ICA WITH REFERENCE

Wei Lu, Jagath C. Rajapakse

School of Computer Engineering
Nanyang Technological University, Singapore
email:{p141035107, asjagath}@ntu.edu.sg

ABSTRACT

We present a novel approach to extract a subset of independent sources from multidimensional observations when some a priori information that can be incorporated to the learning algorithm as reference is available. The constrained independent component analysis (cICA) is extended to use new constraints, and a Newton-like learning algorithm is proposed to give an optimal solution to the constrained optimization problem. The convergence and the effect of parameters of the learning algorithm are analyzed. Simulations with the mixtures of deterministic and random signals and synthetic fMRI data demonstrate the efficacy and accuracy of the proposed algorithm.

Keywords: ICA, constrained independent component analysis (cICA), ICA with reference

1. INTRODUCTION

Some source separation problems like signal detection and noise cancellation often expect to estimate a desired single source or an interesting subset of sources from mixtures thereof, for example, the problems in speech extraction and financial analysis [1]. Conventionally, second-order methods, like the minimum mean square error (MMSE) technique, were often used to detect, extract and recognize the desired signals [2], but their applications are limited due to the use of only 2nd-order statistics. Using kurtosis or negentropy as the contrast function, one-unit ICA algorithms have been proposed to separate a single source from a set of mixtures of independent sources [3]. However, the extraction of a particular source by using these algorithms is always determined by the contrast function or sometimes arbitrary due to local minima [4]. Some researchers had introduced asymmetric information using sparse decomposition of signals [4] or fourth-order cumulants [5] into the algorithms to solve such problems, however, they all still have some drawbacks, such as the system reliability and complexity.

Let us denote the time varying observed signal by $\mathbf{x}(t) = (x_1(t) \ \cdots \ x_n(t))^{\mathrm{T}}$ and the source signal consisting of independent components (ICs) by $\mathbf{c}(t) = (c_1(t) \ \cdots \ c_m(t))^{\mathrm{T}}$. The linear ICA assumes that the signal $\mathbf{x}(t)$ is a linear mixture of ICs:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{c}(t) \tag{1}$$

where the matrix $\mathbf{A}$ of size $n \times m$ represents linear memoryless mixing channels. It often assumes that $n = m$ (complete ICA); we hold this assumption in this paper for simplicity, which can be relaxed without losing generality. The time index t is omitted in the following for simplicity of equations.

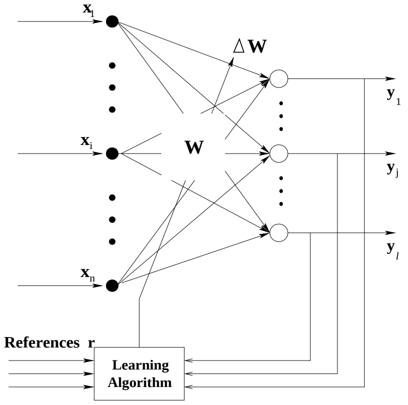


Fig. 1. Illustration of a neural network extracting multiple desired independent sources from the input mixtures $\mathbf{x} = (x_1 \cdots x_n)^{\mathrm{T}}$ using reference signals $\mathbf{r}$. $\mathbf{W}$ denotes the weight matrix and $\mathbf{y} = (y_1 \cdots y_l)^{\mathrm{T}}$ the output signals.

This paper presents a general approach to extract one or several desired independent sources from the observations $\mathbf{x}$ with minimal knowledge of $\mathbf{A}$ and original sources $\mathbf{c}$. The algorithm incorporates *a priori* information of the sources in the form of a rough template which is referred to as the *reference* signal, $\mathbf{r} = (r_1 \ \cdots \ r_l)$. The technique of constrained independent component analysis (cICA) [6] is adopted to systematically introduce a measure of the closeness between the output and the reference into the ICA contrast function. Such extraction is expressed as a constrained optimization problem which is solved by a neural network algorithm using a Newton-like learning. In this paper, first, the algorithm of extracting one desired source is introduced, and then extended to the situation when a subset of sources is desired. Fig. 1 shows the neural network with multiple neurons; the output $\mathbf{y}$ is given by

$$\mathbf{y} = \mathbf{W}\mathbf{x}, \tag{2}$$

where $\mathbf{W} = [\mathbf{w}_1 \ \cdots \ \mathbf{w}_l]^{\mathrm{T}}$ is the matrix containing l weight vectors $\mathbf{w}_j = (w_{j1} \ w_{j2} \ \cdots \ w_{jn})^{\mathrm{T}}$ which need to be learned. In a single-source extraction, the output $y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$ where $\mathbf{w}$ is a vector and r denotes the reference.

