Biometrics 70, 224–236
March 2014

DOI: 10.1111/biom.12111

An Evaluation of Independent Component Analyses with an Application to Resting-State fMRI

Benjamin B. Risk,[1,*] David S. Matteson,[1] David Ruppert,[1] Ani Eloyan,[2] and Brian S. Caffo[2]

[1]Department of Statistical Science, Cornell University, 301 Malott Hall, Ithaca, New York, U.S.A.
[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, U.S.A.
[*]*email:* bbr28@cornell.edu

Summary. We examine differences between independent component analyses (ICAs) arising from different assumptions, measures of dependence, and starting points of the algorithms. ICA is a popular method with diverse applications including artifact removal in electrophysiology data, feature extraction in microarray data, and identifying brain networks in functional magnetic resonance imaging (fMRI). ICA can be viewed as a generalization of principal component analysis (PCA) that takes into account higher-order cross-correlations. Whereas the PCA solution is unique, there are many ICA methods–whose solutions may differ. Infomax, FastICA, and JADE are commonly applied to fMRI studies, with FastICA being arguably the most popular. Hastie and Tibshirani (2003) demonstrated that ProDenICA outperformed FastICA in simulations with two components. We introduce the application of ProDenICA to simulations with more components and to fMRI data. ProDenICA was more accurate in simulations, and we identified differences between biologically meaningful ICs from ProDenICA versus other methods in the fMRI analysis. ICA methods require nonconvex optimization, yet current practices do not recognize the importance of, nor adequately address sensitivity to, initial values. We found that local optima led to dramatically different estimates in both simulations and group ICA of fMRI, and we provide evidence that the global optimum from ProDenICA is the best estimate. We applied a modification of the Hungarian (Kuhn-Munkres) algorithm to match ICs from multiple estimates, thereby gaining novel insights into how brain networks vary in their sensitivity to initial values and ICA method.

Key words: Group ICA; Hungarian algorithm; Nonconvex optimization; Permutation problem; ProDenICA; Stability analysis.

1. Introduction

In independent component analysis (ICA), multivariate observations are linearly transformed to minimize dependencies between variables resulting in so-called independent components (ICs). The goal of ICA is to identify both the mixing matrix and the ICs, and the problem is not identifiable if more than one component has a Gaussian distribution (Comon, 1994). ICA has diverse applications including artifact removal in electrophysiology (Iriarte et al., 2003), extracting gene expression features in microarray data (Kong et al., 2008), facial recognition (Bartlett, Movellan, and Sejnowski, 2002), and separating mixed audio signals (Bell and Sejnowski, 1995). In addition, it has been used in thousands of studies to identify brain networks from functional magnetic resonance imaging (fMRI) (Beckmann, 2012). In fMRI studies, the blood oxygen level dependent (BOLD) signal is an aggregate measure of neural activity across many brain networks that is measured across time. In spatial ICA, the BOLD signal is decomposed into a mixing matrix containing the temporal loadings of ICs and into ICs representing spatial networks. The spatial networks may capture distinct functionalities (e.g., somatomotor, auditory, or visual network), physiological processes (e.g., breathing, heart-beating), and/or artifacts (e.g., head movement) (Damoiseaux et al., 2006).

Networks and their loadings estimated via an ICA contribute to our understanding of the human brain. Recently, there has been a collaborative effort to make a large amount of resting-state fMRI (rs-fMRI) publicly available (Biswal et al., 2010). The BOLD signals in rs-fMRI are measured in subjects who are assigned no particular task, which contrasts with experimental (i.e., task-based) fMRI. Group ICA can be used to combine data from hundreds of subjects (Calhoun et al., 2001). Coupled with basic biological assumptions regarding spatial contiguity of networks and association with paradigm-related fMRI, group ICA can greatly facilitate the evaluation of resting-state brain networks. The resulting weight matrices of group ICA are often used in inference, for example to compare diseased and non-diseased populations. ICA has been used to identify abnormalities and biomarkers of disorders including Alzheimer's disease (Celone et al., 2006), major depression (Veer et al., 2010), and schizophrenia (Jafri et al., 2008). ICA will likely play an integral role in the Human Connectome Project, which seeks to create a database of all neurological pathways to further our understanding of disease, brain development, and aging (Beckmann, 2012). Many ICA methods exist, and disentangling the differences between methods could improve the ability to use ICA for clinical application and biomarker development.

ICA is a semiparametric problem with a finite dimensional matrix parameter and infinite dimensional IC distributions. Since the IC distributions are latent, one challenge is to find an estimator that is accurate for a wide variety of source distributions. Parametric ICA methods such as information maximization (Infomax) (Bell and Sejnowski, 1995) and FastICA (Hyvarinen, 1999) assume a parametric source distribution and/or properties of higher order moments to derive comparatively simple algorithms. Infomax assumes a distribution (typically logistic), and FastICA assumes a quasi-likelihood function (typically the negative of the hyperbolic cosine). Both methods are commonly used in fMRI studies in part because they are fast even for large datasets. Although these algorithms work well for a variety of IC distributions, a large mismatch between the assumed densities and the true densities can result in inconsistent estimates of ICs (Cardoso, 1998). A semiparametric approach to modeling ICs, product density ICA via tilted Gaussians (ProDenICA), outperformed FastICA in simulations for a large class of IC distributions (Hastie and Tibshirani, 2003). Other methods using nonparametric estimation of the IC densities have also been developed (Chen and Bickel, 2006; Eloyan and Ghosh, 2013). These are similar to ProDenICA but typically computationally more expensive.

From a biological perspective, a voxel (volumetric pixel) might be a non- or primary contributor to a network, suggesting the use of mixture distributions for some networks in spatial ICA of fMRI (Guo, 2011). Two to three mixtures of normals for an IC has been found to work well in task-based fMRI, where voxels can be regarded as activated by the experiment or their fluctuations may correspond to background noise (Beckmann and Smith, 2004). Simulation studies with two ICs found that FastICA performed poorly when densities were a mixture of normals, while semiparametric methods performed well (Hastie and Tibshirani, 2003; Eloyan and Ghosh, 2013). However, the performance of FastICA, Infomax, and ProDenICA has not been evaluated in dimensions typically found in fMRI applications (e.g., 20 components). Moreover, ProDenICA has not been applied to ICA of fMRI. This suggests a need to determine whether ProDenICA outperforms FastICA and Infomax in simulations with higher dimensions and whether ProDenICA differs from other methods when applied to fMRI.

ICA methods require nonconvex optimization, yet fMRI toolboxes that address the issue of sensitivity to initial values are problematic, and most statistical packages do not address the issue whatsoever. A method called Icasso uses agglomerative hierarchical clustering on absolute correlations to match ICs from multiple starting values (Himberg, Hyvärinen, and Esposito, 2004), and the centroids of tight clusters from multiple initializations are regarded as the best estimates for reliable ICs from fMRI (Correa, Adali, and Calhoun, 2007). ICs that do not tightly cluster spatially are excluded from subsequent analyses. Consequently, this approach may mistakenly exclude ICs from analyses of neurological disorders simply because they have local optima.

The global maximum usually corresponds to the best estimate of the true mixing matrix when an ICA method is statistically consistent (e.g., Matteson and Tsay 2013); unfortunately, some ICA methods are not consistent for many IC distributions. For the FastICA estimator, the set of local maxima of the expected value of the objective function contains the true unmixing matrix under certain conditions relating the true and hypothesized densities, which is referred to as *local consistency* (Hyvarinen, 1999). However, the FastICA objective function does not provide a way to identify which local maximum corresponds to a consistent estimator (if one exists). To investigate, we simulate distributions with large samples sizes and present examples where a local maximum corresponds to the true unmixing matrix but the global maximum is spurious. We also conduct simulations with five, ten, and twenty components and show none of the local optima of FastICA or Infomax that we located are close to the true unmixing matrix, thereby identifying reasonable distributions for which these estimators perform poorly.

It is well known that ICs are only identifiable up to scaled permutations, which is sometimes called the *permutation problem*. As a result, ICs from different initializations or methods are difficult to compare. In contrast to fMRI studies relying upon Icasso (e.g., Correa et al., 2007) or upon matching by highest absolute correlation (e.g., Guo, 2011), we optimally match components from different methods via a modification of the Hungarian (Kuhn-Munkres) algorithm (Tichavsky and Koldovsky, 2004). This allows a more detailed comparison of ICs within each method that vary due to initialization, as well as a comparison of ICs between methods that vary due to their assumptions and dependency measures.

To quantify the practical impacts of initialization and choice of methodology, we consider a large collection of rs-fMRI data from multiple data collection centers worldwide on children and adolescents with Attention Deficit Hyperactive Disorder (ADHD) (Milham et al., 2012). This data was made publicly available in a competition on automated diagnosis of ADHD, and two of the authors of this article were part of the declared winning team (Eloyan et al., 2012). Here, we use the dataset as a source of multi-subject, multi-site rs-fMRI. We use the group ICA of Calhoun et al. (2001), which is easily adapted to any ICA algorithm and to multi-site rs-fMRI. We evaluate the impact of initial values and compare the mixing matrices and group ICs estimated using FastICA, Infomax, joint approximate diagonalization of eigenmatrices (JADE; Cardoso and Souloumiac, 1993), and ProDenICA.

