Approach and Applications of Constrained ICA

Wei Lu, *Member, IEEE,* and Jagath C. Rajapakse, *Senior Member, IEEE*

***Abstract*—This paper presents the technique of constrained independent component analysis (cICA) and demonstrates two applications, less-complete ICA, and ICA with reference (ICA-R). The cICA is proposed as a general framework to incorporate additional requirements and prior information in the form of constraints into the ICA contrast function. The adaptive solutions using the Newton-like learning are proposed to solve the constrained optimization problem. The applications illustrate the versatility of the cICA by separating subspaces of independent components according to density types and extracting a set of desired sources when rough templates are available. The experiments using face images and functional MR images demonstrate the usage and efficacy of the cICA.**

***Index Terms*—Constrained optimization, ICA with reference (ICA-R), independent component analysis (ICA), less-complete ICA.**

I. Introduction

INDEPENDENT COMPONENT ANALYSIS (ICA) is a technique that transforms a multivariate random signal into a signal having components that are mutually independent in the complete statistical sense [1], [2]. Let the time varying observed signal be $\mathbf{x} = (x_1, x_2, \ldots x_n)^{\mathrm{T}}$ and the desired signal consisting of independent components (ICs) be $\mathbf{c} = (c_1, c_2, \ldots c_m)^{\mathrm{T}}$. The classical ICA assumes that the signal $\mathbf{x}$ is an instantaneous linear mixture of ICs or independent sources $c_i, i = 1, 2, \ldots m$ and, therefore, $\mathbf{x} = \mathbf{A}\mathbf{c}$ where the matrix $\mathbf{A}$ of size $n \times m$ represents the linear memoryless mixing channels. The goal of the ICA is to obtain an $m \times n$ demixing matrix $\mathbf{W}$ to recover all the ICs of the observed signal, with minimal knowledge of $\mathbf{A}$ and $\mathbf{c}$. The recovered signal after the ICA, say $\mathbf{y} = (y_1, y_2, \ldots y_m)^{\mathrm{T}}$, is given by $\mathbf{y} = \mathbf{W}\mathbf{x}$. For simplicity, we address the case of *complete* ICA, in which $m = n$, in this manuscript.

The wide applicability of the ICA, including blind source separation [3]–[5], feature extraction [6], [7], and signal detection [8], etc., has resulted in many batch and adaptive algorithms [1], [9]–[13] for the efficient realization of the ICA. Bell and Sejnowski [10], and Lee *et al.* [12] made use of nonlinearities in the neuronal activation functions to maximize the output entropy or the input likelihood, and Amari *et al.* used a cumulant-based approximation to minimize the Kullback–Leibler (KL) divergence of the output [13]. Oja *et al.* [14] explored the nonlinear principal component analysis (PCA) [15] to perform blind source separation. The fastICA algorithm proposed by Hyvärinen [16] provides an efficient method to perform ICA.

Most existing ICA algorithms estimate same number of ICs as the observed mixtures [1], [10], [12], [13], [17] though, often in real applications, the number of components needed to be recovered is less. For example, although the number of components of interest in many biomedical signals are a few, the observed data consist of a large number of components related to noise and artifacts. The classical ICA is only capable of separating the full space of ICs because the inverse matrix $\mathbf{W}^{-1}$, being present in the learning rule, is valid only for a square matrix $\mathbf{W}$. Therefore, an additional process for dimension reduction is needed either at the input or at the output of the classical ICA for subspace extraction.

Some one-unit ICA algorithms were proposed to extract all ICs one by one with a deflation process [18], [19], but the inefficiency of such techniques and the arbitrary order of extraction remain as major drawbacks. The nonlinear PCA algorithm [20] and fastICA algorithm [16] provided possible ways to extract the subspace in parallel manner. However, the extraction of signals at the global optimum by using these algorithms is always determined by the contrast function adopted. Furthermore, the additional process of prewhitening or compulsive decorrelation may distort the original data.

The motivation of the constrained independent component analysis (cICA) is to provide a systematic and flexible method to incorporate more assumptions and prior information, if available, into the contrast function so the ill-posed ICA is converted to a better-posed problem, facilitating more applications. As seen later, constraints may be adopted to reduce the dimensionality of the output of the ICA. Incorporation of prior information, such as statistical properties or rough templates of the sources, avoids local minima and increases the quality of the separation. The authors first introduced the cICA in a brief paper [21]. How indeterminacy in ICA can be resolved was presented earlier in [22]. Here, the cICA approach is presented in detail as a general framework to introduce additional constraints to the ICA, and two other applications are illustrated.

The manuscript is organized as follows. Section II describes the cICA framework and general Newton learning rules. How the cICA technique is used for the extraction of a subspace of sources and for the extraction of a set of desired ICs when reference signals are available are described in Sections III and IV, respectively. Section V demonstrates two cICA applications with real images and fMRI data. Section VI discusses the efficacy of the cICA technique and provides a conclusion.

Manuscript received May 15, 2002; revised August 25, 2003. This work was supported by the Ministry of Education and Agency of Science, Technology, and Research (A*Star), Singapore, under A*Star Grant 022 101 0017.

W. Lu was with the School of Computer Engineering, Nanyang Technological University, Singapore 639798. He is now with the Sony Singapore Research Laboratory, Singapore 117684 (e-mail: wei.lu@ap.sony.com).

J. C. Rajapakse is with the School of Computer Engineering, Nanyang Technological University, Singapore 639798 (e-mail: asjagath@ntu.edu.sg).

Digital Object Identifier 10.1109/TNN.2004.836795

II. cICA

The cICA deals with the following constrained minimization problem:

$$\begin{aligned}&\text{minimize } \mathcal{C}(\mathbf{y})\\ &\text{subject to } \mathbf{g}(\mathbf{y}:\mathbf{W})\le\mathbf{0} \quad \text{and/or} \quad \mathbf{h}(\mathbf{y}:\mathbf{W})=\mathbf{0}\end{aligned} \tag{1}$$

where $\mathcal{C}(\mathbf{y})$ represents an ICA contrast function, and $\mathbf{g}(\mathbf{y}:\mathbf{W})=(g_1(\mathbf{y}:\mathbf{W}),g_2(\mathbf{y}:\mathbf{W}),\dots g_u(\mathbf{y}:\mathbf{W}))^{\mathrm{T}}$ and $\mathbf{h}(\mathbf{y}:\mathbf{W})=(h_1(\mathbf{y}:\mathbf{W}),h_2(\mathbf{y}:\mathbf{W}),\dots h_v(\mathbf{y}:\mathbf{W}))^{\mathrm{T}}$ define the vectors of u inequality and v equality constraints, respectively.

The method of Lagrange multipliers [23] is adopted to search for the optimal solution. The corresponding augmented Lagrangian function $\mathcal{L}$ is given by

$$\begin{aligned}\mathcal{L}(\mathbf{W},\boldsymbol{\mu},\boldsymbol{\lambda},\mathbf{z})=\mathcal{C}(\mathbf{y})+\boldsymbol{\mu}^{\mathrm{T}}\hat{\mathbf{g}}(\mathbf{y}:\mathbf{W})+\frac{1}{2}\gamma\|\hat{\mathbf{g}}(\mathbf{y}:\mathbf{W})\|^2\\ +\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{h}(\mathbf{y}:\mathbf{W})+\frac{1}{2}\gamma\|\mathbf{h}(\mathbf{y}:\mathbf{W})\|^2\end{aligned} \tag{2}$$

where $\hat{\mathbf{g}}(\mathbf{y}:\mathbf{W})=(\hat{g}_1(\mathbf{y}:\mathbf{W}),\hat{g}_2(\mathbf{y}:\mathbf{W}),\dots\hat{g}_u(\mathbf{y}:\mathbf{W}))^{\mathrm{T}}$ in which $\hat{g}_p(\mathbf{y}:\mathbf{W})=g_p(\mathbf{y}:\mathbf{W})+z_p^2=0, p=1,2,\dots u$, transform the original inequality constraints into equality constraints with a vector of slack variables $\mathbf{z}=(z_1,z_2,\dots z_u)^{\mathrm{T}}$; $\boldsymbol{\mu}=(\mu_1,\mu_2,\dots\mu_u)^{\mathrm{T}}$ and $\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\dots\lambda_v)^{\mathrm{T}}$ are the vectors of positive Lagrange multipliers corresponding to inequality and equality constraints, respectively. $\gamma(>0)$ is the penalty parameter and $\|\cdot\|$ denotes the Euclidean norm. The quadratic penalty term $(1/2)\gamma\|\cdot\|^2$ ensures that the minimization problem holds the condition of local convexity: the Hessian matrix of $\mathcal{L}$ is positive–definite. The inequality constraints are further translated to eliminate slack variables $\mathbf{z}$ by using the procedure explained in [23]. So, the cICA objective function is given by

$$\mathcal{L}(\mathbf{W},\boldsymbol{\mu},\boldsymbol{\lambda})=\mathcal{C}(\mathbf{y})+\mathcal{G}(\mathbf{y}:\mathbf{W},\boldsymbol{\mu})+\mathcal{H}(\mathbf{y}:\mathbf{W},\boldsymbol{\lambda}) \tag{3}$$

