ can be presented succinctly as follows:

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Let $S_i^{(M)} = W_i^{(M-1)}X_i$, for each $i = 1, \ldots, I$</td>
</tr>
<tr>
<td>2.</td>
<td>For each IC $q$ construct the set of midpoints $M_{q1}, M_{q2}, \ldots, M_{qp}$ of the bins and the corresponding counts $c_{q1}, c_{q2}, \ldots, c_{qp}$.</td>
</tr>
<tr>
<td>3.</td>
<td>For each $q = 1, \ldots, Q$ construct the set of means $M_{q(M)} \supseteq M_{q(M-1)}$ and the variance component $\sigma_q$.</td>
</tr>
<tr>
<td>4.</td>
<td>By using MDE estimate $(\theta_{q1}^{(M)}, \ldots, \theta_{qL}^{(M)})$.</td>
</tr>
<tr>
<td>5.</td>
<td>For each $i = 1, \ldots, I$ compute the gradient $l' (\hat{W}^{(M)}_i)$ and Hessian matrix $l'' (\hat{W}^{(M)}_i)$.</td>
</tr>
<tr>
<td>6.</td>
<td>For each $i = 1, \ldots, I$ update the unmixing matrix $\hat{W}^{(M+1)}_i = \hat{W}^{(M)}_i - l'' (\hat{W}^{(M)}_i)^{-1}l' (\hat{W}^{(M)}_i)$.</td>
</tr>
<tr>
<td>7.</td>
<td>$\delta = \max</td>
</tr>
</tbody>
</table>

One of the striking differences of our method compared with other group ICA algorithms is that, at each iteration, only one $Q \times V$ dimensional subject-specific matrix is loaded in the memory to compute the update for $W$. In addition, the densities are estimated using a binning algorithm and hence the increase in the sample size does not affect the speed of the density estimation part of the algorithm.

3. Results

3.1 Simulation results

To illustrate the performance of the proposed method, we conducted simulation studies where data were generated using distributions of different shapes. Three different cases were used in the study. The results are compared with the commonly used group ICA algorithm by Calhoun and others (2001a), which is based on fastICA (Hyvärinen and Oja, 1997).

Suppose that the number of subjects is $I = 3$. First, suppose that the number of underlying sources $Q = 2$. The data are generated by the ICA model $X_i = A_i S$ with $T = 2$ and $V = 2000$. We further assume that

$$A_1 = \begin{pmatrix} 0.75 & 0.25 \\ 0.5 & -0.5 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 0 \\ 0.5 & -0.5 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 0.5 \\ 0.75 & 1 \end{pmatrix},$$

with det($A_1$) = -0.5, det($A_2$) = -0.5, and det($A_3$) = 0.625. The shapes of the densities used for generating values of the ICs are shown in Figure 1. For each of the densities in Figure 1, two ICs are
Population independent component analysis

<table>
<thead>
<tr>
<th>Distribution</th>
<th>HDICA FICA</th>
<th>HDICA FICA</th>
<th>HDICA FICA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laplace</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gamma</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bimodal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weibull</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multimodal</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1. The boxplots of the Amari errors computed by using the proposed HDICA and fastICA for \( Q = 2 \). Each plot corresponds to one run of 100 simulations. The densities used for generating the two ICs are shown above each plot. Each boxplot shows the distribution of the 100 Amari errors obtained by HDICA or fastICA, respectively. The shapes of the densities used to generate data for the underlying sources for simulation studies are shown on the right ((a)–(f)).

generated resulting in six sets of simulation runs. For instance, in case (a) the rows of matrix \( S \) are generated by \( S(1, \cdot) \sim \text{Gamma}(4, 0.25) \) and \( S(2, \cdot) \sim \text{Gamma}(2, 2) \). Finally, the individual mixture matrices \( X_i \) are constructed by using \( A_1, A_2, \) and \( A_3 \) above as \( X_i = A_i S \).