In Section 2, we describe the noise-free ICA model and characterize the objective functions used by FastICA, Infomax, JADE, and ProDenICA. We formalize group ICA as a noisy ICA model with a known number of components, and then we discuss a canonical ordering of ICs and the matching algorithm. In Section 3, we demonstrate the existence of spurious global optima in the FastICA and Infomax objective functions—but not ProDenICA—for simulations with large sample sizes and two components. We also show that the FastICA, Infomax, and ProDenICA algorithms are sensitive to initial values for 5, 10, and 20 components, and that ProDenICA is the most accurate. In Section 4, we conduct a group ICA of the ADHD-200 sample using the four methods. In Section 5, we conclude that multiple starting values are necessary and that ProDenICA may be more reliable in fMRI studies.

2. ICA Methods

2.1. *The Noise-Free ICA Model*

Let $\mathbf{Z}_v$ be a random vector in $\mathbb{R}^Q$ with finite second moments. Without loss of generality, assume $\mathrm{E}\,\mathbf{Z}_v = 0$ and $\mathrm{E}\,\mathbf{Z}_v\mathbf{Z}_v' = \mathbf{I}$, where $\mathbf{Z}_v'$ is the transpose of $\mathbf{Z}_v$. Let the mixing matrix, $\mathbf{A}$, be a $Q \times Q$ matrix of full rank, and denote the unmixing matrix as $\mathbf{W}$, which is equal to $\mathbf{A}^{-1}$. Let $\mathbf{S}_v \in \mathbb{R}^Q$ be a random vector in which the components are mutually independent with $\mathrm{E}\,\mathbf{S}_v = 0$ and $\mathrm{E}\,\mathbf{S}_v\mathbf{S}_v' = \mathbf{I}$. The noise-free ICA model is

$$\mathbf{Z}_v = \mathbf{A}\mathbf{S}_v. \tag{1}$$

We observe V identically distributed samples of $\mathbf{Z}_v$. Then the goal is to estimate $\mathbf{W}$, which we can then use to estimate the ICs. We briefly describe four methods to estimate $\mathbf{W}$ below.

2.2. *Mutual Information, Maximum Likelihood, and Infomax ICA*

Minimization of mutual information (MI) provides a unifying framework for a variety of ICA methods, including maximum likelihood (ML), Infomax, and negentropy (Cardoso, 1997, 1998). MI measures the Kullback–Leibler divergence between a joint density (assumed to be known) and the product of its marginal densities. Let $F_{\mathbf{S}}$ denote the joint distribution of a random vector $\mathbf{S} \in \mathbb{R}^Q$, and suppose $F_{\mathbf{S}}$ is absolutely continuous with density $f_{\mathbf{S}}(s)$. Let F_{S_q} denote the marginal distribution of the qth component of $\mathbf{S}$ and $f_{S_q}(s)$ the corresponding density. Let $\Theta = \{\mathbf{s} \in \mathbb{R}^Q : f_{\mathbf{S}}(\mathbf{s}) > 0\}$. MI is defined as

$$\mathcal{K}\left(F_{\mathbf{S}}; \prod_{q=1}^{Q} F_{S_q}\right) = \int_{\mathbf{s}\in\Theta} \log\left(\frac{f_{\mathbf{S}}(\mathbf{s})}{\prod_{q=1}^{Q} f_{S_q}(s_q)}\right) f_{\mathbf{S}}(\mathbf{s})\mathrm{d}\mathbf{s}. \tag{2}$$

Then, $S_1, \ldots, S_Q$ are mutually independent if and only if their MI is equal to zero.

Suppose we have the noise-free ICA model in (1) with $\mathbf{W}$ denoting the true unmixing matrix and $F_{\mathbf{S}} = \prod_{q=1}^{Q} F_{S_q}$. Let $\mathcal{O}$ be the set of $Q \times Q$ orthogonal matrices, and let $\mathcal{P}$ be the set of $Q \times Q$ signed permutation matrices. Define the equivalence relation $\mathbf{A} \cong \mathbf{B}$ if there exists some $\mathbf{P}_{\pm} \in \mathcal{P}$ such that $\mathbf{A} = \mathbf{P}_{\pm}\mathbf{B}$. Then,

$$\mathbf{W} \cong \operatorname*{argmin}_{\mathbf{O}\in\mathcal{O}} \mathcal{K}\left(F_{\mathbf{OZ}}; \prod_{q=1}^{Q} F_{\mathbf{o}_q'\mathbf{Z}}\right),$$

where $\mathbf{o}_q'$ is the qth row of $\mathbf{O}$. Let $\mathcal{H}(\mathbf{S})$ denote the differential entropy,

$$\mathcal{H}(\mathbf{S}) = -\int_{\mathbf{s}\in\mathbb{R}^Q} \{\log f(\mathbf{s})\} f(\mathbf{s})\mathrm{d}\mathbf{s}, \tag{3}$$

and note that the MI is equal to the sum of the marginal entropies less the joint entropy, $\mathcal{K}\left(F_{\mathbf{S}}; \prod_{q=1}^{Q} F_{S_q}\right) = \sum_{q=1}^{Q} \mathcal{H}(S_q) - \mathcal{H}(\mathbf{S})$.

If the true joint density of the ICs is known, we can define the objective function for identically distributed observations $\mathbf{z}_1, \ldots, \mathbf{z}_V$ as

$$\mathcal{J}_{MI}(\mathbf{O}) = -\sum_{v=1}^{V}\sum_{q=1}^{Q} \log f_{S_q}(\mathbf{o}_q'\mathbf{z}_v) + \sum_{v=1}^{V} \log f_{\mathbf{S}}(\mathbf{O}\mathbf{z}_v).$$

Since $\sum_{v=1}^{V} \log f_{\mathbf{S}}(\mathbf{O}\mathbf{z}_v)$ is invariant to rotations $\mathbf{O}$, we obtain

$$\widehat{\mathbf{W}} = \operatorname*{argmin}_{\mathbf{O}\in\mathcal{O}} -\sum_{v=1}^{V}\sum_{q=1}^{Q} \log f_{S_q}(\mathbf{o}_q'\mathbf{z}_v). \tag{4}$$

From (4), it is clear that the MI criterion is equal to the negative of the ML criterion.

In practice, the densities f_{s_q} are not known, so most ML ICA methods assume a parametric density $f_{s_q}^*$. In particular, the Infomax criterion is equal to the ML criterion in which the information transfer function described in Bell and Sejnowski (1995) equals the (assumed) common cumulative distribution function of the ICs (Cardoso, 1997). The information transfer function is most commonly taken to be the logistic distribution, $F_{s_q}^*(x) = 1/(1 + e^{-x})$ for $x \in \mathbb{R}$, and $\widehat{\mathbf{W}}$ is not restricted to $\mathcal{O}$ (Bell and Sejnowski, 1995). Let $\mathcal{B}$ be the set of full rank $Q \times Q$ matrices, and let $\mathbf{B} \in \mathcal{B}$ with rows $\boldsymbol{b}_q'$. Then the infomax objective function is

$$\mathcal{J}_{Info}(\mathbf{B}) = V\log|\det\mathbf{B}| + \sum_{v=1}^{V}\sum_{q=1}^{Q}\{\log f_s^*(\boldsymbol{b}_q'\mathbf{z}_v)\}.$$

The wrong $F_{\mathbf{S}}^*$ may still result in a consistent estimator of $\mathbf{W}$ and successfully recover ICs from a variety of distributions, although the use of $F_{\mathbf{S}}^* \neq F_{\mathbf{S}}$ always results in some loss of efficiency. Cardoso (1998) provided a heuristic treatment of the consistency of ICA estimators, where $\widehat{\mathbf{W}}$ may be inconsistent when there is a large mismatch between the hypothesized and true IC distributions. Here, we investigate the accuracy of estimators via simulations with large sample sizes, which is suggestive of consistency properties; a formal consistency analysis is beyond the scope of this article. We modify the Infomax algorithm from Bell and Sejnowski (1995) as described in Web Appendix A.1. Our R code is available in the Supplementary Materials.

2.3. *Negentropy and the FastICA Algorithm*

The FastICA algorithm is based on maximizing the sum of the marginal negentropies. Under the constraint of orthogonal ICs, maximizing negentropy is equal to minimizing MI (Hyvarinen, 1999). Using the notation from (3), negentropy is defined as

$$\mathcal{I}(\mathbf{X}) = \mathcal{H}(\mathbf{Y}) - \mathcal{H}(\mathbf{X}),$$

where $\mathbf{Y} \sim \mathcal{N}(0, \mathrm{E}\,\mathbf{X}\mathbf{X}^T)$. Note that for $\mathbf{X} \sim (0, \mathbf{I})$, $\mathcal{H}(\mathbf{Y}) = Q\mathcal{H}(Y)$ with $Y \sim \mathcal{N}(0, 1)$. Then the MI for linear transformations in (2) equals

$$\mathcal{K}\left(F_{\mathbf{S}}; \prod_{q=1}^{Q} F_{S_q}\right) = \mathcal{I}(\mathbf{OZ}) - \sum_{q=1}^{Q} \mathcal{I}(\mathbf{o}_q'\mathbf{Z}).$$

Since multivariate negentropy is invariant to orthogonal rotations, it follows that minimizing MI is equal to maximizing the negentropy of the marginals.

