Published in final edited form as:

Magn Reson Imaging. 2011 November ; 29(9): 1288–1303. doi:10.1016/j.mri.2011.07.017.

Fixed-point Algorithms for Constrained ICA and their Applications in fMRI Data Analysis

Ze Wang*

Department of Psychiatry, School of Medicine, Department of Bioengineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA 19104

Abstract

Constrained independent component analysis (CICA) eliminates the order ambiguity of standard ICA by incorporating prior information into the learning process to sort the components intrinsically. However, the original CICA (OCICA) and its variants depend on a learning rate, which is not easy to be tuned for various applications. To solve this problem, two learning-rate free CICA algorithms were derived in this paper using the fixed-point learning concept. A complete stability analysis was provided for the proposed methods, which also made a correction to the stability analysis given to OCICA. Variations for adding constraints either to the components or the associated time courses were derived too. Using synthetic data, the proposed methods yielded a better stability and a better source separation quality in terms of higher SNR and smaller performance index (PI) than OCICA. For the artificially generated brain activations, the new CICA demonstrated a better sensitivity/specificity performance than standard univariate general linear model (GLM) and standard ICA. OCICA showed a similar sensitivity/specificity gain but failed to converge for several times. Using fMRI data acquired with a well-characterized sensorimotor task, the proposed CICA yielded better sensitivity than OCICA, standard ICA, and GLM in all the target functional regions in terms of either higher t-values or larger suprathreshold cluster extensions using the same significance threshold. In addition, they were more stable than OCICA and standard ICA for analyzing the sensorimotor fMRI data.

Keywords

independent component analysis; constrained independent component analysis; augmented Lagrangian, fixed-point iteration; Newton-like gradient; general linear model; fMRI

Introduction

Independent component analysis (ICA) [17] is an advanced signal processing method designed to separate mutually independent components (ICs) from their observed mixtures without using any prior information of them. As a multivariate and data-driven method, ICA has gained popularity in many signal processing fields [17], including fMRI data analysis [4, 24, 30]. Without need of prior modeling for brain response, ICA bestows a large freedom to analyzing fMRI data with various kinds of tasks or even the null hypothesis data [2, 11]. However, standard ICA does not extract ICs with a fixed order [6], resulting in a problem of IC comparisons across different runs or across different subjects. Constrained ICA (CICA) [13, 20] provides a solution for this problem. In this approach, prior information contained in constraints is integrated into the source separation process, so that the decomposed ICs

*Correspondence: 3900 Chestnut Street, Treatment Research Center, Department of Psychiatry, University of Pennsylvania, Philadelphia, PA 19104, Tel: (215) 222-3200, Fax: (215) 386-6770, zewang@mail.med.upenn.edu, redhatw@gmail.com..

are sorted intrinsically. This idea was first proposed by Luo *et al.* [23] but was explicitly defined by Lu and Rajapakse who also provided a CICA algorithm based on a more advanced ICA method [20, 21]. For this reason, the conceptual framework of CICA proposed by Lu and Rajapakse will be employed in this paper, though a similar approach, the semi-blind ICA, has later been proposed by Calhoun *et al* [1].

One problem of the original CICA (OCICA) algorithm proposed by Lu and Rajapakse [20, 21] is that it uses a learning rate to control the updating gradient at each iteration. This learning-rate is not easy to tune for various applications, and a bad choice of it could completely destroy the algorithm convergence [3, 17]. For fMRI data analysis, the OCICA was directly adopted to extract temporally independent components using temporal constraints. As extracting spatially independent components represents another major interest of ICA-based fMRI data analysis [5, 24], CICA should be adapted to have these two capabilities, as well as applying either spatial or temporal constraints. Moreover, the stability analysis for CICA given in [21] was incomplete and was based on a non-established condition, so a complete stability analysis is still required to prove the convergence of CICA.

We have recently proposed a fixed-point CICA in a conference abstract [31], which does not need a learning rate. The purpose of this paper is to provide full derivations, comprehensive evaluations, and detailed convergence analysis for the fixed-point CICA algorithms. Method evaluations were conducted using synthetic 1D signal and 2D fMRI data and sensorimotor fMRI data.

Theory

ICA with constraints and the OCICA algorithm

Assuming the observed signal $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ to be a linear mixture of some unknown mutually independent sources $\mathbf{s} = (s_1, s_2, \dots, s_m)^T$: $\mathbf{x} = \mathbf{A}\mathbf{s}$, where $\mathbf{A}$ is an unknown $n \times m$ matrix ($m \leq n$), standard ICA [6] is to seek an $m \times n$ unmixing matrix $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m)^T$, such that $\mathbf{s}$ can be approximately recovered by $\mathbf{y} = \mathbf{W}\mathbf{x}$, with $\mathbf{y} = (y_1, y_2, \dots, y_m)^T$. In general, ICA can be formed and solved as an optimization problem to minimize an object function $O(\cdot)$ that measures the independence of the output components. One such function used in FastICA [16, 17] and OCICA is:

$$O(\mathbf{y}) = - \sum_i \mathcal{J}(y_i) \quad (1)$$

where $\mathcal{J}(\cdot)$ is the negentropy and can be approximated by [15]:

$$\mathcal{J}(y_i) \approx c[E\{G(y_i)\} - E\{G(v)\}]^2 \quad (2)$$

where c is a positive constant, $G(\cdot)$ is a nonquadratic function, and v is a Gaussian variable with a zero mean and a unit variance. This approximation was used in the OCICA and will be used in our proposed methods as well.

Due to the blindness to $\mathbf{s}$ and $\mathbf{A}$, ICA is up to two ambiguities [17]: the scale ambiguity and the order ambiguity. Since signal shape is usually more informative than the amplitude in practice, the scale ambiguity is generally less problematic than the order ambiguity. To eliminate the order ambiguity, prior information was incorporated to constrain the directions of IC learning in CICA [20, 21]. Denote the constraint function by $\mathbf{q}(\mathbf{y}) = (q_1(y_1), \dots,$

$q_M(y_M))^T$ ($M \leq m$), and convert them into equality constraints: $q_i(y_i) + z_i^2 = 0$ ($i = 1, \dots, M$, and z_i is a slack variable). OCICA was solved using an augmented Lagrangian function [3]:

$$\mathcal{Q}(\mathbf{W}, \mu, \mathbf{z}, \lambda) = O(\mathbf{y}) + \mu^T (\mathbf{q}(\mathbf{y}) + \mathbf{z}^2) + \frac{1}{2} \gamma \|\mathbf{q}(\mathbf{y}) + \mathbf{z}^2\|^2 + \lambda^T \mathbf{h}(\mathbf{y}) + \frac{1}{2} \gamma \|\mathbf{h}(\mathbf{y})\|^2 \quad (3)$$

where $\mu = (\mu_1, \dots, \mu_M)^T$ and $\lambda = (\lambda_1, \dots, \lambda_N)^T$ are two sets of Lagrange multipliers, γ is the scalar penalty parameter, $\mathbf{z} = (z_1, \dots, z_i, \dots, z_M)^T$, and $\|\cdot\|$ means Euclidean norm. $\mathbf{h}(\mathbf{y}) = (h_1(\mathbf{y}), \dots, h_N(\mathbf{y}))^T$ are the N equality constraints: $E\{(\mathbf{W}\mathbf{x})^2\} = \mathbf{I}$ used to facilitate an easy ICA solution [17]. The last item $1/2\gamma\|\cdot\|^2$ is the penalty term to ensure that the local convexity assumption holds at a solution of the optimization problem [14]. Since the optimal

value of $\mathbf{z}$ can be obtained as $(\mathbf{z}^*)^2 = \max \left\{ 0, -\frac{1}{\gamma} [\mu + \gamma \mathbf{q}(\mathbf{y})] \right\}$ via maximizing

$\mathcal{Q}_1(\mathbf{W}, \mu, \mathbf{z}) = \mu^T (\mathbf{q}(\mathbf{y}) + \mathbf{z}^2) + \frac{1}{2} \gamma \|\mathbf{q}(\mathbf{y}) + \mathbf{z}^2\|^2$ wrt $\mathbf{z}^2$, the augmented Lagrangian function in Eq. 3 can be rewritten as:

$$\mathcal{Q}_{\mathbf{z}^*}(\mathbf{W}, \mu, \lambda) = O(\mathbf{y}) + \mathcal{Q}_{1\mathbf{z}^*}(\mathbf{W}, \mu) + \lambda^T \mathbf{h}(\mathbf{y}) + \frac{1}{2} \gamma \|\mathbf{h}(\mathbf{y})\|^2 \quad (4)$$

where $\mathcal{Q}_{1\mathbf{z}^*} = \mathcal{Q}(\mathbf{W}, \mu, \mathbf{z}^*)$ (see Eq. A.4). Without introducing any ambiguity, $\mathcal{Q}_{\mathbf{z}^*}$ and $\mathcal{Q}$ will be used interchangeably hereafter in this paper, and so will be $\mathcal{Q}_{1\mathbf{z}^*}$ and $\mathcal{Q}_1$. A Newton-like gradient method can be used to solve the optimization problem of CICA [20]:

$$\mathbf{W}_{k+1} = \mathbf{W}_k - \eta \nabla_{\mathbf{W}} \mathcal{Q} / \nabla_{\mathbf{W}} \nabla_{\mathbf{W}} \mathcal{Q} \quad (5)$$

