


A Unifying Objective Function of Independent Component Analysis for Ordering Sources by Non-Gaussianity

Yoshitatsu Matsuda , *Member, IEEE*, and Kazunori Yamaguchi

Abstract—The independent component analysis (ICA) is a widely used method for solving blind separation problems. The ICA assumes that the sources are independent of each other and extracts them by maximizing their non-Gaussianity as the objective function. There are the two types of non-Gaussianity of the sources (the super-Gaussian type with the positive kurtosis and the sub-Gaussian one with the negative kurtosis). In this paper, we propose a new objective function unifying the two types of non-Gaussianity naturally, which is derived by applying the Gaussian approximation to the distribution of sources in the second-order polynomial feature space. The proposed objective function [called the adaptive ICA function (AIF)] is a simple form given as a summation of weighted fourth-order statistics, where the weights are adaptively estimated by the current kurtoses. The first practical advantage of the AIF is that it can extract the sources one by one in the descending order of the criterion of non-Gaussianity. It can solve the permutation ambiguity problem. The second and more important advantage is that it can estimate the number of non-Gaussian sources by the Akaike information criterion irrespective of the specific form of their distributions. In order to utilize the above-mentioned advantages of the AIF, we construct a new algorithm named the ordering ICA by extending the fast ICA. Experimental results verify that the ordering ICA can estimate the number of non-Gaussian sources correctly in both artificial and real data sets.

Index Terms—Blind source separation, higher order statistics, independent component analysis (ICA), signal denoising.

I. INTRODUCTION

THE independent component analysis (ICA) is a widely used method for solving blind source separation problems [1], [2] and extracting features from given signals [3]. It assumes that the source signals are statistically independent of each other and are given according to non-Gaussian distributions. The linear model of the ICA is given as

$$\mathbf{x} = \mathbf{A}\mathbf{s} \quad (1)$$

where $\mathbf{x} = (x_i)$ is the N -dimensional observed signal. $\mathbf{A} = (a_{ij})$ and $\mathbf{s} = (s_i)$ are the $N \times N$ invertible mixing matrix and the N -dimensional source, respectively. Here, only the

mixture \mathbf{x} can be observed and the others are unknown. The ICA can estimate \mathbf{A} and \mathbf{s} by regarding the non-Gaussianity of the source s as the objective function and maximizing it. The objective function of the ICA (often called the contrast function) can be given without knowing the original source distribution in advance. Actually, simple higher order statistics such as kurtosis are effective as the objective function irrespective of the accurate form of the source distribution. This “simplicity” is one of the most significant advantages of the ICA. However, there are two different types of the non-Gaussian distributions, which are classified by the sign of the kurtosis of the distribution. They are called the super-Gaussian distribution with the positive kurtosis and the sub-Gaussian one with the negative kurtosis. Though both of them are non-Gaussian, they diverge from the Gaussian distribution in the opposite directions. Therefore, the simplicity of the ICA in the previous works is generally deteriorated when both the types coexist in the sources. For example, the extended InfoMax [4] switches different objective functions in the progress of the estimation of sources. The fast ICA [5] implicitly switches the minimization and the maximization of an objective function. There is no “smooth” objective function in these methods. Though joint approximation diagonalization of eigen-matrices (JADE) [6] utilizes a smooth objective function consisting of the squares of higher order statistics, its optimization process is relatively complicated and time-consuming when there are a large number of signals.

In this paper, we propose a new objective function of the ICA unifying the super- and sub-Gaussian distributions as possible as “natural,” which is named the adaptive ICA function (AIF). The AIF is continuous with respect to almost all the source distributions and does not include any non-linear transformation of statistics. It is derived by applying the Gaussian approximation to the second-order polynomial feature space of sources where the fourth-order statistics are naturally involved. The AIF has several significant practical advantages over the previous objective functions. The AIF can extract all the sources in the descending order of a criterion of the non-Gaussianity no matter whether they are super- or sub-Gaussian. It can resolve the permutation ambiguity, which has been known to be one of the inherent problems of the ICA. Moreover, it enable us to estimate the number of non-Gaussian sources by a simple Gaussianity test, which is based on the Akaike information criterion (AIC) in the feature space and does not depend on any specific form of the source distributions.