Approximations to marginal negentropy can take the form (Hyvarinen, 1999)

$$\mathcal{I}(X) \propto \left[\mathrm{E}\left\{G\left(X\right)\right\} - \mathrm{E}\left\{G\left(Y\right)\right\}\right]^2, \tag{5}$$

where G is a non-quadratic function referred to as the "nonlinear function." A common choice is $G(x) = \frac{1}{\alpha}\log\{\cosh(\alpha x)\}$ for $1 \leq \alpha \leq 2$. Then for observations $v = 1, \ldots, V$, define the objective function

$$\mathcal{J}_{FastICA}\left(\mathbf{O}\right) = \sum_{q=1}^{Q}\left[\frac{1}{V}\sum_{v=1}^{V} G\left(\mathbf{o}_q'\mathbf{z}_v\right) - \mathrm{E}\left\{G\left(Y\right)\right\}\right]^2, \tag{6}$$

where $\mathrm{E}\left\{G\left(Y\right)\right\}$ is a known constant. This is maximized using an approximative Newton algorithm, or *fixed-point algorithm* (Hyvarinen, 1999). The fixed-point algorithm assumes a diagonal Hessian matrix, which allows for faster rates of convergence than the Infomax algorithm and fewer computations than an exact Newton algorithm. It can also be derived as a stochastic gradient ascent algorithm for quasi-MLE, where the derivative of G equals the score function (Hyvärinen and Oja, 2000). We implement FastICA using the R package of that name by Marchini, Heaton, and Ripley (2010) with the log cosh nonlinearity, $\alpha = 1$, and the symmetric estimation scheme.

2.4. *ProDenICA*

ProDenICA combines semiparametric estimation of the IC distributions with a fixed-point algorithm (Hastie and Tibshirani, 2003). The joint density of independent ICs is modeled as the product of tilted Gaussians, $f_{\mathbf{S}}(\mathbf{s}) = \prod_{q=1}^{Q}\phi(s_q)e^{g_q(s_q)}$. Here, ϕ is a standard normal density and $g_q(s_q)$ is estimated with cubic B-splines. Let $h_q(x)$ denote the second derivative of $g_q(x)$. The objective function is a penalized log likelihood,

$$\mathcal{J}_{ProDen}(\mathbf{O}) = \sum_{q=1}^{Q}\frac{1}{V}\left(\sum_{v=1}^{V}\log\phi(\mathbf{o}_q^T\mathbf{z}_v) + g_q(\mathbf{o}_q^T\mathbf{z}_v)\right) \tag{7}$$

$$-\log\int\phi(x)e^{g_q(x)}\mathrm{d}x - \lambda\int\{h_q(x)\}^2\mathrm{d}x, \tag{8}$$

where the first penalty enforces the constraint that $\phi(x)e^{g_q(x)}$ integrates to one, and the second is a roughness penalty.

This objective function is maximized by alternately estimating g_q, which is found using an application of generalized additive models, and updating $\mathbf{O}$ with one-step of the fixed-point algorithm used in FastICA. Since it is the log likelihood ratio of the tilted Gaussian to Gaussian, g_q is used as an estimate of marginal negentropy in the fixed-point algorithm. Thus, ProDenICA adapts to the IC distributions while minimizing dependencies. We implement ProDenICA using the R package of that name by Hastie and Tibshirani (2010), and we describe solutions to computational issues that arose when using ProDenICA in Web Appendix A.2.

2.5. *JADE*

For mutually independent random variables, the cross cumulants of all orders are equal to zero. JADE seeks a rotation of whitened data that approximately diagonalizes the fourth-order cross-cumulant tensor (Cardoso and Souloumiac, 1993). The JADE algorithm requires all but one of the excess kurtoses to be non-zero, and it is based on necessary, but not sufficient, conditions for independence. An important difference betwen JADE and other algorithms is that it does not require initialization. We implement JADE using the R package of that name by Nordhausen et al. (2011).

2.6. *A Group ICA Model*

We estimate group ICs using the approach proposed by Calhoun et al. (2001), which involves a two-stage dimension reduction via the singular value decomposition prior to applying a noise-free ICA. Let s_{vq}, $q \in 1, \ldots, Q$, denote mutually independent random variables and $\mathbf{s}_v = [s_{v1}, \ldots, s_{vQ}]'$. We assume $\mathbf{s}_v$ are iid F for $F \in \mathcal{F}$, in which $\mathcal{F}$ is the class of Q-variate non-Gaussian mean zero distributions with covariance equal to the identity matrix. Let $\mathbf{M}^{(m)}$ be a $T_r \times Q$ matrix of mixing weights for the ICs for the mth subject. Our probabilistic spatial group ICA model is

$$\mathbf{x}_v^{(m)} = \mathbf{M}^{(m)}\mathbf{s}_v + \epsilon_v^{(m)}, \tag{9}$$

where $\epsilon_v^{(m)}$ has mean zero and is the error that is not explained by the group ICs.

Suppose $\mathbf{X}^{(m)}$ is a $V \times T_r$ matrix where each column corresponds to a three-dimensional snapshot of the BOLD signal that has been vectorized, and suppose the data have been centered such that both rows and columns have zero mean. Now, consider the singular value decomposition (SVD) of observations from subject m: $\mathbf{X}^{(m)} = \widehat{\mathbf{U}}^{(m)}\widehat{\mathbf{D}}^{(m)}\widehat{\mathbf{V}}^{(m)}$. Let $\widehat{\mathbf{U}}_Q^{(m)}$ denote the first Q left singular vectors and $\widehat{\mathbf{Z}}_Q^{(m)} = \sqrt{V}\widehat{\mathbf{U}}_Q^{(m)}$, where $\sqrt{V}$ standardizes $\widehat{\mathbf{Z}}_Q^{(m)}$ to have sample covariance equal to the identity matrix.

We can align the voxels across subjects from multiple sites, while in general we cannot align time courses in rs-fMRI. Consequently, we concatenate the data matrices $\widehat{\mathbf{Z}}_Q^{(m)}$ across subjects into a matrix $\mathbf{Y}$ with dimensions $V \times MQ$. Next, a second SVD is performed, and the first Q^* left singular vectors are retained and multiplied by $\sqrt{V}$. Here, we let $Q^* = Q$. This results in a whitened data matrix $\widehat{\mathbf{Z}}$ with dimensions $V \times Q$. Applying the methods described in Section 2, we now find a linear transformation $\widehat{\mathbf{W}}$ that results in group ICs $\widehat{\mathbf{S}}$ that minimize a measure of dependence. Thus, the multi-subject ICA problem is reduced to the noise-free ICA model in (1).

Note that we can estimate $\mathbf{M}^{(m)}$ for each subject using standard multivariate regression, such that for a given $\mathbf{S}$, we use least squares to solve $\mathbf{X}^{(m)} = \mathbf{S}\mathbf{M}^{(m)\prime} + \mathbf{E}^{(m)}$, where $\mathbf{E}^{(m)}$ is the $V \times T_r$ matrix of residuals not accounted for by the group components. With this approach, any ICA method can be applied to fMRI from multiple subjects and sites.

2.7. *Canonical Form for ICA and Matching ICs*

The ICA model as presented in (1) is only identifiable on an equivalence class of signed permutations since both $\mathbf{W}$ and $\mathbf{S}$ are unknown (Section 2.2). Eloyan and Ghosh (2013) demonstrate that the ICA model is uniquely identified if $\mathrm{E}\,\mathbf{s}_1^3 > \cdots > \mathrm{E}\,\mathbf{s}_Q^3 \geq 0$. Since $\mathrm{E}\,\mathbf{s}_q = 0$ and $\mathrm{E}\,\mathbf{s}_q^2 = 1$, this is the same as assuming the skewnesses are distinct and positive. Then we define a canonical form for the ICs:

Definition 1. *Let $\widehat{\gamma}_q$ denote the sample skewness for the qth IC. Then the canonical form for $\widehat{\mathbf{S}}$ is the signed permutation that results in $\widehat{\gamma}_1 > \cdots > \widehat{\gamma}_Q \geq 0$.*

In fMRI, assuming positive skewness is biologically plausible because voxels that have very positive BOLD signals may be considered primary contributors to a network, and in practice, many IC densities have large skewnesses.

We matched ICs from multiple estimates using a modification of the Hungarian algorithm proposed by Tichavsky and Koldovsky (2004). For two estimates $\widehat{\mathbf{S}}_{(1)}$ and $\widehat{\mathbf{S}}_{(2)}$, let $\widehat{\mathbf{s}}_i^{(1)}$ and $\widehat{\mathbf{s}}_j^{(2)}$ be the ith and jth columns, respectively. Let $||\cdot||$ denote the L2 norm. Let $\mathbf{C}$ be the cost matrix with elements defined by an auxiliary metric $c_{i,j} = \min(||\widehat{\mathbf{s}}_i^{(1)} - \widehat{\mathbf{s}}_j^{(2)}||, ||\widehat{\mathbf{s}}_i^{(1)} + \widehat{\mathbf{s}}_j^{(2)}||)$, which accounts for the sign ambiguity. Let $\mathcal{S} = \{\sigma : \sigma = \{\sigma(1), \ldots, \sigma(Q)\}$ be the set of all permutations $\{1, \ldots, Q\}$. The Hungarian algorithm is used to find the permutation σ^* such that

$$\sigma^* = \underset{\sigma \in \mathcal{S}}{\operatorname{argmin}} \sum_{i=1}^{Q} c_{i,\sigma(i)}.$$

Details are described in Web Appendix B. R code implementing the canonical form and the matching algorithm is available in the Supplementary Materials.

3. Simulation Study

3.1. *Convexity and Accuracy for $Q = 2$*

We simulated pairs of identically distributed ICs for eighteen distributions that were used in previous ICA studies (Bach and Jordan, 2003; Hastie and Tibshirani, 2003) including the t-distribution, exponential, double exponential, uniform, a mixture of exponentials, and various symmetric and asymmetric mixtures of normals (Web Appendix A.3; Web Figure 1).