where $\nabla_{\mathbf{W}}$ and $\nabla_{\mathbf{W}} \nabla_{\mathbf{W}}$ mean the first and second derivative respectively, η is the learning rate. This algorithm involves an inversion of the Hessian matrix, which is usually computationally intensive even when an approximation was used [20, 21]. Nevertheless, this matrix inversion can be avoided [19] upon using data prewhitening that is now a general preprocessing step in ICA [17]. We can then get $\nabla_{\mathbf{W}} \mathcal{Q}_1 = [\nabla_{\mathbf{W}_1} \mathcal{Q}_1, \dots, \nabla_{\mathbf{W}_m} \mathcal{Q}_1]$ and $\nabla_{\mathbf{W}} \nabla_{\mathbf{W}} \mathcal{Q}_1 = [\nabla_{\mathbf{W}_1} \nabla_{\mathbf{W}_1} \mathcal{Q}_1, \dots, \nabla_{\mathbf{W}_m} \nabla_{\mathbf{W}_m} \mathcal{Q}_1]$ with:

$$\nabla_{\mathbf{W}_p} \mathcal{Q}_1 = \mu_p \nabla_{\mathbf{W}_p} q_p(y_p) = \mu_p E \{ \mathbf{x} q_p'(y_p) \} \quad (6a)$$

$$\nabla_{\mathbf{W}_p} \nabla_{\mathbf{W}_p} \mathcal{Q}_1 = \mu_p E \{ \mathbf{x}^T \mathbf{x} q_p''(y_p) \} \approx \mu_p E \{ \mathbf{x}^T \mathbf{x} \} E \{ q_p''(y_p) \} = \mu_p E \{ q_p''(y_p) \} \quad (6b)$$

Since the maxima of $\mathcal{J}(\mathbf{y}_i)$ is approached at certain optima of $E\{G(\mathbf{y}_i)\}$, the second item in Eq. 2 can be ignored during the process of pursuing the optimal $\mathbf{W}$ [16]. Also assuming that $G(\cdot)$ is second-order differentiable, we can get the first and second derivative of O wrt $\mathbf{W}$ as:

$$\nabla_{\mathbf{W}} O = -\widehat{c} \nabla_{\mathbf{W}} [E\{G(\mathbf{W}\mathbf{x})\}] = -\widehat{c} E \{ \mathbf{x} G'(\mathbf{W}\mathbf{x}) \} \quad (7a)$$

$$\nabla_{\mathbf{W}} \nabla_{\mathbf{W}} O = -\widehat{c} E \{ \mathbf{x}^T \mathbf{x} G''(\mathbf{W}\mathbf{x}) \} - 2c E^2 \{ \mathbf{x} G'(\mathbf{W}\mathbf{x}) \} \quad (7b)$$

$$\approx -\bar{c}E\{G''(\mathbf{W}\mathbf{x})\} + 2cE^2\{\mathbf{x}G'(\mathbf{W}\mathbf{x})\} \quad (7c)$$

where

$$\bar{c} = 2c[E\{G(\mathbf{W}\mathbf{x})\} - E\{G(v)\}] \quad (8)$$

By inserting Eqs. 6 and 7 and the first and second derivatives of the last two items in Eq. 4 into Eq. 5, we can get:

$$\mathbf{W}^+ = \mathbf{W} - \eta \frac{\bar{c}E\{\mathbf{x}G'(\mathbf{W}\mathbf{x})\} - \mu^T E\{\mathbf{x}\mathbf{q}'(\mathbf{y})\} - 2\lambda E\{\mathbf{y}\mathbf{x}^T\}}{\bar{c}E\{G''(\mathbf{W}\mathbf{x})\} + 2cE^2\{\mathbf{x}G'(\mathbf{W}\mathbf{x})\} - \mu^T E\{\mathbf{q}''(\mathbf{y})\} - 2\lambda} \quad (9)$$

followed by a normalization step:

$$\mathbf{W} = \mathbf{W}^+ / \|\mathbf{W}^+\| \quad (10)$$

and an orthogonalization [16] to prevent different weight vector $\mathbf{w}_i$ from converging to the same optima [16], which could also reduce the accumulated residual errors in the sequential component extractions [34]:

$$\mathbf{W} = (\mathbf{W}\mathbf{W}^T)^{-1/2} \mathbf{W} \quad (11)$$

The Lagrangian multipliers μ and λ are updated with an ascent iteration process [3]:

$$\mu_{k+1} = \mu_k + \gamma (\mathbf{q}(\mathbf{y}) + (\mathbf{z}^*)^2) = \max\{0, \mu_k + \gamma \mathbf{q}(\mathbf{y})\} \quad (12a)$$

$$\lambda_{k+1} = \lambda_k + \gamma \mathbf{h}(\mathbf{y}) \quad (12b)$$

Eqs. 9, 10, 11, and 12 together constitute a new version of the OCICA method. Although this is a simplified and improved version, we will use it as the OCICA in this paper in order to get a fair comparison with the methods to be derived below, because all of these methods are dependent on the same data prewhitening and post-learning weight normalization processes.

Fixed-point CICA algorithms

As in [16], c in Eq. 2 can be set to be 1, and $[E\{G(\mathbf{W}\mathbf{x})\} - E\{G(v)\}]$ can be treated as a constant, so Eq. 7 can be simplified as:

$$\nabla_{\mathbf{w}} O \approx -\bar{c} E\{\mathbf{x}G'(\mathbf{W}\mathbf{x})\} \quad (13a)$$

$$\nabla_{\mathbf{w}} \nabla_{\mathbf{w}} O \approx -\bar{c} E\{G''(\mathbf{W}\mathbf{x})\} \quad (13b)$$

where $\bar{c} = \text{sign}(E\{G(\mathbf{W}\mathbf{x})\} - E\{G(y)\})$. Therefore, the Newton-like CICA algorithm Eq. 5 can be further simplified as:

$$\mathbf{W}^+ = \mathbf{W} - \eta \frac{\bar{c} E\{\mathbf{x}G'(\mathbf{W}\mathbf{x})\} - \mu^T E\{\mathbf{x}\mathbf{q}'(y)\} - 2\lambda E\{\mathbf{y}\mathbf{x}^T\}}{\bar{c} E\{G''(\mathbf{W}\mathbf{x})\} - \mu^T E\{\mathbf{q}''(y)\} - 2\lambda} \quad (14)$$

As mentioned above, the separating matrix $\mathbf{W}$ is forced to be orthogonal and normalized after each iteration, so that $\mathbf{W}\mathbf{W}^T = \mathbf{I}$. A simple Lagrangian function can be formed to minimize O with this constraint: $\mathcal{L} = O + \lambda^T (\mathbf{W}\mathbf{W}^T - \mathbf{I})$. Subsequently, the gradient $\Delta\mathbf{W} \propto \nabla_{\mathbf{W}} \mathcal{L} = \nabla_{\mathbf{W}} O + 2\lambda^T \mathbf{W}$. According to the Karush-Kuhn-Tucker (KKT) condition [22, 29], at a stable point of this optimization problem, this gradient turns to be 0. Because λ is a scalar, this KKT condition indicates $\mathbf{W} \propto \nabla_{\mathbf{W}} O$ and subsequently $\Delta\mathbf{W} \propto \mathbf{W}$. In other words, the updating gradient of $\mathbf{W}$ must fall in the same direction of $\mathbf{W}$ around the stable point, and they will only differ by a scaling factor. This is the basic idea of the fixed-point ICA learning algorithm proposed by Hyvärinen et al. [15, 16]. The same idea can be used here to derive a fixed-point CICA through replacing $\mathbf{W}$ by the calculated gradient at each iteration, and canceling the scaling factor out during the post normalization step (Eq. 10). Three approaches can be used to derive the fixed-point CICA algorithms.

The first approach is similar to that used to derive the FastICA in [15]. Since $\nabla_{\mathbf{W}} \nabla_{\mathbf{W}} \mathcal{L}$ becomes diagonal due to the approximations used in Eq. 13b, multiplying $\mathbf{W}$ with $\nabla_{\mathbf{W}} \nabla_{\mathbf{W}} \mathcal{L}$ will not affect $\mathbf{W}$ after normalization. Resultantly, a fixed-point CICA algorithm (CICA1) can be obtained by multiplying the right side of Eq. 15 with the denominator item in the 2nd item of Eq. 14 and then updating $\mathbf{W}$ with the multiplication result:

$$\mathbf{W}^+ = \bar{c} E\{\mathbf{x}G'(\mathbf{W}\mathbf{x})\} - \bar{c} E\{G''(\mathbf{W}\mathbf{x})\} \mathbf{W} - \mu^T (E\{\mathbf{x}\mathbf{q}'(y)\} - E\{\mathbf{q}''(y)\} \mathbf{W}) \quad (15)$$

As compared to the simplified OCICA in Eq. 9, λ is canceled out in CICA1, so only one Lagrangian multiplier is required in CICA1. The complete CICA1 learning process consists of Eqs. 15, 10, 11, and 12a.

