Second-Order Approximation of Minimum Discrimination Information in Independent Component Analysis

YunPeng Li

***Abstract*—Independent Component Analysis (ICA) is intended to recover the mutually independent sources from their linear mixtures, and $FastICA$ is one of the most successful ICA algorithms. Although it seems reasonable to improve the performance of $FastICA$ by introducing more nonlinear functions to the negentropy estimation, the original fixed-point method (approximate Newton method) in $FastICA$ degenerates under this circumstance. To alleviate this problem, we propose a novel method based on the second-order approximation of minimum discrimination information (MDI). The joint maximization in our method is consisted of minimizing single weighted least squares and seeking unmixing matrix by the fixed-point method. Experimental results validate its efficiency compared with other popular ICA algorithms.**

***Index Terms*—Independent component analysis, minimum discrimination information, second-order approximation, FastICA, weighted least squares.**

I. Introduction

Independent Component Analysis (ICA) has been widely used in diverse fields, such as machine learning, signal processing, and stats. Given the m dimensional observed mixtures $\mathbf{x} = (x_1, \ldots, x_m)^T$, independent component analysis models them as the linear combination of m independent sources $\mathbf{s} = (s_1, \ldots, s_m)^T$,

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

where the mixing matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$. Given N independent identically distributed samples of $\mathbf{x}$, the goal of ICA is to recover the unknown sources $\mathbf{s}$ and estimate the mixing matrix $\mathbf{A}$. We model the recovering process as,

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

where the sources' estimation $\mathbf{y} = (y_1, \ldots, y_m)^T$ is the scaling and permutation of sources $\mathbf{s}$, and $\mathbf{W} \in \mathbb{R}^{m\times m}$ is called the unmixing matrix. In general, one assumes that $\mathrm{E}(\mathbf{s}) = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{s}) = \mathbf{I}$. It has been shown that $\mathbf{W}$ is identifiable up to scaling and permutation of its rows if at most one s_i is Gaussian [1]. Since there often exists centering and whitening preprocessing stages for the observation $\mathbf{x}$, $\mathbf{W}$ is restricted to be an orthonormal matrix $\mathbf{W}\mathbf{W}^T = \mathbf{I}$.

Many approaches have been proposed for ICA in the past researches [2], [3], and the most two popular types of ICA algorithms seem to be the maximum likelihood estimation and the contrast function approaches. Parametric maximum likelihood estimation (density matching) [4]–[7] is used to infer the parametric model in ICA by specifying the distributions for the component s_i. Unfortunately, the performances of these parametric methods are highly dependent on the prior assumptions on the unknown sources. To keep the components of $\mathbf{s}$ unspecified, several nonparametric ICA [8]–[11] are proposed at the cost of high computation burden or the difficult selection of tuning parameters. In the contrast function approaches, several criteria are chosen to represent the measure of independence or non-Gaussianity, for example, the mutual information [1], [5], the nonlinear decorrelation [12], [13], higher-order moments [14], [15], and the entropy [16]–[18]. For other recent approaches, see also [19]–[23].

$FastICA$ [16] is one of the most successful ICA algorithms, whose contrast function (approximation of negentropy [24]) is defined as the expectation of a single nonlinear function. $FastICA$ enjoys low computation and fast convergence due to the efficient fixed-point method [16], which is equivalent to the approximate Newton method without calculating the inverse of the Hessian matrix. Unfortunately, $FastICA$ usually fails when there is a great mismatching between its single nonlinear function and the unknown sources' distributions. Although we can introduce more nonlinear functions to improve the negentropy estimation (suggested by research [24]), the original fixed-point method in $FastICA$ fails (Hessian matrix can not be approximated to a diagonal matrix).

In this paper, we present a novel ICA algorithm $MDIICA$ based on the second-order approximation of minimum discrimination information (MDI) [25]. Although conceptually, our work seems similar to the approximation of negentropy used in $FastICA$, they are quite different in derivations. In addition, the fixed-point method can be directly applied to our method at the negligible cost of computation, when more nonlinear functions are required to improve the negentropy estimation. In Section II, we explain the difficulties in $FastICA$, when its contrast function is composed of several nonlinear functions. Section III presents the derivations of our novel ICA algorithm.