The Amari error (introduced by Amari, 1998) is used for evaluating the performance of the ICA methods. It is given by

\[
AE(A, \hat{W}) = \frac{1}{2Q} \sum_{i=1}^{Q} \left( \sum_{j=1}^{Q} \frac{|p_{ij}|}{\max_k |p_{ik}|} - 1 \right) + \frac{1}{2Q} \sum_{j=1}^{Q} \left( \sum_{i=1}^{Q} \frac{|p_{ij}|}{\max_k |p_{kj}|} - 1 \right),
\]

where \( P = A \hat{W} \) and \( \hat{W} \) is the estimated unmixing matrix (normalized to have column norm of 1) for each subject by each method. The Amari error is invariant to matrix permutations and sign changes, and is a value between 0 and \( Q - 1 \). The Amari error can be computed for each subject in a single simulation run. We obtain the common Amari error as the mean of the errors computed for all three subjects in this case.

For each of the six sets of simulations, fastICA and our proposed HDICA are applied to estimate the unmixing matrices \( \hat{W}_i \). The boxplots of the 100 Amari errors (each as an average for three subjects) computed for each case by using fastICA and HDICA are shown in Figure 1. When the distributions of the
underlying sources are symmetric and unimodal \((t, \text{ Laplace})\), we observe that our method is competitive with fastICA. However, for the other distributions our method seriously outperforms fastICA in terms of minimizing the Amari error.

For the second scenario suppose that the number of underlying sources is \(Q = 4\) with \(V = 5000\) voxels. Again, the number of subjects is \(I = 3\). The mixing matrices used are given by

\[
A_1 = \begin{pmatrix}
2 & 1 & 2 & 3 \\
3 & 3 & 1 & 0.5 \\
1 & 2 & 2 & 4 \\
4 & 3 & 2 & 1
\end{pmatrix}, \quad A_2 = \begin{pmatrix}
2 & 3 & 2 & 1 \\
3 & 4 & 1 & 0.5 \\
3 & 2 & 3 & 4 \\
2 & 3 & 3 & 1
\end{pmatrix}, \quad A_3 = \begin{pmatrix}
1 & 2 & 2 & 1 \\
3 & 4 & 1 & 0.5 \\
3 & −1 & 3 & 4 \\
2 & 1 & 3 & 1
\end{pmatrix},
\]

with \(\det(A_1) = −8\), \(\det(A_2) = 7.5\), and \(\det(A_3) = 47\). The densities used for generating the values of the ICs \(S_q, q = 1, \ldots, Q\) for four different cases are chosen as follows:

<table>
<thead>
<tr>
<th>Case</th>
<th>IC 1</th>
<th>IC 2</th>
<th>IC 3</th>
<th>IC 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>(\text{Laplace}(\mu, \sigma))</td>
<td>(\text{Laplace}(\mu, \sigma))</td>
<td>(\text{Laplace}(\mu, \sigma))</td>
<td>(\text{Laplace}(\mu, \sigma))</td>
</tr>
<tr>
<td>(b)</td>
<td>(\text{Gamma}(a, b))</td>
<td>(\text{Gamma}(a, b))</td>
<td>(\text{Laplace}(\mu, \sigma))</td>
<td>(\text{Laplace}(\mu, \sigma))</td>
</tr>
<tr>
<td>(c)</td>
<td>(\text{Gamma}(a, b))</td>
<td>(\text{Gamma}(a, b))</td>
<td>(\text{MixNorm}(\mu, \sigma))</td>
<td>(\text{MixNorm}(\mu, \sigma))</td>
</tr>
<tr>
<td>(d)</td>
<td>(t \ (d.f))</td>
<td>(\text{Gamma}(a, b))</td>
<td>(\text{Weibull}(a, b))</td>
<td>(\text{MixNorm}(\mu, \sigma))</td>
</tr>
</tbody>
</table>