The second approach (CICA2) is to update $\mathbf{W}$ directly with the Newton-like learning gradient (the second item in Eq. 15) and dropping the diagonal denominator for the same reason as described above. The corresponding learning procedure, which is called CICA2 in this paper, is:

$$\mathbf{W}^+ = -\bar{c} E\{\mathbf{x}G'(\mathbf{W}\mathbf{x})\} + \mu^T E\{\mathbf{x}\mathbf{q}'(y)\} \quad (16)$$

The third approach [31] is to remove the last item in the right side of Eq. 4. That item is usually added to enhance the local convexity so that the optimization problem is guaranteed to converge [3], but can be ignored here because of the additional normalization step involved in ICA. As we will discuss later, this simplification-based CICA will still converge to an optimum point. After removing the last term, the Lagrangian object function becomes:

$$\mathcal{L}_{\mathbf{Z}^*}(\mathbf{W}, \mu) = O(y) + \mathcal{L}_1(\mathbf{W}, \mu) + \lambda^T (\mathbf{W}\mathbf{W}^T - \mathbf{I}) \quad (17)$$

The optimal $\mathbf{W}$ can be approached by setting the first derivative of $\mathcal{L}_{\mathbf{Z}^*}$ wrt $\mathbf{W}$ to zero, resulting in the following equation:

$$2\lambda^T \mathbf{W} = \nabla_{\mathbf{w}} O + \nabla_{\mathbf{w}} \mathcal{L}_1 \quad (18)$$

As the preceding scaling factor 2λ will be removed due to the weight normalization after each learning procedure, the above equation can be simplified as:

$$\mathbf{W} = \nabla_{\mathbf{w}} O + \nabla_{\mathbf{w}} \mathcal{L}_1 \quad (19)$$

which is exactly the same as CICA2 as described above in Eq. 16. However, in order to make the algorithm converge for both super- and sub- Gaussian sources as discussed in the following stability analysis, Eq. 16 needs to be changed as:

$$\mathbf{W}^+ = E\{\mathbf{x}G'(\mathbf{W}\mathbf{x})\} \pm \mu^T E\{\mathbf{x}\mathbf{q}'(\mathbf{y})\} \quad (20)$$

the plus sign + before μ is for using the distance-based closeness measure q_1 , and the negative sign - is for using the correlation-based closeness measure q_2 .

Closeness measures

As in [20, 21], the closeness can be measured by distance of correlation to the reference. To facilitate the method convergence and allow $\mathbf{y}$ (or $\mathbf{W}$) to evolve along both the same or the opposite direction of $\mathbf{r}$, two functions can be used: $\mathbf{q}_1(\mathbf{y}) = 1/4(\min(E\{(\mathbf{y} - \mathbf{r})^4\}, E\{(\mathbf{y} + \mathbf{r})^4\}) - \xi^4)$ and $\mathbf{q}_2 = 1/4(\xi^4 - E\{(\mathbf{y}^T \mathbf{r})^4\})$, where ξ^2 is the desired closeness.

Stability analysis

The stability analysis given in [21] assumes $\lambda > 0$, which is not a classical KKT condition and can not be proven. In the following, we give a full derivation of the stability analysis for CICA1 and CICA2, which can be easily adapted for OCICA too. The standard KKT conditions can be described as: $\nabla_{\mathbf{w}} \mathcal{L}_{z^*}(\mathbf{W}^*, \mu^*) = 0$, $\mathbf{q}(\mathbf{y}^*) \leq 0$, $\mu^* \geq 0$ and $\mu^{*T} \mathbf{q}(\mathbf{y}^*) = 0$, which represent the necessary conditions for whether the solution $\mathbf{y}^*$ is an optimum of CICA. Since CICA1 is still based on the object function $\mathcal{Q}$ in Eq. 4, the corresponding Hessian matrix of the object function can be approximated by:

$$\nabla_{\mathbf{w}} \nabla_{\mathbf{w}} \mathcal{Q} = -\bar{c} E\{G''(\mathbf{W}\mathbf{x})\} + \mu^T E\{\mathbf{q}''(\mathbf{y})\} + 2\lambda + 2\gamma \mathbf{h}(\mathbf{y}) + 4\gamma E\{\mathbf{W}\mathbf{W}^T\} \approx -\bar{c} E\{G''(\mathbf{W}\mathbf{x})\} + \mu^T E\{\mathbf{q}''(\mathbf{y})\} + 2\lambda + 4\gamma \quad (21)$$

Eq. 12b and $E\{\mathbf{W}\mathbf{W}^T\} = \mathbf{I}$ have been used in the above derivation. The first item in the Hessian matrix, $-\bar{c} E\{G''(\mathbf{W}\mathbf{x})\}$ can be ensured to be no less than 0 by choosing the appropriate nonlinear function $G(\cdot)$ as described in [21]. For the two closeness functions used in CICA (see the following sub-section), $\mu^T E\{\mathbf{q}''(\mathbf{y})\} \geq 0$ is satisfied. If we choose a large positive value for γ , which is true in Lagrangian multiplier-based optimization [3], we can then make sure $2\lambda + 4\gamma > 0$ and $\nabla_{\mathbf{w}} \nabla_{\mathbf{w}} \mathcal{Q} > 0$. Different from [21], $\lambda > 0$ is not required here. This relation is further guaranteed in CICA1 for any positive γ since λ is canceled out after taking the fixed-point manipulation (Eq. 15), which is equivalent to $\lambda = 0$. The approximation of $E\{\mathbf{W}\mathbf{W}^T\} = \mathbf{I}$ does not change the validity of this condition since $E\{\mathbf{W}\mathbf{W}^T\} > 0$ and we can always adjust γ if that approximation is not true. In summary, CICA1 has a positive definite Hessian matrix, ensuring that any solution of the KKT conditions, $\mathbf{y}^*$, is an optimum of the CICA optimization problem. Due to the use of a Newton-like gradient updating process as well as the modified closeness measures, CICA1 is also guaranteed to have an at least quadratic convergence rate [26].

Similar to the convergence analysis for FastICA given in [27], the stability of CICA2 requires the following assumption and the full analysis is given in the Appendix B.

Assumption 1: Functions $g(\cdot) = G'(\cdot)$ and $q'(\cdot)$ are odd, twice differentiable, which satisfies:

$$E\{s_i g(s_i)\} \pm \mu E\{s_i q'(s_i)\} \neq 0 \quad (22)$$

for all sources s_i , $i = 1, \dots, m$.

$g(\cdot)$ is ensured to be odd since it is the derivative of the even nonlinear function $G(\cdot)$. For the two closeness functions described in last sub-section, $q'(\cdot)$ is also guaranteed to be odd. The 2 order differentiability can be easily verified for the functions used in CICA. With CCCM as the closeness measure, $E\{s_i g(s_i)\} - \mu E\{s_i q'(s_i)\} > 0$ and then Eq. 22 is satisfied. With MSE as the closeness measure, $E\{s_i g(s_i)\} + \mu E\{s_i q'(s_i)\} > 0$, which also makes Eq. 22 true.

Since the constraints themselves may not be optima at all, a too small closeness threshold will destroy the convergence of any CICA algorithm. As proposed in [21], the closeness threshold ξ can be initialized with a small value to avoid any local maxima, and gradually increased so that the algorithm converges at the global maximum corresponding to the desired IC. When the threshold becomes large, the inequality constraints will be automatically satisfied, and μ will be 0, and the convergence performance of CICA1 will be the same as that of FastICA, which has been fully proven in [16, 27]. For CICA2, Assumption 1 will be automatically true for either of the closeness measure functions, and its convergence can be proven following the analysis in Appendix B as well.

CICA for fMRI data analysis, temporal constraints, and statistical inference

For fMRI data analysis, ICA is usually applied along the temporal domain to extract some independent spatial component maps. Using the above CICA on fMRI data will require spatial brain activation patterns that are generally to be discovered during data analysis. A more widely available constraint is the temporal behavior of the brain activity since many fMRI experiments do have explicit design paradigms. This kind of temporal constraints acts on the mixing matrix $\mathbf{A}$, or equivalently the separating weight $\mathbf{W}$. The above described algorithms should be accordingly adjusted by replacing $\mathbf{q}(\mathbf{y})$ with $\mathbf{q}(\mathbf{W})$ in the CICA learning process.

As a data-driven method, ICA does not have a statistical framework for assessing the extracted components. A widely used approach to threshold the component map is to convert it to a z map, and applying a standard z significance threshold [24]. That method can only tell whether a voxel value in the component map is significantly different from 0 or not. To assess whether these patterns are invoked by the tasks or the stimuli, a statistical inference can be achieved by relating the component map associated TC to the design function. Since CICA is right designed to select the interested components, population level inference can be easily made using a standard one-sample t-test.

Materials and methods

Method implementations

The FastICA (<http://www.cis.hut.fi/projects/ica/fastica>) package was modified to implement all the CICA algorithms in Matlab (Mathworks, Natick MA) scripts. The proposed CICA1 and CICA2 were compared to OCICA and FastICA (fICA) for performance evaluations. $G(\cdot)$ was chosen to be the log cosh function [16], and the constant c was set to be 1. The termination criteria for the ICA learning procedure was set to be $\|\nabla \mathbf{W}\| < 1e-6$. For CICA,

the Lagrangian multiplier μ was initialized to be 1, and the positive scalar parameter γ for the quadratic penalty item in Eq. 3 was set to $0.1 \times 4^{k-1}$ (k is the iteration index). As suggested in [20, 21], the learning rate η in OCICA was initialized to be 1 and decreased by 10 times after every 200 iterations if the learning process did not converge. All input data were centered and dewhitened using PCA through singular vector decomposition [28]. And the number of ICs was estimated after PCA decomposition so that the remained components represented no less than 99% of the original data's variance. For each ICA decomposition run, a maximum of 1000 iterations was allowed. If an algorithm did not meet the convergence criteria as described above, a new decomposition run will be started.