2. PREVIOUS METHODS

2.1. Second-Order Approach

There are many second-order approaches using either minimum mean square error (MMSE) or maximum correlation (MC) criteria to extract a signal close to the reference [2]. In MMSE approaches, a least-mean-square (LMS) method is used to estimate the desired output. At the minimum of MSE, we have the optimum weight $\mathbf{w}^*$ given by

$$\mathbf{w}^* = \mathbf{R}_{\mathbf{xx}}^{-1} E\{r\mathbf{x}\} \tag{3}$$

where $\mathbf{R}_{\mathbf{xx}}$ is the covariance matrix of the signal $\mathbf{x}$ and $E\{\cdot\}$ is the temporal expectation. The corresponding output $y^* = \mathbf{w}^{*\mathrm{T}}\mathbf{x}$ gives a signal close to the reference in second-order statistics.

In the following, we show that the second-order methods are insufficient to recover an independent source from input $\mathbf{x}$. In order to extract one of the ICs, say c_i, we should have the optimum weight vector $\mathbf{w}^*$ such that the vector

$$\mathbf{p}^* = \mathbf{A}^{\mathrm{T}}\mathbf{w}^* \tag{4}$$

is a canonical base vector, $\mathbf{p}^* = \pm\alpha\mathbf{e}_i$, where α is a constant and $\mathbf{e}_i$ is a vector whose elements are zero except that the ith element is 1. Substituting Eq. (3) into Eq. (4) and using $\mathbf{x} = \mathbf{A}\mathbf{c}$, the vector

$$\mathbf{p}^* = \mathbf{R}_{\mathbf{cc}}^{-1} E\{r\mathbf{c}\} \tag{5}$$

where $\mathbf{R}_{\mathbf{cc}} = E\{\mathbf{c}\mathbf{c}^{\mathrm{T}}\}$. Because sources $\mathbf{c}$ are presumed to have components that are mutually independent in statistical sense, the covariance matrix $\mathbf{R}_{\mathbf{cc}}$ is a diagonal matrix with diagonal elements, $\{\mathbf{R}_{\mathbf{cc}}\}_{jj} = \sigma_{c_j}^2$ $(\forall j = 1, \cdots, n)$ which is the variance of the corresponding component. Thus, the inverse of $\mathbf{R}_{\mathbf{cc}}$ is also a diagonal matrix with diagonal elements $\beta_j = \frac{1}{\{\mathbf{R}_{\mathbf{cc}}\}_{jj}} = \frac{1}{\sigma_{c_j}^2}$. So, the vector $\mathbf{p}^*$ can be written as

$$\mathbf{p}^* = (\beta_1 E\{rc_1\} \;\; \beta_2 E\{rc_2\} \;\; \cdots \;\; \beta_n E\{rc_n\})^{\mathrm{T}}. \tag{6}$$

Note that the reference signal is not identical to the original source. In order that $\mathbf{p}^*$ is a canonical base vector, $E\{rc_j\}$ for all $j \neq i$ must be zero, and $E\{rc_i\}$ is not equal to zero. In general, it is impossible for such conditions to be true because we assume that only the component c_i has the biggest correlation with the reference r, but any sources in $\mathbf{c}$ may have a non-zero correlation with the reference. Therefore, as long as more than one IC have non-zero correlations with the reference, the extraction of statistically independent sources cannot be achieved by second-order statistical techniques [7] even if a reference signal is available.

2.2. One-Unit ICA

Higher-order statistics have been adopted to estimate independent sources in ICA [7]. Recently, instead of estimating the whole ICA model consisting of matrix $\mathbf{A}$ and ICs $\mathbf{c}$ as in classical ICA, one-unit ICA simply finds one weight vector $\mathbf{w}$ so that the product $\mathbf{w}^{\mathrm{T}}\mathbf{x}$ equals to one of the ICs [3]. The negentropy $J(y)$ is defined as the natural information-theoretic contrast function of one-unit ICA [7]:

$$J(y) = H(y_{\mathrm{Gaus}}) - H(y) \tag{7}$$

where y_{Gaus} is a Gaussian random variable with the same variance as the output signal y, and $H(\cdot)$ denotes the differential entropy. Maximizing the negentropy produces an independent component [8]. Hyvärinen introduced a flexible and reliable approximation of the negentropy [9]:

$$J(y) \approx \rho[E\{G(y)\} - E\{G(\nu)\}]^2 \tag{8}$$

where ρ is a positive constant, $G(\cdot)$ can be any non-quadratic function, ν is a Gaussian variable having zero mean and unit variance. Some practical functions were suggested for $G(\cdot)$ [3]:

$$G_1(y) = \log\cosh(a_1 y)/a_1 \tag{9}$$

$$G_2(y) = \exp(-a_2\frac{y^2}{2})/a_2 \tag{10}$$

$$G_3(y) = y^4/4 \tag{11}$$

where $1 \leq a_1 \leq 2$ and $a_2 \approx 1$. G_1 is a good general purpose function, G_2 and G_3 are better suited for super-Gaussian and sub-Gaussian signals, respectively [3].