First, we examined the objective functions for two components for each distribution. We defined $\mathbf{W}$ using the Givens parameterization with $\theta_{true} = \pi/6$. For each distribution, we conducted one simulation with a very large sample size (V=131,072), such that inaccuracies would be suggestive of consistency issues rather than chance variability or small-sample bias. We evaluated the objective functions on a grid for $\theta \in [0, \pi/2]$ with mesh size $\pi/100$. Then for each estimator we estimated $\hat{\theta}_i$ using $N = 25$ equally spaced starting values in $[0, \pi/2]$.

For FastICA and ProDenICA, there are distributions for which the objective functions include local maxima (Figure 1). For the symmetric, unimodal, and super-Gaussian (having positive excess kurtosis) distributions a, b, and d (t-distribution with df $= 3$, double exponential, and t-distribution with df $= 5$, respectively), the global maximum for each method correctly identifies θ_{true}, and there are no complications owing to local maxima. In contrast, the asymmetric mixture of two normals in distribution k contains a local maximum for both FastICA and ProDenICA. Thus, even when $Q = 2$, local maxima can be an issue.

It also appears that Infomax and FastICA typically and occasionally, respectively, identify the wrong optima, while the global maxima for ProDenICA correctly identify θ_{true}. The global maxima is associated with θ_{true} for all methods for distributions a, b, d, and e, which are all super-Gaussian, unimodal distributions. For sub-Gaussian distributions f through r, the minimum of Infomax, rather than a maximum, usually appears to correspond to θ_{true} (the one exception is distribution q, which has the largest kurtosis among all sub-Gaussian distributions examined). This is indicative of the Infomax estimator being inaccurate for sub-Gaussian distributions (see Lee, Girolami, and Sejnowski, 1999). FastICA misidentifies θ_{true} for distributions j and k, which are asymmetric mixtures of normals. For these distributions, a local maximum is associated with θ_{true}, but the global maximum suggests that the FastICA estimator is not consistent. Additionally, $theta_{true}$ in distribution r is associated with a minimum of the FastICA objective function instead of a maximum, which suggests the FastICA method may not be locally consistent for some mixtures of normals.

3.2. *Convexity and Accuracy for $Q = 5$, 10, and 20*

To examine convexity and accuracy in higher dimensions, we conducted 100 simulations of the ICA model in (1) for $Q = 5$, 10 and 20 randomly chosen (with replacement) distributions from those in Web Figure 1. We used 25 initial values generated via latin hypercube sampling of the rotation angles for each simulation, as described in the Web Appendix A.3. We used the minimum distance (d_{MD}) measure introduced in Ilmonen et al. (2010) and defined in Web Appendix A.4. Let $\widehat{\mathbf{W}}_{(i)}$ denote the unmixing matrix estimated from the ith initial value, $i = 1, \ldots, N$. We then examined $d_{\mathrm{MD}}(\widehat{\mathbf{W}}_{(i)}, \mathbf{W})$.

From these simulations, the methods ordered from most to least accurate were ProDenICA, FastICA, JADE, and Infomax (Figure 2). The MD measure tended to increase as the number of components increased, although this is partly owing to the manner in which d_{MD} scales with dimension.

Infomax was inaccurate in part because it performs poorly for sub-Gaussian distributions, and fourteen of the eighteen distributions in Web Figure 1 are sub-Gaussian. We also investigated the performance of the methods when all ICs had a logistic distribution, which is the best-case scenario for Infomax. Using ten components and the simulation design described above, the means ± standard errors of d_{MD} for FastICA, Infomax, JADE, and ProDenICA were 0.273 ± 0.007, 0.263 ± 0.005, 0.377 ± 0.008, and 0.350 ± 0.014. Not surprisingly, in the unlikely situation where the IC distributions are known, there is a benefit to using the true likelihood in (4) rather than the semi-parametric likelihood in (7).

For FastICA, two issues are clear from Figure 2: there are many stationary points, and in most instances, there exist stationary points that are closer than the global minimum to the true unmixing matrix. Regarding the first issue, comparing the negentropy approximations (6) from many initial values would eliminate the use of estimators to the right of the global maximum. The second issue is more problematic. Ideally, we would like to identify the left-most stationary point as our estimate rather than the global maximum. This left-most point represents an *empirical oracle* since it is only known when θ_{true} is known. Given the local consistency properties of FastICA (Hyvarinen, 1999), it is not surprising that there

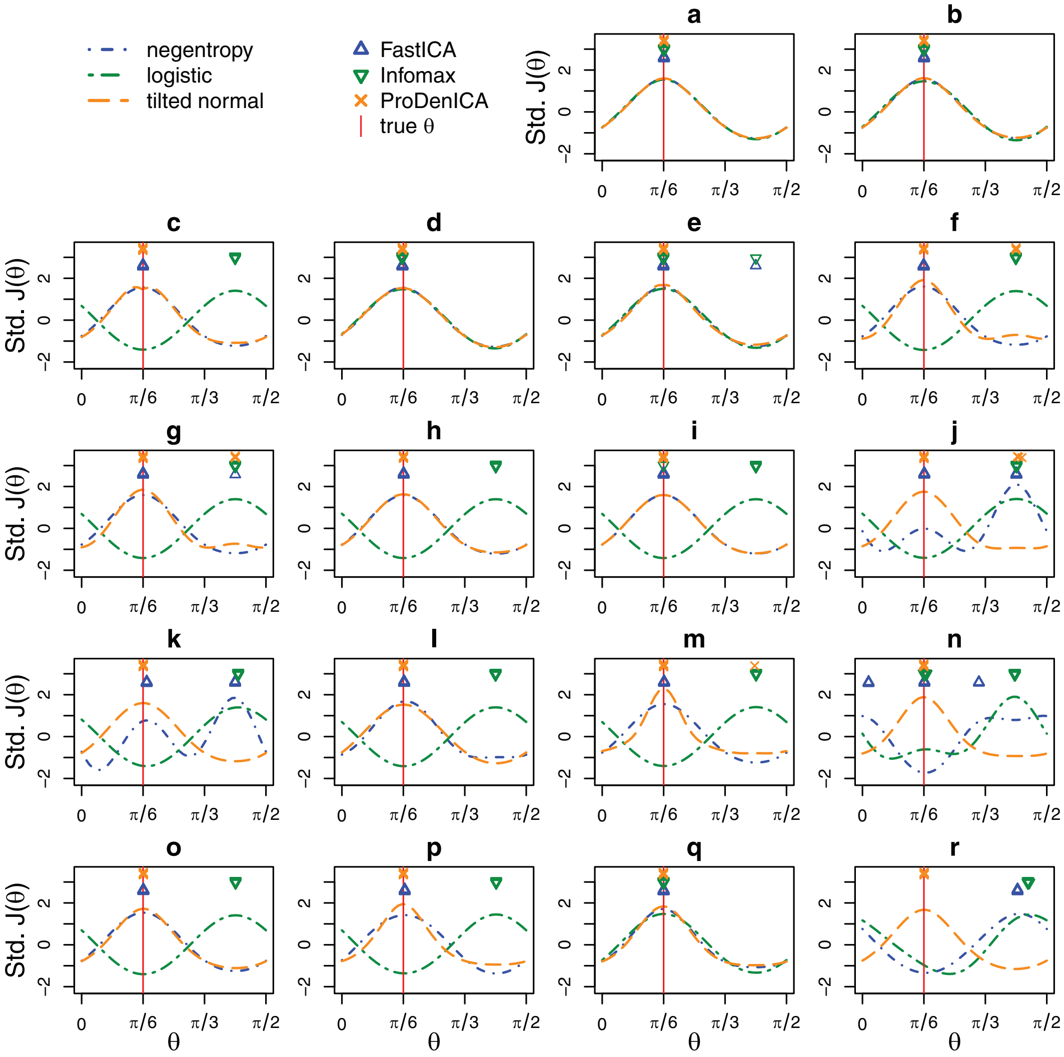

Figure 1. Objective functions (standardized $J(\theta)$; lines) for $V = 131,072$ and $Q = 2$ from distributions *a-r* (Web Figure 1) using the angular (Givens) parameterization with $\theta_{true} = \pi/6$ and $\theta \in [0, \pi/2]$ and parameter estimates (characters; y-value chosen for display purposes) from 25 initial values equally spaced in $[0, \pi/2]$. This figure appears in color in the electronic version of this article.

exist local maxima that are closer to the true unmixing matrix than the FastICA solution. But in FastICA, the left-most gray point is often still a poor estimate of **W**.

In contrast to the other methods, the ProDenICA global maximum usually corresponded to the left-most gray point, and ProDenICA clearly dominated all other estimators (Figure 2). ProDenICA is computationally more expensive than other methods (Web Appendix A.5; Web Table 1). For $Q = 20$ and $V = 1,024$, ProDenICA, Infomax, FastICA, and JADE took approximately 9 minutes, 25 seconds, 7.5 seconds, and 4 seconds, respectively.