Manuscript received May 16, 2017; revised November 29, 2017 and February 9, 2018; accepted February 10, 2018. Date of publication March 13, 2018; date of current version October 16, 2018. This work was supported in part by Grant-in-Aid for Young Scientists (KAKENHI) under Grant 26730013 and in part by JSPS KAKENHI under Grant 17H01837. (*Corresponding author: Yoshitatsu Matsuda.*)

The authors were with the Department of General Systems Studies, The University of Tokyo, Tokyo 113-8654, Japan (e-mail: matsuda@graco.c.u-tokyo.ac.jp; yamaguch@graco.c.u-tokyo.ac.jp).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TNNLS.2018.2806959

This paper is organized as follows. In Section II, the related works are explained. Section III describes the derivation of the AIF and its mathematical properties including the criterion of non-Gaussianity and the Gaussianity test. Section IV proposes a new algorithm using the AIF named “the ordering ICA.” Section V shows the experimental results on both artificial and real data sets. Finally, this paper is concluded in Section VI.

This paper is an elaborated and extended version of our previous papers [7]–[9]. The new contributions of this paper consist of the following four parts. First, we unify the fragments in our previous papers and clarify the significance of the AIF by explicitly stating its relation to the classical works on the ICA in Section II. Second, we propose a new criterion for the Gaussianity test by utilizing the AIC in Section II-C. Though a different criterion has been proposed in [9] by the Fisher information, its scale parameter needs to be set empirically. On the other hand, the new criterion is deductively derived from the AIC and does not include any arbitrary parameter. The proposed Gaussianity test can estimate the number of non-Gaussian sources quite accurately as shown in Section V. This is the most significant contribution in this paper. Third, we show that the fast ICA using kurtosis can maximize the AIF locally in Section IV. Then, we propose a new efficient algorithm for maximizing the AIF globally by multiple executions of the fast ICA. A stochastic gradient algorithm was employed in our previous papers so that the convergence was generally slow. Fourth, the numerical experiments are carried out for various data sets in Section V. Though only the natural images were employed as the sources in the previous papers, we employ the three additional data sets: generalized Gaussian signals, sine waves, and voices. We also investigate the experimental results extensively for generalized Gaussian signals by using various settings and various evaluations.

II. RELATED WORKS

A. Objective Function of ICA

The objective function of the ICA is a criterion measuring the independency among the estimated sources. As its maximization can extract independent sources, it is an essential part of the ICA. Here, the previously proposed objective functions are explained in brief. The following notation is introduced:

$$\mathbf{y} = \mathbf{W}\mathbf{x} \quad (2)$$

where $\mathbf{y} = (y_i)$ is an N -dimensional estimated source and $\mathbf{W} = (w_{ij})$ is an estimated separating matrix. When \mathbf{W} is the inverse of \mathbf{A} , the accurate estimation of the source is achieved. In addition, it is assumed for simplicity that \mathbf{x} and \mathbf{y} are prewhitened. Then, \mathbf{A} and \mathbf{W} are assumed to be orthonormal, because the independent source \mathbf{s} is whitened. Though there are various objective functions without the orthonormality constraint (for example, the original InfoMax [10]), they are omitted in this section, because we focus on the effects of the form of the source distributions. In addition, we focus on only the non-Gaussianity-based objective functions so that we omit the cross-correlation-based methods. Then, the previously

proposed objective functions of the ICA are roughly classified into the following three types.