However, when using only the negentropy as one-unit contrast function, theoretically one cannot obtain an IC other than the one having the maximum negentropy among the sources. Therefore, the present one-unit ICA method cannot be used to produce a desired independent source. Moreover, the learning algorithm [3] is not guaranteed to converge to the global maximum at all times because the local convergent point depends on the initial weight vector $\mathbf{w}_0$, the learning rates and other factors [4].

3. ICA WITH REFERENCE

In many blind signal separation problems, one may only want to reliably obtain a particular desired component or a set of desired sources, and automatically discard the rest of uninteresting signals or noises. At some instances, a trace of the desired signals is available, for example, the On-Off stimulation scheme of an fMRI experiments [10].

In this section, a variation to the classical ICA is proposed, in which only a set of desired independent components (ICs) is extracted incorporating the prior information as reference signals; we refer this technique as *ICA with reference*. These reference signals carry some information of the desired sources but not identical to the corresponding desired signals. In this section, first, we present the technique of the one-unit ICA with a reference, and then extend to the ICA with multi-reference.

3.1. One-Unit ICA with a Reference

Our goal is to obtain a learning algorithm that satisfies the following two conditions simultaneously: (1) the estimated output is one of the ICs mixed in the input signal, and (2) the extracted IC is the closest one to the reference signal r in some distance criteria.

Suppose that the contrast function of one-unit ICA is given by the negentropy function $J(y)$ that may have m local or global optimum solutions $\mathbf{w}_i$ $(i = 1, \cdots, m)$ to give the output identical to each independent source c_i $(i = 1, \cdots, m)$. The closeness between the estimated output y and the reference r is measured by some norm, $\varepsilon(y, r)$, where the closest one has the minimum value. Assuming that one of ICs is the one and only one closest to reference r, we can write the inequality relationship:

$$\varepsilon(\mathbf{w}^{*\mathrm{T}}\mathbf{x}, r) < \varepsilon(\mathbf{w}_1^{\mathrm{T}}\mathbf{x}, r) \leq \cdots \leq \varepsilon(\mathbf{w}_{m-1}^{\mathrm{T}}\mathbf{x}, r) \qquad (12)$$

where the optimum vector $\mathbf{w}^*$ corresponds to the desired output. Thus, there exists a threshold $\xi \in \Upsilon = \left[\varepsilon(\mathbf{w}^{*\mathrm{T}}\mathbf{x}, r),\ \varepsilon(\mathbf{w}_1^{\mathrm{T}}\mathbf{x}, r)\right)$ such that the closeness $\varepsilon(y, r)$ is less than or equal to the threshold parameter ξ. i.e. $g(\mathbf{w}) = \varepsilon(y, r) - \xi \leq 0$, only when $y = \mathbf{w}^{*\mathrm{T}}\mathbf{x}$, but not with any other vectors $\mathbf{w}_i$ $(\neq \mathbf{w}^*)$.

While treating the formula $g(\mathbf{w})$ as a feasible constraint to one-unit ICA contrast function $J(y)$, we model our problem in the framework of the constrained independent component analysis (cICA) [6]:

$$\begin{array}{ll} \text{maximize} & J(y) \approx \rho\left[E\{G(\mathbf{w}^{\mathrm{T}}\mathbf{x})\} - E\{G(\nu)\}\right]^2 \\ \text{subject to} & g(\mathbf{w}) \leq 0,\ h(\mathbf{w}) = E\{y^2\} - 1 = 0 \end{array} \qquad (13)$$

The equality constraint $h(\mathbf{w})$ is included to ensure that the contrast function $J(y)$ and the weight vector $\mathbf{w}$ are bounded. Because one and only one IC satisfies the conditions defined in this problem (13), the problem is solved by a global convergent algorithm.

By introducing a slack variable z, we transform the inequality constraint into an equality, $g(\mathbf{w}) + z^2 = 0$. By explicitly manipulating the optimum z^*, the augmented Lagrangian function $\mathcal{L}_1(\mathbf{w}, \mu, \lambda)$ for the problem in Eq. (13) is given by:

$$\begin{aligned} \mathcal{L}_1(\mathbf{w}, \mu, \lambda) = J(y) - \tfrac{1}{2\gamma}[\max{}^2\{\mu + \gamma g(\mathbf{w}), 0\} - \mu^2] \\ -\lambda h(\mathbf{w}) - \tfrac{1}{2}\gamma \left\|h(\mathbf{w})\right\|^2 \end{aligned} \qquad (14)$$

where μ and λ are the Lagrange multipliers for constraints $g(\mathbf{w})$ and $h(\mathbf{w})$, respectively, γ is the scalar penalty parameter, $\|\cdot\|$ denotes the Euclidean norm, and $\frac{1}{2}\gamma\left\|\cdot\right\|^2$ is the penalty term to ensure that the optimization problem is held at the condition of local convexity assumption.