where $\mathcal{G}(\mathbf{y}:\mathbf{W},\boldsymbol{\mu})=(1/2\gamma)\sum_{p=1}^{u}\{(\max\{0,\mu_p+\gamma g_p(\mathbf{y}:\mathbf{W})\})^2-\mu_p^2\}$ denotes the term corresponding to inequality constraints; $\mathcal{H}(\mathbf{y}:\mathbf{W},\boldsymbol{\lambda})=\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{h}(\mathbf{y}:\mathbf{W})+(1/2)\gamma\|\mathbf{h}(\mathbf{y}:\mathbf{W})\|^2$ is the term corresponding to the equality constraints. By partially differentiating the objective function in (3), the gradient of $\mathcal{L}(\mathbf{W},\boldsymbol{\mu},\boldsymbol{\lambda})$ is given by

$$\begin{aligned}\nabla_{\mathbf{W}}\mathcal{L}(\mathbf{W},\boldsymbol{\mu},\boldsymbol{\lambda})=\nabla_{\mathbf{W}}\mathcal{C}(\mathbf{y})+\nabla_{\mathbf{W}}\mathcal{G}(\mathbf{y}:\mathbf{W},\boldsymbol{\mu})\\ +\nabla_{\mathbf{W}}\mathcal{H}(\mathbf{y}:\mathbf{W},\boldsymbol{\lambda})\end{aligned} \tag{4}$$

where the matrix $\nabla_{\mathbf{W}}\mathcal{C}(\mathbf{y})=[\nabla_{\mathbf{w}_1}\mathcal{C}(\mathbf{y})\ \nabla_{\mathbf{w}_2}\mathcal{C}(\mathbf{y})\cdots\nabla_{\mathbf{w}_k}\mathcal{C}(\mathbf{y})]^{\mathrm{T}}$ denotes the gradient of $\mathcal{C}(\mathbf{y})$ with respect to $\mathbf{W}$, and the two terms $\nabla_{\mathbf{W}}\mathcal{G}(\mathbf{y}:\mathbf{W},\boldsymbol{\mu})=[\nabla_{\mathbf{w}_1}\mathcal{G}(\mathbf{y}:\mathbf{W},\boldsymbol{\mu})\ \nabla_{\mathbf{w}_2}\mathcal{G}(\mathbf{y}:\mathbf{W},\boldsymbol{\mu})\cdots\nabla_{\mathbf{w}_k}\mathcal{G}(\mathbf{y}:\mathbf{W},\boldsymbol{\mu})]^{\mathrm{T}}$ with elements $\nabla_{\mathbf{w}_i}\mathcal{G}(\mathbf{y}:\mathbf{W},\boldsymbol{\mu})=\sum_{p=1}^{u}\mu_p\nabla_{\mathbf{w}_i}g_p(\mathbf{y}:\mathbf{W})$, and $\nabla_{\mathbf{W}}\mathcal{H}(\mathbf{y}:\mathbf{W},\boldsymbol{\lambda})=[\nabla_{\mathbf{w}_1}\mathcal{H}(\mathbf{y}:\mathbf{W},\boldsymbol{\lambda})\ \nabla_{\mathbf{w}_2}\mathcal{H}(\mathbf{y}:\mathbf{W},\boldsymbol{\lambda})\cdots\nabla_{\mathbf{w}_k}\mathcal{H}(\mathbf{y}:\mathbf{W},\boldsymbol{\lambda})]^{\mathrm{T}}$ with elements $\nabla_{\mathbf{w}_i}\mathcal{H}(\mathbf{y}:\mathbf{W},\boldsymbol{\lambda})=\sum_{q=1}^{v}\lambda_q\nabla_{\mathbf{w}_i}h_q(\mathbf{y}:\mathbf{W})$. Furthermore, the Hessian matrix is obtained by partially differentiating the gradient (4) with respect to the demixing matrix $\mathbf{W}$ in their vector forms

$$\nabla^2_{\mathrm{Vec}(\mathbf{W}^{\mathrm{T}})}\mathcal{L}(\mathbf{W},\boldsymbol{\mu},\boldsymbol{\lambda})=\frac{\partial\left(\nabla_{\mathrm{Vec}(\mathbf{W}^{\mathrm{T}})}\mathcal{L}\right)^{\mathrm{T}}}{\partial(\mathrm{Vec}(\mathbf{W}^{\mathrm{T}}))}$$

where $\mathrm{Vec}(\cdot)$ is an operator which cascades the columns of the matrix from left to right to form a column vector. Thus, the Newton-like algorithm for $\mathbf{W}$ in vector form is obtained

$$\Delta\mathrm{Vec}(\mathbf{W}^{\mathrm{T}})=-\eta\left(\nabla^2_{\mathrm{Vec}(\mathbf{W}^{\mathrm{T}})}\mathcal{L}\right)^{-1}\nabla_{\mathrm{Vec}(\mathbf{W}^{\mathrm{T}})}\mathcal{L} \tag{5}$$

where η equals to one or may be decreased gradually to ensure stable convergence, and the matrix form of $\mathbf{W}$ is obtained by applying the inversed $\mathrm{Vec}(\cdot)$ operation to (5).

There are several methods for learning Lagrange multipliers. Here, for simplicity and keeping their positivity, we use the gradient ascent method as given here

$$\Delta\boldsymbol{\mu}=\max\{-\boldsymbol{\mu},\gamma\mathbf{g}(\mathbf{y}:\mathbf{W})\} \tag{6}$$

$$\Delta\boldsymbol{\lambda}=\gamma\mathbf{h}(\mathbf{y}:\mathbf{W}) \tag{7}$$

where γ is the learning rate. In the learning algorithm, the $\mathbf{W}$ is initialized to a uniform random matrix, the $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ to zeros. Then, it is updated according to (5)–(7) until convergence.

The statistical properties (e.g., consistency, asymptotic variance, robustness) of the cICA algorithm depend on the choice of the contrast function and the constraints involved in the objective function [24]. Any function whose optimization enables the estimation of the ICs can be treated as the ICA contrast function. However, the constraints that are used to define or restrict the properties of the ICs should not infringe the independence criteria. This can be confirmed by verifying the formulated cICA objective functions with the ICA equivariant properties defined by Cardoso and Laheld [25]. The nonholonomic orthogonal constraint introduced by Amari *et al.* [26] in ICA is a good example that preserves the variance of the recovered signal and also hold the ICA equivariant properties. However, the constraints do not always satisfy this requirement, then we need to choose suitable parameters and adjust them properly to keep both independence criteria and constraint properties valid all the time. We will show how these can be achieved in the following two cICA applications.

The first application chooses the uncorrelatedness as the constraints of marginal negentropies to solve less-complete ICA problem, which are well-matched with the statistical independence criteria. The ICA with reference (ICA-R) will show the importance of proper adjustment of parameters in a parametric constraint to ensure the quantity of the restriction within a suitable range for desired solutions.

III. Less-Complete ICA

The less-complete ICA is the problem that transforms an observed signal into a signal with less number of components, which is a subspace of the original ICs mixed in the observed signals. In this section, we presents an algorithm to perform

less-complete ICA within the framework of cICA. Let the components of the output $\mathbf{y} = (y_1, y_2, \ldots y_l)^{\mathrm{T}}$ be mutually independent and correspond to $l(<m)$ original sources mixed in the observations, in which the demixing matrix $\mathbf{W}$ is $l \times m$ while the mixing $\mathbf{A}$ is an $m \times m$ square matrix. Cao and Liu proved [27] that a classical ICA algorithm solving the less-complete ICA problem is l-row decomposable (independent) but the outputs may not consist of the original ICs for any $l < m$ when the mixing matrix is of full column rank. The criterion of statistical independence is insufficient for extracting a subset of original sources.

Recently, some one-unit ICA algorithms were augmented to extract all ICs mixed in the input signals, one by one, with a deflation process [18], [19], [28], which could possibly be treated as a sequential solution to extract a subspace of ICs. Cichocki *et al.* claimed that the enhanced nonlinear PCA with a whitening process was able to extract less number of ICs than the sources [20]. However, the necessary preprocessing (whitening) stage results in failure of separation due to data distortion when ill-conditioned mixing matrices or weak sources are involved. The fastICA algorithm proposed by Hyvärinen [16] can separate a subset of ICs in parallel manner, but the interference caused by an explicit decorrelation process after each iteration is too rigid to orient the learning process toward the correct convergence trend.