Manuscript received September 28, 2021; revised November 29, 2021; accepted December 2, 2021. Date of publication December 14, 2021; date of current version January 28, 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xun Chen.

The author is with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: liyp18@tsinghua.org.cn).

Digital Object Identifier 10.1109/LSP.2021.3135193

In Section IV, we compare our method with other known algorithms in both simulation and real data experiment. We conclude our contributions in Section V.

II. Difficulties in $FastICA$

In this section, we firstly review the contrast function (expectation of a single nonlinear function) and the fixed-point method used in $FastICA$, then we reveal the difficulties when more nonlinear functions are used.

To separate the sources from their mixtures, $FastICA$ maximizes the approximation of negentropy [24] $J(\mathbf{w})$

$$\max_{\mathbf{w}} \quad J(\mathbf{w}) = \frac{1}{2}\sum_{k=1}^{p} \mathrm{E}\{G_k(\mathbf{w}^T\mathbf{x})\}^2 \quad \text{s.t.} \quad \mathbf{w}^T\mathbf{w} = 1 \tag{3}$$

where $\mathbf{w}$ is the row of the unmixing matrix $\mathbf{W}$, $\{G_k(.)\}_{k=1}^p$ are p nonlinear functions. These nonlinear functions should satisfy the following constraints [24],

$$\int \phi(x)G_i(x)G_j(x)dx = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{4}$$

$$\int \phi(x)G_i(x)x^k dx = 0 \quad k = 0, 1, 2 \tag{5}$$

where $\phi(.)$ is the standard Gaussian function. In practice, one can take any set of linearly independent functions $\{\overline{G}_k(.)\}_{k=1}^p$, and apply Gram-Schmidt orthonormalization on the selected set to satisfy the assumptions above [24]. Fortunately, such computation can be simplified if a single nonlinear function is used. When $p = 1$ in $FastICA$, the maximal of $J(\mathbf{w}) = \frac{1}{2}\mathrm{E}\{G_1(\mathbf{w}^T\mathbf{x})\}^2$ are obtained at certain optima of $\mathrm{E}\{G_1(\mathbf{w}^T\mathbf{x})\}$, the equivalent optimization problem in $FastICA$ becomes,

$$\max_{\mathbf{w},\lambda} / \min_{\mathbf{w},\lambda} \quad \mathcal{L}(\mathbf{w}, \lambda) = \mathrm{E}\{G_1(\mathbf{w}^T\mathbf{x})\} - \lambda(\mathbf{w}^T\mathbf{w} - 1) \tag{6}$$

where λ is the Lagrange multiplier, and the gradient $\nabla\mathcal{L}(\mathbf{w}, \lambda)$ is

$$\nabla\mathcal{L}(\mathbf{w}, \lambda) = \mathrm{E}\{G_1'(\mathbf{w}^T\mathbf{x})\mathbf{x}\} - 2\lambda\mathbf{w} = \mathbf{0} \tag{7}$$

It is easy to find that $\lambda = \frac{1}{2}\mathrm{E}\{G_1'(\mathbf{w}^T\mathbf{x})\mathbf{w}^T\mathbf{x}\}$, and the Hessian matrix $\nabla^2\mathcal{L}(\mathbf{w}, \lambda)$ is

$$\begin{aligned} \nabla^2\mathcal{L}(\mathbf{w}, \lambda) &= \mathrm{E}\{G_1''(\mathbf{w}^T\mathbf{x})\mathbf{x}\mathbf{x}^T\} - 2\lambda\mathbf{I} \\ &\approx \mathrm{E}\{G_1''(\mathbf{w}^T\mathbf{x})\}\mathrm{E}\{\mathbf{x}\mathbf{x}^T\} - 2\lambda\mathbf{I} \\ &\approx \mathrm{E}\{G_1''(\mathbf{w}^T\mathbf{x})\}\mathbf{I} - 2\lambda\mathbf{I} \end{aligned} \tag{8}$$

Thanks to the whitening stage ($\mathrm{E}\{\mathbf{x}\mathbf{x}^T\} = \mathbf{I}$) on the observation $\mathbf{x}$, the approximation in (8) is available. Thus, the Hessian matrix is diagonal and easy to be inverted, and the fixed-point method (approximate Newton method) can be directly applied in $FastICA$.