To find the maximum of $\mathcal{L}_1$, $\mathbf{w}$ can be adapted using a Newton-like learning method:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta\left({\mathcal{L}_1}''_{\mathbf{w}_k^2}\right)^{-1} {\mathcal{L}_1}'_{\mathbf{w}_k},$$

where k is an iteration index, η is a positive learning rate added to avoid the uncertainty in convergence, ${\mathcal{L}_1}'_{\mathbf{w}}$ is the first derivative of $\mathcal{L}_1$ with respect to $\mathbf{w}$:

$${\mathcal{L}_1}'_{\mathbf{w}} = \bar{\rho}E\{\mathbf{x}G'_y(y)\} - \tfrac{1}{2}\mu E\{\mathbf{x}g'_y(\mathbf{w})\} - \lambda E\{\mathbf{x}y\} \qquad (15)$$

where $\bar{\rho} = \pm\rho$ whose positive or negative sign coincident with $E\{G(y)\} - E\{G(\nu)\}$, $G'_y(y)$ and $g'_y(\mathbf{w})$ are the first derivatives of $G(y)$ and $g(\mathbf{w})$ with respect to y. To simplify the inversion, the Hessian matrix ${\mathcal{L}_1}''_{\mathbf{w}^2}$ is approximated as the product of a scalar value and the input covariance:

$${\mathcal{L}_1}''_{\mathbf{w}^2} = s(\mathbf{w})\mathbf{R}_{\mathbf{xx}}, \qquad (16)$$

where the scalar $s(\mathbf{w}) = \bar{\rho}E\{G''_{y^2}(y)\} - \frac{1}{2}\mu E\{g''_{y^2}(\mathbf{w})\} - \lambda$, the covariance matrix $\mathbf{R}_{\mathbf{xx}} = E\{\mathbf{x}\mathbf{x}^{\mathrm{T}}\}$, $G''_{y^2}(y)$ and $g''_{y^2}(\mathbf{w})$ are the second derivatives. Then, the approximate Newton learning is given by

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta\mathbf{R}_{\mathbf{xx}}^{-1}{\mathcal{L}_1}'_{\mathbf{w}_k}/s(\mathbf{w}_k) \qquad (17)$$

The optimum multipliers μ^* and λ^* are found by iteratively updating them based on a gradient-ascent method [6]:

$$\mu_{k+1} = \max\{0, \mu_k + \gamma g(\mathbf{w}_k)\}, \qquad (18)$$

$$\lambda_{k+1} = \lambda_k + \gamma h(\mathbf{w}_k). \qquad (19)$$

The expectation in the equations can be estimated using all the samples of the input $\mathbf{x}$.

Following the learning algorithm presented above, the network is able to achieve the local maximum at the optimum point, defined by Kuhn-Tucker (KT) triple $(\mathbf{w}^*, \mu^*, \lambda^*)$, that satisfies the first-order conditions: ${\mathcal{L}_1}'_{\mathbf{w}}(\mathbf{w}^*, \mu^*, \lambda^*) = \mathbf{0}$; $h(\mathbf{w}^*) = 0$; $g(\mathbf{w}^*) \leq 0$; $\lambda^* > 0$, $\mu^* \geq 0$ and $\mu^* g(\mathbf{w}^*) = 0$.

Suppose that the network is in a local maximum and perturbed by a small vector ϵ to $\mathbf{w}^*$. By a truncated Taylor series expansion:

$$\begin{aligned} \mathcal{L}_1(\mathbf{w}^* + \epsilon, \mu^*, \lambda^*) \approx \mathcal{L}_1(\mathbf{w}^*, \mu^*, \lambda^*) + \epsilon^{\mathrm{T}}{\mathcal{L}_1}'_{\mathbf{w}^*}(\mathbf{w}^*, \mu^*, \lambda^*) \\ + \tfrac{1}{2}\epsilon^{\mathrm{T}}{\mathcal{L}_1}''_{\mathbf{w}^{*2}}(\mathbf{w}^*, \mu^*, \lambda^*)\,\epsilon. \end{aligned} \qquad (20)$$

The second term is equal to 0 at the local maximum. For the system to converge, the third term should always be negative for small ϵ. In other words, the Hessian matrix ${\mathcal{L}_1}''_{\mathbf{w}^{*2}}$ should be negative definite, which is known as the second-order condition. For simplicity, let us consider the approximated ${\mathcal{L}_1}''_{\mathbf{w}^{*2}}$ in Eq. (16) although our examination is valid for the original Hessian as well.