4. Group ICA of Resting-State fMRI

4.1. *Resting-State fMRI Dataset*

Data were selected for analysis from the ADHD-200 Data Sample (Milham et al., 2012), which consists of rs-fMRI data from children and adolescents (ages 7-21) from eight sites comprising 491 typically developing subjects and 285 with ADHD (Web Table 2). The number of time slices recorded varied by site from 76 to 261. We restricted our analysis to subjects that were right-hand dominant with no history of drug therapy and to images with no quality control flags. This resulted in 206 typically developing and 78 ADHD children and adolescents from four sites (Web Table 2). Data were registered and masked using the MNI 152 T1 3 mm template. Processing scripts were based on the 1000 Functional Connectome project's (Biswal et al., 2010) processing scripts (available at `http://www.nitrc.org/frs/?group_id=296`). We aggregated adjacent voxels to result in $6 \times 6 \times 6$ mm voxels. Additional information is provided in Web Appendix C.1.

To determine the number of components, rs-fMRI studies frequently fix the number of ICs at twenty, which is sufficient to capture the most frequently observed large-scale resting-state networks (Smith et al., 2009). Task-based fMRI studies sometimes use a probabilistic PCA (PPCA) prior to ICA to

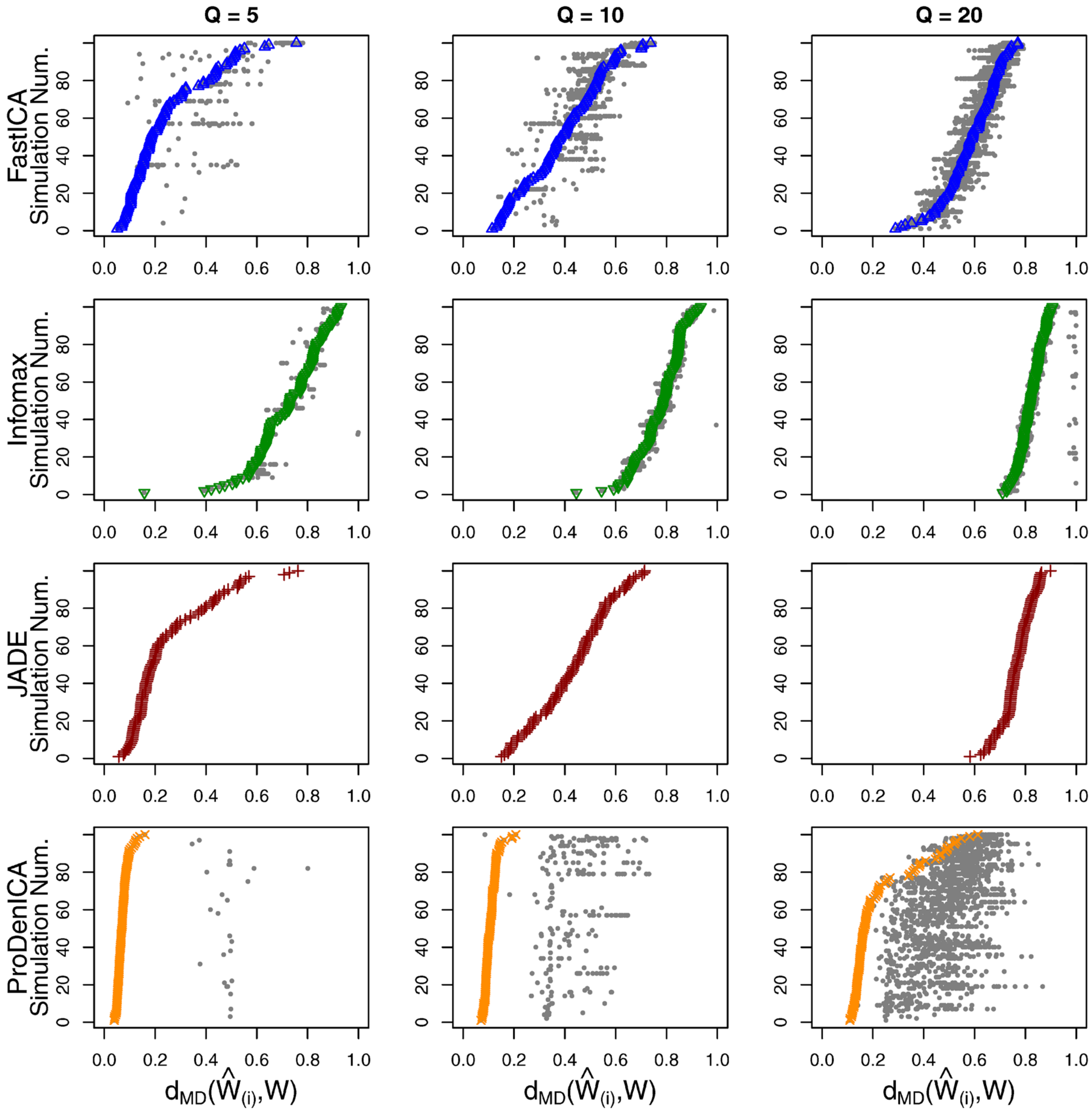

Figure 2. Simulations using $Q = 5$, 10, or 20 from randomly chosen distributions with $V = 1024$. For k = FastICA, Infomax, and ProDenICA, the results from 25 initial values for 100 simulations are depicted: small gray points correspond to stationary points ($\widehat{\mathbf{W}}^k_{(i)}, i = 1, \ldots, 25$), and symbols correspond to the global maximum ($\widehat{\mathbf{W}}^k_{(0)}$). For each method k, simulations are sorted from lowest to highest $d_{\mathrm{MD}}(\widehat{\mathbf{W}}^k_{(0)}, \mathbf{W})$. The JADE algorithm is not initialized with multiple values. This figure appears in color in the electronic version of this article.

determine the number of ICs (Beckmann and Smith, 2004). The signal-to-noise ratio is smaller in rs-fMRI than task-based fMRI, and a low signal-to-noise ratio can be problematic for PPCA. Consequently, we followed previous studies and let $Q = Q^* = 20$.

4.2. *Differences Within Algorithms*

We examined the sensitivity to initialization of FastICA, Infomax, and ProDenICA on group ICA of the ADHD-200 dataset. We generated $N = 1{,}000$ initial values using latin hypercube sampling of the Givens rotation angles. We created a dissimilarity matrix with entries $d_{\mathrm{MD}}(\widehat{\mathbf{W}}^k_{(i)}, \widehat{\mathbf{W}}^k_{(j)})$ for the kth method, $i \neq j \in 1, \ldots, 1000$, and $Q = 20$. We also created a dissimilarity matrix for each IC. Define

$$\widehat{\mathbf{W}}^k_{(0)} = \operatorname*{argmax}_{i \in 1,\ldots,N} \mathcal{J}(\widehat{\mathbf{W}}^k_{(i)}).$$

Let $\widehat{\mathbf{S}}^k_{(0)}$ be the estimated ICs associated with $\widehat{\mathbf{W}}^k_{(0)}$ and ordered as in Definition 1. Define $\widehat{\mathbf{S}}^k_{(i)}$, $i = 1, \ldots, N$, to be the ICs associated with $\widehat{\mathbf{W}}^k_{(i)}$ that have been matched to $\widehat{\mathbf{S}}^k_{(0)}$. The dissimilarity matrix for the qth IC has entries $||\widehat{\mathbf{S}}^k_{(i),q} - \widehat{\mathbf{S}}^k_{(j),q}||_2$, in which $\widehat{\mathbf{S}}^k_{(i),q}$ is the qth column of $\widehat{\mathbf{S}}^k_{(i)}$. We then used classical