A sequential mode implementation of each CICA was given and tested in the following corresponding listing. The batch mode implementations can be easily obtained based on the equations presented above.

Implementation of CICA2 was nearly the same as that of CICA1 in Listing 2 except that in CICA2, line 8 of Listing 2 should be replaced by “update $\mathbf{w}_i$ using Eq. 20”. In all CICA implementations, $\mathbf{q}_2(\cdot)$ was used as the closeness function. The closeness threshold ε was set to 0.9, and reduced by 0.2 after each of the first 2 iterations, then reduced by 0.1 after each of the next 3 iterations, then reduced by 0.02 after each iteration until it was less than 0, then μ for the constraint-related learning terms became 0 and the effects of constraints were completely released.

Evaluations using 1D synthetic data

Thirty experiments were performed using randomized mixing matrixes and source signals (each with 2000 samples). Fig. 1A shows the first 400 time points of the 6 source signals, S1, S2, S3, S4, S5, and S6 used in one of the 30 experiments. S2 is an ECG signal; S4 is a brain activation function generated by convolving a boxcar function with a canonical hemodynamic reference function (HRF) [9]; S5 is a Gaussian noise which was randomly generated each time; S6 is a 1/f noise which was regenerated at each time of simulation. During each experiment, S2 was randomly extracted from a segment of a long ECG signal; S1, S3, and S4 were generated by randomly shifting the full waves as shown in Fig. 1A by ± 100 points. At each experiment, the regenerated 6 sources were mixed 5 times using 6 randomly generated mixing matrices. The mixed data were then separated using the 4 ICA algorithms as described above. For CICA, the 4 reference signals shown in Fig. 1B were used as the constraints to extract the signals of interest. Those references were generated to have the same major frequency item as in their corresponding sources so that they were correlated to the sources. They were not chosen to be the same as the sources in order to test the capability of CICA for source sorting using constraints with large shape deviations from the sources of interest.

After decomposition, the fICA extracted ICs were sorted based on their correlations to the original signals; their signs were also flipped if the correlations were negative. For CICA, only the signs of the ICs were flipped if their correlations to the constraints were negative. The accuracy of each IC was measured using the signal-to-noise-ratio (SNR) in dB: $\text{SNR} = 10 \log_{10}(\sigma^2/\text{MSE})$, where σ^2 represents the variance of the source signal and MSE is the mean square error between the extracted IC and original source signal. The performance of each implemented algorithm for source separation was measured with the convergence time and an aggregate performance index (PI):

$$\text{PI} = \frac{1}{m} \left(\sum_{i=1}^m \text{rPI}_i + \sum_{j=1}^m \text{cPI}_j \right) \quad (23)$$

where $rPI_i = \sum_{j=1}^m \frac{|p_{ij}|}{\max_k p_{ik}} - 1$, and $cPI_j = \sum_{i=1}^m \frac{|p_{ij}|}{\max_k p_{kj}} - 1$, and p_{ij} denotes the (i,j)-th element of the permutation matrix $\mathbf{P} = \mathbf{W}\mathbf{A}$.

Evaluations with 2D synthetic data

Synthetic fMRI data were generated from 6 source components ($\mathbf{S}$) and the associated 6 time courses ($\mathbf{A}$) as shown in Fig. 2 by $\mathbf{X} = \mathbf{AS} + \mathbf{v}$, where $\mathbf{v}$ is random Gaussian noise. The Gaussian noise added here was to simulate the noise in real fMRI data after preprocessing and should not be confused as one of the target sources. Signal intensities were increased in several brain regions in the grey matter to simulate brain activations in the first 3 components: S1, S2, and S3. S1 was created to simulate activation in response to a standard “on-off” fMRI stimulus series as shown in the associated time course (tc1) in Fig. 2, which is a box function convolved with a canonical HRF [9]. S2 and S3 were generated to simulate some transient activations. S4 was related to resting brain activity that associated with a slowly fluctuating time course (tc4). S5 was related to head motion as the abrupt motion can be perceived easily around the 40th timepoint in tc5. S6 was simulating a quickly fluctuating noise component with CSF involved. Gaussian noise was added to simulate the additive noise in real fMRI environment. The same signal mixing process was repeated 50 times with different Gaussian noise. The resultant 50 pseudo-fMRI time series were then fed into the above mentioned 4 ICA methods to separate the source components using the same parameter settings. The image masks shown in Fig. 2C were used as spatial constraints for the source components that to be separated using CICA. The references were chosen to be not the same as the sources but just with similar spatial locations for them to be correlated with the sources. After decomposition, the fICA extracted ICs were sorted based on their correlations to the original signals; their signs were also flipped if the correlations were negative. For CICA, only the signs of the ICs were flipped if their correlations to the constraints were negative. PI and SNR were collected to measure the algorithm performance.

The same pseudo-fMRI data were used to validate the performance of the proposed CICA methods when temporal constraints were used. Giving the fact that the stimulus paradigm is usually the only available reference for the temporal behavior of brain activations, temporally constrained ICAs were only used to extract the first source component: S1, as S1 was generated to represent brain activation in response to the pseudo stimulus paradigm. The box function of the pseudo stimulus paradigm (as shown in Fig. 6.TREF) was used as the temporal constraint, which was subsequently converted into the constraint for the separating weight $\mathbf{W}$ using the same dewhitening matrix as used for data dewhitening. Corresponding adjustments in Eqs. 15, 20, and 9 were made by replacing $\mathbf{q}(\mathbf{y})$ with $\mathbf{q}(\mathbf{W})$ in order to implement the temporally constrained ICA learning process. Other configurations were the same as in the spatially constrained cases.

GLM was also applied to the random noise contaminated synthetic fMRI data using TC1 in Fig. 2A as the regressor. The statistical parametric maps were collected. To compare the sensitivity and specificity of CICA and GLM for brain activation detection, receiver operator characteristic (ROC) [25] curves were calculated by moving the threshold from the minimum to the maximum of the signal intensity of the extracted ICs or the statistical parametric maps generated by GLM and then calculating the true positive rate and false positive rate from the suprathresholded (using the current threshold) voxel as described in [33]. The true activated voxels were defined to be those within the two big blobs in s1 shown in Fig. 2A. The area under curve (AUC) was used as the surrogate score for the efficacy of each method.

Evaluations with sensorimotor fMRI data

Imaging experiments were performed using the same protocol as mentioned in previous work [35]. Briefly, both T1-weighted images and BOLD fMRI data were obtained on a 3T Siemens Trio whole body MR scanner. Seventeen young healthy subjects were scanned with written informed consent obtained before scanning following an Institutional Review Board approved protocol. The acquisition parameters for BOLD were: 30 interleaving slices, slice thickness=3 mm, flip angle=75°, TR/TE=3000/30 msec, in plane FOV=220 mm. The stimuli comprised 5 pairs of a 48 sec 8.3 Hz reversing black and white checkerboard and a 48 sec black screen block (baseline). Each participant was also asked to perform a self-paced left hand only fingertapping task during the visual stimuli.

All data preprocessing was performed using SPM (Wellcome Department of Cognitive Neurology, London, UK, <http://www.fil.ion.ucl.ac.uk/spm>) based batch scripts [32]. Functional images of each subject were motion corrected, coregistered with the structural images, and smoothed with an isotropic Gaussian filter (FWHM=6 mm). The structural image was normalized to the Montreal Neurological Institute/International Consortium for Brain Mapping (MNI/ICBM) 152 standard brain using SPM5. Standard GLM was performed on the spatially smoothed sensorimotor fMRI data using the canonical HRF function [9] convolved design function as the reference. The temporal autocorrelated noise was considered using the AR(1) model [8], and the high frequency noise was lowpassed with a cutoff of 1/128 Hz. The 3 directional translation and 3 directional rotation movement time course estimated during image realignment were included as nuisance covariates in the GLM model. Statistical parametric maps (SPMs) of the task minus baseline contrast were collected at each voxel using a standard t-test, which were subsequently normalized to the MNI space using the same normalization parameters estimated from the T1 images and finally a group analysis was performed by running a one-sample t-test on these spatially normalized parametric maps.

For CICA, the realigned and smoothed fMRI series were filtered using the same GLM model as described above but without the HRF convolved reference function in the model in order to preserve the task effects for the successive activation detection. The filtered images were obtained from the residual images after running the GLM estimation. CICA1, CICA2, OCICA, and fICA were used to analyze the preprocessed fMRI data. The box-car design function was used as the temporal constraint in all CICAs for extracting the activation component. Both OCICA and fICA were applied to each subject's data first, and the number of runs that they needed to converge was recorded. Then, the proposed CICAs were run on the same data for the same number of times in order to validate their stability. The extracted IC map of interest was sign flipped if the associated TC was negatively correlated to the box-car function. For fICA, the IC of interest was identified as the one whose TC presented the highest correlation to the box-car function. The extracted ICs using the constraints were then spatially normalized to the MNI standard space and a one-sample t-test was performed to assess the group level statistical significance for these revealed spatially distributed activation patterns. The same ICA process was run several times until it converged.