The concise of $FastICA$ is due in large part to the selection of a single nonlinear function. Unfortunately, these advantages are diminished, when more nonlinear functions are required. To use more nonlinear functions, the Gram-Schmidt orthonormalization in (4) can not be omitted anymore, and the original fixed-point method fails in $FastICA$. The equivalent optimization problem concerning (3) develops into the following form

$$\max_{\mathbf{w},\lambda} \quad \mathcal{L}(\mathbf{w}, \lambda) = \frac{1}{2}\sum_{k=1}^{p} \mathrm{E}\{G_k(\mathbf{w}^T\mathbf{x})\}^2 - \lambda(\mathbf{w}^T\mathbf{w} - 1) \tag{9}$$

the gradient $\nabla\mathcal{L}(\mathbf{w}, \lambda)$ becomes

$$\nabla\mathcal{L}(\mathbf{w}, \lambda) = \sum_{k=1}^{p} \mathrm{E}\{G_k(\mathbf{w}^T\mathbf{x})\}\mathrm{E}\{G_k'(\mathbf{w}^T\mathbf{x})\mathbf{x}\} - 2\lambda\mathbf{w} = \mathbf{0} \tag{10}$$

where $\lambda = \frac{1}{2}\sum_{k=1}^{p} \mathrm{E}\{G_k(\mathbf{w}^T\mathbf{x})\}\mathrm{E}\{G_k'(\mathbf{w}^T\mathbf{x})\mathbf{w}^T\mathbf{x}\}$, and the structure of Hessian matrix $\nabla^2\mathcal{L}(\mathbf{w}, \lambda)$ becomes complex

$$\begin{aligned} \nabla^2\mathcal{L}(\mathbf{w}, \lambda) = &\sum_{k=1}^{p} \mathrm{E}\{G_k'(\mathbf{w}^T\mathbf{x})\mathbf{x}\}\mathrm{E}\{G_k'(\mathbf{w}^T\mathbf{x})\mathbf{x}^T\} \\ &+ \sum_{k=1}^{p} \mathrm{E}\{G_k(\mathbf{w}^T\mathbf{x})\}\mathrm{E}\{G_k''(\mathbf{w}^T\mathbf{x})\mathbf{x}\mathbf{x}^T\} - 2\lambda\mathbf{I} \end{aligned} \tag{11}$$

When more nonlinear functions are required to improve the negentropy estimation ($p > 1$), the inversion of Hessian matrix cannot be simplified compared with the diagonal approximation in (8), and the efficient fixed-point method in $FastICA$ fails.

III. Proposed Method

Given the expectations of linearly independent functions $\overline{\mathbf{G}}(y_i) = (\overline{G}_1(y_i), \ldots, \overline{G}_p(y_i))^T$, minimum discrimination information [25] is aimed at determining the distribution $p_i(y_i)$, which is closest to the prior $p_0(y_i)$ in KL divergence.

$$\begin{aligned} \min_{p_i} \quad & KL(p_i||p_0) = \int p_i(y_i) \log \frac{p_i(y_i)}{p_0(y_i)} dy_i \\ \text{s.t.} \quad & \int p_i(y_i)G_k(y_i)dy_i = c_k \quad k = 1, 2, \ldots, p \\ & \int p_i(y_i)dy_i = 1 \end{aligned} \tag{12}$$