The first type is based on the entropy of the estimated source, which is often called InfoMax. The entropy of each y_i (denoted by $H(y_i)$) is given as

$$H(y_i) = E(-\log p(y_i)) \quad (3)$$

where $E()$ is the expectation operator and $p(y_i)$ is the probability density function of y_i . Though almost all non-Gaussian approximations of $H(y_i)$ can extract arbitrary non-Gaussian sources [11], the following two approximations are widely used [5], namely, the kurtosis approximation:

$$H(y_i) \simeq \pm E(y_i^4) \quad (4)$$

and the hyperbolic approximation:

$$H(y_i) \simeq \pm E(\log \cosh y_i). \quad (5)$$

The former is simple and efficient. Though the latter is relatively complicated and inefficient, it is robust to outliers (their differences are discussed further in the last paragraph of this section). Then, the objective function is given as follows:

$$\Omega_1 = \sum_i \pm E(y_i^4) \text{ or } \sum_i \pm E(\log \cosh y_i). \quad (6)$$

The sign of \pm for each i depends on whether $p(y_i)$ is super-Gaussian or sub-Gaussian. In other words, the final form of Ω_1 cannot be decided in advance. Therefore, the objective function is not smooth in the optimization process. For example, the extended InfoMax [4] switches two different hyperbolic functions depending on the currently estimated kurtoses in the stochastic gradient-based optimization. Reference [12] switches the signs of the kurtoses in the pairwise rotation-based optimization. The fast ICA [5] implicitly switches the maximization and minimization of the objective functions by the fixed-point method. In summary, though the entropy-based approach gives simple objective functions so that various efficient optimization methods are applicable, there is no unifying, smooth function when both sub-Gaussian and super-Gaussian sources exist. Only the asymptotic analysis around the solutions is available.

The second type of the objective function is based on rotated cumulant matrices (also known as the joint diagonalization approach). Here, let c_{ijkl} and $\mathbf{C}_{kl} = (c_{ijkl})$ be the fourth-order cumulant on (x_i, x_j, x_k, x_l) and the kl th cumulant matrix, respectively. The rotated cumulant matrix $\tilde{\mathbf{C}}_{kl} = (\tilde{c}_{ijkl})$ is defined by $\tilde{\mathbf{C}}_{kl} = \mathbf{W}\mathbf{C}_{kl}\mathbf{W}^T$. It is proved in [13] that any $\tilde{\mathbf{C}}_{kl}$ is diagonal when \mathbf{W} is the inverse of \mathbf{A} . Therefore, the following objective function can be derived:

$$\Omega_2 = - \sum_{k,l \leq k} \sum_{i,j < i} \tilde{c}_{ijkl}^2. \quad (7)$$

Ω_2 can be maximized by pairwise rotations [6]. Ω_2 unifies the sub-Gaussian and super-Gaussian sources naturally. One of the serious disadvantages of Ω_2 is that its optimization is too much time-consuming when the number of signals N is large (for example, even $N = 50$ is intractable).

The third type is based on the Edgeworth expansion of the source distribution [14]. It gives the following quite simple objective function:

$$\Omega_3 = (E(y_i^4) - 3)^2 \quad (8)$$

which is the sum of the squares of the kurtoses. Ω_3 can be maximized by pairwise optimizations [14], [15]. Removing the uncertainty of the signs in Ω_1 by squaring each term, it gives a smooth objective function. Though it is a quite simple form unifying the sub-Gaussian and super-Gaussian sources, it is hardly used in practice, because its performance is generally low. One of the reasons of its low performance is probably that the squaring operation overestimates the influence of the sources with high kurtoses and fails to extract the sources with low kurtoses.

In this paper, we propose a new continuous objective function different from the above-mentioned three types, which is a weighted sum of kurtoses and unifies sub-Gaussian and super-Gaussian sources naturally without switching multiple functions. One of the most significant advantages of this non-switching function is that it can naturally include many Gaussian noises in the sources. It is known to be hard for the switching functions to manage multiple Gaussian noises appropriately and to estimate the number of non-Gaussian sources. It is essentially because the Gaussian noises are presumed to be excluded in the derivation of such switching functions. On the other hand, our proposed function can estimate the number of non-Gaussian sources accurately by a simple Gaussianity test (see Sections II-C and V).

