FEATURE-SELECTIVE ICA AND ITS CONVERGENCE PROPERTIES*

Yi-Ou Li and Tülay Adalı

Department of CSEE
University of Maryland Baltimore County
Baltimore, MD 21250

Vince D. Calhoun

Olin Neuropsychiatry Research Center
Institute of Living, Hartford, CT 06106
School of Medicine, Yale University
New Haven, CT 06520

ABSTRACT

We present a projection-based framework for a feature-selective independent component analysis (FS-ICA) scheme and study its convergence property for two ICA algorithms, FastICA and Infomax. As examples, we implement bandpass filter as the feature-selective filter to improve the estimation of a bandpass signal from the mixtures and a periodic task-related time course embedded in the functional Magnetic Resonance Imaging (fMRI) data. Hence, we demonstrate that the proposed method can incorporate $a\ priori$ information into ICA to effectively improve estimation of the underlying components of practical interest, such as periodic time courses and smooth brain activation areas in fMRI data.

1. INTRODUCTION

ICA is a technique to estimate statistically independent components from their linear mixtures. Most ICA algorithms are derived by forming a linear demixing model, defining a measure of statistical independence and performing numerical optimization of the independence measure based on the given observations. In this framework, each component is treated as a random variable and the independence measure used by ICA algorithm is a statistical measure such as the higher order statistics, negentropy and mutual information. However, in most application scenarios, the contextual information is encoded in the sample space of the components, in other words, the order by which the samples are indexed depends on the nature of the underlying components. For example, a noise signal with a Laplace distribution is by no means the same as a speech signal that has the same distribution in amplitude although they are identical so far as the statistical measures are concerned. A speech signal is a bandpass signal with slowly changing envelop while the noise signal does not assume any temporal pattern.

*THIS RESEARCH IS SUPPORTED IN PART BY THE NIH UNDER GRANT 1 R01 EB 000840-02.

Methods using the temporal structure to achieve blind source separation of signals have been developed in, $e.g.$, [1, 2]. However, these methods can not estimate components with identical autocovariances [3]. It is desirable to use the contextual information in the components together with their statistical independence to perform ICA estimation. The challenge here is to find a way to combine the features in sample space with the ICA learning algorithm that is based on a statistical measure. One possibility is that, in the ICA estimation process, a feature-selective filtering [4] is imposed on the sequence of component estimates and the filtering effect is projected onto the demixing vectors. The projection is achieved by obtaining linear estimator of the demixing vector based on the filtered component estimate and the ICA model. This projection transfers the feature-enhanced variation in the sample space to the space of the demixing vectors where the numerical optimization is carried out. This feature-selective projection during the sequential ICA estimation process hence biases the estimates to the components whose features match the characteristics of the feature-selective filter. Detailed discussion on this effect is given in the following sections.

In the next section, we outline the method of feature-selective projection within a general ICA framework. Section 3 studies its convergence behavior for two widely used ICA algorithms, FastICA [5] and Infomax [6]. In Section 4, we show simulation results of this projection based FS-ICA in separation of independent waveforms with different temporal patterns and temporal ICA of fMRI data [7]. We conclude our work with a discussion about this scheme.

2. FEATURE-SELECTIVE PROJECTION IN ICA

The generative model of ICA can be stated as $\mathbf{x}[t] = \mathbf{A}\mathbf{s}[t]$ where $\mathbf{s}[t]$ is a vector containing the components with sample data indexed by t, $\mathbf{A}$ is a nonsingular mixing matrix and $\mathbf{x}[t]$ is the resulting observations. The estimation is most typically performed in batch mode on all the observed samples. Therefore, we can drop the index t and expand each

component as a row vector of data arranged in its natural order, $e.g.$, time sequence, image contrast from sequentially located pixels etc. Accordingly, the generative model becomes: $\mathbf{X} = \mathbf{AS}$ where $\mathbf{X}$ and $\mathbf{S}$ are matrices whose rows contain the data from each observation and component respectively. The task of ICA is then to find a demixing matrix $\mathbf{W}$ such that the original components are recovered as $\hat{\mathbf{S}} = \mathbf{WX}$.