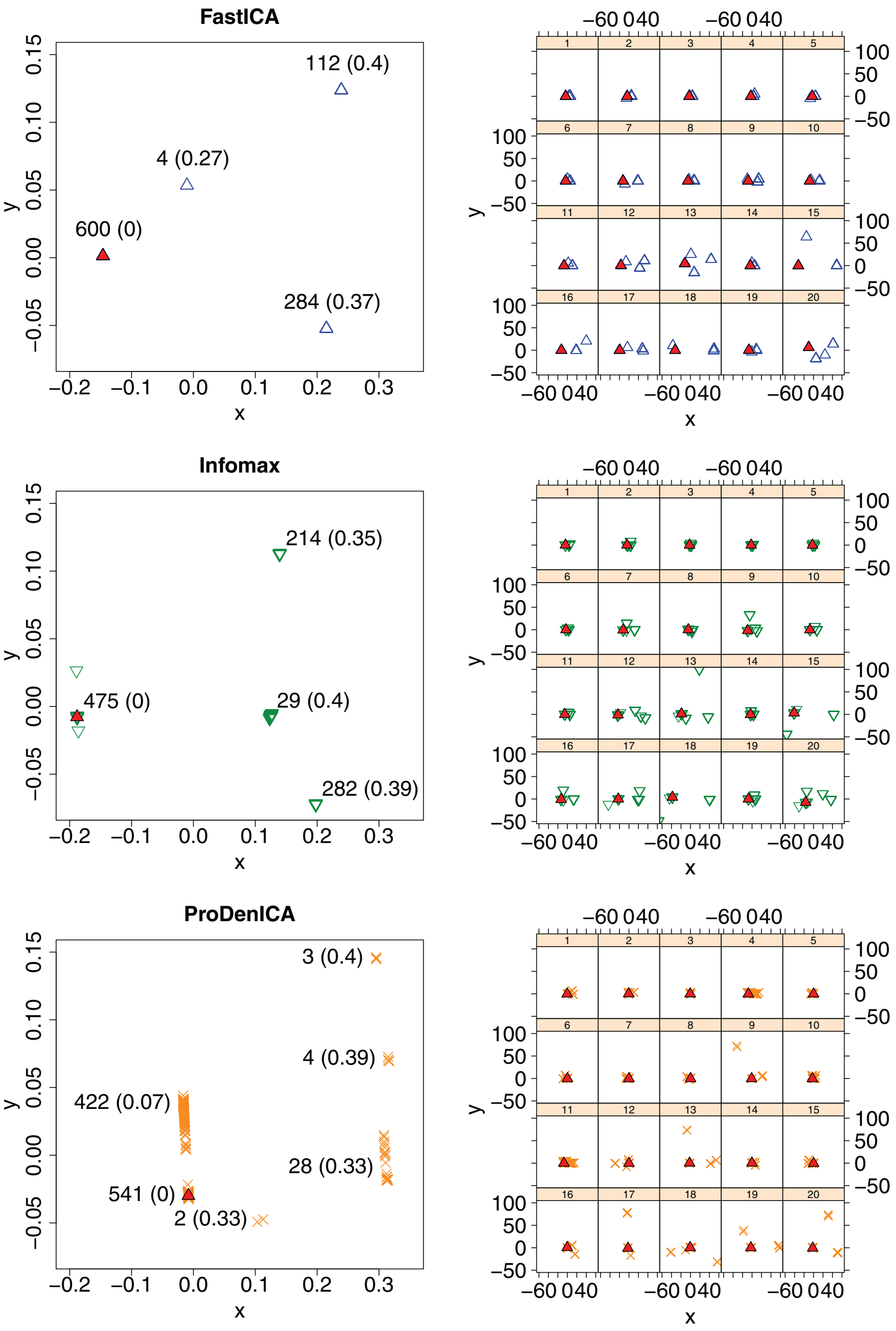

Figure 3. Multidimensional scaling of $d_{\mathrm{MD}}(\widehat{\mathbf{W}}^k_{(i)}, \widehat{\mathbf{W}}^k_{(j)})$ with the number of points in each basin and the average d_{MD} from the basin to $\widehat{\mathbf{W}}^k_{(0)}$ in parentheses, where k indexes method and $i \neq j \in 1, \dots, 1000$ (left), and $||\widehat{\mathbf{S}}^k_{(i),q} - \widehat{\mathbf{S}}^k_{(j),q}||_2$ for $q = 1, \dots, 20$ (right). The coordinates of $\widehat{\mathbf{W}}^k_{(0)}$ and $\widehat{\mathbf{S}}^k_{(0),q}$, $q = 1, \dots 20$, are depicted by solid triangles. This figure appears in color in the electronic version of this article.

multidimensional scaling (Torgerson, 1952) with two dimensions to visualize the dissimilarities among unmixing matrices.

In estimates of the mixing matrix, there were four basins of attraction for both FastICA and Infomax (Figure 3). For ProDenICA, there were two major basins of attraction and four smaller basins. In all methods, the basin with the most points contained the argmax, along with 60%, 47.5%, and 54.1% of estimates for FastICA, Infomax, and ProDenICA, respectively. The remaining 40% of FastICA estimates had an MD (relative to $\widehat{\mathbf{W}}^k_{(0)}$) of approximately 0.38; 52.5% of Infomax estimates had an MD of approximately 0.37; and 42% of ProDenICA estimates had an MD of 0.07 and another 3.5% had an MD of approximately 0.34.

In estimates of the individual ICs, it is clear that some ICs were nearly identical for most starting values (e.g., ICs 1 through 6 for all methods; recall that the ICs are ordered

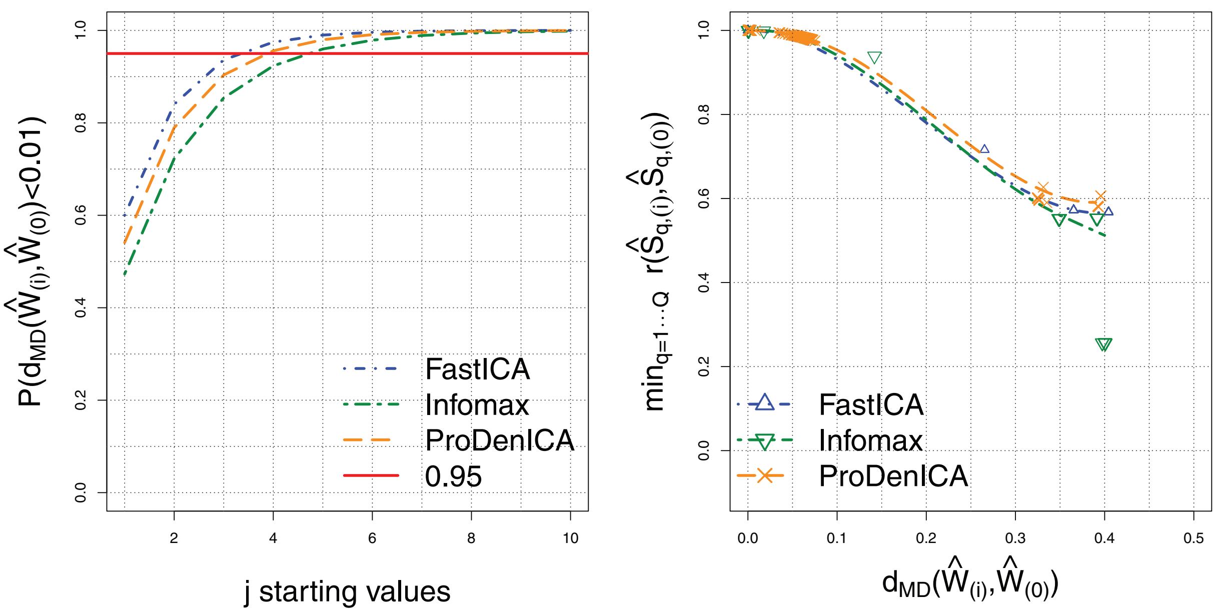

Figure 4. The probability of obtaining $\widehat{\mathbf{W}}^k_{(0)}$ when using j initial values for $k =$ FastICA, Infomax, or ProDenICA (left). The relationship between $d_{\mathrm{MD}}(\widehat{\mathbf{W}}^k_{(i)}, \widehat{\mathbf{W}}^k_{(0)})$ and the Pearson correlation between $\widehat{\mathbf{S}}^k_{(i),q}$ and $\widehat{\mathbf{S}}^k_{(0),q}$ (lines are from a loess smoother), where for each initial value, the symbol denotes the minimum correlation $r(\widehat{\mathbf{S}}^k_{(i),q}, \widehat{\mathbf{S}}^k_{(0),q})$ with $q = 1, \ldots, 20$ versus $d_{\mathrm{MD}}(\widehat{\mathbf{W}}^k_{(i)}, \widehat{\mathbf{W}}^k_{(0)})$ (right). This figure appears in color in the electronic version of this article.

by decreasing skewness; see Figure 3), while others were more sensitive to initialization in all methods (e.g., ICs 13, 17, 18, and 20), and some were sensitive in some methods but not others (e.g., IC 15 was sensitive in FastICA and Infomax, but not ProDenICA; IC 19 was sensitive in ProDenICA, but not FastICA or Infomax). Overall, the estimation of ICs with the largest skewnesses and kurtoses (where kurtosis was generally higer in lower-numbered ICs) tended to be more stable than those that were more nearly symmetric with lower kurtoses (see Web Figure 2).

We estimated the probability of obtaining $\widehat{\mathbf{W}}^k_{(0)}$ using j starting values. Consider the probability of obtaining an initial value that is close to the argmax, $P(d_{\mathrm{MD}}(\widehat{\mathbf{W}}^k_{(0)}, \widehat{\mathbf{W}}^k_{(i)}) < \delta) \geq \epsilon$, when using j starting values. We chose δ such that $\{d_{\mathrm{MD}} < \delta\}$ is the event that we have found the global maximum (within some numerical tolerance). Here, we let $\delta = 0.01$. Now recall the hypergeometric distribution, $P(X = x|N, m_k, j) = \{\binom{m_k}{x}\binom{N-m_k}{j-x}\}/\binom{N}{j}$, where N is the total number of starting values ($N = 1{,}000$), m_k is the number of times $\widehat{\mathbf{W}}^k_{(i)}$ was within δ of $\widehat{\mathbf{W}}^k_{(0)}$, j is the number of starting values for which we wish to calculate the probability of getting within δ, and x is the number of times that $\widehat{\mathbf{W}}^k_{(i)}$ is within δ of $\widehat{\mathbf{W}}^k_{(0)}$ when using j starting values. We calculated $P(X > 0|N = 1{,}000, m_k, j)$ for $j \in 1, \ldots, 10$. We also calculated $\underset{q=1,\ldots,Q}{min}\, r(\widehat{\mathbf{S}}^k_{(i),q}, \widehat{\mathbf{S}}^k_{(0),q})$, where r is the Pearson correlation, and examined the relationship of this minimum correlation to d_{MD}.

In our application, FastICA, Infomax, and ProDenICA required 4, 5, and 4 initial values, respectively, to have a ¿0.95 probability of obtaining the argmax (Figure 4). When comparing ICs from different initializations, $\min\limits_{q=1,\ldots,Q} r(\widehat{\mathbf{S}}_{(i),q}, \widehat{\mathbf{S}}_{(0),q})$ was on average approximately 0.60 and as low as 0.25 for $d_{\mathrm{MD}} > 0.3$. Overall, a minimum distance measure less than 0.10 for some pair of $\widehat{\mathbf{W}}^k_{(i)}$ and $\widehat{\mathbf{W}}^k_{(0)}$ translated to a minimum correlation (over q) between $\widehat{\mathbf{S}}_{(i),q}$ and $\widehat{\mathbf{S}}_{(0),q}$ of at least 0.95.