The prior $p_0(y_i)$ used in ICA is the standard Gaussian $\phi(y_i)$, and the solution to the above optimization is Gibbs distribution with prior,

$$p_i(y_i) = \frac{\phi(y_i)e^{f_i(y_i)}}{\int \phi(y_i)e^{f_i(y_i)}dy_i} \tag{13}$$

where $f_i(y_i) = \boldsymbol{\beta}_i^T\overline{\mathbf{G}}(y_i)$, $\boldsymbol{\beta}_i$ contains the coefficients concerning nonlinear functions. The form of $p_i(y_i)$ in (13) is similar to the exponentially tilted Gaussian (used in $ProDenICA$ [8]), and $\int \phi(y_i)e^{f_i(y_i)}dy_i$ is the partition function. We then substitute the $p_i(y_i)$ in $KL(p_i||\phi)$ to obtain the minimum discrimination information $KL^{min}(p_i||\phi)$,

$$\begin{aligned} &KL^{min}(p_i||\phi) \\ &= \int \frac{\phi(y_i)e^{f_i(y_i)}}{\int \phi(x)e^{f_i(y_i)}dy_i} \log \frac{\frac{\phi(y_i)e^{f_i(y_i)}}{\int \phi(y_i)e^{f_i(y_i)}dy_i}}{\phi(y_i)} dy_i \\ &= \int \phi(y_i)e^{f_i(y_i)}f_i(y_i)dy_i - \int \phi(y_i)e^{f_i(y_i)}dy_i + 1 \end{aligned} \tag{14}$$

The last expression in (14) can be easily proved by noticing that $KL^{min}(p_i||\phi)$'s invariance of the scale of partition function $\int \phi(y_i)e^{f_i(y_i)}dy_i$ and the maximal value (of the last expression) is obtained when $\int \phi(y_i)e^{f_i(y_i)}dy_i = 1$. We will maximize $KL^{min}(p_i||\phi)$ as the contrast function in our ICA method. In the $MDIICA$, several assumptions are likely to be true:

1) Since $KL(p_i||\phi)$ works as the measure concerning the departure from standard Gaussian, $KL^{min}(p_i||\phi)$ is a lower-bound for KL divergence and the maximization of $KL^{min}(p_i||\phi)$ may lead to the maximization of true $KL(p_i||\phi)$;
2) Owing to the definition of minimum discrimination information, it is reasonable to deem that the unknown $p_i(y_i)$ in $KL^{min}(p_i||\phi)$ is close to the standard Gaussian $\phi(y_i)$.

Although the two assumptions are partly similar to the maximum entropy used in $FastICA$ [2], [24], we will show their main differences in the rest of the section.

Maximizing the total minimum discrimination information $\sum_{i=1}^{m} KL^{min}(p_i||\phi)$ can be viewed as a joint maximization over the unmixing matrix $\mathbf{W}$ and the density distributions of sources' estimation $\mathbf{y}$, fixing one argument and maximizing over the other. The optimization problem in our ICA algorithm is

$$\max_{\{\mathbf{w}_i, p_i\}_{i=1}^{m}} \sum_{i=1}^{m} KL^{min}(p_i||\phi) \quad \text{s.t.} \quad \mathbf{W}^T\mathbf{W} = \mathbf{I} \tag{15}$$

where $\mathbf{w}_i$ is the i_{th} row in the unmixing matrix $\mathbf{W}$. The joint maximization in (15) consists of two iterative stages:

- $\max_{\{p_i\}_{i=1}^{m}} \sum_{i=1}^{m} KL^{min}(p_i||\phi)$. Fixing $\mathbf{W}$, each p_i is estimated by minimizing the weighted least squares concerning the approximation of MDI.
- $\max_{\{\mathbf{w}_i\}_{i=1}^{m}} \sum_{i=1}^{m} KL^{min}(p_i||\phi)$. Given p_i, $\mathbf{W}$ is restricted to be orthonormal and is calculated via the fixed-point method [8], [16].

Similar joint maximization has been used in past researches concerning projection pursuit [26], [27] and ICA [8], [11].