Results

CICAs for 1D source separation

All algorithms converged for all 150 source separation trials when log cosh was used as the nonlinear function $G(\cdot)$. The average convergence time of CICA1, CICA2, OCICA, and fICA were 0.246, 0.237, 0.507, and 0.028 sec, respectively. Although the time saving was moderate, the differences were significant across different methods ($P < 0.000001$). The proposed CICAs still converged in another 150 separation trials when the kurtosis was used

as the nonlinear function. By comparison, OCICA and fICA both did not converge for a few times. In the following, the performance validation results were based on the first nonlinear function used in the ICA algorithms. Fig. 3 shows the aggregate PI, and mean SNR of each ICA method for separating the 4 sources of interest from their randomly mixed observations. As compared to OCICA and fICA, the proposed 2 CICA algorithms yielded significantly lower PI value. OCICA did not improve the aggregate PI as compared to fICA. CICA1 and CICA2 yielded the same PI. As shown in Fig. 3B, CICA1 and CICA2 performed the same and both yielded higher SNR than OCICA and fICA for separating s2, s3, and s4. CICA1 and CICA2 also demonstrated higher SNR than OCICA for extracting s1. OCICA performed better than fICA only for separating the 3rd source. All the difference reported above were statistically significant at $P < 0.005$ (one-way ANOVA).

Fig. 4 shows the first 400 points of the signals extracted from one randomly mixed dataset using CICA and fICA. The 4 signals of interest extracted by CICA1 (Fig. 4A) and CICA2 (Fig. 4B) were almost identical to the original sources as displayed in Fig. 1A. The most obvious performance differences were found in the 3rd and 4th sources, Y3 and Y4 extracted using OCICA and fICA, though minor variations still presented in Y1 and Y2 in Fig. 4C and 4D.

CICAs for 2D source separation

All tested ICA algorithms, including the spatially constrained CICA1, CICA2, OCICA, and fICA converged for all 50 source separation trials for those 50 different random noise contaminated 2D source data. Fig. 5 shows the aggregate PI, and SNR of each ICA method for separating the 6 sources. Using the same spatial constraints as shown in Fig. 2C, CICA1 and CICA2 showed lower PI values than OCICA and fICA, and OCICA yielded lower PI than fICA. No PI difference was found between CICA1 and CICA2. Source extraction SNR was shown in Fig. 5B; spatial constraints were used in all the 3 CICA methods. CICA1 and CICA2 demonstrated higher SNR for extracting all the 6 sources than OCICA and fICA. OCICA presented higher SNR for extracting S1, S2, S3, and S5 than fICA. No performance difference was found between CICA1 and CICA2. Both CICA1 and CICA2 yielded slightly increased (not significant) CC for extracting S1 as compared to OCICA when the same temporal constraint was used (results not shown here). However, the latter did not converge for 2 times out of the 50 source separation runs. CICA with temporal constraint still outperformed fICA though only the first component was tested. The significance level used for reporting the differences in all comparisons was $P < 0.005$ (one-way ANOVA).

Fig. 6 shows the 2D source separation results using fICA and CICA with either spatial or temporal constraints. To save space, only the separation results for the 1st source were displayed here. As shown in Fig. 6A (AT), BS (BT), ASTC (ATTC), BSTC (BTTC), the IC maps and TCs extracted by the proposed CICA methods using either spatial (or temporal) constraint were almost the same. When converged (as shown in Fig. 6), OCICA still yielded a good quality source separation, while fICA did not perform well for separating S1 from the same noise contaminated mixed signal. By comparing both the IC maps and the TCs in Fig. 6, we can find that CICA with spatial constraints demonstrated slightly better spatial map quality in terms of less background noise while CICA with temporal constraints yielded a slightly better temporal quality in terms of less temporal fluctuation in the TC.

In ROC analysis, both the proposed CICAs and OCICA (when converged) yielded the same AUC values (mean $\pm$ std = $0.9998464 \pm 5e-05$), better than GLM-based method ($0.9905818 \pm 6e-04$). However, fICA yielded the lowest AUC ($0.9543832 \pm 5.48E-02$) in all compared methods.

CICA for detecting sensorimotor brain activations

OCICA and fICA converged at the first time when they were applied to 15 subjects' data, but both did not converge for 2 subjects until the 3rd run and OCICA did not converge until the 2nd run. The proposed CICA methods converged at the 1st run for analyzing all subjects' data, and also converged at the 2nd and 3rd time when they were applied to those 2 particular subjects' data for which OCICA and fICA did not converge at the 1st run. OCICA also converged at the 3rd run for those 2 subjects. Fig. 7 shows the statistical parametric maps of the 3 CICA methods, fICA, and GLM-based group level analyses using the sensorimotor fMRI data. At the same significance level ($t=4.67$, $P=0.05$ with the false detection rate based multiple comparison correction [10]), the proposed 2 CICA methods demonstrated higher sensitivity than OCICA in terms of higher t -values in all regions activated by the sensorimotor task. The 2 new CICA methods also yielded larger suprathreshold clusters in thalamus and left Insula than those revealed by OCICA. The peak locations of the group level statistical parametric maps of CICA1, CICA2, and OCICA based analyses are in the same location in both visual cortex (Vis) and motor cortex (Mot), with the value of 12.72/12.73/11.57 (CICA1/CICA2/OCICA) for Vis, and 12.87/12.87/12.62 (CICA1/CICA2/OCICA) for Mot. All CICA methods outperformed fICA for sensorimotor task activated brain activations in terms of larger suprathreshold clusters in all activated brain regions and also larger peak- t values in both Vis and Mot (9.53/8.05 in Vis/Mot in fICA-based group level statistical parametric map). As compared to GLM-based method, all CICA yielded higher sensitivity in terms of both peak t -value (12.12/8.5 in Vis/Mot in GLM based group analysis result) and cluster extensions in all activated brain regions except the left Insula.

Discussion

Two fixed-point CICA algorithms were proposed in order to improve the stability of the OCICA method. CICA1 is based on a Newton-like gradient learning process; CICA2 is using a regular gradient descent approach. Algorithms for applying constraints to either the components or to the mixing time courses were derived especially for 2D source separations. No learning rate was used in both proposed algorithms. 1D randomly mixed synthetic data and 2D random noise contaminated synthetic data were used to evaluate the performance of the proposed algorithms with comparisons to the OCICA and standard ICA. The 1D dataset was used to validate the algorithms for separating randomly mixed signal. The 2D synthetic data were simulating fMRI data and were used to validate the proposed algorithms for source separation in the case of randomly generated noise; as well as their performance for source separation with either spatial constraints or temporal constraints. In both 1D and 2D source separation experiments, the proposed methods were demonstrated to be better than OCICA in terms of higher stability (converged for all experiments for both 1D and 2D source separations), higher source separation quality, and lower PI. All CICA methods outperformed fICA for source separation with higher CC and lower PI due to the use of constraints. No performance difference was observed between CICA1 and CICA2. The proposed CICA methods presented higher convergence rate than OCICA. Because fICA does not include any additional constraints, it showed the fastest convergence rate when converged. The ROC analysis showed that CICA yielded better sensitivity/specificity performance than GLM for brain activation detection. However, fICA was demonstrated to be even less effective than GLM for the pseudo-activation detection, which may partly due to the interference of noise added to the mixed signal. Another reason may be that FastICA does not yield the best separation results for fMRI data as compared to other ICA algorithms [7]. The additive noise could contribute significant background fluctuations, and the sub-optimal fMRI data source separation could make ICA converge to a local optima, which subsequently contributes component split or cross-contaminations. Among the 50 times

source separations, the fICA identified ICs had large background noise or contaminations from other sources as shown in Fig. 6.D, which markedly degraded the sensitivity and specificity performance for the activation detection. By attracting the separating weight to the correct direction in the first several iterations, CICA provides a practical solution to deal with both the noise interference and the local optima problem.

A detailed stability analysis was provided for the proposed CICA methods. It is worth to note that the stability analysis for the OCICA [21] was based on an additional KKT condition $\lambda > 0$, which is not necessary to be true. However, the convergence of the OCICA is still satisfied if the stability analysis is based on the classical KKT condition and the convexity condition as described in this paper.

Random noise is usually not explicitly modeled in standard ICA since it can be treated as a regular IC if there is only one Gaussian noise source. However, noise added after source mixing could drastically degrade the source separation quality of regular ICA as demonstrated in Fig. 6.D. By contrast, CICA is much less sensitive to this additive noise as shown in Fig. 6. An interesting finding in the additive noise contaminated 2D source separation simulations is that better separation quality for IC can be achieved by using CICA with constraints for the ICs (the so-called spatial constraints in the 2D data separation simulations) and better separation quality for TC can be achieved by using CICA with constraints for the TCs (the so-called temporal constraints in the 2D data separation simulations), suggesting gaining further quality improvement for both ICs and the associated TCs using constraints for both of them if we do have the constraints.