Negentropy has been used to separate ICs from their mixtures [16], [29], [30] because the sources considered in the ICA usually have non-Gaussian distributions. The negentropy of a signal $y, J(y)$, is given by

$$J(y) = H(y_{\mathrm{G}}) - H(y) \tag{8}$$

where y_{G} is a Gaussian random variable having the same variance as the signal y. The negentropy is always nonnegative since Gaussian signals have the maximum entropy of zero [31]. Extended to multiple neurons, maximizing the marginal negentropies projects the input data onto a low-dimensional subspace and finds for the structure of non-Gaussianity in the projection [29]. Hence, the contrast function for the less-complete ICA is defined as

$$\mathcal{J}(\mathbf{y}) = -\sum_{i=1}^{l} J(y_i) \tag{9}$$

where $J(y_i)$ is estimated by a flexible and reliable approximation given in [32]

$$J(y_i) \approx \rho(E\{f_i(y_i)\} - E\{f_i(\nu)\})^2 \tag{10}$$

where ρ is a positive constant, $f_i(\cdot)$ is a nonquadratic function, and ν is a Gaussian variable having zero mean and unit variance. Optimization of (9), individually, yielding a non-Gaussian signal for each output could cause different outputs estimate the same independent source. Therefore, the uncorrelation among estimated ICs is introduced as constraints to prevent such a scenario [16] such that $h_{ij}(y_i, y_j) = (E\{y_i y_j\})^2 = 0, \forall i, j = 1, 2, \ldots l; i \neq j$. The problem of less-complete ICA can then be modeled as minimizing the contrast function in (9) with the equality constraint term given by $\mathbf{h}(\mathbf{y}) = (h_{11}(y_1), h_{12}(y_1, y_2), \ldots h_{1l}(y_1, y_l), h_{21}(y_2, y_1) \ldots h_{ll}(y_l))^{\mathrm{T}}$ where $h_{ii}(y_i) = (E\{y_i^2\} - 1)^2, \forall i = 1, 2, \ldots l$, added to restrict each output have unit variance.

Following (3), the less-complete ICA objective function $\mathcal{L}_1(\mathbf{W}, \boldsymbol{\lambda})$ is given by

$$\mathcal{L}_1(\mathbf{W}, \boldsymbol{\lambda}) = \mathcal{J}(\mathbf{y}) + \mathcal{H}(\mathbf{y}, \boldsymbol{\lambda}) \tag{11}$$

where $\mathcal{H}(\mathbf{y}, \boldsymbol{\lambda}) = \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{h}(\mathbf{y}) + (1/2)\gamma\|\mathbf{h}(\mathbf{y})\|^2$ and the concatenated vector, $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^{\mathrm{T}}, \boldsymbol{\lambda}_2^{\mathrm{T}}, \ldots \boldsymbol{\lambda}_l^{\mathrm{T}})^{\mathrm{T}}$ in which $\boldsymbol{\lambda}_i = (\lambda_{i1}, \lambda_{i2}, \ldots \lambda_{il})^{\mathrm{T}}$. The gradient of $\mathcal{L}_1$ is given by

$$\nabla_{\mathbf{W}}\mathcal{L}_1(\mathbf{W}, \boldsymbol{\lambda}) = E\{\nabla_{\mathbf{y}}\mathcal{J}(\mathbf{y})\mathbf{x}^{\mathrm{T}}\} + 2\nabla_{\sum_{\mathbf{yy}}}\mathcal{H}(\mathbf{y}, \boldsymbol{\lambda})E\{\mathbf{y}\mathbf{x}^{\mathrm{T}}\} \tag{12}$$

where the vector $\nabla_{\mathbf{y}}\mathcal{J}(\mathbf{y}) = (\mathcal{J}'_{y_1}(\mathbf{y}), \mathcal{J}'_{y_2}(\mathbf{y}), \ldots \mathcal{J}'_{y_l}(\mathbf{y}))^{\mathrm{T}}$ with $\mathcal{J}'_{y_i}(\mathbf{y}) = -\hat{\rho}_i f'_{y_i}(y_i)$ in which $\hat{\rho}_i = 2\rho(E\{f_i(y_i)\} - E\{f_i(\nu)\})$, and the matrix $\nabla_{\sum_{\mathbf{yy}}}\mathcal{H} = [\nabla_{\sum_{y_1\mathbf{y}}}\mathcal{H}\ \nabla_{\sum_{y_2\mathbf{y}}}\mathcal{H}\ \cdots\ \nabla_{\sum_{y_l\mathbf{y}}}\mathcal{H}]^{\mathrm{T}}$ with the ith row, $\nabla_{\sum_{y_i\mathbf{y}}}\mathcal{H} = (\mathcal{H}'_{\sum_{y_iy_1}}, \mathcal{H}'_{\sum_{y_iy_2}}, \ldots \mathcal{H}'_{\sum_{y_iy_l}})^{\mathrm{T}}$ where $\mathcal{H}'_{\sum_{y_iy_1}} = 2\lambda(E\{y_iy_j\} - \delta_{ij})$, and δ_{ij} is zero when $i \neq j$, and equals to one, otherwise. To simplify the matrix inversion and keep the stability of the learning, let us approximate the Hessian matrix as

$$\nabla^2_{\mathrm{Vec}(\mathbf{W}^{\mathrm{T}})}\mathcal{L}_1(\mathbf{W}, \boldsymbol{\lambda}) \approx \mathbf{D}_1 \otimes \sum_{\mathbf{xx}} \tag{13}$$

where $\mathbf{D}_1$ is a diagonal matrix in which the diagonal is a vector $\mathbf{d}(\mathbf{y}, \boldsymbol{\lambda}) = (d_1(y_1, \lambda_{11}), d_2(y_2, \lambda_{22}), \ldots d_l(y_l, \lambda_{ll}))^{\mathrm{T}}$ with elements $d_i(y_i, \lambda_{ii}) = 4\lambda_{ii}(3E\{y_i^2\} - 1) - E\{\hat{\rho}_i f''_{y_i}(y_i)\}$ that can be further simplified to $d_i(y_i, \lambda_{ii}) = 8\lambda_{ii} - E\{\hat{\rho}_i f''_{y_i}(y_i)\}$; $\sum_{\mathbf{xx}}$ is the covariance matrix of input $\mathbf{x}$ and $\otimes$ denotes the Kronecker product of two matrices. Applying the relational properties: $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$ and $\mathrm{Vec}(\mathbf{ABC}) = (\mathbf{C}^{\mathrm{T}} \otimes \mathbf{A})\mathrm{Vec}(\mathbf{B})$, the learning in (5) becomes

$$\Delta\mathrm{Vec}(\mathbf{W}^{\mathrm{T}}) = -\eta\mathrm{Vec}\left(\sum_{\mathbf{xx}}^{-1}\nabla_{\mathbf{W}^{\mathrm{T}}}\mathcal{L}_1(\mathbf{W}, \boldsymbol{\lambda})\mathbf{D}_1^{-1}\right). \tag{14}$$

By transforming the vector form back to the matrix

$$\Delta\mathbf{W} = -\eta(\boldsymbol{\Phi}(\mathbf{y}, \boldsymbol{\lambda}) + \boldsymbol{\Omega}(\mathbf{y}, \boldsymbol{\lambda}))\sum_{\mathbf{xx}}^{-1} \tag{15}$$

where the matrices $\boldsymbol{\Phi}(\mathbf{y}, \boldsymbol{\lambda}) = [\boldsymbol{\phi}_1(y_1, \lambda_{11})\ \boldsymbol{\phi}_2(y_2, \lambda_{22}) \cdots \boldsymbol{\phi}_l(y_l, \lambda_{ll})]^{\mathrm{T}}$ with elements, $\boldsymbol{\phi}_i(y_i, \lambda_{ii}) = E\{\mathcal{J}'_{y_i}(\mathbf{y})\mathbf{x}\}/d_i(y_i, \lambda_{ii})$ and $\boldsymbol{\Omega}(\mathbf{y}, \boldsymbol{\lambda}) = [\boldsymbol{\omega}_1(\mathbf{y}, \boldsymbol{\lambda}_1)\ \boldsymbol{\omega}_2(\mathbf{y}, \boldsymbol{\lambda}_2) \cdots \boldsymbol{\omega}_l(\mathbf{y}, \boldsymbol{\lambda}_l)]^{\mathrm{T}}$ with elements, $\boldsymbol{\omega}_i(\mathbf{y}, \boldsymbol{\lambda}_i) = 2E\{\mathbf{x}\mathbf{y}^{\mathrm{T}}\}\nabla_{\sum_{y_i\mathbf{y}}}\mathcal{H}(\mathbf{y}, \boldsymbol{\lambda})/d_i(y_i, \lambda_{ii})$. The Lagrange multipliers $\boldsymbol{\lambda}$ are learned using the updating rule in (7) with $\mathbf{h}$ defined as above.