Note again that the kurtosis approximation is sensitive to the outliers, because it drastically amplifies extreme values by its fourth power. On the other hand, other widely used approximations are more robust to the outliers, because they can constrain the effects of extreme values. For example, the hyperbolic approximation $\log \cosh$ converges to the linear order for sufficiently large values. This lack of robustness is known to be one of the serious problems in the kurtosis approximation. As our proposed function is based on the kurtosis, it is not expected to be so robust to outliers. Nevertheless, the experimental results in Section V will show that our proposed function is much more useful than the hyperbolic approximation in the estimation of the number of non-Gaussian sources. It may be promising to utilize a robust method of the ICA after the number of non-Gaussian sources is identified by our method.

B. Permutation Ambiguity in ICA

As shown in Section II-A, the objective functions of the ICA are generally insensitive to the permutation of sources. In other words, they are invariant with respect to the permutation of the rows of \mathbf{W} . Though some methods such as the fast ICA can extract the sources one by one (known as the deflation approach), their ordering is not unique [16]. Therefore, the permutation ambiguity has been regarded as an unavoidable problem in the ICA. Many of the solutions for fixing the permutation utilize the characteristics of specific applications other than non-Gaussianity (for example, [17] for audio signals and [18] for radio waves). On the other hand, a method

based on the fast ICA in [16] extracts the sources in a unique order of non-Gaussianity. However, this method needs to know the numbers of sub-Gaussian sources and super-Gaussian ones in advance. In this paper, we propose a new method extracting the sources one by one in a unique order without using any prior knowledge.

C. Estimation of the Number of Non-Gaussian Sources in ICA

The number of non-Gaussian sources (denoted by K) is often less than the number of signals N (the so-called undercomplete case). The estimation of K is usually carried out by utilizing principal component analysis (PCA) in the prewhitening phase. PCA can select the principal components by neglecting the minor components whose variances are below a given small threshold. However, the PCA-based estimation is useful only when the variances of Gaussian sources are assumed to be relatively small. PCA cannot directly detect whether a source is Gaussian or not. Another approach has been proposed, which employs an information criterion for deciding the number of sources [19]. One of the widely used information criteria is the AIC [20]. The AIC is defined as the difference between the degree of freedom of the parameters and the maximum log-likelihood. In the similar way as in (3), the AIC of the ICA is given as

$$\text{AIC} = - \sum_{i,t} \log p(\hat{y}_{it}) + V \quad (9)$$

where \hat{y}_{it} is the optimum value of y_i of the t th sample and $V \leq NK$ is the degree of freedom of the separating matrix \mathbf{W} . The number of sources can be estimated by selecting K so that the AIC is minimal. However, this approach strongly depends on the form of each source distribution $p(\hat{y}_{it})$. In this paper, we propose a simple Gaussianity test by the AIC in the second-order polynomial feature space. The test is carried out easily and does not depend on any specific form of source distributions. In our previous work [9], the Fisher information was utilized for the Gaussianity test. However, it was hard to control the significance level suitably. Though we proposed another idea applying the AIC to the ICA in [21], it was used only for deriving a threshold for Givens rotations in JADE.

D. Other Related Methods in ICA

The kernel method using the feature mapping is a widely used technique in machine learning (for example, a support vector machine and kernel PCA) [22]. This technique has also been applied to the ICA (named the kernel ICA in [23]). The kernel ICA uses the wide classes of feature mapping for estimating sources more accurately. On the other hand, we use the Gaussian approximation in only the second-order polynomial feature space for constructing a new simple objective function of the ICA. The efficient FastICA [24] switches more than two functions for separating more accurate sources. However, it did not consider the Gaussian noises. More flexible models of sources have been proposed in [25] and [26], which can include the Gaussian noises as the sources. However, their objective functions are too complicated to optimize easily.