${\mathcal{L}_1}''_{\mathbf{w}^{*2}}$ is always negative definite when the scalar $s(\mathbf{w}^*)$ is negative and the input covariance matrix $\mathbf{R}_{\mathbf{xx}}$ is nonsingular. The latter condition is true in most cases that a large number of sample points of signals is available. Even when it is singular or near-singular if inputs have a small number of samples, $\mathbf{R}_{\mathbf{xx}}$ can be transformed to a non-singular matrix by applying a whitening process, e.g. PCA. Consider $G(\cdot)$ functions given in Eqs. (9), (10) and (11), the scalar value $s(\mathbf{w}^*)$ is always negative when we use the general-purpose functions $G_1(y)$, but is negative only for super-Gaussian signal when using $G_2(y)$ and for sub-Gaussian signal when using $G_3(y)$. Therefore, our algorithm is converged locally when one uses the function G_1 in general cases, the function G_2 for super-Gaussian or G_3 for sub-Gaussian signal.

The value of ξ is critical to the convergence of the algorithm. Given a suitable $\xi \in \Upsilon$, one and only one desired IC is defined in the constrained optimization problem, and hence the algorithm is converged globally to produce the particular IC at the output. If ξ is beyond the upper bound of the range Υ, the algorithm has more than one convergent point. If ξ is too small, the algorithm may not converge because the constraint $g(\mathbf{w}) \gg 0$ causes the algorithm to be unstable. In practice, the algorithm better uses a small ξ initially to avoid going to a local optimum, and then gradually increases it to converge at the global maximum.

3.2. ICA with Multi-Reference

When a set of corresponding reference signals is available, the problem can be easily extended to extract several desired independent sources simultaneously. Every output corresponds to a unique independent source different from others as their individual constrained optimizations, defined in Eq. (13), are able to produce the globally optimal solution with the output component distinguishable from others. The problem for ICA with multi-reference can be written as:

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{l} J(y_i) \\ \text{subject to} & \mathbf{g}(\mathbf{W}) \leq \mathbf{0}, \;\; \mathbf{h}(\mathbf{W}) = 0 \end{array}$$

where l is the number of desired independent sources to be extracted, $\mathbf{g}(\mathbf{W}) = (g_1(\mathbf{w}_1) \cdots g_l(\mathbf{w}_l))^{\mathrm{T}}$ in which each $g_i(\mathbf{w}_i) = \varepsilon_i(y_i, r_i) - \xi_i$ for $\forall i = 1, \cdots, l$, and $\mathbf{h}(\mathbf{W}) = (h_1(\mathbf{w}_1) \cdots h_l(\mathbf{w}_l))^{\mathrm{T}}$ containing $h_i(\mathbf{w}_i) = E\{y_i^2\} - 1$ for $\forall i = 1, \cdots, l$. The corresponding augmented Lagrangian function $\mathcal{L}_2(\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ is given by:

$$\begin{aligned} \mathcal{L}_2 = & \sum_{i=1}^{l} \left(J(y_i) - \frac{\max^2\{\mu_i + \gamma_i g_i(\mathbf{w}_i), 0\} - \mu_i^2}{2\gamma_i} \right) \\ & -\boldsymbol{\lambda}^{\mathrm{T}} \mathbf{h}(\mathbf{W}) - \frac{1}{2} \gamma^{\mathrm{T}} \|\mathbf{h}(\mathbf{W})\|^2 \end{aligned}$$

where $\boldsymbol{\mu} = (\mu_1 \cdots \mu_l)^{\mathrm{T}}$ and $\boldsymbol{\lambda} = (\lambda_1 \cdots \lambda_l)^{\mathrm{T}}$ are two sets of the Lagrange multipliers for inequality and equality constraints, respectively, and $\boldsymbol{\gamma} = (\gamma_1 \cdots \gamma_l)^{\mathrm{T}}$ are the parameters to form the penalty terms.

ICA with multi-reference combines individual one-unit ICA with reference together and learns their weight vectors simultaneously. A Newton-like learning algorithm is extensively derived to learn the weight matrix $\mathbf{W}$:

$$\mathbf{W}_{k+1} = \mathbf{W}_k - \eta \langle \bar{\mathbf{s}}(\mathbf{W}) \rangle {\mathcal{L}_2}'_{\mathbf{W}} \mathbf{R}_{\mathbf{xx}}^{-1} \tag{21}$$