A. Stability Analysis

With the Newton-like learning, theoretically, the algorithm is able to produce the optimum output, say $\mathbf{y}^*$, to extract the desired ICs, defined by Kuhn–Tucker (KT) optimum double $(\mathbf{W}^*, \boldsymbol{\lambda}^*)$ that satisfies the first-order conditions:

$\nabla_{\mathbf{W}}\mathcal{L}(\mathbf{W}^*,\boldsymbol{\lambda}^*) = \mathbf{0}; \mathbf{h}(\mathbf{y}^*) = \mathbf{0}$ and $\boldsymbol{\lambda}^* > \mathbf{0}, *$ indicates the optimum value. Let us examine the stability at the global optima by testing that the Hessian matrix $\nabla^2_{\mathrm{Vec}(\mathbf{W}^{*\mathrm{T}})}\mathcal{L}_1$ is positive–definite. For simplicity, consider the approximated Hessian matrix in (13), although the examination is valid for the original Hessian matrix as well. It is always positive–definite when the input covariance matrix $\sum_{\mathbf{xx}}$ is nonsingular and the nonzero elements of $\mathbf{D}_1, d(y_i^*,\lambda_{ii}^*), \forall i = 1,2,\ldots l$ are positive. The former condition is true in most cases, especially when a large number of samples of the signal are available. Let us consider two terms $d^*_{\mathcal{H}}(y_i^*,\lambda_{ii}^*) = 4\lambda_{ii}^*(3E\{y_i^{*2}\}-1)$ and $d^*_J(y_i^*) = -E\{\hat{\rho}_i^* f''_{y_i}(y_i^*)\}$ in the latter elements $d(y_i^*,\lambda_{ii}^*)$ as follows.

Because the variance of y_i^* approaches to one, $d^*_{\mathcal{H}}(y_i^*,\lambda_{ii}^*)$ is equal to $8\lambda_{ii}^*$, always positive as the Lagrange multipliers λ_{ii}s hold the positivity property by following Lagrange multiplier learning methods given in (7) and [23]. In order to keep the learning algorithm stable, the terms $d_i(y_i,\lambda_{ii})$, corresponding to the uncorrelation constraints in the learning rule, are approximated by $8\lambda_{ii}$.

The positivity of $d^*_J(y_i^*)$ is subject to Gaussianity of the signals. We propose the following nonlinear function $f(y)$ in (10) to approximate the negentropies of the super-Gaussian and sub-Gaussian signals, respectively, [33]

$$f_{\mathrm{sup}}(y) = \frac{1}{a}\log\cosh(ay) - \frac{a}{2}y^2 \tag{16}$$

$$f_{\mathrm{sub}}(y) = \frac{b}{4}y^4 \tag{17}$$

where $a, b \in \mathbf{R}^+$. It can be easily shown that $-f''_{\mathrm{sup}}(y) \geq 0$ and $-f''_{\mathrm{sub}}(y) \leq 0$, and the value of $\hat{\rho}_i^* = 2\rho(E\{f_i(y_i^*)\} - E\{f_i(\nu)\})$ is always positive for the optimal solution with super-Gaussian distribution and negative with sub-Gaussian distribution. Therefore, their product $-\hat{\rho}_i^* f''_{y_i}(y_i^*) \geq 0$ for all signals. Thus, the convergence of the learning in (15) is always stable because when one uses the above nonlinear functions, appropriately.

IV. ICA-R

The subspace of the ICs, extracted in the less-complete ICA is always determined by the contrast function and the nonlinear functions adopted. Additional conditions have previously been incorporated by using sparse decomposition of signals [34] or fourth-order cumulants [35] into the contrast function to find the global optimum separating the desired components. However, if the desired number of sources is unknown or the desired sources cannot be categorized according to their density types, the subset recovered will not be useful. On the other hand, additional knowledge available on the sources or mixing channels in some applications can be treated as *a priori* constraints in the cICA to guide the separation of desired components. A particular instance is treated in this section, where the traces of the interesting sources, which are referred to as the *reference* signals, are available. Reference signals carry some information to distinguish the desired components but are not identical to the corresponding sources. A new cICA framework, the ICA-R, is formed by minimizing the less-complete ICA objective function provided that the extracted ICs are the closest to the corresponding reference signals.

Let the reference signal be $\mathbf{r} = (r_1, r_2, \ldots r_l)^{\mathrm{T}}$ and the closeness between the estimated output y_i and the corresponding reference r_i be measured by some norm $\varepsilon(y_i, r_i)$. The minimum value of $\varepsilon(y_i, r_i)$, with all outputs, indicates that the estimated output y_i is the desired IC closest to the reference signal r_i. If the desired IC is the one and only one closest to the reference $r_i, \varepsilon(y_i^*, r_i) < \varepsilon(y_i^\circ, r_i)$ where y_i^* denotes the output producing the desired IC which is closest to r_i and y_i° denotes the output having the next closest value of the norm. Thus, a constraint can be defined for the desired output component to have the closeness measure with r_i, that is less than or equal to a threshold parameter $\xi_i : \varepsilon(y_i, r_i) - \xi_i \leq 0$ only when $y_i = y_i^*$, and none of the other $m-1$ sources corresponds to the reference r_i if the threshold ξ_i is chosen in the scalar range $\Upsilon_i = [\varepsilon(y_i^*, r_i), \varepsilon(y_i^\circ, r_i))$.

With the additional constraint of restricting the closeness measure, the problem of the ICA-R can be modeled as minimizing $\mathcal{L}_1$ subjected to constraints $\mathbf{g}(\mathbf{y}) \leq \mathbf{0}$ where the inequality constraint term $\mathbf{g}(\mathbf{y}) = (g_1(y_1), g_2(y_2), \ldots g_l(y_l))^{\mathrm{T}}$ with $g_i(y_i) = \varepsilon(y_i, r_i) - \xi_i, \forall i = 1, \ldots l$. The ICA-R objective function is given by

$$\mathcal{L}_2(\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \mathcal{L}_1(\mathbf{W}, \boldsymbol{\lambda}) + \mathcal{G}(\mathbf{y}, \boldsymbol{\mu}) \tag{18}$$

where the inequality constraints term $\mathcal{G}(\mathbf{y}, \boldsymbol{\mu}) = (1/2\gamma)\sum_{i=1}^{l}\{(\max\{0, \mu_i + \gamma g_i(y_i)\})^2 - \mu_i^2\}$. Also, the gradient of (18) is obtained

$$\nabla_{\mathbf{W}}\mathcal{L}_2(\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = E\{\nabla_{\mathbf{y}}\mathcal{J}(\mathbf{y})\mathbf{x}^{\mathrm{T}}\} + E\{\nabla_{\mathbf{y}}\mathcal{G}(\mathbf{y}, \boldsymbol{\mu})\mathbf{x}^{\mathrm{T}}\} + 2\nabla_{\sum_{\mathbf{yy}}}\mathcal{H}(\mathbf{y}, \boldsymbol{\lambda})E\{\mathbf{y}\mathbf{x}^{\mathrm{T}}\} \tag{19}$$

where the terms containing the vector $\nabla_{\mathbf{y}}\mathcal{J}(\mathbf{y})$ and the matrix $\nabla_{\sum_{\mathbf{yy}}}\mathcal{H}(\mathbf{y}, \boldsymbol{\lambda})$ are as given in (12) and the additional term $\nabla_{\mathbf{y}}\mathcal{G}(\mathbf{y}, \boldsymbol{\mu}) = (\mathcal{G}'_{y_1}(\mathbf{y}, \boldsymbol{\mu}), \mathcal{G}'_{y_2}(\mathbf{y}, \boldsymbol{\mu}), \ldots \mathcal{G}'_{y_l}(\mathbf{y}, \boldsymbol{\mu}))^{\mathrm{T}}$ with elements, $\mathcal{G}'_{y_i}(\mathbf{y}, \boldsymbol{\mu}) = \mu_i g'_{y_i}(y_i)$. The Hessian matrix is approximated as

$$\nabla^2_{\mathrm{Vec}(\mathbf{W}^{\mathrm{T}})}\mathcal{L}_2(\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \approx \mathbf{D}_2 \otimes \sum_{\mathbf{xx}} \tag{20}$$

where $\mathbf{D}_2$ is a diagonal matrix having diagonal elements given by the vector $\mathbf{d}(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ with elements $d_i(y_i, \mu_i, \lambda_{ii}) = E\{\mu_i g''_{y_i}(y_i)\} + 8\lambda_{ii} - E\{\hat{\rho}_i f''_{y_i}(y_i)\}$. By applying the relational properties same as in Section III, the Newton-like learning in (5) is then obtained. By further transforming the vector back to the matrix form

$$\Delta\mathbf{W} = -\eta(\boldsymbol{\Phi}(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \boldsymbol{\Psi}(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\lambda}) + \boldsymbol{\Omega}(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\lambda}))\sum_{\mathbf{xx}}^{-1} \tag{21}$$

where the matrices $\boldsymbol{\Phi}(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = [\boldsymbol{\phi}_1(y_1, \mu_1, \lambda_{11}) \; \boldsymbol{\phi}_2(y_2, \mu_2, \lambda_{22}) \; \cdots \; \boldsymbol{\phi}_l(y_l, \mu_l, \lambda_{ll})]^{\mathrm{T}}$ with elements, $\boldsymbol{\phi}_i(y_i, \mu_i, \lambda_{ii}) = E\{\mathcal{J}'_{y_i}(\mathbf{y})\mathbf{x}\}/d_i(y_i, \mu_i, \lambda_{ii}), \boldsymbol{\Psi}(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = [\boldsymbol{\psi}_1(y_1, \mu_1, \lambda_{11}) \; \boldsymbol{\psi}_2(y_2, \mu_2, \lambda_{22}) \; \cdots \; \boldsymbol{\psi}_l(y_l, \mu_l, \lambda_{ll})]^{\mathrm{T}}$ with $\boldsymbol{\psi}_i(y_i, \mu_i, \lambda_{ii}) = E\{\mathcal{G}'_{y_i}(\mathbf{y}, \boldsymbol{\mu})\mathbf{x}\}/d_i(y_i, \mu_i, \lambda_{ii})$, and $\boldsymbol{\Omega}(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = [\boldsymbol{\omega}_1(\mathbf{y}, \mu_1, \boldsymbol{\lambda}_1) \; \boldsymbol{\omega}_2(\mathbf{y}, \mu_2, \boldsymbol{\lambda}_2) \; \cdots$

$\boldsymbol{\omega}_l(\mathbf{y}, \mu_l, \boldsymbol{\lambda}_l)]^{\mathrm{T}}$ with elements, $\boldsymbol{\omega}_i(\mathbf{y}, \mu_i, \boldsymbol{\lambda}_i) = 2E\{\mathbf{x}\mathbf{y}^{\mathrm{T}}\}\nabla_{\sum_{y_i\mathbf{y}}} \mathcal{H}(\mathbf{y}, \boldsymbol{\lambda})/d_i(y_i, \mu_i, \lambda_{ii})$. The Lagrange multipliers are updated based on (6) and (7) with corresponding $\mathbf{g}$ and $\mathbf{h}$.