The Amari errors are computed as above. The boxplots of the Amari errors for each of the four cases computed using the resulting estimates by HDICA and fastICA are presented in Figure 2. Here again, the performance of HDICA is comparable with that of fastICA using the Amari error criterion. In case (a), where the true underlying densities have Laplace distribution, the results obtained by HDICA are similar to that of fastICA. As has been observed before, when the underlying sources have non-symmetric or multimodal densities as in (b)–(d), HDICA is outperforming fastICA in minimizing the Amari error.
Fig. 3. The boxplots of the Amari errors (in the left) computed by using the proposed HDICA and fastICA for $Q = 8$. The plot corresponds to one run of 100 simulations. The densities used for generating the eight ICs are presented in Figure 1. Each boxplot shows the distribution of the 100 Amari errors obtained by HDICA or fastICA, respectively. The scatterplots (center and right) for two components showing the difference of the estimated components from the true data in one of the simulation runs with $Q = 8$. The x-axis shows the true component and the y-axis shows the absolute difference of the true component with the estimated component (by HDICA (red diamonds) and by fastICA (blue circles)).

Finally, suppose the number of components $Q = 8$ with $V = 5000$ voxels. The number of subjects is $I = 3$. Each of the eight ICs was generated by using one of the distributions given in Figure 1, at least one IC was generated using each of the distributions. The elements of the mixing matrices $A_1, A_2, A_3$ are generated before running the simulation study by using independent standard normal draws. The boxplots of the Amari errors obtained by using fastICA and the proposed HDICA for 100 simulation runs are shown in Figure 3. Here again, our method performs significantly better than fastICA.

The scatterplots in Figure 3 show the absolute difference of the true ICs with the estimates found by using HDICA (in red diamonds) and fastICA (in blue circles) for one of the datasets generated for the last simulation example with $Q = 8$. Two of the components generated using Gamma (center) and multimodal (right) densities are plotted. As can be seen from this figure, the ICs are almost completely recovered by HDICA. In other words, the absolute difference of the estimate by HDICA with the truth is less variable than the absolute difference of the estimate found by fastICA with the truth.

The simulations suggest that the proposed method, including data reduction steps that remove information, results in an algorithm that is as good as the most popular group ICA algorithm. More importantly, however, is that the HDICA algorithm is immediately scalable to hundreds or thousands of subjects.

### 3.2 Application to the 1000 functional connectomes fMRI dataset

To illustrate our algorithm, we use one of the freely available large fMRI datasets. The 1000 Functional Connectomes Project dataset (Biswal and others, 2010) is aimed at discovering the science behind brain function. As discussed in Section 1, applying the current group ICA methods to a large number of subjects is computationally impossible for regular computers. As discussed, the motivation of this paper is to develop a group ICA algorithm that can be applied to use for any number of subjects without excessive data reduction steps. Because there are several covariates in these data, we illustrate our method for a subset of 150 subjects from the NITRC collection that have similar characteristics. To apply the model to the complete dataset, further analysis is needed to exclude possible effects, such as those associated with the data collection site. Our method does not require the concatenation of the subject-specific data matrices. Moreover, at each step of the algorithm only one of the subject-specific data matrices is loaded into the memory. The HDICA algorithm is linear in the number of subjects and can be scaled up relatively easily.
Fig. 4. Motor (top row) and default (second row) networks computed for $I = 50$ subjects using HDICA and fastICA. The white (red) color corresponds to the highest intensities in the brain network followed by the black (blue) as the intensities decrease. 3D representations of motor (third row) and visual (bottom row) networks computed for $I = 50$ subjects using the proposed HDICA algorithm. The networks (in black/red) are overlaid on a gray template.

The scans are collected when the subjects are in the rs. Each dataset is a 4D array of intensities. Each subject was in the scanner for 2.2–20 min. For the subset used in this analysis, the number of time points was $T = 123$. Note that even if the number of time points varies among the subjects, the algorithm can still be applied, since the first PCA step will reduce the dimensions of the datasets to the same $Q$. The scans are collected using a 3T scanner. Standard image processing was performed to register the data to the MNI standard brain space. However, no smoothing was done on the data before applying group ICA.

The MNI152_T1_3mm_brain_mask.nii mask was used to extract the background of the images and obtain the voxels that are in the actual brain. For each time point, the 3D array is vectorized to obtain a $V$-dimensional vector of intensities that are then concatenated over time. Hence, we obtain a $T \times V$-dimensional matrix $X_i$ for each subject. The HDICA algorithm is then applied using these $X_i$ matrices.