Since the noise environment and source features might be quite different in real fMRI from what was simulated in the synthetic datasets, sensorimotor fMRI data were used to further validate the proposed methods for analyzing fMRI data. The proposed 2 CICA algorithms were able to converge at the first time when they were applied to each subject's data. As compared to OCICA, the proposed methods demonstrated increased sensitivity for detecting brain activations in response to the sensorimotor and visual task. As in the simulations, all CICA methods demonstrated better activation detection sensitivity than fICA due to the use of constraints. Both CICA1 and CICA2 performed nearly the same for detecting both visual and sensorimotor activations. CICA including the proposed new CICA methods and OCICA (when converged) yielded much higher sensitivity for sensorimotor brain activation detection than the GLM-based univariate method. The aforementioned ROC analysis results and the demonstrated superiority of CICA for the well-characterized sensorimotor task activation detection suggest that the proposed CICA methods can be employed as sensitive tools for brain activation in fMRI. One reason for this sensitivity gain is the multivariate data processing in ICA. As each input sample is treated as a single unit, ICA-based fMRI data analysis is searching the spatial patterns as an entity rather than pursuing a collection of discrete patterns separately at each voxel like in the univariate GLM. Because the spatially coherent activation patterns are more robust to noise interference than each single voxel if the noise is spatially independent which is a quite common assumption taken in the real MRI field [12], assessing the patterns as a whole should then gain a sensitivity increase as compared to the way that assesses the patterns at each isolated voxel. However, standard ICA especially fICA might not be able to fully demonstrate this benefit as partly shown in Fig. 7. A right solution is then to incorporate prior information into the ICA learning process as in CICA.

The sensorimotor data analysis results (Fig. 7) showed that all CICA methods and fICA yielded less sensitivity than GLM in left Insula in terms of lower peak t-value and small cluster size, which actually reflected a caveat of a multivariate fMRI data analysis method that it could induce a sensitivity loss to some focal brain activations if they are not coherent

with the overall spatially connected activation pattern. Actually, we have identified in [31] that the left insula was activated less coherently than the other three regions, resulting in a less sensitive detection using ICA and CICA. Nevertheless, combining prior information and the data-driven property of ICA, CICA yielded better sensitivity for brain activation detection in the target visual and motor cortex in a well characterized sensorimotor fMRI experiment. In reality, resting fMRI data analysis is one of the hottest ICA application in fMRI. Since several resting networks have been repeatedly identified in the literature, it is possible to build templates for those networks using CICA and retrospectively use them in CICA for general resting state analysis. However, either topic needs additional extensive future work and should be reported separately.

One issue of CICA is that adding constraints could possibly affect the blindness feature of ICA. Similar to [20, 21], this potential bias is minimized by quickly releasing the constraints, so that the constraints will be totally inactive after several iterations and the rest of iterations will be the same as those in standard ICA. From this point of view, adding constraints does not change the data-driven property of ICA. Including constraints does not force ICA to produce the desired ICs if there does not exist any separating weights or ICs close to the constraints, but CICA through the proposed CICA1, CICA2, or OCICA (when an appropriate learning rate is used) is guaranteed to converge if some local optima exist regardless of whether they are close to the constraints or not. All the simulations demonstrated that the proposed CICA even the OCICA (when converged) were capable to identify the sources when the corresponding constraints were chosen to be a binary pulse or box-car function or binary image mask which are correlated to but with large deviations from the latent sources.

Although the proposed methods do not use any learning rate, they are still using two other parameters, the Lagrangian multiplier μ and the scalar penalty parameter γ . Different settings could potentially give different convergence rate. But we did not observe significant difference when the initial value of μ was set to be any number between 0.1 to 5 and when two additional different updating rules: $0.1 \times 5^{k-1}$ and 4^{k-1} were used for γ . More details about these two parameters can be found in [3].

For CICA itself, we noticed that there is another approach called semi-blind ICA [1], which incorporates the prior information by adding an additional adjusting step to the mixing matrix to attract the separating weights to the desired directions. A possible issue is that appending an additional adjusting step is not guaranteed to keep the stability of the original ICA. The two factors for controlling the additional adjustment may not be easy to be tuned for different applications too. By contrast, the convergence of the proposed CICA methods was fully proven. As the focus of this paper was CICA method development, a direct comparison between CICA and semi-blind ICA would be beyond the scope of this paper and was not provided. We also noticed that a CICA algorithm very similar to CICA2 proposed in this paper has recently been published by Lin *et al.* [18] using a Lagrangian similar to Eq. 17. Although they did not provide a stability analysis for their method, they had demonstrated the superiority of their improved CICA (similar to CICA2 in this paper) to the improved OCICA proposed by the same group [19].

Acknowledgments

This research was supported by NIH grants: R21DC011074, R03DA023496, RR02305, R21DA026114, R01DA025906 and NS045839.

APPENDIX

A The KKT condition and derivations of $\mathcal{L}_{1z^*}$

According to KKT condition [29], μ and $\mathbf{q}$ should be orthogonal at the local optimal point:

$$\mu^T \mathbf{q}(\mathbf{y}) = 0 \quad (\text{A.1})$$

Considering $\gamma > 0$, we can obtain the following equations using the optimal solution of $(\mathbf{z}^*)^2$ and A.1:

$$\mu^T (\mathbf{z}^*)^2 = 0 \quad (\text{A.2})$$

$$\mu^T (\mathbf{q}(\mathbf{y}) + (\mathbf{z}^*)^2) = 0 \quad (\text{A.3})$$

Using Eqs. A.1, A.2, and A.3, we can get the following classical equation of $\mathcal{L}_{1z^*}$ (simplified as $\mathcal{L}_{1z}$) as described in [3]:

$$\begin{aligned} \mathcal{L}_{1z}(\mathbf{W}, \mu) &= \frac{\gamma}{2} (\mathbf{q}(\mathbf{y}) + (\mathbf{z}^*)^2) = \frac{1}{2\gamma} (\max\{\mu + \gamma \mathbf{q}(\mathbf{y}), 0\} - \mu)^2 \\ &= \frac{1}{2\gamma} (\max^2\{\mu + \gamma \mathbf{q}(\mathbf{y}), 0\} + \mu^2 - 2 \max\{\mu^2 + \gamma \mu^T \mathbf{q}(\mathbf{y}), 0\}) \\ &= \frac{1}{2\gamma} (\max\{\mu + \gamma \mathbf{q}(\mathbf{y}), 0\} - \mu)^2 \end{aligned} \quad (\text{A.4})$$

B Proof of convergence of CICA2

As in [27], let's assume the source signals s_i are zero mean, unit variance, so that they are whitened because of their independence. Since the input data $\mathbf{x}$ is first whitened and normalized before ICA decomposition to obtain vector $\mathbf{z} = \mathbf{V}\mathbf{x}$ for which $E\{\mathbf{z}\mathbf{z}^T\} = \mathbf{I}$, we can have

$$\mathbf{z} = \mathbf{V}\mathbf{x} = \mathbf{V}\mathbf{A}\mathbf{s} \quad (\text{B.1})$$

which is actually the real input to the CICA2 learning process as described in Eq. 20. It is easy to see that $(\mathbf{V}\mathbf{A})^T (\mathbf{V}\mathbf{A}) = \mathbf{I}$. If the mixing matrix $\mathbf{A}$ were known, the sources could be estimated as $\mathbf{s} = (\mathbf{V}\mathbf{A})^T \mathbf{z}$. To facilitate an easy derivation of the convergence stability, a simplifying linear transformation is made in [27]

$$\mathbf{U} = \mathbf{W}(\mathbf{V}\mathbf{A}) \quad \bar{\mathbf{U}} = \bar{\mathbf{W}}(\mathbf{V}\mathbf{A}) \quad (\text{B.2})$$

where $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_m)^T$. Combining with Eq. B.1, we get $\mathbf{W}\mathbf{z} = \mathbf{W}(\mathbf{V}\mathbf{A})\mathbf{s} = \mathbf{U}\mathbf{s}$ and $\mathbf{w}_i^T \mathbf{z} = \mathbf{u}_i^T \mathbf{s}$.

Replacing $\mathbf{x}$ with its whitened version $\mathbf{z}$ in Eq. 20, and left multiplying both sides of Eq. 20 with $(\mathbf{V}\mathbf{A})^T$, we get

$$\bar{\mathbf{u}}_i = E\{\mathbf{s}g(\mathbf{u}_i \mathbf{s})\} \pm \mu^T E\{\mathbf{s}q'(\mathbf{u}_i \mathbf{s})\} \quad (\text{B.3})$$

Note that this derivation has considered $s = (\mathbf{VA})^T \mathbf{z}$.

By right multiplying both sides of Eq. 11 with $(\mathbf{VA})$, we can get [27]

$$\mathbf{W}(\mathbf{VA}) = \left(\bar{\mathbf{W}}(\mathbf{VA})(\mathbf{VA})^T \bar{\mathbf{W}} \right)^{-1/2} \bar{\mathbf{W}}(\mathbf{VA}) \quad (\text{B.4})$$

which is equivalent to say [27]

$$\mathbf{U} = \left(\bar{\mathbf{U}} \bar{\mathbf{U}}^T \right)^{-1/2} \bar{\mathbf{U}} \quad (\text{B.5})$$

Therefore, Eqs. B.3 and B.5 form the general CICA2 algorithm in the transformed space.

Based on Assumption 1, we have *Lemma 1*: Matrix $\mathbf{U} = \mathbf{D} = \text{diag}(\pm 1, \dots, \pm 1)$ is a fixed point of Eqs. B.3 and B.5.