A. Second-Order Approximation of Minimum Discrimination Information

To simplify the integral in (14), we construct a grid of L (500) values y_i^{*l} with Δ step, and let the corresponding frequency q_i^{*l} be

$$q_i^{*l} = \sum_{j=1}^{N} \mathbb{I}(y_i^j \in (y_i^{*l} - \Delta/2, y_i^{*l} + \Delta/2])/N \tag{16}$$

where $\mathbb{I}(.)$ is the indicator function, and $y_i^j = \mathbf{w}_i^T\mathbf{x}_j$ ($j = 1, \ldots, N$). Thus, the original $KL^{min}(p_i||\phi)$ in (14) is converted to the following form

$$KL^{min}(p_i||\phi) = \sum_{l=1}^{L} \{q_i^{*l} f_i(y_i^{*l}) - \Delta\phi(y_i^{*l})e^{f_i(y_i^{*l})}\} + 1 \tag{17}$$

This is similar to the generative additive models [28] used in $ProDenICA$ [8], which can be solved by a sequence of iterative reweighted least squares (IRLS) [29]. However, we don't adopt that strategy in our method, we utilize the definition of minimum discrimination information to cut down the computation burden instead. Since $p_i(y_i)$ is close to the standard Gaussian $\phi(y_i)$ and the partition function $\int \phi(y_i)e^{f_i(y_i)}dy_i$ is equal to 1 at the maximal point of $KL^{min}(p_i||\phi)$, we can conclude that $f_i(y_i)$ is close to the zero. The second-order approximation of $p_i(y_i)$ is

$$p_i(y_i) = \phi(y_i)e^{f_i(y_i)} \approx \phi(y_i)(1 + f_i(y_i) + \frac{1}{2}f_i^2(y_i)) \tag{18}$$

Substituting (18) into (17), the original maximization of $KL^{min}(p_i||\phi)$ becomes equivalent to the minimization of weighted least squares below

$$\min_{f_i} \quad \sum_{l=1}^{L} \Delta\phi(y_i^{*l}) \left(f_i(y_i^{*l}) - \frac{q_i^{*l} - \Delta\phi(y_i^{*l})}{\Delta\phi(y_i^{*l})} \right)^2 \tag{19}$$

Two nonlinear functions $\overline{\mathbf{G}}(y_i) = (\overline{G}_1(y_i), \overline{G}_2(y_i))^T$ have been used in projection pursuit [30] and negentropy estimation [24], and they are appropriate to be the basis functions in most cases.

$$\overline{G}_1(y_i) = y_i e^{-\frac{y_i^2}{2}} \quad \overline{G}_2(y_i) = e^{-\frac{y_i^2}{2}} \tag{20}$$

Since the size of the nonlinear basis used in (19) is constant and we can solve the weighted least squares efficiently in linear time. Compared with $ProDenICA$ and $FastICA$, our method has successfully replaced a sequence of iterative reweighted least squares (in $ProDenICA$) [8] with a single weighted least squares, and the complex constraints in (4)(5) are avoided.

B. Fixed-Point Method

Given the fixed $p_i(y_i)$, the partition function $\int \phi(y_i)e^{f_i(y_i)} dy_i = 1$ and the minimum discrimination information in (14) develops into the following form

$$KL^{min}(p_i||\phi) \quad = \int \phi(y_i)e^{f_i(y_i)}f_i(y_i)dy_i = \mathrm{E}\{f_i(\mathbf{w}_i^T\mathbf{x})\} \tag{21}$$

Different from $FastICA$, we can apply the fixed-point method to $KL^{min}(p_i||\phi)$ directly no matter how many nonlinear functions are used. The additive model (21) is easier to be optimized compared with the sum of quadratic terms in (3).

Algorithm 1: Estimating the *Unmixing Matrix* $\mathbf{W}$ by Fixed-Point Method.