4.3. *Differences Between Algorithms*

We matched $\widehat{\mathbf{W}}^k_{(0)}$, for k indexing Infomax, JADE, and ProDenICA, to the canonically ordered results from FastICA. We also compared each method to the SVD, which represents a baseline for understanding the impact of the additional rotation via ICA. We compared unmixing matrices using three measures: (1) the MD measure, d_{MD}; (2) the Amari measure (Amari, Cichocki, and Yang, 1996); and (3) the Frobenius norm between matched unmixing matrices.

FastICA and Infomax had very similar results, while ProDenICA and JADE differed from each other and from FastICA and Infomax (Web Table 4). All ICA solutions were substantially different from the SVD solution. The measures between ICA unmixing matrices were all substantially smaller than between random matrices (see Web Appendix C.2).

We compared estimated ICs between methods using Pearson correlations, where all methods were matched to the canonically ordered FastICA. We used Kolmogorov–Smirnov (KS) two-sample tests to examine differences in the CDFs of matched ICs. We did not formally test for equality in distribution because IC samples (i.e., values at different voxels) were spatially dependent. Nonetheless, we calculated FDR-adjusted p-values (Benjamini and Hochberg, 1995) as

Table 1
Pearson correlation between matching ICs for each method from the rs-fMRI study

Method1	Method2	IC 1	IC 2	IC 3	IC 4	IC 5	IC 6	IC 7	IC 8	IC 9	IC 10
SVD	FastICA	0.51	0.47	0.33	0.35	0.43	0.44	0.48	0.46	0.49	0.41
SVD	Infomax	0.51	0.48	0.35	0.38	0.43	0.42	0.48	0.46	0.49	0.43
SVD	JADE	0.53	0.43	0.40	0.37	0.44	0.38	0.50	0.47	0.53	0.40
SVD	ProDenICA	0.49	0.44	0.36	0.48	0.44	0.45	0.41	0.47	0.48	0.41
FastICA	Infomax	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FastICA	JADE	0.96	0.98	0.98	0.99	0.99	0.99	0.98	0.99	0.94	0.98
FastICA	ProDenICA	0.99	0.97	0.98	0.83	0.99	0.99	0.98	0.98	0.96	0.98
Infomax	JADE	0.97	0.97	0.98	0.99	1.00	0.99	0.98	0.99	0.94	0.98
Infomax	ProDenICA	0.99	0.97	0.98	0.85	0.99	0.99	0.99	0.98	0.97	0.98
JADE	ProDenICA	0.96	0.97	0.95	0.83	0.99	0.97	0.96	0.99	0.96	0.96

Method1	Method2	IC 11	IC 12	IC 13	IC 14	IC 15	IC 16	IC 17	IC 18	IC 19	IC 20
SVD	FastICA	0.51	0.61	0.51	0.35	0.39	0.27	0.71	0.59	0.36	0.46
SVD	Infomax	0.51	0.60	0.52	0.32	0.40	0.27	0.74	0.59	0.36	0.48
SVD	JADE	0.55	0.67	0.21	0.31	0.26	0.27	0.70	0.70	0.42	0.25
SVD	ProDenICA	0.42	0.61	0.23	0.44	0.33	0.17	0.79	0.40	0.32	0.18
FastICA	Infomax	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
FastICA	JADE	0.99	0.95	0.63	0.97	0.92	0.80	0.95	0.96	0.78	0.74
FastICA	ProDenICA	0.82	0.93	0.69	0.95	0.94	0.89	0.90	0.89	0.89	0.69
Infomax	JADE	0.98	0.96	0.60	0.97	0.92	0.80	0.96	0.96	0.78	0.74
Infomax	ProDenICA	0.80	0.93	0.66	0.94	0.94	0.87	0.90	0.89	0.90	0.69
JADE	ProDenICA	0.83	0.94	0.87	0.96	0.94	0.83	0.85	0.85	0.74	0.95

a measure of the difference between ICs, as described in Web Appendix C.2. Lastly, we estimated the density of ICs for each method using Gaussian kernels (Web Appendix C.2).

The Pearson correlations were high for most ICs but not all (Table 1), and the shapes of the estimated densities across the four methods were similar for most distributions with some notable exceptions (Web Figure 2). In contrast, the KS statistics often indicated differences in the distributions of ICs by method (Web Table 4). Overall, $r(\widehat{\mathbf{S}}^k_{(0),q}, \widehat{\mathbf{S}}^l_{(0),q}) > 0.95$ in 78/120 comparisons (excluding SVD) and $r(\widehat{\mathbf{S}}^k_{(0),q}, \widehat{\mathbf{S}}^l_{(0),q}) < 0.80$ in 12/120 comparisons. Some ICs were highly correlated for all methods (e.g., ICs 1–3, 5–10, and 14), while for other ICs, ProDenICA and JADE had relatively low correlations with FastICA and Infomax (e.g., ICs 13 and 20), and occasionally, ProDenICA differed from all other methods (e.g., IC 11) or JADE differed from all other methods (e.g., IC 19). In the KS tests, FDR-adjusted $p \leq 0.01$ in 72/120 comparisons. In some cases, $p \leq 0.01$ even though the ICs were highly correlated (e.g., IC 3). For FastICA and Infomax, $p > 0.05$ for all ICs except IC 4. In cases with low correlations, differences in the density plots were often visible (e.g., in IC 13, ProDenICA was less peaked; also see ICs 3, 4, 18, and 19). Sometimes correlations were high, but KS-statistics and density plots indicated differences between ProDenICA and other methods (e.g., IC 12), or differences between JADE, ProDenICA, and FastICA/Infomax (e.g., IC 3).

A visual comparison of the spatial configuration of the group ICs revealed that moderate correlations, e.g., less than 0.80, were sometimes associated with large differences. For each IC, we used thresholding and retained 2.5% of voxels corresponding to the most positive values. We visually associated our ICs with networks from Damoiseaux et al. (2006) and present images for selected ICs (Web Figure 3). IC 13 has strong lateralization in FastICA and Infomax but is nearly symmetric in JADE and ProDenICA. IC 20 appears to contain areas associated with memory and has strong lateralization in all methods, but FastICA and Infomax suggest a different spatial configuration than ProDenICA and JADE. ICs 13 and 20 were previously noted to be sensitive to initialization (Section 4.2). These networks may be ignored in fMRI studies that use Icasso despite the fact that they do not appear to be artifactual. Parts of the visual cortex are contained in IC4, in which all correlations were greater than 0.80 and the methods look similar, although ProDenICA shows some deviations. IC 3 contains parts of the default network, an area associated with day-dreaming that is often examined in rs-fMRI studies, and the spatial configuration was similar across methods.

We also estimated ICs from a single individual randomly chosen from the ADHD-200 Data Sample. The spatial configuration in individual ICs is less pronounced than in the corresponding group ICs (Web Figure 4). The default network (IC 3) in the individual IC is very similar to the group IC, and similarities between IC 4 and IC 20 are also apparent, while IC 13 did not appear to be recovered in this individual.

5. Discussion

There is a collaborative effort to share rs-fMRI data from multiple sites in order to improve sample sizes, as in the 1,000 Functional Connectomes Project, the ADHD-200 Sample, and the Autism Brain Imaging Dataset (ABIDE). Thus, there is an urgent need to evaluate whether widely used ICA methods effectively recover resting-state networks, or whether more robust, but typically computationally more expensive, methods produce different results. We have applied a semiparametric method, ProDenICA, to an analysis of rs-fMRI data and demonstrated that multiple initial values are necessary to identify the argmax. In contrast to other fMRI studies, we applied the Hungarian algorithm to match ICs from multiples estimates, and thereby gained novel insights into how some brain networks are more sensitive to initial values than others, and how some brain network estimates varied little by ICA method while others differ. Given the results from simulations and the fact that IC distributions are rarely, if ever, known in practice, we suggest the use of ICA methods that are effective for a wide range of IC distributions and methods wherein the argmax estimates from multiple initializations correspond to the best estimate. Thus, we suggest ProDenICA be used over FastICA, Infomax, or JADE.

The few studies that considered the impact of starting values on ICA estimation suggested that spurious optima were rarely a problem or excluded ICs that were sensitive to initial values from further analyses; however, we found that ICs that were not sensitive to initialization were the exception and not the rule. In an application of ICA to signal processing, Tichavsky et al. (2005)Tichavsky, Koldovsky, and Oja (2005) claimed that approximately 1-100 cases in 10,000 initializations produced estimates from spurious stationary points, and that these cases could be recognized by extremely low signal to interference ratios. In our simulations, local optima were nearly always problematic for twenty components (Figure 2). Furthermore, in our fMRI study, spurious optima were found in 40% of initializations for FastICA, 52.5% for Infomax, and 46% for ProDenICA (Figure 3).