A. Stability Analysis

The rest of this section presents an analysis on the convergence stability and the selection of the parameters of the ICA-R. The algorithm is able to reach the minimum of the objective function with Newton-like learning and produce the optimum output $\mathbf{y}^*$ to extract the desired ICs defined by the KT triple $(\mathbf{W}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ that satisfies the first-order conditions: $\nabla_{\mathbf{W}}\mathcal{L}(\mathbf{W}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) = \mathbf{0}, \mathbf{h}(\mathbf{y}^*) = \mathbf{0}, \mathbf{g}(\mathbf{y}^*) \leq \mathbf{0}, \boldsymbol{\lambda}^* > \mathbf{0}, \boldsymbol{\mu}^* \geq \mathbf{0}$ and $\boldsymbol{\mu}^{*\mathrm{T}}\mathbf{g}(\mathbf{y}^*) = \mathbf{0}$. The convergence of the ICA-R learning depends on the thresholds, $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots \xi_l)^{\mathrm{T}}$. In single-source extraction, the desired IC is determined by the value of the single threshold, say ξ. Any component c_j other than the desired source, having the closeness measure $\varepsilon(c_j, r) \leq \xi$ corresponds to the local optima. By selecting a suitable value for ξ, as in the range Υ_i, one and only one desired IC is obtained as the optimum output at the global minimum. However, if ξ is selected beyond the upper bound of the range, the output may produce a different IC. If ξ is too small, the corresponding output, say y, is unable to produce any desired IC because the corresponding constraint $g(y) \gg 0$ causes the learning become unpredictable. In practice, the algorithm uses a small ξ initially to avoid any local minima and then gradually increases its value to converge at the global minimum corresponding to the one and only one desired IC. This discussion can be extended to the cases of multiple outputs $(l > 1)$ to select and adjust the thresholds, individually. If all reference signals are different from one another and the thresholds are adjusted properly, the algorithm reaches the global minimum extracting all the desired ICs.

The algorithm converges stably to the minimum point when the Hessian matrix is positive definite. Since the latter two items of (19) were examined for positivity in the last section, here, we consider only the first item $\mu_i g''_{y_i}(y_i)$ corresponding to the closeness between the outputs and reference signals. It is important to select a suitable distance as the closeness measure $\varepsilon(y_i, r_i)$, which is the key function in the inequality constraints. A common and simple measure is the mean square error (MSE): $\varepsilon(y_i, r_i) = E\{(y_i - r_i)^2\}$, which requires both y_i and r_i normalized to have same mean and variance. Alternatively, correlation can also be used such that $\varepsilon(y_i, r_i) = 1/(E\{y_i r_i\})^2$; both the output and reference signals need to be normalized, so the value of correlation is bounded. The selection of the closeness measure may be different from one output to another, depending on what form the reference signals are available. Using either the MSE or correlation for $\varepsilon(y_i, r_i)$, it could be easily shown that the term $\mu_i^* g''_{y_i^*}(y_i^*)$ is always positive for positive Lagrange multipliers μ_is. Therefore, extending the stability analysis in the last section, the convergence of the learning algorithms in (21) is always stable when one uses the nonlinear functions $f_{\mathrm{sup}}(y)$ for super-Gaussian or $f_{\mathrm{sub}}(y)$ for sub-Gaussian signals and the MSE or correlation as the closeness measure.

V. Experiments and Results

Two experiments using face images and fMRI data are presented to demonstrate the aforementioned applications of the cICA.

A. Extracting Face Images

When analyzing images, the accuracies of the recovered ICs, compared to the sources, were expressed using the peak signal-to-noise ratio (PSNR) in dB: $\mathrm{PSNR} = 10\log_{10}(\mathrm{P}^2/\mathrm{MSE})$, where P is the peak intensity of the image. The performance of the source separation was measured by a performance index (PI) of the permutation error

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

where $\mathrm{rPI}_i = \sum_{j=1}^{m}(|p_{ij}|/\max_k |p_{ik}|) - 1$ and $\mathrm{cPI}_j = \sum_{i=1}^{m}(|p_{ij}|/\max_k |p_{kj}|) - 1$ in which p_{ij} denotes the (i, j) th element of the permutation matrix $\mathbf{P} = \mathbf{WA}$. The term rPI_i gives the error of the separation of the output component y_i with respect to the sources, and cPI_j measures the degree of the source c_j appearing multiple times at the output. PI is zero when all ICs are perfectly separated.

Two face images p_1 and p_2, a scenic picture p_3, and a white noise image p_4, each with size of 233×327 pixels were transformed to have zero means and unit variances and mixed using a 4×4 random mixing matrix to produce four mixture images. The normalized kurtosis (κ_4) of the original images were $-0.82, -1.59, 1.70$, and -0.04, respectively, indicating that two face images are distributed sub-Gaussianly and the scenic picture is distributed super-Gaussianly. The image mixtures were processed in order to separate the two face images, using the less-complete ICA, classical ICA with preprocessing of PCAs dimension reduction, classical ICA with postselection from all outputs, nonlinear PCA, and fastICA approaches. The original and mixture images, and the recovered face images are shown in Figs. 1 and 2, respectively.

The less-complete ICA Newton-like learning rule (15), with the nonlinear function $f_{\mathrm{sub}}(y)$ given in (17) recovered only the two face images, without separating the other images. The traces of shadows of other images, appeared in the recovered images are due to the existing dependence among original images and the mismatch between the assumed density functions and the properties of the face images. The inputs were reduced to two by applying the PCA before performing the classical ICA. The recovered images with this approach were severely corrupted, maybe due to the preprocessing. Alternatively, two face images were postselected from the four components recovered by the classical ICA. The outputs were similar to those recovered by the less-complete ICA algorithm, but the algorithm needed more iterations and produced images with poorer quality. The face images recovered by the nonlinear PCA were also poor in quality because of the artifacts introduced in the prewhitening process. The fastICA could not separate the face images; instead, it produced a mixture of the two face images and another of the scenic

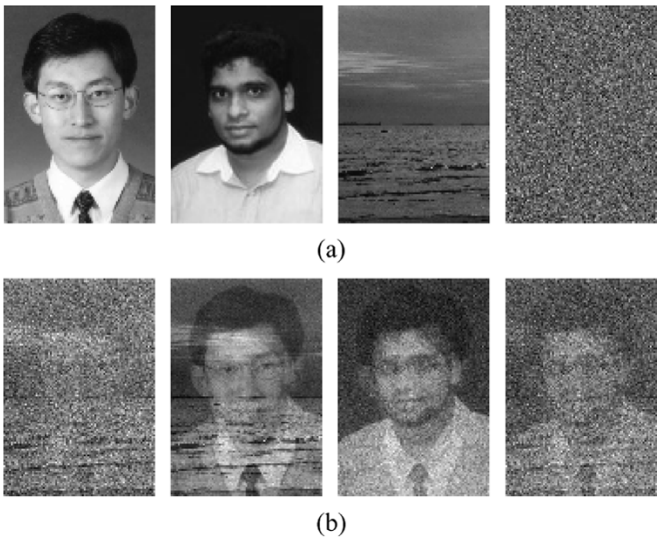


Fig. 1. Generation of mixtures of gray images. (a) Original face images, scenic image, and white noise. (b) Four mixture images.