1: **for** $i = 1$ to m **do**
2: $\quad\mathbf{w}_i \leftarrow \mathrm{E}\{\mathbf{x}f_i'(\mathbf{w}_i^T\mathbf{x})\} - \mathrm{E}\{f_i''(\mathbf{w}_i^T\mathbf{x})\}\mathbf{w}_i$
3: **end for**
4: $\mathbf{W} \leftarrow (\mathbf{w}_1, \ldots, \mathbf{w}_m)^T$
5: symmetric decorrelation
6: $\mathbf{W} \leftarrow (\mathbf{W}\mathbf{W}^T)^{-\frac{1}{2}}\mathbf{W}$

IV. Experiments and Results

A. Implementation Details

Two experiments are implemented to test the performance of the proposed method. The first experiment has been conducted in the past researches [8], [13], [17], where the independent sources' components s_i are chosen from 18 distributions [13] in Fig. 1. The second experiment is designed to validate our method with real signals [31].

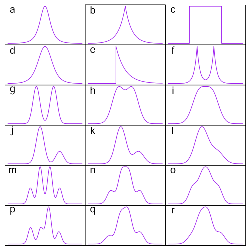

Fig. 1. Eighteen probability density functions of sources.

TABLE I
ICA METHODS USED IN THE EXPERIMENTS

Methods	Symbols	Parameters	Sources
$FastICA$ [16]	F-G0	$G0 = \frac{y_i^4}{4}$	$ProDenICA$ package [32]
$FastICA$	F-G1	$G1 = \log cosh(y_i)$	$ProDenICA$ package
$ProDenICA$ [8]	PICA	Gfunc=GPois	$ProDenICA$ package
$MDIICA$	MICA	equation (20)	
$EFICA$ [18]	EICA	default	$EFICA$ package [33]
$WeICA$ [22]	WICA	/	our own implementation

Several existing algorithms are chosen for comparisons, the implementation details are presented in Table I. $EFICA$ [18] is a statistically efficient version of the FastICA, and $WeICA$ [22] is the recent ICA algorithm based on the weighted second moments. It might not be appropriate to compare the CPU elapsed time between $EFICA$ (Matlab implementation) and other ICA algorithms (R implementation), thus we denote the EFICA's CPU elapsed time with superscript *.

The separation performance of ICA is measured by the value of Amari metrics $d(\mathbf{W}, \mathbf{W}_0)$[34], which is equal to zero if and only $\mathbf{W}$ and $\mathbf{W}_0$ are equivalent,

$$d(\mathbf{W}, \mathbf{W}_0) = \frac{1}{2m} \sum_{i=1}^{m} \left(\frac{\sum_{j=1}^{m} |r_{ij}|}{\max_j |r_{ij}|} - 1 \right) + \frac{1}{2m} \sum_{j=1}^{m} \left(\frac{\sum_{i=1}^{m} |r_{ij}|}{\max_i |r_{ij}|} - 1 \right) \tag{22}$$

where $r_{ij} = (\mathbf{W}\mathbf{W}_0^{-1})_{ij}$, $\mathbf{W}_0$ is the known truth.

B. Experiments With Simulated Signals

For each distribution in Fig. 1, a pair of independent components ($N = 1000$) is generated as sources in two-dimensional ICA, then they are mixed by a random invertible matrix to produce the mixtures $\mathbf{x}$. This experiment is replicated 100 times for each distribution, the average Amari metrics and CPU elapsed time are recorded in Fig. 2. $FastICA$ works well only when its single nonlinear function is close to the unknown sources' distribution, otherwise its separation performance or negentropy estimation might degenerate due to the unwanted density mismatching. As can be seen in Fig. 2, F-G0 and F-G1 perform well in symmetric distributions (f, g, h, i) thanks to their symmetric nonlinear functions (F-G0: $G0 = \frac{y_i^4}{4}$, F-G1: $G1 = \log cosh(y_i)$) used in negentropy estimation, whereas their separation performance deteriorates in the cases of non-symmetric or nontrivial distributions (j, k, l, p, q, r, n). Compared with the single nonlinear function used in F-G0 and F-G1, the nonparametric estimation used in PICA and nonlinear functions

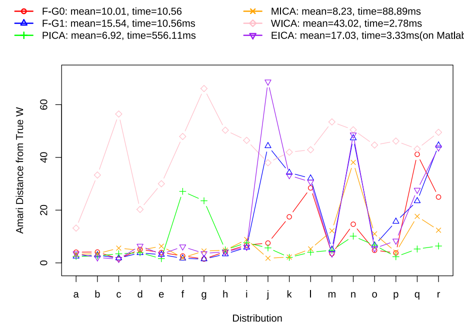

Fig. 2. Average Amari metrics (multiplied by 100) for two-component ICA. The overall mean Amari metrics and CPU Elapsed time (ms) is recorded in the legend.