III. OBJECTIVE FUNCTION

A. Derivation

Here, we derive a new objective function of the ICA by applying the Gaussian approximation to the distribution of sources in the second-order polynomial feature space. The detailed derivation is shown in [7]. We assume that each independent source s_i is normalized [namely, $E(s_i) = 0$ and $E(s_i s_j) = \delta_{ij}$ (the Kronecker delta)]. $\varphi_2(\mathbf{s}) = (s_i s_j)$ for $(i \leq j)$ denotes the $N(N+1)/2$ -dimensional vector of the sources in the second-order polynomial feature space. The Gaussian approximation needs only the mean and the covariance of $\varphi_2(\mathbf{s})$. The mean of $s_i s_j$ (denoted by μ_{ij}) is given as

$$\mu_{ij} = E(s_i s_j) = \delta_{ij}. \quad (10)$$

The covariance between $s_i s_j$ and $s_k s_l$ (denoted by $v_{ij,kl}$) is given as

$$v_{ij,kl} = E(s_i s_j s_k s_l) - E(s_i s_j) E(s_k s_l) = \begin{cases} \alpha_i + 2 & (i = j = k = l) \\ 1 & (i = k, j = l, i \neq j) \\ 0 & (\text{otherwise}) \end{cases} \quad (11)$$

where α_i is an unknown parameter estimating the kurtosis of s_i (denoted by $\kappa_i = E(s_i^4) - 3$). The vector of the parameters is denoted by $\boldsymbol{\alpha} = (\alpha_i)$. Then, the Gaussian approximation of the conditional distribution of $\varphi_2(\mathbf{s})$ given $\boldsymbol{\alpha}$ is given as

$$P_{\text{source}}(\varphi_2(\mathbf{s})|\boldsymbol{\alpha}) = \prod_{i,j>i} g_{ij}(s_i s_j) \prod_i g_{ii}(s_i^2|\alpha_i) \quad (12)$$

where $g_{ij}(u)$ is a Gaussian distribution given as

$$g_{ij}(u) = \begin{cases} \frac{1}{\sqrt{2\pi}(\alpha_i + 2)} \exp\left(\frac{-(u-1)^2}{2(\alpha_i + 2)}\right) & (i = j) \\ \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right) & (i < j). \end{cases} \quad (13)$$

The ICA estimates the separating matrix \mathbf{W} so that the estimator of source $\mathbf{y} = \mathbf{W}\mathbf{x}$ is “nearest” to the true source \mathbf{s} . By regarding $\varphi_2(\mathbf{y}) = (y_i y_j)$ as the estimator of $\varphi_2(\mathbf{s})$, the estimation of \mathbf{W} is formulated by using a maximum likelihood estimation of $P_{\text{source}}(\varphi_2(\mathbf{y})|\boldsymbol{\alpha})$. The linear transformation from $\varphi_2(\mathbf{x}) = (x_i x_j)$ to $\varphi_2(\mathbf{y})$ is given as

$$\varphi_2(\mathbf{y}) = \mathbf{E}(\mathbf{W} \otimes \mathbf{W})\mathbf{D}\varphi_2(\mathbf{x}) \quad (14)$$

where \otimes denotes the Kronecker product. \mathbf{E} and \mathbf{D} are the elimination binary matrix extracting only the elements with $i \leq j$ [namely, $\varphi_2(\mathbf{y}) = \mathbf{E}(\mathbf{y} \otimes \mathbf{y})$] and the duplication matrix duplicating the symmetric elements [namely, $\mathbf{x} \otimes \mathbf{x} = \mathbf{D}\varphi_2(\mathbf{x})$]. Therefore, the conditional distribution of $\varphi_2(\mathbf{x})$ given \mathbf{W} and $\boldsymbol{\alpha}$ is given as

$$P(\varphi_2(\mathbf{x})|\mathbf{W}, \boldsymbol{\alpha}) = P_{\text{source}}(\varphi_2(\mathbf{y})|\boldsymbol{\alpha})|\mathbf{E}(\mathbf{W} \otimes \mathbf{W})\mathbf{D}| \quad (15)$$

where $|\mathbf{U}|$ denotes the determinant of any square matrix \mathbf{U} . The following equation holds (its proof is in Appendix A):

$$|\mathbf{E}(\mathbf{W} \otimes \mathbf{W})\mathbf{D}| = |\mathbf{W}|^{N+1}. \quad (16)$$