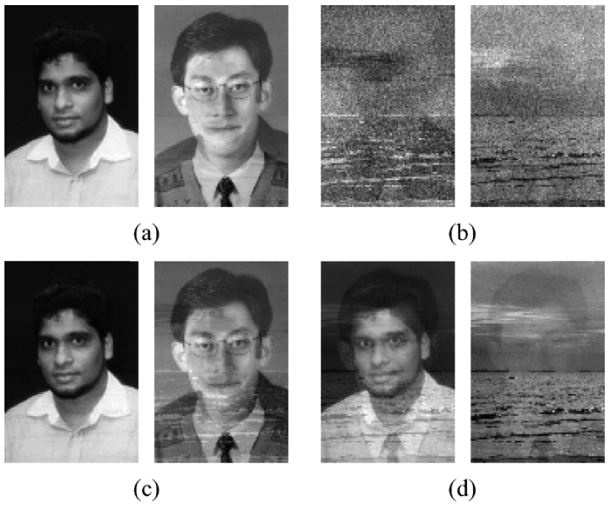


Fig. 2. Two face images separated from the four mixture images by (a) the less-complete ICA. (b) Classical ICA with a prewhitening. (c) Nonlinear PCA. (d) FastICA.

picture as it could not distinguish the signals according to super- or sub-Gaussianity. A comparison of the PSNRs of the outputs and the PIs of the converged algorithms is given in Table I. The present algorithm is superior over the other approaches in face image recovery as seen by the clearer images and indicated by better PSNR and PI values.

B. Extracting Task-Related Components From fMRI Data

FMRI data carry spatio-temporal information of the brain activated by some functional task, which are always confounded by physiological signals, the electronic noise of the scanners, and other environmental artifacts [36]–[39]. The detection of the activated brain voxels is often done by either using the correlation analysis [40] or using the statistical parametric mapping (SPM) [41]. Prior to this analysis, fMRI data are required

TABLE I
COMPARISON OF THE RESULTS OF RECOVERING TWO FACE IMAGES FROM FOUR IMAGE MIXTURES BY DIFFERENT APPROACHES. p_1 AND p_2 DENOTE THE FACE IMAGES, p_3 THE SCENIC IMAGE, AND p_4 THE NOISE IMAGE. THE PEAK SIGNAL-TO-NOISE RATIOS (PSNRS) ARE INDICATED ONLY FOR THE RECOVERED IMAGES. THE TOTAL PERFORMANCE INDEX (PI) INDICATES THE QUALITY OF THE SEPARATION

Algorithm	PSNR (dB)				Total PI
	p_1	p_2	p_3	p_4	
less-complete ICA	16.86	40.34	-	-	0.31
ICA with Preprocess : PCA Dimension Reduction	-	-	13.31	18.83	1.29
ICA with Postprocess : Desired Outputs Selection	16.79	30.64	-	-	1.10
Nonlinear PCA	10.74	34.62	-	-	0.53
FastICA	-	25.70	-	20.49	0.47

to preprocess using appropriate denoising or smoothing filters and correct for artifacts. Recently, ICA has been treated as a data-driven approach for the analysis of fMRI data, which intends to separate the components of task-related brain activation from other components due to interferences and artifacts [38]. Therefore, preprocessing of fMRI data, that could affect the data, is unnecessary when using the ICA for the detection of brain activation.

The classical ICA approach presumes that the task-related activation are independent of nontask related components in spatial-domain and results in all the independent component maps; postprocessing is therefore necessary to identify the components of brain activation [42]. The spatial independence of the effect of heartbeat, respiration, and blood flow in the brain is questionable as their influence is common to most regions of the brain. Since these physiological signals have frequencies different from functional tasks, it is more appropriate to assume that these signals mixed in fMRI data are independent rather in time-domain. Because of the huge computational requirement to process all the components, the assumption of the temporal independence of the sources in the classical ICA is practically prohibitive. On the other hand, the ICA-R technique could effectively use the input stimuli as the reference signals to produce only the few components of brain activation and therefore presume the temporal independence of the sources mixed in fMRI.

In what follows, we present two analyzes of the fMRI data obtained in visual task and event-related working memory experiments performed at 3.0 Tesla Medspec 30/100 scanner at the MRI Center of the Max-Planck-Institute of Cognitive Neuroscience, Leipzig, Germany.

1) Visual Task: An 8-Hz alternating checker board pattern with a central fixation point was projected on an LCD system; five subjects were asked to fixate on the point of stimulation. A series of 64 FLASH images were acquired at three axial levels at the visual cortex of the brain while the subjects were performing alternatively the stimulation and rest tasks. The parameters of the scanning protocols and more detailed experimental can be found in Rajapakse *et al.* [36].

The brain activation was detected using the ICA with postselection, the ICA-R, and the SPM technique. The ON-OFF stimulation was used as the reference signal and the correlation was

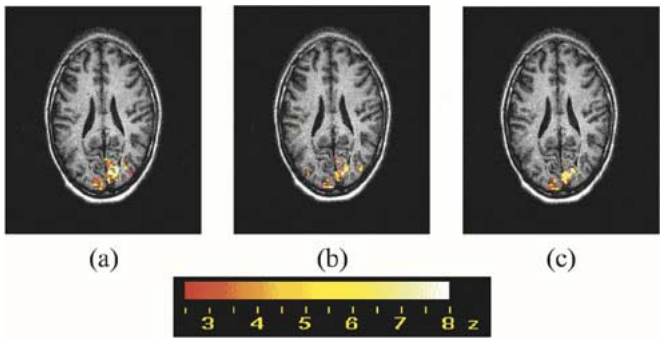


Fig. 3. Task-related activation at the second-axial slice of a representative subject performing the visual experiment, detected by the (a) SPM, (b) classical ICA with postselection, and (c) ICA-R techniques. The significance values (z-values) of the activated voxels are shown color-coded.

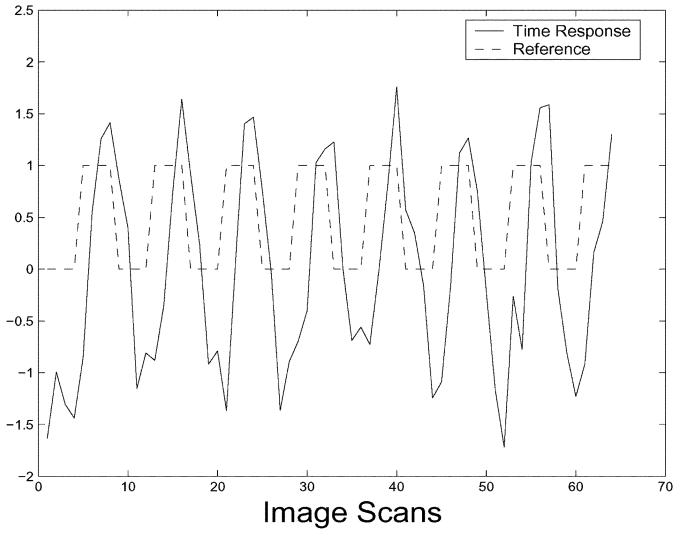


Fig. 4. Input stimulus used as the reference, and the task-related time response from the activated brain voxels detected by the ICA-R algorithm at the second-axial slice from a representative subject performing the visual task.

used as the closeness measure for the ICA-R. Unlike the ICA-R, the SPM technique required preprocessing of the fMRI data in order to remove noise, and artifacts due to subjects' head motion; without preprocessing, the detected activation maps looked noisy and unrealistic. In the classical ICA approach, the corresponding activation maps were selected from the 64 independent component maps by using a correlation analysis between the time responses of the components and the stimuli. The detected activation at the second-axial level of a representative subject, by using the SPM, the classical ICA with postselection, and ICA-R are shown in Fig. 3(a)–(c). Activations detected by the classical ICA and ICA-R were similar to that of the SPM, but the detected activation by the ICA-R contained less spurious noise than others. The ICA-R required much less computation resources and time since the task-related component map was detected without any pre- or postprocessing, in a single process. Fig. 4 shows the input stimulation used as the reference signal and the time response from a representative activated brain region detected using the ICA-R technique.

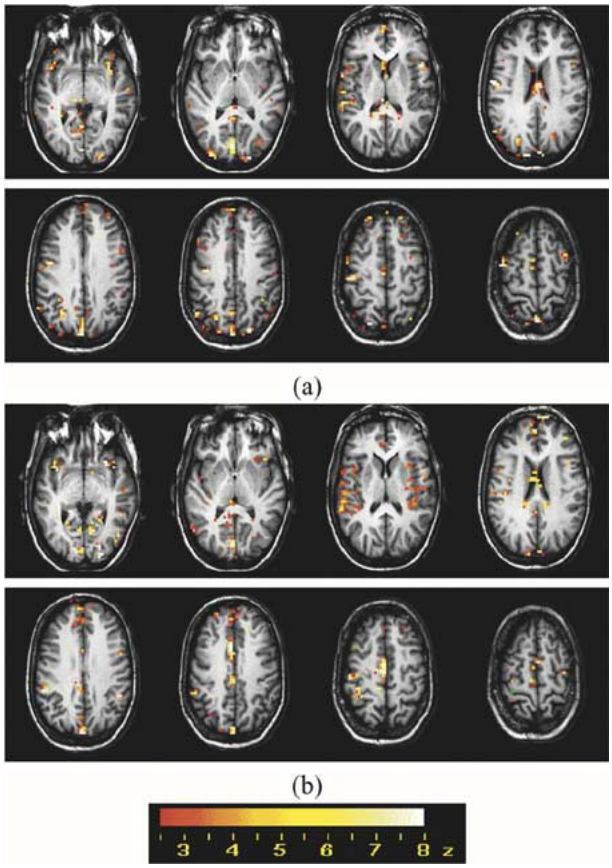


Fig. 5. Activation detected by ICA-R technique at eight axial levels of the brain of a representative subject performing the event-related working memory experiment for (a) the cue phase and (b) the probe phase. The significance values (z-values) of the activated voxels are shown color-coded.