TABLE II
ICA EXPERIMENTS ON ICS IMAGES (100 REPLICATIONS)

Mean	F-G0	F-G1	PICA	MICA	WICA	MICA_4	EICA
Amari metrics	40.79	55.12	19.4	48.83	30	28.96	42.55
Elapsed time(ms)	225.26	134.66	1683.41	401.16	1226.42	741.37	29.50*
Standard deviation	**F-G0**	**F-G1**	**PICA**	**MICA**	**WICA**	**MICA_4**	**EICA**
Amari metrics	7.36	3.17	3.57	10.13	0.0	0.0	5.07
Elapsed time(ms)	7.52	4.62	37.78	51.22	14.03	67.50	18.88*

used in MICA are more flexible in these nontrivial cases. The CPU elapsed time required by MICA (88.89 ms) is much less than PICA's (556.11 ms), as a result of transforming a sequence of IRLS (in PICA) into a single weighted least squares (in MICA).

C. Experiment With Real Data

We design an image separation experiment in this subsection, where the three gray-scale images (depicting a forest road, cat, and sheep) used are from the *ICS* package [31]. We vectorize them to arrive into a $130^2 \times 3$ data matrix, then mix them by a random invertible matrix (100 replications). The results are recorded in Table II. As can be seen in Table II, F-G0, F-G1, MICA, EICA suffer from the large Amari metrics, whereas PICA acquires the best separation performance with the longest CPU elapsed time. To improve the separation performance of MICA, we introduce another two nonlinear functions ($G0, G1$ in Table I) to the negentropy estimation, then an efficient version MICA_4 is immediately available. The average Amari metrics of MICA_4 is the second lowest and the corresponding CPU elapsed time is 44% of PICA's.

V. CONCLUSION

In this paper, we propose a novel ICA algorithm based on the second-order approximation of minimum discrimination information. Our algorithm alleviates the difficulties in $FastICA$ when more nonlinear functions are required in negentropy estimation. In addition, we reduce the computation burden by transforming a sequence of IRLS in $ProDenICA$ [8] into a single weighted least squares. The proposed method is concise and efficient, several experiments validate its performance compared with other ICA methods.

References

[1] P. Comon, "Independent component analysis, a new concept," *Signal Process.*, vol. 36, pp. 287–314, 1994.

[2] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, 1st ed. Hoboken, NJ, USA: Wiley, 2001.

[3] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, 1st ed. New York, NY, USA: Academic, 2010.

[4] D. Pham, P. Garrat, and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach," in *Proc. Eur. Signal Process. Conf.*, 1992, pp. 771–774.

[5] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.

[6] D. J. C. MacKay, "Maximum likelihood and covariant algorithms for independent component analysis," Tech. Rep., 1996.

[7] J.-F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Process. Lett.*, vol. 4, no. 4, pp. 112–114, Apr. 1997.

[8] T. Hastie and R. Tibshirani, "Independent components analysis through product density estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 15, 2003, pp. 665–672.

[9] A. Samarov and A. Tsybakov, "Nonparametric independent component analysis," *Bernoulli*, vol. 10, no. 4, pp. 565–582, 2004.

[10] A. Chen and P. J. Bickel, "Efficient independent component analysis," *Ann. Statist.*, vol. 34, no. 6, pp. 2825–2855, 2006.

[11] R. Samworth and M. Yuan, "Independent component analysis via nonparametric maximum likelihood estimation," *Ann. Statist.*, vol. 40, no. 6, pp. 2973–3002, 2012.

[12] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture," *Signal Process.*, vol. 24, no. 1, pp. 1–10, 1991.