Proof. Since the i -th row vector of $\mathbf{D}$ is $\mathbf{d}_i^T = (0, \dots, \pm 1, \dots, 0)$, we have $\mathbf{d}_i^T \mathbf{s} = \pm s_i$, $g(\mathbf{d}_i^T \mathbf{s}) = g(\pm s_i)$ and $q'(\mathbf{d}_i \mathbf{s}) = q'(\pm s_i)$, which are independent of other sources $\mathbf{s}_j, j \neq i$. Consequently, $E\{\mathbf{s}_j g(\mathbf{d}_i \mathbf{s})\} = 0$ and $E\{\mathbf{s}_j q'(\mathbf{d}_i \mathbf{s})\} = 0$ for $i \neq j$. From Eq. B.3, we have

$$\bar{\mathbf{d}}_i = (0, \dots, E\{\mathbf{s}_i g(\pm s_i)\} \pm \mu^T E\{\mathbf{s}_i q'(\pm s_i)\}, \dots, 0)^T \quad (\text{B.6})$$

For $\mathbf{d}_i^T = (0, \dots, 1, \dots, 0)$, the above equation becomes

$$\bar{\mathbf{d}}_i = (0, \dots, E\{\mathbf{s}_i g(s_i)\} - \mu^T E\{\mathbf{s}_i q'(s_i)\}, \dots, 0)^T = c_i \mathbf{d}_i \quad (\text{B.7})$$

where $c_i = E\{\mathbf{s}_i g(s_i)\} - \mu^T E\{\mathbf{s}_i q'(s_i)\}$. For $\mathbf{d}_i^T = (0, \dots, -1, \dots, 0)$, we can have

$$\bar{\mathbf{d}}_i = (0, \dots, -(E\{\mathbf{s}_i g(s_i)\} - \mu^T E\{\mathbf{s}_i q'(s_i)\}), \dots, 0)^T = c_i \mathbf{d}_i \quad (\text{B.8})$$

as well because of the oddness of both $g(\cdot)$ and $q'(\cdot)$.

Therefore, we have $\bar{\mathbf{d}}_i = c_i \mathbf{d}_i$ for both positive and negative cases, which can be also written in matrix form

$$\bar{\mathbf{D}} = \mathbf{KD} \quad (\text{B.9})$$

where $\mathbf{K} = \text{diag}(c_1, \dots, c_m)$. According to Assumption 1, all c_i are nonzero. We can then get

$$\left(\bar{\mathbf{D}} \bar{\mathbf{D}}^T \right)^{-1/2} \bar{\mathbf{D}} = (\mathbf{KDD}^T \mathbf{K})^{-1/2} \mathbf{KD} = \mathbf{K}^{-1} \mathbf{KD} = \mathbf{D} \quad (\text{B.10})$$

which means that $\mathbf{D}$ satisfies Eq. B.5 and then completes the proof.

Lemma 2: With Assumption 1, the fixed point stated in Lemma 1 is asymptotically stable and the convergence rate of CICA2 is at least two.

Proof. Assume the fixed point $\mathbf{D}$ in Lemma 1 is perturbed by:

$$u_{ij} = d_{ij} + \epsilon_{ij} \quad (\text{B.11})$$

where d_{ij} are the elements of $\mathbf{D}$, and ϵ_{ij} are small numbers for all i, j , which are limited by $|\epsilon_{ij}| < \epsilon$ and ϵ is a small positive number. Denote the row vectors of the perturbation matrix (ϵ_{ij}) as $\mathbf{e}_i^T$, then $\mathbf{u}_i = \mathbf{d}_i + \mathbf{e}_i$. Inserting this disturbed vector into Eq. B.3, we can get

$$\begin{aligned} \bar{\mathbf{u}}_i &= E \{ \mathbf{sg}(\mathbf{d}_i \mathbf{s} + \mathbf{e}_i \mathbf{s}) \} - \mu^T E \{ \mathbf{sq}'(\mathbf{d}_i \mathbf{s} + \mathbf{e}_i \mathbf{s}) \} \\ &= E \{ \mathbf{sg}(\pm \mathbf{s}_i + \mathbf{e}_i \mathbf{s}) \} - \mu^T E \{ \mathbf{sq}'(\pm \mathbf{s}_i + \mathbf{e}_i \mathbf{s}) \} \end{aligned} \quad (\text{B.12})$$

According to Taylor series expansion theory, $g(\pm \mathbf{s}_i + \mathbf{e}_i \mathbf{s})$ and $q'(\pm \mathbf{s}_i + \mathbf{e}_i \mathbf{s})$ can be expanded at the points of $\pm \mathbf{s}_i$ through:

$$g(\pm \mathbf{s}_i + \mathbf{e}_i \mathbf{s}) = \pm g(\mathbf{s}_i) + g'(\mathbf{s}_i) \mathbf{e}_i^T \mathbf{s} + O(\epsilon^2) \quad (\text{B.13a})$$

$$q'(\pm \mathbf{s}_i + \mathbf{e}_i \mathbf{s}) = \pm q'(\mathbf{s}_i) + q''(\mathbf{s}_i) \mathbf{e}_i^T \mathbf{s} + O(\epsilon^2) \quad (\text{B.13b})$$

$O(\epsilon^2)$ collects all the remaining terms that are two and above order in the elements of $\mathbf{e}_i$. It is then bounded by ϵ^2 times a scalar constant. We then have

$$\bar{\mathbf{u}}_i \pm E \{ \mathbf{sg}(\mathbf{s}_i) \} \mp \mu^T E \{ \mathbf{sq}'(\mathbf{s}_i) \} + E \{ \mathbf{sg}'(\mathbf{s}_i) \mathbf{e}_i^T \mathbf{s} \} - \mu^T E \{ \mathbf{sq}''(\mathbf{s}_i) \mathbf{e}_i^T \mathbf{s} \} + O(\epsilon^2) \quad (\text{B.14})$$

Because

$$E \{ \mathbf{sg}'(\mathbf{s}_i) \mathbf{e}_i^T \mathbf{s} \} - \mu^T E \{ \mathbf{sq}''(\mathbf{s}_i) \mathbf{e}_i^T \mathbf{s} \} = \pm \left(E \{ g'(\mathbf{s}_i) \mathbf{s}_i^2 \} - \mu^T E \{ q''(\mathbf{s}_i) \mathbf{s}_i^2 \} \right) \epsilon_{ii} \mathbf{d}_i \quad (\text{B.15})$$

we have

$$\bar{\mathbf{u}}_i = \left(E \{ \mathbf{sg}(\mathbf{s}_i) \} - \mu^T E \{ \mathbf{sq}'(\mathbf{s}_i) \} \right) \mathbf{d}_i \pm \left(E \{ g'(\mathbf{s}_i) \mathbf{s}_i^2 \} - \mu^T E \{ q''(\mathbf{s}_i) \mathbf{s}_i^2 \} \right) \epsilon_{ii} \mathbf{d}_i + O(\epsilon^2) \quad (\text{B.16})$$

This can be written in matrix form

$$\bar{\mathbf{U}} = \mathbf{KD} + \widehat{\mathbf{K}}\mathbf{D} + O(\epsilon^2) \quad (\text{B.17})$$

where $\widehat{\mathbf{K}}$ is the diagonal matrix whose diagonal elements are

$\left(E \{ g'(\mathbf{s}_i) \mathbf{s}_i^2 \} - \mu^T E \{ q''(\mathbf{s}_i) \mathbf{s}_i^2 \} \right) \epsilon_{ii}$. With some trivial algebraic manipulations and based on $\mathbf{D}^2 = \mathbf{I}$, we can get [27]:

$$\bar{\mathbf{U}} \bar{\mathbf{U}}^T = \mathbf{K}^2 (\mathbf{I} - 2\widehat{\mathbf{K}}\mathbf{K}^{-1}) + O(\epsilon^2) \quad (\text{B.18})$$

Since $-2\widehat{\mathbf{K}}\mathbf{K}^{-1}$ is first order in ϵ , using the 1st order Taylor series expansion on $(\mathbf{I} - 2\widehat{\mathbf{K}}\mathbf{K}^{-1})^{-1/2}$, we can have

$$\left(\bar{\mathbf{U}} \bar{\mathbf{U}}^T\right)^{-1/2} = \mathbf{K}^{-1} (\mathbf{I} + \widehat{\mathbf{K}}\mathbf{K}^{-1}) + O(\epsilon^2) \quad (\text{B.19})$$

and

$$\begin{aligned} \left(\bar{\mathbf{U}} \bar{\mathbf{U}}^T\right)^{-1/2} \bar{\mathbf{U}} &= \mathbf{K}^{-1} (\mathbf{I} + \widehat{\mathbf{K}}\mathbf{K}^{-1}) (\mathbf{K} - \widehat{\mathbf{K}}) \mathbf{D} + O(\epsilon^2) \\ &= \mathbf{D} + O(\epsilon^2) \end{aligned} \quad (\text{B.20})$$

This equation proves the stability of the fixed point $\mathbf{D}$ of CICA2 with second-order convergence. The proof can also be adapted to prove the convergence of the fixed point solution of CICA1 with minor math manipulations as well as changes to Assumption 1.