where $\bar{\mathbf{s}}(\mathbf{W})$ is a vector equals to $(\frac{1}{s_1(\mathbf{w}_1)} \cdots \frac{1}{s_l(\mathbf{w}_l)})^{\mathrm{T}}$ in which $s_i(\mathbf{w}_i) = \bar{\rho}_i E\{G''_{y_i^2}(y_i)\} - \frac{1}{2}\mu_i E\{g''_{y_i^2}(\mathbf{w}_i)\} - \lambda_i$ for $\forall i = 1, \cdots, l$ obtained from the Hessian matrix ${\mathcal{L}_2}''_{\mathbf{W}^2}$, $\langle \cdot \rangle$ represents a diagonal matrix whose off-diagonal elements are all zeros and the diagonal is given by the vector inside and the gradient ${\mathcal{L}_2}'_{\mathbf{W}}$ is given as

$${\mathcal{L}_2}'_{\mathbf{W}} = \langle \bar{\boldsymbol{\rho}} \rangle E\{G'_{\mathbf{y}}(\mathbf{y})\mathbf{x}^{\mathrm{T}}\} - \frac{1}{2} \langle \boldsymbol{\mu} \rangle E\{\mathbf{g}'_{\mathbf{y}}(\mathbf{W})\mathbf{x}^{\mathrm{T}}\} - \langle \boldsymbol{\lambda} \rangle E\{\mathbf{y}\mathbf{x}^{\mathrm{T}}\}$$

where $G'_{\mathbf{y}}(\mathbf{y})$ and $\mathbf{g}'_{\mathbf{y}}(\mathbf{W})$ are the first derivatives of $G(\mathbf{y})$ and $\mathbf{g}(\mathbf{W})$ with respect to the corresponding y_i in $\mathbf{y}$. The learning of Lagrange multipliers $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ are also based on the gradient-ascent method

$$\boldsymbol{\mu}_{k+1} = \max\{\mathbf{0}, \boldsymbol{\mu}_k + \langle \gamma \rangle \mathbf{g}(\mathbf{W}) \tag{22}$$

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \langle \gamma \rangle \mathbf{h}(\mathbf{W}) \tag{23}$$

Although the constraints define that one neuron can produce one particular IC different from others, in practice, improper values of ξ_i can cause different neurons converge to the same independent source. Instead of the exact adjustment for the threshold which is impossible with little knowledge we have about the sources, it may be desired to postprocess the weight vectors by decorrelating them in each learning iteration to prevent the different neurons estimating the same independent source [11] as follows:

$$\mathbf{W} = [\mathbf{W}\mathbf{R}_{\mathbf{xx}}\mathbf{W}^{\mathrm{T}}]^{-\frac{1}{2}} \mathbf{W} \tag{24}$$

where the inverse square root $[\mathbf{W}\mathbf{R}_{\mathbf{xx}}\mathbf{W}^{\mathrm{T}}]^{-\frac{1}{2}}$ is obtained from the eigenvalue decomposition of $\mathbf{W}\mathbf{R}_{\mathbf{xx}}\mathbf{W}^{\mathrm{T}} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$ as $[\mathbf{W}\mathbf{R}_{\mathbf{xx}}\mathbf{W}^{\mathrm{T}}]^{-\frac{1}{2}} = \mathbf{U}\mathbf{D}^{-\frac{1}{2}}\mathbf{U}$ with the simple calculation of $\mathbf{D}^{-\frac{1}{2}}$ [3]. The decorrelation process in Eq. (24) helps the global convergence achieved in this multi-unit network when each unit reaches its global optimum.

The selection of the closeness measure depends on what form the reference signals are available. A common measure of closeness between the estimated output and the reference is the mean square error (MSE) given by $\varepsilon(y_i, r_i) = E\{(y_i - r_i)^2\}$. This measure requires both y_i and r_i normalized to have zero-means and unit-variances. Alternatively, correlation can also be used as a closeness measure: $\varepsilon(y_i, r_i) = -E\{y_i r_i\}$. Both the output and reference are normalized so that the value of correlation is bounded. The proper choice of closeness function helps us easily choose a threshold ξ_i and make the algorithm more robust and globally convergent.

4. EXPERIMENTS

We demonstrate our technique with experiments using synthetic data and compare the accuracy and effectiveness of the algorithm with the second-order method and one-unit ICA. The accuracy of the extracted ICs was measured with

Desired Output	c_2	c_3	c_4	c_5
SNR(dB)	27.75	33.18	23.65	32.92
PI	0.07	0.05	0.10	0.02

Table 1. Signal-to-noise ratio (SNR) of the output and the performance indices (PI) of the network in individually extracting the desired sources, c_2, c_3, c_4 and c_5, by using the present algorithm.