Within one iteration of ICA algorithm, the feature-selective projection can be carried out in the following procedure [4]:
(i) Restoration: $\hat{\mathbf{s}}(k) = \mathbf{X}^T\mathbf{w}(k)$ where k is the iteration index of the ICA algorithm;
(ii) Filtering: $\hat{\mathbf{s}}'(k) = \mathbf{H}\hat{\mathbf{s}}(k)$ where filtering is expressed as premultiplication of the signal vector with the convolution matrix $\mathbf{H}$ defined by the feature-selective filter;
(iii) Projection: $\mathbf{w}'(k) = (\mathbf{XX}^T)^{-1}\mathbf{X}\hat{\mathbf{s}}'(k)$, $i.e.$, the least squares solution of $\min_{\mathbf{w}}\|\hat{\mathbf{s}}'(k) - \mathbf{X}^T\mathbf{w}\|^2$.

When the observations are prewhitened, $i.e.$, $\mathbf{XX}^T = n\mathbf{I}$ where n is the dimension of the sample space and $\mathbf{I}$ is the identity matrix, we can rewrite the expression of $\mathbf{w}'(k)$ in (iii) as

$$\mathbf{w}'(k) = (\mathbf{XX}^T)^{-1}\mathbf{ASHS}^T\mathbf{A}^T\mathbf{w}(k) = \mathbf{A}\Sigma_{ss'}\mathbf{A}^T\mathbf{w}(k) \tag{1}$$

where we use the ICA generative model $\mathbf{X} = \mathbf{AS}$ and define $\Sigma_{ss'} \equiv \frac{1}{n}\mathbf{SHS}^T = \frac{1}{n}\mathbf{SS'}^T$ as the sample correlation matrix of the original components and the filtered ones, $i.e.$, the feature correlation matrix.

Since $\mathbf{A}$ is assumed to be a nonsingular matrix in the ICA model, we can premultiply $\mathbf{A}^{-1}$ to both sides of (1) to obtain:

$$\mathbf{A}^{-1}\mathbf{w}'(k) = \Sigma_{ss'}\mathbf{A}^T\mathbf{w}(k). \tag{2}$$

Also by the whitening condition, we have

$$\mathbf{XX}^T = \mathbf{ASS}^T\mathbf{A}^T = n\mathbf{A}\Sigma_s\mathbf{A}^T = n\mathbf{I} \tag{3}$$

where $\Sigma_s \equiv \frac{1}{n}\mathbf{SS}^T$ is the sample correlation matrix of the true components. Because of the independence assumption, Σ_s is a diagonal matrix with the sample correlation $r_{s_i} = \frac{1}{n}\mathbf{s}_i{}^T\mathbf{s}_i$ on its main diagonal.

From the last equality in (3), we have $\mathbf{A}^T = \Sigma_s^{-1}\mathbf{A}^{-1}$. Substituting $\mathbf{A}^T$ into (2), we obtain

$$\mathbf{A}^{-1}\mathbf{w}'(k) = \overline{\Sigma}_{ss'}\mathbf{A}^{-1}\mathbf{w}(k) \tag{4}$$

where $\overline{\Sigma}_{ss'} \equiv \Sigma_{ss'}\Sigma_s^{-1}$. Assuming that the feature-selective filtering on each individual component does not introduce correlation across different components and the filtering gain is 1, $\Sigma_{ss'}$ is a diagonal matrix with the feature correlation factors $r_{s_is'_i} = \frac{1}{n}\mathbf{s}_i{}^T\mathbf{s}'_i, i = 1, 2, ..., m$ on its main diagonal. From here and onwards, m represents the dimension of the demixing vectors. Therefore, $\overline{\Sigma}_{ss'}$ is a diagonal matrix with the $normalized$ feature correlation factor of each component $r_{s_is'_i}/r_{s_i} \in [0, 1]$ on its main diagonal. Now we define a new vector $\mathbf{z}(k) \equiv \mathbf{A}^{-1}\mathbf{w}(k)$ and rewrite (4) as

$$\mathbf{z}'(k) = \overline{\Sigma}_{ss'}\mathbf{z}(k). \tag{5}$$

Hence, the feature-selective projection is equivalent to premultiplying the transformed demixing vector $\mathbf{z}(k)$ with the normalized feature correlation matrix $\overline{\Sigma}_{ss'}$ within each iteration of the ICA algorithm.