Consequently, the expectation of the log-likelihood of $P(\varphi_2(\mathbf{x})|\mathbf{W}, \boldsymbol{\alpha})$ is given as

$$\begin{aligned} E(\log P(\varphi_2(\mathbf{x})|\mathbf{W}, \boldsymbol{\alpha})) &= -\frac{\sum_{i,j>i} E(y_i^2 y_j^2)}{2} - \frac{\sum_i \log(\alpha_i + 2)}{2} \\ &\quad - \sum_i \frac{E((y_i^2 - 1)^2)}{2(\alpha_i + 2)} + (N+1) \log |\mathbf{W}| \end{aligned} \quad (17)$$

where some constants are neglected. Note that every y_i depends on \mathbf{W} . Finally, (17) is rewritten as the following objective function $\Psi(\mathbf{W}, \boldsymbol{\alpha})$:

$$\begin{aligned} \Psi(\mathbf{W}, \boldsymbol{\alpha}) &= -\sum_i \log(\alpha_i + 2) + 2(N+1) \log |\mathbf{W}| \\ &\quad - \sum_{i,j} \left(\frac{1 - \delta_{ij}}{2} + \frac{\delta_{ij}}{\alpha_i + 2} \right) E((y_i y_j - \delta_{ij})^2) \end{aligned} \quad (18)$$

where the constant factor of 1/2 was removed. $\Psi(\mathbf{W}, \boldsymbol{\alpha})$ is called the AIF in this paper.

B. Mathematical Properties

Here, various mathematical properties of the AIF are shown and proved. Though the derivation of the AIF needs the “imprecise” approximation (for example, $s_i s_j$ never obey the Gaussian distribution actually), we show that the AIF $\Psi(\mathbf{W}, \boldsymbol{\alpha})$ is an appropriate objective function of the ICA *irrespective of its derivation*. The more details of the proofs are shown in [7]–[9].

1) *Adaptive Estimation of Kurtosis*: The optimal $\hat{\alpha}_i$ for a given \mathbf{W} satisfies

$$\frac{\partial \Psi}{\partial \alpha_i} \Big|_{\alpha_i = \hat{\alpha}_i} = -\frac{1}{\hat{\alpha}_i + 2} + \frac{E((y_i^2 - 1)^2)}{(\hat{\alpha}_i + 2)^2} = 0. \quad (19)$$

Thus, $\hat{\alpha}_i$ is given analytically as

$$\hat{\alpha}_i = E((y_i^2 - 1)^2) - 2 \quad (20)$$

which is the unique optimum of $\Psi(\mathbf{W}, \boldsymbol{\alpha})$ with respect to α_i . It is worth noting that the optimal $\hat{\alpha}_i$ is given as

$$\hat{\alpha}_i = E(y_i^4) - 3 \quad (21)$$

under the normalization constraint $E(y_i^2) = 1$. In other words, $\hat{\alpha}_i$ can be regarded as the adaptive estimator of the kurtosis. We can define the following objective function $\Phi(\mathbf{W})$ depending on only \mathbf{W} as:

$$\begin{aligned} \Phi(\mathbf{W}) &= \Psi(\mathbf{W}, \hat{\boldsymbol{\alpha}}) \\ &= -\sum_i \log E((y_i^2 - 1)^2) + 2(N+1) \log |\mathbf{W}| \\ &\quad - \frac{\sum_{i,j \neq i} E(y_i^2 y_j^2)}{2}. \end{aligned} \quad (22)$$

The optimal \mathbf{W} of $\Phi(\mathbf{W})$ is equivalent to that of $\Psi(\mathbf{W}, \boldsymbol{\alpha})$ if $\boldsymbol{\alpha}$ is optimal. Therefore, we often substitute $\Phi(\mathbf{W})$ for $\Psi(\mathbf{W}, \boldsymbol{\alpha})$ in the following analysis.

2) *Gaussian Case*: In the simplest case where all the sources are Gaussian with the mean 0, Theorem 1 holds.

Theorem 1: When every s_i is Gaussian with the mean 0 and \mathbf{A} is invertible, $\Phi(\mathbf{W})$ of (22) is maximal if and only if \mathbf{y} is whitened.