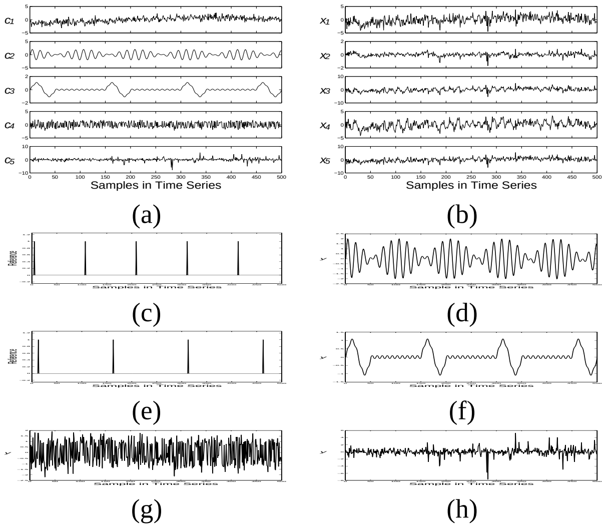
the signal-to-noise ratio (SNR) in dB, given by $\mathrm{SNR} = 10\log_{10}\left(\frac{\sigma^2}{\mathrm{mse}}\right)$ where σ^2 denotes the variances of the signal and mse denotes the mean square error between the original and extracted signals [6]. The performance of the network was measured by a performance index (PI): $\mathrm{PI}_i = \sum_{j=1}^{n} \frac{|p_{ij}|}{\max_k |p_{ik}|} - 1$, where p_{ij} is the jth element of the vector $\mathbf{p}_i = \mathbf{A}^{\mathrm{T}}\mathbf{w}_i$ [6].

4.1. Mixtures of Random and Deterministic Signals

Five zero-mean and unit-variance independent sources: a Gaussian noise signal, c_1, two periodic deterministic signals, c_2 and c_3, and two random signals, sub-Gaussian c_4 and super-Gaussian c_5, were randomly mixed to obtain five mixtures. The reference for the random signal was simulated by applying an operation $\mathrm{sign}(\cdot)$ to roughly gave the signs of most data samples of the desired source. The reference for the deterministic signal was simulated by a series of impulses having the same period as the desired source. The algorithm of one-unit ICA with reference ran using the MSE as the closeness measure to extract each deterministic and random signal, individually, with the corresponding reference. As expected, the network converged to produce the output signals identical to the desired sources in all cases. Table 1 shows the results of extracting the desired sources by using the present technique: the high SNRs and low PIs indicate good performance of the algorithm. The output waveforms are displayed in Fig. 2.

The results of the second-order method with the same experiment settings were compared. The low SNRs with average value of 3.3 dB and poor PIs with average of 1.0 indicated the failure of this method to separate the desired sources. Also, the previously proposed one-unit ICA algorithm was also run with these mixtures and always produced a signal identical to c_5 irrespective of the reference because the super-Gaussian source c_5 had the maximum negentropy among all sources.

The present algorithm was also used to extract the Gaussian noise with a signed reference closest to c_1. The trained network produced an output identical to the Gaussian signal c_1 with SNR 10.88dB and PI 0.37. The classical ICA algorithms are unable to deal with Gaussian signals because such a signal does not have statistical properties higher than 2nd-order. With a trace of the source, our algorithm can extract even Gaussian signals as the closeness between the reference and signal is measured in 2nd-order statistics.

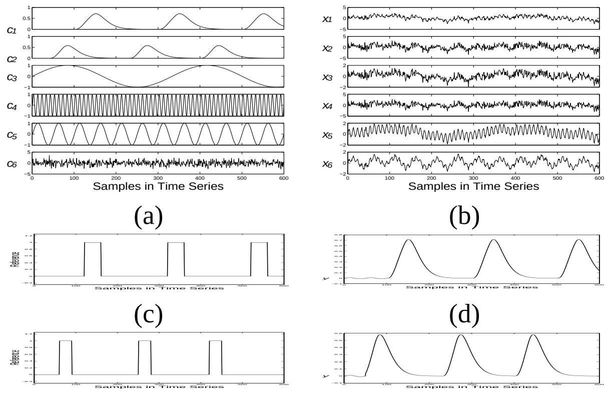


(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 2. Signals of sources, mixture inputs, outputs and references in the experiments to extract the desired signals c_2, c_3, c_4 and c_5, individually: (a) five independent sources: Gaussian noise (c_1), periodic deterministic signals (c_2 and c_3), random signals (c_4 and c_5), (b) the mixture inputs, (c) and (d) the reference and the output for extracting the desired signal c_2, respectively, (e) and (f) the reference and the output for extracting the desired signal c_3, respectively, (g) and (h) the outputs extracted by using the references of c_4 and c_5, respectively.