3. CONVERGENCE BEHAVIOR OF PROJECTION BASED FS-ICA

3.1. FastICA with feature-selective projection

An important principle of FastICA algorithm is that all the demixing vectors $\mathbf{w}_i$ are kept orthogonal and of unit norm during the updates, $i.e.$, the demixing matrix $\mathbf{W}$ satisfies $\mathbf{WW}^T = \mathbf{I}$. This indicates that $\mathbf{W}^T = \mathbf{W}^{-1} = \mathbf{A}$. In this case, the whitening condition in (3) implies $\Sigma_s = \frac{1}{n}\mathbf{WXX}^T\mathbf{W}^T = \mathbf{I}$ and equation (5) becomes

$$\mathbf{z}'(k) = \Sigma_{ss'}\mathbf{z}(k).$$

Hence, the effect of feature-selective projection is weighting each element z_i in the transformed demixing vector $\mathbf{z}$ by the feature correlation factor $r_{s_is'_i}$. Hereon, the iteration index k is dropped for simple expression.

Under this setting, Hyvarinen [5] shows that the convergence criterion for the estimation of one component is that $\mathbf{z}$ converges to a unit vector with only one nonzero element. Suppose we estimate the component s_j whose transformed demixing vector is $\mathbf{z} = [z_1, z_2, ..., z_m]^T$. When s_j assumes the specified feature, we have $r_{s_is'_i} < r_{s_js'_j} \approx 1, \forall i \neq j$, and this correlation factor will accelerate the decrease of $z_i, \forall i \neq j$ to zero. Therefore, the convergence of each z_i is enhanced by the ratio $r_{s_js'_j}/r_{s_is'_i}$.

3.2. Infomax with feature-selective projection

A substantial difference between the Infomax and the FastICA algorithms is that Infomax optimizes a measure of ensemble independence among all the components. Therefore, it estimates all the independent components simultaneously. This makes the analysis on the estimation of one particular component less straightforward. However, we can study the extreme case such that the feature-selective projection is applied to the entire demixing matrix. In this case, (5) becomes

$$\mathbf{Z}'(k) = \mathbf{Z}(k)\overline{\Sigma}_{ss'} \tag{6}$$

where $\mathbf{Z}$ is the transformed demixing matrix with each vector $\mathbf{z}$ as its column.

For an Infomax-type updating rule with the $tanh$ nonlinearity, Amari $et\ al.$ [8] shows that the demixing matrix

asymptotically converges to $\mathbf{W}^o = \mathbf{A}^{-1}$. Hence, $\mathbf{Z}$ converges to

$$\mathbf{Z}^o = \mathbf{W}^o \mathbf{A}^{-T} = \mathbf{A}^{-1}\mathbf{A}^{-T} = \Sigma_s$$

where we use the relation $\mathbf{A}^{-T} = \mathbf{A}\Sigma_s$ from (3). In other words, each row of $\mathbf{Z}$ converges to a vector with only one nonzero element, r_{s_i}, corresponding to component s_i. Now, we plug in the feature-selective projection as defined in (6) and assume that component s_j has the specified feature, then we have $r_{s_i s_i'}/r_{s_i} < r_{s_j s_j'}/r_{s_j} \approx 1, \forall i \neq j$ and

$$\mathbf{Z}'(k) = \mathbf{Z}(k)\ \mathrm{diag}[r_{s_1 s_1'}/r_{s_1}, ..., r_{s_m s_m'}/r_{s_m}].$$

Similar to the observation in the FastICA case, the convergence of the transformed demixing vector $\mathbf{z}$ for s_j is improved by the weighting matrix $\overline{\Sigma}_{ss'}$. For other components, this weighting matrix is not going to help with the convergence, therefore, the feature selective projection has to be applied selectively on certain demixing vectors and in a controlled manner [4] to preserve the general performance of the ICA algorithm.