2) Event-Related Working Memory Task: A subspan set of sizes 3, 4, 5, and 6 letters comprising of consonants of the alphabet excluding the letter Y was presented visually for 3 s (cue phase), followed by a probe letter (probe phase) that appeared after a variable delay length. Subjects had to decide if the probe letter belonged to the previously presented subspan set. All trial combinations were presented randomly in a single run; in each trial, a series of 3456 slices at eight axial levels of the brain were obtained using an echo-planar imaging (EPI) protocol. More details about the experiments are available in [43].

The images were analyzed to obtain the activation maps of the cue and probe phases by ICA-R technique using the two stimuli, periodic cue phase and nonperiodic probe phase, as reference and the correlation as the closeness measure. To find and display voxels contributing significantly, the activation maps were scaled to z-values [38] and those voxels having absolute z-values greater than 2.5 were considered as activated voxels. Z-statistical scores of the detected activated voxels were superimposed on to the corresponding anatomical slices for visualization. The activation maps from a representative subject are shown in Fig. 5. The maps showed the activation in the anticipated regions of the brain and were similar to those obtained with the SPM technique [43]. The detected time responses from

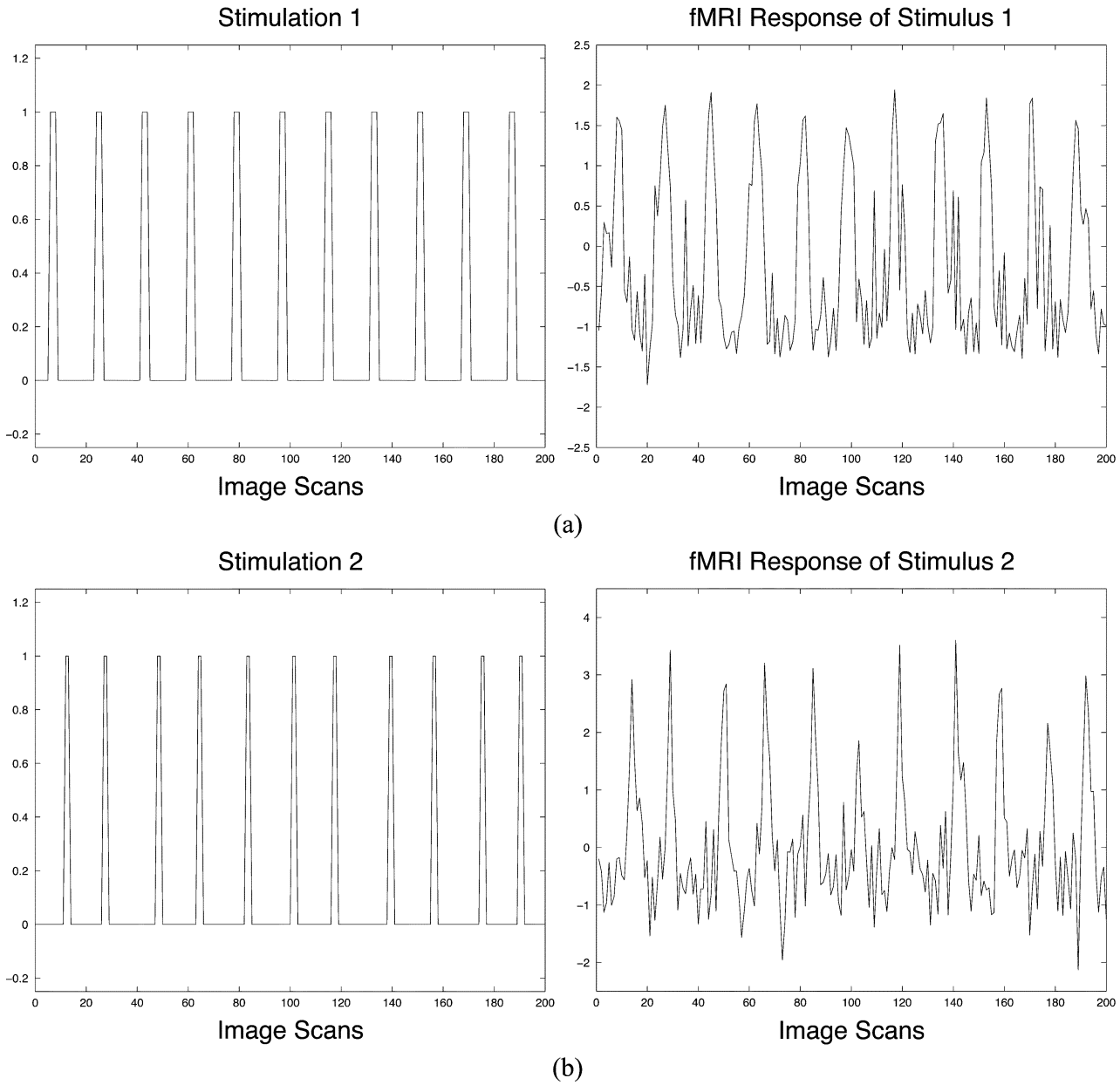


Fig. 6. fMRI time responses of representative activated brain regions from a subject performing the event-related working memory experiment, detected by the ICA-R technique, with the corresponding stimuli for (a) the cue phase and (b) the probe phase.

representative activated brain regions are shown with the input stimuli (references) in Fig. 6. The classical ICA could not be used to detect the brain activation because of the demand for huge memory to process all 3456 components.

VI. Discussion and Conclusion

The cICA is a general framework to incorporate additional knowledge into the ICA. We demonstrated two applications of the cICA: the less-complete ICA and the ICA-R.

The less-complete ICA provides an approach to dimensionality reduction in the ICA; Newton-like learning was derived using the negentropies of the estimated signals as the contrast function and provided at least quadratic learning speed with stable convergence. With suitable nonlinear functions, $f(\cdot)$, chosen, the less-complete ICA algorithm is capable of separating the ICs according to their density types, such as the Gaussianity, which was useful in separating face images from their mixtures with other images. However, the less-complete ICA algorithm is insufficient to facilitate the applications involving an unknown number of sources having close densities.

The ICA-R algorithm extends the less-complete ICA better extract an interesting subset of ICs if the information available on the ICs can be formulated as reference signals. Reference signals carry traces of information of the desired signals and need not to be exact to the sources. The ICA-R reaches the global minimum producing exactly the original sources if the closeness measures and thresholds are properly selected. The ICA-R facilitates the analysis of fMRI data as the input stimuli can effectively be used as the reference signals to produce only the task-related brain activation, discarding all other interferences,

artifacts and noises. In contrast, the classical ICA produces all the components mixed in the fMRI data, which number could be so large that makes ICA practically prohibitive. Importantly, the ICA-R enables the use of temporal independence of signal sources in fMRI. The computational resources and time required by the ICA-R are much less than those required by the classical ICA.

In summary, the cICA is a semiblind approach and has many advantages and applications over the classical ICA methods when useful constraints are incorporated into the contrast function. But not all kinds of constraints can be used in cICA because some may infringe the ICA equivariant properties. The constraints should be selected and formulated properly in the sense of being consistent with the independence criteria. When the less-complete ICA problem is considered, the constraints of uncorrelatedness between each pair of outputs are just suitable for both compatibility with independence criteria and solving the ambiguities of duplicate ICs. The success of ICA-R algorithm is achieved while we have the proper selection and adjustment of the threshold ξ for the constraints of closeness to reference signals. From these two cICA example applications, enhancement of the quality of separation, reduction of dimensionality, and stable and fast convergence of the ICA were seen feasible with the incorporation of choosing right types of constraints.

ACKNOWLEDGMENT

The visual task data and event-related working memory data were provided by Dr. F. Kruggel at the Max-Planck-Institute of Cognitive Neuroscience, Leipzig, Germany.

REFERENCES

[1] P. Comon, "Independent component analysis: A new concept?," *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[2] S. Amari and A. Cichocki, "Adaptive blind signal processing—Neural network approaches," *Proc. IEEE*, vol. 86, no. 10, pp. 2026–2048, Oct. 1998.

[3] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," in *Proc. Inst. Electr. Eng. Radar and Signal Processing*, vol. 140, Dec. 1993, pp. 362–370.

[4] J. Karhunen, A. Hyvärinen, R. Vigario, J. Hurri, and E. Oja, "Applications of neural blind separation to signal and image processing," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'97)*, Munich, Germany, 1997, pp. 131–134.

[5] S. Makeig, T. Jung, A. Bell, D. Ghahremani, and T. Sejnowski, "Blind separation of auditory event-related brain responses into independent components," *Proc. Nat. Acad. Sci.*, vol. 94, pp. 10979–10984, 1997.