4.2. Synthetic fMRI Time-Series Data

Multiple input stimuli have often been used in fMRI experiments [10]. FMRI time responses from activated brain voxels are always confounded by physiological signals such as cardiac, respiratory, and blood flow, and the electronic noise of the scanners [10]. To simulate this situation, two fMRI time responses from activated voxels $c_1(t) = r_1(t) * h(t)$ and $c_2(t) = r_2(t) * h(t)$, where the input stimulus functions $r_1(t)$ and $r_2(t)$ convolved with a gamma hemodynamic response function $h(t)$, three sinusoid functions c_3, c_4 and c_5 with frequencies 0.03Hz, 1 Hz and 0.2 Hz to represent periodic activations of blood flow, cardiac and respiratory interferences, respectively, and a random Gaussian noise c_6 were generated. They were mixed to mimic the time series generated in an fMRI experiment and to generate six inputs to our algorithm. The function $r_1(t)$ and $r_2(t)$ with On-Off scheme were used as the references for two outputs, respectively. This experiment used the correlation to measure the closeness between the reference and the sources. The algorithm converged in 10 iterations with PI of 0.052. The outputs gave two signals very close to the original fMRI time responses c_1 and c_2 with SNR of 28.74dB and 30.82dB, respectively. The extracted waveforms are shown in Fig. 3. Normal one-unit ICA algorithm was not able to separate either of two fMRI response signals in this experiment: instead, it produced the output signals identical to sinusoid



(a) (b)

(c) (d)

(e) (f)

Fig. 3. Waveforms of sources, mixture inputs, the reference and the output in the simulation of extracting multiple fMRI time-domain response signals. (a) FMRI activation responses, c_1 and c_2, three sinusoid sources, c_3, c_4 and c_5, and a Gaussian noise c_6, (b) six mixtures, (c) the input stimulation-1 as the reference to extract the activation signal c_1, (d) the extracted activation identical to c_1, (e) the input stimulation-2 as the reference to extract the activation signal c_2 and (f) the extracted activation identical to c_2.

sources c_4 or c_5 at most times because their histograms had higher negentropies than others.

5. CONCLUSIONS

This paper presented a novel algorithm to extract one or several desired independent components, in a one-step process, by using reference signals that carry some a priori information of the desired sources. The problem was formulated in the cICA framework, and then a Newton-like learning algorithm was derived and its robustness was analyzed. The algorithm was able to extract independent periodic or non-periodic sources having any distributions, including Gaussians, close to the reference signals in real time when the closeness measures were properly chosen and the parameters were properly adjusted.

The experiments demonstrated the advantages and superiority of our algorithm compared to earlier methods. The second-order methods, which use only 2nd-order statistics, failed to extract the independent sources close to the reference. The present technique accurately extracted the independent sources with high SNR and low PI because the negentropy, a higher-order statistical property, was used as the contrast function. The source extracted by previous one-unit ICA was always determined by the negentropy. The present algorithm can simultaneously extract several desired sources if the available information can be introduced as reference signals. An extensive application of the present algorithm for fMRI data analysis is presented elsewhere [12].

6. REFERENCES

[1] A. D. Back, “A first application of independent component analysis to extracting structure from stock returns,” *Neural Systems*, vol. 8, no. 4, pp. 473–484, 1997.

[2] J.S. Goldstein, J.R. Guerci, and I.S. Reed, “An optimal generalized theory of signal representation,” in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 1999, vol. 3, pp. 1357–1360.

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

[4] M. Zibulevsky and Y. Zeevi, “Extraction of single source from multichannel data using sparse decomposition,” *Technical Report*, 2001.

[5] J. Luo, B. Hu, X. T. Ling, and R. W. Liu, “Principal independent component analysis,” *IEEE Trans. on Neural Networks*, vol. 10, no. 4, pp. 912–917, 1999.

[6] W. Lu and J. C. Rajapakse, “Constrained independent component analysis,” in *Advances in Neural Information Processing Systems 13 (NIPS2000)*, MIT Press, 2000, pp. 570–576.

[7] P. Comon, “Independent component analysis: A new concept?,” *Signal Processing*, vol. 36, pp. 287–314, 1994.

[8] M. Girolami and C. Gyfe, “Extraction of independent signal sources using a deflationary exploratory projection pursuit network with lateral inhibition,” *Vision, Image and Signal Processing, IEE Proceedings*, vol. 144, no. 5, pp. 299–306, October 1997.

[9] A. Hyvärinen, “New approximations of differential entropy for independent component analysis and projection pursuit,” in *Advances in Neural Information Processing Systems 10 (NIPS*97)*, 1998, pp. 273–279.

[10] 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.

[11] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo, “A class of neural networks for independent component analysis,” *IEEE Transactions on Neural Networks*, vol. 8, no. 3, pp. 487–504, 1997.

[12] J. C. Rajapakse and W. Lu, “Extracting task-related components in functional MRI,” *Submitted to ICA 2001*, 2001.