[13] F. Bach and M. Jordan, "Kernel independent component analysis," *J. Mach. Learn. Res.*, vol. 3, pp. 1–48, 2003.

[14] J.-F. Cardoso, "Source separation using higher order moments," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 1989, vol. 4, pp. 2109–2112.

[15] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," *IEE Proc. F. - Radar Signal Process.*, vol. 140, no. 6, pp. 362–370, 1993.

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

[17] E. Learned-Miller and J. Fisher III, "ICA using spacings estimates of entropy," *J. Mach. Learn. Res.*, vol. 4, pp. 1271–1295, 2003.

[18] Z. Koldovsky, P. Tichavsky, and E. Oja, "Efficient variant of algorithm fastica for independent component analysis attaining the Cramér-Rao lower bound," *IEEE Trans. Neural Netw.*, vol. 17, no. 5, pp. 1265–1277, Sep. 2006.

[19] P. Ilmonen and D. Paindaveine, "Semiparametrically efficient inference based on signed ranks in symmetric independent component models," *Ann. Statist.*, vol. 39, no. 5, pp. 2448–2476, 2011.

[20] D. S. Matteson and R. S. Tsay, "Independent component analysis via distance covariance," *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 623–637, 2017.

[21] P. Ablin, J.-F. Cardoso, and A. Gramfort, "Faster independent component analysis by preconditioning with hessian approximations," *IEEE Trans. Signal Process.*, vol. 66, no. 15, pp. 4040–4049, Aug. 2018.

[22] P. Spurek, J. Tabor, L. Struski, and M. Smieja, "Fast independent component analysis algorithm with a simple closed-form solution," *Knowl.-Based Syst.*, vol. 161, pp. 26–34, 2018.

[23] A. Podosinnikova, A. Perry, A. S. Wein, F. Bach, A. d'Aspremont, and D. Sontag, "Overcomplete independent component analysis via SDP," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, vol. 89, 2019, pp. 2583–2592.

[24] A. Hyvärinen, "New approximations of differential entropy for independent component analysis and projection pursuit," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 1998, pp. 273–279.

[25] D. V. Gokhale and S. Kullback, "The minimum discrimination information approach in analyzing categorical data: The minimum discrimination information," *Commun. Statist.-Theory Methods*, vol. 7, no. 10, pp. 987–1005, 1978.

[26] J. H. Friedman, W. Stuetzle, and A. Schroeder, "Projection pursuit density estimation," *J. Amer. Stat. Assoc.*, vol. 79, no. 387, pp. 599–608, 1984.

[27] J. Friedman, "Exploratory projection pursuit," *J. Amer. Stat. Assoc.*, vol. 82, pp. 249–266, 1987.

[28] T. Hastie and R. Tibshirani, *Generalized Additive Models*, 1st ed. Boca Raton, FL, USA: Routledge, 1990.

[29] R. Wolke and H. Schwetlick, "Iteratively reweighted least squares: Algorithms, convergence analysis, and numerical comparisons," *SIAM J. Sci. Stat. Comput.*, vol. 9, no. 5, pp. 907–921, 1988.

[30] D. Cook, A. Buja, and J. Cabrera, "Projection pursuit indexes based on orthonormal function expansions," *J. Comput. Graph. Statist.*, vol. 2, no. 3, pp. 225–250, 1993.

[31] K. Nordhausen, H. Oja, and D. E. Tyler, "Tools for exploring multivariate data: The package ICS," *J. Stat. Softw.*, vol. 28, no. 1, pp. 1–31, 2008.

[32] T. Hastie and R. Tibshirani, "ProDenICA: Product density estimation for ICA using tilted Gaussian density estimates," 2010. [Online]. Available: https://CRAN.R-project.org/package=ProDenICA

[33] Z. Koldovský, "Efica algorithm," 2009. [Online]. Available: https://asap.ite.tul.cz/downloads/the-efica-algorithm/

[34] S.-i. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind signal separation," in *Proc. Adv. Neural. Inform. Proc. Sys.*, 1999, vol. 8, pp. 757–763.