4. SIMULATIONS

4.1. Temporal waveform separation experiment

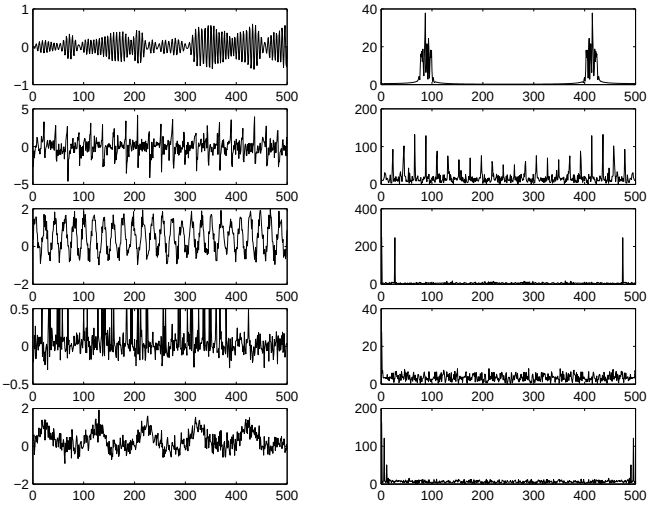

Fig. 1. Simulated independent waveforms (Left) and their spectra (Right)

Figure 1 shows five waveforms (From top to bottom: $s_1, s_2, ..., s_5$) we create as the independent source signals and their frequency spectra. Five hundred samples are generated for each signal with a common sampling frequency f_s. We select s_1 as the signal of interest and design a bandpass FIR filter with a passband from $0.3f_s$ to $0.4f_s$ as the feature-selective filter to match its bandlimited characteristics. The kurtosis value $\tilde{\kappa}$ and the bandpass feature correlation $r_{ss'}$ of each signal are listed in Table 1.

Table 1. Statistics of the components

	s_1	s_2	s_3	s_4	s_5
$\tilde{\kappa}$	0.51	0.95	-1.20	1.97	-0.54
$r_{ss'}$	0.99	0.49	0.16	0.48	0.26

Table 2. Result of waveform separation

	ICA		FS-ICA	
	Infomax	FastICA	Infomax	FastICA
$r_{s_1\hat{s}_1}$	0.82±0.03	0.85±0.07	1.00±0.00	0.99±0.01
$r_{s_2\hat{s}_2}$	0.89±0.05	0.91±0.09	0.98±0.01	0.97±0.02
$r_{s_3\hat{s}_3}$	0.69±0.02	0.98±0.02	0.70±0.01	0.99±0.01
$r_{s_4\hat{s}_4}$	0.99±0.00	0.97±0.03	0.99±0.00	0.98±0.02
$r_{s_5\hat{s}_5}$	0.71±0.04	0.84±0.15	0.73±0.01	0.98±0.02
n_1	-	12±8	-	6±3
n	54±5	71±28	100±1	74±46

The estimation results are shown in Table 2 and the correlation between the true and the estimated signals, $r_{s_i\hat{s}_i}$, is calculated for the evaluation of performance. The result shows that the ICA algorithms incorporated with the feature-selective projection lead to better estimation of s_1. Since the feature-selective filtering introduces a perturbation in the ICA iterations, the total step count n for convergence of the FS-ICA algorithms increases. However, when the convergence of each individual component can be observed, $e.g.$, in FastICA, we see from the step count n_1 that the incorporation of the feature-selective projection accelerates the estimation of s_1. This observation verifies the conclusion in the previous section.

4.2. Temporal ICA of fMRI data

The temporal ICA model for fMRI data analysis takes the different time courses, $e.g.$, task-related, physiology-related, as the independent components and assumes that the time courses are mixed spatially by the brain activation maps. The fMRI data we used for this example is acquired from a visual activation experiment where the visual stimulus is applied periodically [9]. This indicates that the corresponding visual task-related time course assumes similar temporal periodicity. Specifically, the frequency characteristics of the feature-selective filter can be obtained by multiplying the spectrum of the experimental paradigm with the Fourier transform of the hemodynamic response function [10]. Therefore, we impose this periodicity feature on the estimate of the visual task-related time course and incorporate the cor-

Table 3. TICA result of fMRI data

	ICA		FS-ICA	
	Infomax	FastICA	Infomax	FastICA
$r_{t\hat{t}}$	0.63±0.01	0.73±0.03	0.74±0.04	0.74±0.03
n_1	-	29±11	-	28±13
n	111±10	75±27	132±29	68±21

responding feature-selective projection in the ICA estimation process.