References

1. Semi-blind ICA of fMRI: A method for utilizing hypothesis-derived time courses in a spatial ica analysis. *NeuroImage*. 2005; 25(2):527–538. [PubMed: 15784432]
2. Beckmann CF, DeLuca M, Devlin JT, Smith SM. Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2005; 360(1457):1001–1013.
3. Bertsekas, DP. *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific; Belmont, Mass: 1996.
4. Calhoun V, Adali T, McGinty V, Pekar J, Watson T, Pearlson G. fMRI activation in a visual-perception task: Network of areas detected using the general linear model and independent components analysis. *NeuroImage*. 2001; 14:1080–1088. [PubMed: 11697939]
5. Calhoun V, Adali T, Pearlson G, Pekar J. Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. *Hum. Brain Map*. 2001; 13:43–53.
6. Comon P. Independent component analysis-a new concept? *Signal Processing*. 1994; 36:287–314.
7. Daubechies I, Roussos E, Takerkart S, Benharrosh M, Golden C, D'Ardenne K, Richter W, Cohen JD, Haxby J. Independent component analysis for brain fMRI does not select for independence. *Proceedings of the National Academy of Sciences*. 2009; 106(26):10415–10422.
8. Friston K, Holmes A, Poline J, Grasby PJ, Williams SCR, Frackowiak R, Turner R. Analysis of fmri time series revisited. *NeuroImage*. 1995; 2:45–53. [PubMed: 9343589]
9. Friston K, Worsley K, Frackowiak R, Mazziotta J, Evans A. Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*. 1994; 1:214–220.
10. Genovese C, Lazar N, Nichols T. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*. 2002; 870–878. [PubMed: 11906227]
11. Greicius MD, Srivastava G, Reiss AL, Menon V. Default-mode network activity distinguishes Alzheimer's disease from healthy aging: Evidence from functional MRI. *Proceedings of the National Academy of Sciences of the United States of America*. 2004; 101(13):4637–4642. [PubMed: 15070770]

12. Haacke, EM.; Brown, RW.; Thompson, MR.; Venkatesan, R.; Haacke, EM.; Brown, RW.; Thompson, MR.; Venkatesan, R. Magnetic Resonance Imaging: Physical Principles and Sequence Design. 1 edition. Wiley-Liss; 1999.
13. Hesse C, James C. The fastica algorithm with spatial constraints. Signal Processing Letters, IEEE. Nov.; 2005 12(11):792–795.
14. Hestenes MR. Multiplier and gradient methods. J.O.T.A. 1969; 4:303–320.
15. Hyvärinen A. New approximation of differential entropy for independent component analysis and projection pursuit. Advances in Neural Information Processing Systems. 1998; 10:273–279.
16. Hyvärinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. on Neural Network. 1999; 10(3):626–634.
17. Hyvärinen, A.; Karhunen, J.; Oja, E. Independent Component Analysis. John Wiley & Sons; New York: 2001.
18. Lin Q-H, Liu J, Zheng Y-R, Liang H, Calhoun VD. Semiblind spatial ica of fmri using spatial constraints. Human Brain Mapping. 2010 page Articles online in advance of print.
19. Lin Q-H, Zheng Y-R, Yin F-L, Liang H, Calhoun VD. A fast algorithm for one-unit ICA-R. Information Sciences. 2007; 177:1265–1275.
20. Lu, W.; Rajapakse, JC. Constrained independent component analysis. In: Leen, TK.; Dietterich, TG.; Tresp, V., editors. Advances in Neural Information Processing Systems. Vol. volume 16. MIT Press; Cambridge, MA: 2000. p. 570-576.
21. Lu W, Rajapakse JC. Approach and applications of constrained ICA. IEEE Trans Neural Netw. 2005; 16(1):203–12. [PubMed: 15732400]
22. Luenberger, DG. Optimization by Vector Space Methods. John Wiley and Sons; 1969.
23. Luo J, Hu B, Ling X-T, Liu R-W. Principal independent component analysis. IEEE Transactions on Neural Networks. 1999; 10(4):912–917. [PubMed: 18252587]
24. McKeown MJ, Sejnowski TJ. Independent component analysis of fMRI data: Examining the assumptions. Human Brain Mapping. 1998; 6(5–6):368–372. [PubMed: 9788074]
25. Metz CE. Basic principle of ROC analysis. Semin Nucl Med. 1978; VIII(4):283–98. [PubMed: 112681]
26. Nocedal, J.; Wright, S. Numerical Optimization. 2nd edition. Springer; New York: 2006.
27. Oja E, Yuan Z. The FastICA Algorithm revisited: convergence analysis. IEEE Transactions on Neural Networks. 2006; 17(6)
28. Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B. Numerical Recipes in C++: the art of scientific computing. 2 edition. Cambridge University Press; 2002.
29. Rockafellar RT. Lagrange multipliers and optimality. SIAM Rev. 1993; 35:183–238.
30. Stone JV, Porrill J, Porter NR, Wilkinson ID. Spatiotemporal independent component analysis of event-related fMRI data using skewed probability density functions. NeuroImage. 2002; 15(2): 407–421. [PubMed: 11798275]
31. Wang, Z. Proc of ISMRM. Stockholm, Sweden: 2010. A fixed-point iteration based constrained independent component analysis and its application in fMRI; p. 3496ISMRM
32. Wang Z, Aguirre GK, Rao H, Wang J, Fernández-Seara MA, Childress AR, Detre JA. Empirical optimization of ASL data analysis using an ASL data processing toolbox: ASLtbx. Magnetic Resonance Imaging. 2008; 26(2):261–269. [PubMed: 17826940]
33. Wang Z, Childress AR, Wang J, Detre JA. Support vector machine learning-based fMRI data group analysis. NeuroImage. 2007; 36(4):1139–1151. [PubMed: 17524674]
34. Wang Z, Lee Y, Fiori S, Leung C, Zhu Y. An improved sequential method for principal component analysis. Pattern Recognition Letters. 2003; 24(9–10):1409–1415.
35. Wang Z, Mechanic-Hamilton D, Pluta J, Glynn S, Detre JA. Function lateralization via measuring coherence laterality. NeuroImage. 2009; 47(1):281–288. [PubMed: 19345736]

Listing 1: Implementation of OCICA in a sequential mode

1. Center and prewhiten the observed signal $\mathbf{x}$;
2. Set ε to a small value to control the learning precision ;
3. Initialize η ;
4. Generate a zero-mean unit-norm Gaussian noise vector $\mathbf{v}$;
5. For the i -th constrained IC:
6. Initialize μ ;
7. Initialize $\mathbf{w}_i$ with a random vector, set its norm to be 1;
8. Repeat the following steps until the norm of $\Delta\mathbf{w}_i$ is less than ε or
 $|1 - \|\mathbf{w}_{i,k}^T \mathbf{w}_{i,k+1}\|| \leq \varepsilon$ (k is the iteration index);
9. Calculate $\hat{\mathbf{c}}$ with Eq. 8;
10. Update γ ;
11. Update $\mathbf{w}_i$ using Eq. 9;
12. Normalize $\mathbf{w}_i$ using Eq. 10 and Eq. 11;
13. Update μ using Eq. 12a;
14. Adjust ε to gradually release the constraints.

Listing 2: Implementation of CICA1 in a sequential mode

1. Center and prewhiten the observed data $\mathbf{x}$;
2. Set ε to a small value to control the learning precision ;
3. For the i -th constrained IC:
4. Initialize μ ;
5. Initialize $\mathbf{w}_i$ with a random vector, set the norm to 1;
6. Repeat the following steps until the norm of $\Delta\mathbf{w}_i$ is less than ε or
 $|1 - \|\mathbf{w}_{i,k}^T \mathbf{w}_{i,k+1}\|| \leq \varepsilon$ (k is the iteration index);
7. Update γ ;
8. Update $\mathbf{w}_i$ using Eq. 15;
9. Normalize $\mathbf{w}_i$ using Eq. 10 and Eq. 11;
10. Update μ using Eq. 12a;
11. Adjust ε to gradually release the constraints.

Figure 1. 1D sources used for algorithm evaluations. A) The 5 source signals to be separated, B) the 4 references used to extract the 4 sources of interest using CICA.

Figure 2.
2D source components, the associated time courses, and the reference maps used in CICA.

Figure 3. Results of 1D source separation using CICAs and fICA. Error bars indicate the standard error.

Figure 4.
The first 400 time points of the 1D source separation results. A), B), C), and D) are the results of CICA1, CICA2, OCICA, and fICA, respectively.

Figure 5. 2D source separation results using CICAs and fICA. The same spatial constraints as shown in Fig. 2C were used in all CICAs. Error bars indicate the standard error. The invisible error bars means the standard error is 0.

Figure 6.

The first source component and the associated time course identified by various ICA algorithms. AS, BS, CS and the associated time course ASTC, BSTC, CSTC were extracted by CICA1, CICA2, and OCICA, respectively using the spatial constraint *sref1* as shown in Fig. 2C. AT, BT, CT and the associated time course ATTTC, BTTC, CTTC were extracted by CICA1, CICA2, and OCICA, respectively using the temporal constraint TREF. D/DTC were extracted by FastICA. E was the correlation map generated by GLM using a HRF convolved box-car function as the regressor. The color bar in the top row is for all spatial ICA maps shown in the figure; the color bar in the bottom row is for the GLM correlation map E.

Figure 7. Group level statistical analysis results of the sensorimotor fMRI data. The method for getting individual level brain activation maps is labeled on the left side of each row of pictures. The threshold for getting the suprathreshold voxels is $t=4.67$ ($P < 0.05$ FDR corrected). The color window is 4.67 – 11.