We argue that evaluating a modest number of randomly chosen initial values and comparing the values of their objective functions is effective and computationally practicable. Tichavsky et al. (2005) proposed a method that imitates a global search for the argmax for a single starting value, although there is no guarantee that it converges to the global maximum. Alternatively, Icasso assumes that cluster centroids accurately characterize ICs. Using cluster centroids produces two sources of error in IC estimates: potential mismatches due to matching via clustering, and error due to the use of cluster centroids instead of the argmax. Furthermore, when multiple estimates of an IC do not tightly cluster, the IC is typically discarded. Consequently, biologically relevant networks may be ignored simply because their local optima are very different from the argmax. In task-based fMRI, Guo (2011) and Beckmann and Smith (2005) suggest using normal mixtures to model activated and inactived voxels, but Figures 1k and 1j indicate that FastICA has spurious optima for certain mixtures. Thus, in some cases, biological networks may be ignored in FastICA studies owing to multiple optima that in turn correspond to diffuse clusters.

Moreover, some biological networks may be mischaracterized owing to the poor performance of FastICA and Infomax in recovering some IC distributions, whereas ProDenICA is more robust to IC distributions. For rs-fMRI, the differences between methods were relatively small according to our similarity measures (Web Table 4), although visual inspection suggests substantive differences (Web Figure 4). In our simulations with two components and very large sample sizes ($V =$ 131,072), Infomax failed for most mixture distributions, and FastICA failed to have a global and/or local maximum at the true unmixing matrix for some assymetric mixture distributions (Figure 1). In contrast, the argmax for ProDenICA with two components corresponded to the true unmixing matrix for all simulated distributions. In simulations with 5, 10, and 20 components, FastICA and Infomax suffered from two problems: oftentimes, an empirical oracle existed that was closer to the true unmixing matrix than the argmax, and secondly, this empirical oracle was inaccurate. JADE was also inaccurate. These issues were resolved in ProDenICA, where the empirical oracle usually corresponded to the argmax, and the argmax was close to the true unmixing matrix (Figure 2). These results suggest the difference between methods may be larger in task-based fMRI where normal mixtures model activated/innactive voxels than observed in the resting-state networks.

One approach to examining brain functioning from fMRI studies is to compare mixing matrices between groups, which is often done by assuming a tensor structure that decomposes sources of group variation and sources of individual variation (Beckmann and Smith, 2005; Guo, 2011). In our application, the use of multi-site data with differing numbers of time points precludes the use of a tensor group structure. Here, we focused on the spatial activation patterns rather than the individual and/or group time courses because an examination of mixing matrices of varying dimensions is not trivial. Future research should investigate methods to compare groups where individuals have varying numbers of time points. For example, converting the temporal patterns of activation (columns of $\mathbf{M}^{(m)}$ in (9)) to the spectral domain may facilitate an examination of the pathophysiology of diseases.

We conclude that the performance of methods differed dramatically in simulations, and the IC estimates in our fMRI application exhibited variability for some, but not all, ICs. Thus, ProDenICA may improve estimates of ICs in fMRI. Additionally, multiple initial values were essential for identifying the argmax in FastICA, Infomax, and ProDenICA.

6. Supplementary Materials

Web Appendices, Figures, Tables, and R Code referenced in Sections 2, 3, 4, and 5 are available at the *Biometrics* website on Wiley Online Library.

Acknowledgements

We thank the co-editor Dr. Russell Millar, an anonymous associate editor, and two anonymous referees for their helpful comments. This study was supported by grants R01EB012547 and P41EB015909 from the National Institute of Biomedical Imaging and Bioengineering and grant R01NS060910 from the National Institute of Neurological Disorders and Stroke.

References

Amari, S. I., Cichocki, A., and Yang, H. H. (1996). A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer (eds), 757–763. Cambridge, MA: MIT Press.

Bach, F. R. and Jordan, M. I. (2003). Kernel independent component analysis. *The Journal of Machine Learning Research* **3**, 1–48.

Bartlett, M. S., Movellan, J. R., and Sejnowski, T. J. (2002). Face recognition by independent component analysis. *IEEE Transactions on Neural Networks* **13**, 1450–1464.

Beckmann, C. F. (2012). Modelling with independent components. *NeuroImage* **62**, 891–901.

Beckmann, C. F. and Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging* **23**, 137–152.

Beckmann, C. F. and Smith, S. M. (2005). Tensorial extensions of independent component analysis for multisubject fMRI analysis. *Neuroimage* **25**, 294–311.

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation* **7**, 1129–1159.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Biswal, B. B., Mennes, M., Zuo, X. N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., et al. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences* **107**, 4734–4739.

Calhoun, V. D., Adali, T., Pearlson, G. D., and Pekar, J. J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping* **14**, 140–151.

Cardoso, J. F. (1997). Infomax and maximum likelihood for blind source separation. *Signal Processing Letters, IEEE* **4**, 112–114.

Cardoso, J. F. (1998). Blind signal separation: statistical principles. *Proceedings of the IEEE* **86**, 2009–2025.

Cardoso, J. F. and Souloumiac, A. (1993). Blind beamforming for non-Gaussian signals. *Radar and Signal Processing, IEEE Proceedings F* **140**, 362–370.

Celone, K. A., Calhoun, V. D., Dickerson, B. C., Atri, A., Chua, E. F., Miller, S. L., DePeau, K., Rentz, D. M., Selkoe, D. J., Blacker, D., et al. (2006). Alterations in memory networks in mild cognitive impairment and Alzheimer's disease: An independent component analysis. *The Journal of Neuroscience* **26**, 10222–10231.

Chen, A. and Bickel, P. J. (2006). Efficient independent component analysis. *The Annals of Statistics* **34**, 2825–2855.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing* **36**, 287–314.

Correa, N., Adali, T., and Calhoun, V. D. (2007). Performance of blind source separation algorithms for fMRI analysis using a group ICA method. *Magnetic Resonance Imaging* **25**, 684–694.

Damoiseaux, J. S., Rombouts, S., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M., and Beckmann, C. F. (2006). Consistent resting-state networks across healthy subjects. *Proceedings of the National Academy of Sciences* **103**, 13848–13853.

Eloyan, A. and Ghosh, S. K. (2013). A semiparametric approach to source separation using independent component analysis. *Computational Statistics & Data Analysis* **58**, 383–396.

Eloyan, A., Muschelli, J., Nebel, M. B., Liu, H., Han, F., Zhao, T., Barber, A., Joel, S., Pekar, J. J., Mostofsky, S., and Caffo, B. S. (2012). Automated diagnoses of attention deficit hyperactive disorder using magnetic resonance imaging. *Frontiers in Systems Neuroscience* **6**, 1–9.

Guo, Y. (2011). A general probabilistic model for group independent component analysis and its estimation methods. *Biometrics* **67**, 1532–1542.

Hastie, T. and Tibshirani, R. (2003). Independent components analysis through product density estimation. In *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer (eds), 649–656. Cambridge, MA: MIT Press.

Hastie, T. and Tibshirani, R. (2010). *ProDenICA: Product Density Estimation for ICA Using Tilted Gaussian Density Estimates.* R package version 1.0.

Himberg, J., Hyvärinen, A., and Esposito, F. (2004). Validating the independent components of neuroimaging time series via clustering and visualization. *Neuroimage* **22**, 1214–1222.

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks* **10**, 626–634.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks* **13**, 411–430.

Ilmonen, P., Nordhausen, K., Oja, H., and Ollila, E. (2010). A new performance index for ICA: Properties, computation and asymptotic analysis. *Latent Variable Analysis and Signal Separation* 229–236.

Iriarte, J., Urrestarazu, E., Valencia, M., Alegre, M., Malanda, A., Viteri, C., and Artieda, J. (2003). Independent component analysis as a tool to eliminate artifacts in EEG: A quantitative study. *Journal of Clinical Neurophysiology* **20**, 249–257.

Jafri, M. J., Pearlson, G. D., Stevens, M., and Calhoun, V. D. (2008). A method for functional network connectivity among spatially independent resting-state components in schizophrenia. *Neuroimage* **39**, 1666–1681.

Kong, W., Vanderburg, C. R., Gunshin, H., Rogers, J. T., and Huang, X. (2008). A review of independent component analysis application to microarray gene expression data. *Biotechniques* **45**, 501.

Lee, T. W., Girolami, M., and Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation* **11**, 417–441.

Marchini, J. L., Heaton, C., and Ripley, B. D. (2010). *FastICA: FastICA Algorithms to Perform ICA and Projection Pursuit.* R package version 1. 1–13.

Matteson, D. S. and Tsay, R. S. (2013). Independent Component Analysis via Distance Covariance. *ArXiv e-prints.*

Milham, M. P., Fair, D., Mennes, M., and Mostofsky, S. H. (2012). The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience* **6**, 62.

Nordhausen, K., Cardoso, J. F., Oja, H., and Ollila, E. (2011). *JADE: JADE and ICA Performance Criteria*. R package version **1**.0–4.

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences* **106**, 13040–13045.

Tichavsky, P. and Koldovsky, Z. (2004). Optimal pairing of signal components separated by blind techniques. *Signal Processing Letters, IEEE* **11**, 119–122.

Tichavsky, P., Koldovsky, Z., and Oja, E. (2005). Asymptotic performance of the FastICA algorithm for independent component analysis and its improvements. In *SP 13th Workshop on Statistical Signal Processing*, 1084–1089. Piscataway, NJ: IEEE.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika* **17**, 401–419.

Veer, I. M., Beckmann, C. F., Van Tol, M. J., Ferrarini, L., Milles, J., Veltman, D. J., Aleman, A., Van Buchem, M. A., van der Wee, N. J., and Rombouts, S. A. (2010). Whole brain resting-state analysis reveals decreased functional connectivity in major depression. *Frontiers in Systems Neuroscience* **4**, 1–10.

Received January 2013. *Revised June* 2013.
Accepted August 2013.