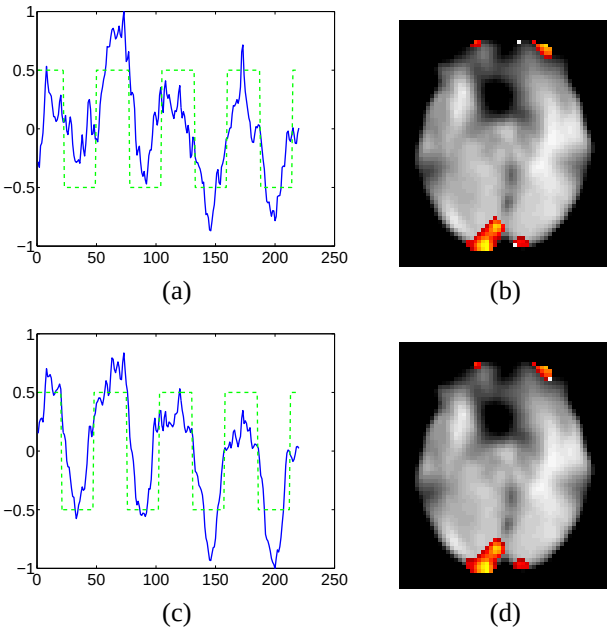

Fig. 2. Estimated visual task-related time courses and the corresponding activation maps of Infomax (a, b) and Infomax with feature-selective projection (c, d)

The correlation between the experimental paradigm and the estimated visual task-related time course, $r_{t\hat{t}}$, is computed and listed in Table 3. Figure 2 shows the estimated time courses and the associated activation maps (Z-scored with $Z > 2.5$). From the results presented we see that when feature-selective projection is incorporated, the estimated visual task-related time course assumes higher correlation with the experimental paradigm and is less noisy compared to that of the original ICA algorithm.

5. DISCUSSION

In this work, we introduce a projection based FS-ICA scheme and study its convergence properties in the context of two ICA algorithms. The method uses controlled feature-selective filtering on the sequential component estimates and projection by the least squares estimator. The design of the feature seletive filter is based on the desired features of the component of interest. Although in our simulations the feature-selective filtering is carried out in the time domain, it is actually defined in a general sample space such as spatial domain for image processing or frequency domain in spectrum analysis.

6. REFERENCES

[1] L. Tong, R.-W. Liu, V. C. Soon, and Y.-F. Huang, "Indeterminacy and identifiability of blind identification," *IEEE Trans. on Circuits and Systems*, vol. 38, pp. 499–509, 1991.

[2] A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique on second order statistics," *IEEE Trans. on Signal Processing*, vol. 45, 1997.

[3] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja, *Independent Component Analysis*, John Wiley & Sons, INC., 2001.

[4] Y.-O. Li, T. Adalı, and V. D. Calhoun, "Independent component analysis with feature selective filtering," in *Proc. MLSP 2004*, Sao Luis, Brazil.

[5] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *Neural Networks*, vol. 10, pp. 626–634, 1999.

[6] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1004–1034, 1995.

[7] V. D. Calhoun, T. Adalı, L. K. Hansen, J. Larsen, and J. J. Pekar, "ICA of functional MRI data An overview," in *Proc. ICA 2003*, Nara, Japan.

[8] S.-I. Amari, T.-P. Chen, and A. Cichocki, "Stability analysis of adaptive blind source separation," *Neural Networks*, vol. 10, no. 8, pp. 1345–1351, 1997.

[9] V. D. Calhoun, T. Adalı, G. D. Pearlson, P. C. M. Van Zijl, and J. J. Pekar, "Independent components analysis of fMRI data in the complex domain," *Magnetic Resonance in Medicine*, vol. 48, pp. 180–192, 2002.

[10] Vince D. Calhoun, *Independent Component Analysis for Functional Magnetic Resonance Imaging*, Ph.D. thesis, University of Maryland Baltimore County, Baltimore, MD, 2002.