[6] A. D. Back, "A first application of independent component analysis to extracting structure from stock returns," *Int. J. Neural Syst.*, vol. 8, no. 4, pp. 473–484, 1997.

[7] J.-H. Lee, H.-J. Jung, T.-W. Lee, and S.-Y. Lee, "Speech feature extraction using independent component analysis," in *Proc. ICASSP*, vol. 2, Seattle, WA, 1998, pp. 1249–1252.

[8] A. Kuh and X. Gong, "Independent component analysis for blind multiuser detections," in *Proc. IEEE Int. Symp. Information Theory*, 2000, p. 246.

[9] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computat.*, vol. 9, no. 7, pp. 1483–1492, 1997.

[10] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computat.*, vol. 7, pp. 1129–1159, 1995.

[11] J. Karhunen, "Neural approaches to independent component analysis and source separation," in *Proc. 4th Eur. Symp. Artificial Neural Networks (ESANN'96)*, Bruges, Belgium, Apr. 24–26, 1996, pp. 249–266.

[12] T.-W. Lee, M. Girolami, and T. Sejnowski, "Independent component analysis using an extended informax algorithm for mixed sub-Gaussian and super-Gaussian sources," *Neural Computat.*, vol. 11, no. 2, pp. 409–433, 1999.

[13] S. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind signal separation," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 1996, vol. 8, pp. 752–763.

[14] E. Oja, "The nonlinear PCA learning rule in independent component analysis," *Neurocomput.*, vol. 17, pp. 25–45, 1997.

[15] J. Karhunen and J. Joutsensalo, "Representation and separation of signals using nonlinear PCA type learning," *Neural Netw.*, vol. 7, no. 1, pp. 113–127, 1994.

[16] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999.

[17] D. T. Pham and P. Garat, "Blind separation of mixture of independent sources through a quasimaximum likelihood approach," *IEEE Trans. Signal Process.*, vol. 45, no. 7, pp. 1712–1725, Jul. 1997.

[18] A. Cichocki, R. Thawonmas, and S. Amari, "Sequential blind signal extraction in order specified by stochastic properties," *Electron. Lett.*, vol. 33, no. 1, pp. 64–65, 1997.

[19] A. Hyvärinen and E. Oja, "Simple neuron models for independent component analysis," *Int. J. Neural Syst.*, vol. 7, no. 6, pp. 671–687, Dec. 1996.

[20] A. Cichocki, J. Karhunen, W. Kasprzak, and R. Vigário, "Neural networks for blind separation with unknown number of sources," *Neurocomput.*, vol. 24, no. 1–3, pp. 55–93, 1999.

[21] W. Lu and J. C. Rajapakse, "Constrained independent component analysis," in *Advances in Neural Information Processing Systems 13 (NIPS2000)*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2000, pp. 570–576.

[22] ——, "Eliminating indeterminacy in ica," *Neurocomput.*, vol. 50, pp. 271–290, 2003.

[23] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. New York: Academic, 1982.

[24] A. Hyvärinen, "Survey on independent component analysis," *Neural Comput. Surveys*, vol. 2, pp. 94–128, 1999.

[25] J. F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Process.*, vol. 44, no. 12, pp. 3017–3030, Dec. 1996.

[26] S. Amari, T. Chen, and A. Cichocki, "Nonholonomic orthogonal constraints in blind source separation," *Neural Computat.*, vol. 12, no. 6, pp. 1463–1484, 2000.

[27] X.-R. Cao and R.-W. Liu, "General approach to blind source separation," *IEEE Trans. Signal Process.*, vol. 44, no. 3, pp. 562–571, Mar. 1996.

[28] F. Herrmann and A. K. Nandi, "Solution of high-dimensional linear separation problems," in *Proc. EUSIPCO'00*, vol. 1, Tempere, Finland, pp. 1489–1492.

[29] M. Girolami and C. Gyfe, "Extraction of independent signal sources using a deflationary exploratory projection pursuit network with lateral inhibition," *Proc. Inst. Electr. Eng. Vision Image, and Signal Processing*, vol. 144, no. 5, pp. 299–306, Oct. 1997.

[30] ——, "Generalized independent component analysis through unsupervised learning with emergent bussgang properties," in *Proc. Int. Conf. Neural Networks*, vol. 3, 1997, pp. 1788–1791.

[31] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1991.

[32] A. Hyvärinen, "New approximations of differential entropy for independent component analysis and projection pursuit," in *Proc. Advances Neural Information Processing Systems 10 (NIPS'97)*, pp. 273–279.

[33] ——, "New approximations of differential entropy for independent component analysis and projection pursuit," in *Proc. Advances Neural Information Processing Systems*, vol. 10, 1998, pp. 273–279.

[34] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition," *Neural Computat.*, vol. 13, no. 4, 2001.

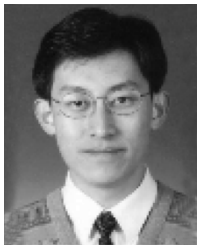
[35] J. Luo, B. Hu, X.-T. Ling, and R.-W. Liu, "Principal independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 4, pp. 912–917, Jul. 1999.

[36] J. C. Rajapakse, F. Kruggel, J. M. Maisog, and D. Y. von Cramon, "Modeling hemodynamic response for analysis of functional MRI time-series," *Human Brain Mapping*, vol. 6, pp. 283–300, 1998.

[37] P. P. Mitra, S. Ogawa, X. Hu, and K. Ugurbil, "The nature of spatiotemporal changes in cerebral hemodynamics as manifested in functional magnetic resoance imaging," *Magnetic Resonance Medicine*, vol. 37, pp. 511–518, 1997.

[38] M. McKeown, S. Makeig, G. Brown, T.-P. Jung, S. Kindermann, A. Bell, and T. Sejnowski, "Analysis of fMRI data by blind separation into independent spatial components," *Human Brain Mapping*, vol. 6, pp. 160–188, 1998.

[39] J. C. Rajapakse and J. Piyaratna, "Bayesian approach to segmentation of statistical parametric maps," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 10, pp. 1186–1194, Oct. 2001.
[40] P. A. Bandettini, A. Jesmanowicz, E. C. Wong, and J. S. Hyde, "Processing strategies for time-course data sets in functional MRI of the human brain," *Magnetic Resonance Medicine*, vol. 30, pp. 161–173, 1993.
[41] K. J. Friston, "Statistical parameter mapping," in *Functional Neuroimaging: Technical Foundations*, R. W. Thatcher, M. Hallett, T. Zeffiro, W. R. John, and M. Huerta, Eds. New York: Academic, 1994.
[42] M. McKeown, S. Makeig, G. Brown, T.-P. Jung, S. Kindermann, and T. Sejnowski, "Spatially independent activity patterns in functional magnetic resonance imaging data during the stroop color-naming task," *Proc. Nat. Acad. Sci. USA*, vol. 95, pp. 803–810, 1998.
[43] F. Kruggel, S. Zysset, and D. Y. von Cramon, "Nonlinear regression of functional MRI data: An item recognition study," *Neuroimage*, vol. 12, no. 2, pp. 173–183, 2000.



Wei Lu (S'00–M'02) received the B.A.Sc. (First Class Honors) and Ph.D. degrees in computer engineering from the Nanyang Technological University, Singapore, in 1999 and 2003, respectively.

In 2002, he joined Singapore Research Laboratory of Sony Electronics, Singapore, where he is presently a Senior Research Engineer. He is author or coauthor of 16 research publications in refereed journals, conference proceedings, and books. He is also principal inventor or coinventor for 13 patent applications in Singapore, the United States, the European Union, and Japan. His current research interests are in statistical data analysis, multimedia signal processing, machine learning, neural networks and artificial intelligence, especially in the advanced area of independent component analysis (ICA). Dr. Lu is a recipient of the Singapore Keppel Corporation Scholarship (1995–98).

Jagath C. Rajapakse (S'91–M'93–SM'00) received the B.Sc.(Eng.) degree (First Class Honors) from the University of Moratuwa, Sri Lanka, in 1985, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the State University of New York at Buffalo in 1989 and 1993, respectively.

He was a Visiting Fellow at the National Institutes of Health Bethesda, MD, from 1993 to 1996, and a then Visiting Scientist at the Max-Planck-Institute of Cognitive Neuroscience, Leipzig, Germany. In 1998, he joined the Nanyang Technological University, Singapore, where he is currently an Associate Professor in the School of Computer Engineering and also the Deputy Director of the BioInformatics Research Center (BIRC). He has been a Guest Researcher at the Institute of Statistical Sciences, Academia Sinica, Taipei, Taiwan, R.O.C., and the Brain Science Institute, RIKEN, Tokyo, Japan. He is author or coauthor for more than 125 research publications in refereed journals, conference proceedings, and books. He also serves on the editorial board of *Neural Information Processing: Letters and Reviews*. His current teaching and research interests are in computational biology, neuroimaging, and machine learning.

Dr. Rajapakse is currently a governing board member of APNNA and a member of AAAS. He was a recipient of the Fulbright Scholarship from 1985 to 1987.