We first chose a subset of $I = 50$ subjects to compare the performance of HDICA and fastICA. Following Biswal and others (2010), we use group ICA to obtain $Q = 20$ components. The slices of two ICs overlaid on the brain template are shown in Figure 4. To find the brain network corresponding to one of the ICs, first the IC vector is back-transformed to the 3D space. Then, the image is thresholded to identify the voxels that have highest (lowest) loadings on the component (e.g. the voxels in the top 5% of the absolute values of IC). The intensities of all other voxels are set to 0. The resulting image is plotted in Figure 4 overlaid on the template brain. The default mode, motor, and visual networks are the most robust networks generally uncovered by rs ICA. We demonstrate them here to show the clear boundaries and sharpness of the results using our variation of ICA on a large collection of subjects. In addition, we did not use smoothing before applying ICA. In most cases, smoothing is used first as one of the preprocessing steps before
Fig. 5. Motor (top) and default (bottom) networks computed for $I = 150$ subjects using the proposed HDICA algorithm. The white (red) color corresponds to the highest intensities in the brain network followed by the black (blue) as the intensities decrease.

applying ICA. We show that, even without smoothing, our algorithm still obtains ICs with sharp boundaries. The correlations of pairs of ICs estimated by fastICA and HDICA that have the highest correlations among all pairs of thresholded ICs were computed. The following is a list of these correlations for some of the brain networks: 0.73 for the visual network, 0.61 for the default network, 0.56 for the frontal cortex, 0.5 for the motor network, and 0.37 for the dorsal attention network.

We observe that the IC maps of the chosen components correspond to well-known brain networks. We also show the 3D renderings of the motor and visual networks estimated by HDICA in Figure 4. Again, we observe sharp boundaries of the estimated networks.

Finally, HDICA was used to compute $Q = 20$ ICs by using fMRI data of $I = 150$ subjects from the Cambridge site (a subset of the NITRC data). The computations took about 5 days on a regular PC. Figure 5 shows two of the ICs overlaid on the brain template. Again, the motor network and the default network are obtained by HDICA.

4. DISCUSSION

In this paper, we present a group ICA algorithm based on non-parametric estimation of the densities of the underlying sources using finite mixtures of continuous densities. The mixing matrix is simultaneously estimated using an iterative optimization algorithm. The proposed algorithm is scalable to large datasets. As a byproduct of the algorithm, we obtain the estimates of the densities of the underlying spatial maps.

We first develop a density estimation method based on binning the data and using a mixture of continuous densities for approximating the histogram of the data. The density estimates are computed using a constrained EM algorithm to satisfy moment constraints for the identifiability of the model. The estimated densities are then used to model the distributions of the underlying spatial maps.

The performance of the proposed algorithm is presented by simulation studies showing that our method performs at least as well as another commonly used method. The algorithm was applied to a set of rs-fMRI data. The method can be used for large groups of fMRI data in different subpopulations to obtain the brain networks and study differences within the subpopulations.
One of the issues in the use of group ICA for fMRI data analysis is the lack of application of group ICA for a large set of subjects. The proposed algorithm can be extended to any number of subjects. The proposed algorithm can be applied to find the brain networks using a larger subset of the NITRC 1000 dataset by, for instance, matching the subjects by center, gender, or other covariates. As a result, population level brain networks can be found as reference networks.

A discussion is pending on the choice of the thresholding of the ICs. The thresholding is usually done by choosing a quantile of the distribution of the ICs and thresholding the values at the quintile. The distributions obtained by HDICA can be used for finding a better estimate of the thresholding value.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

The project described was supported by Grant Number R01EB012547 from the National Institute of Biomedical Imaging and Bioengineering, Grant Number R01NS060910 from the National Institute of Neurological Disorders and Stroke, and NIH Grant Number P41 EB015909 from the National Institute of Biomedical Imaging and Bioengineering.

REFERENCES


Population independent component analysis


[Received October 11, 2011; revised October 15, 2012; accepted for publication November 23, 2012]