See Appendix B for its proof. This theorem guarantees that \mathbf{y} is whitened without any constraints when the sources are Gaussian. The optimal \mathbf{W} is a saddle point, because any orthonormal rotations preserve the optimal condition. It is consistent with the previous ICA methods in the Gaussian case.

3) *Non-Gaussian Case Under Normalization Constraint*: Here, we analyze the cases where almost all the sources are non-Gaussian under the relatively weak normalization constraint [namely, $E(y_i) = 0$ and $E(y_i^2) = 1$]. In addition, we use the following three assumptions.

- 1) The simple linear ICA model $\mathbf{x} = \mathbf{A}\mathbf{s}$ in the real domain is assumed, where every source is normalized and \mathbf{A} is invertible.
- 2) All the sources are non-Gaussian except at most one Gaussian source. In other words, they may include one Gaussian source.
- 3) The kurtosis of any source is not equal to -2 (the theoretical minimum of the kurtosis). In other words, any source does not obey the uniform Bernoulli distribution.

Then, Theorem 2 holds.

Theorem 2: It is assumed that \mathbf{x} is given by $\mathbf{x} = \mathbf{A}\mathbf{s}$ in the real domain, where \mathbf{A} is invertible and \mathbf{s} (with the mean 0) does not include more than one Gaussian signal nor any uniform Bernoulli variable. Then, the solution of the ICA ($\mathbf{W} = \mathbf{A}^{-1}$) is a local maximum of $\Phi(\mathbf{W})$ of (22) under the constraint $E(y_i^2) = 1$.

Its proof consists of the necessary condition and the sufficient one. The necessary condition is equivalent to the Karush–Kuhn–Tucker (KKT) condition of this constrained optimization, which is proved by the first derivatives of $\Phi(\mathbf{W})$ (see Appendix C). The sufficient condition corresponds to the convergence analysis, which shows that the solution is a local maximum in the constrained optimization. The bordered Hessian matrix and the bordered determinantal criterion [27] are used. See Appendix D for its proof. See also [7] for the details of the proofs. This theorem guarantees that the AIF is an appropriate objective function of the ICA. It is worth noting that the optimal variance of y_i under no constraint is not equal to 1 in the non-Gaussian cases. It is easily shown for $N = 1$. In other words, the normalization constraint is necessary in the non-Gaussian cases. Theorem 2 is interesting, because the extraction of the sources needs only the weak normalization constraint instead of the usual orthonormality one. However, we focus on the orthonormality constraint in the following of this paper. It is because the orthonormality constraint guarantees that the global optimality solves the permutation ambiguity and estimates the number of non-Gaussian sources appropriately.

4) *Non-Gaussian Case Under Orthonormality Constraint*: Here, we analyze the non-Gaussian cases under the orthonormality constraint [namely, $E(y_i) = 0$ and $E(y_i y_j) = \delta_{ij}$].

Under this constraint, the AIF of (18) is simplified into

$$\Psi(\mathbf{W}, \boldsymbol{\alpha}) = \sum_i \Psi_i(w_{i1}, \dots, w_{iN}, \alpha_i) \quad (23)$$

where

$$\Psi_i = -\log(\alpha_i + 2) + \left(\frac{1}{2} - \frac{1}{\alpha_i + 2}\right) E(y_i^4 - 1). \quad (24)$$

The proof is described in Appendix E. The local optimality of (23) is guaranteed under the orthonormality constraint, because it is a special case of Theorem 2. The optimal $\hat{\alpha}_i$ is given as the currently estimated kurtosis in the same way as in (21). Then, the following equivalent objective function Φ_i is derived:

$$\Phi_i(w_{i1}, \dots, w_{iN}) = -\log E(y_i^4 - 1) + \frac{E(y_i^4)}{2}. \quad (25)$$

Ψ_i can be replaced with Φ_i , because the maximum point of Ψ_i with respect to (w_{i1}, \dots, w_{iN}) is equal to that of Φ_i . In addition, it is worth noting that Ψ_i is given as follows if α_i is the accurate kurtosis κ_i :

$$\Psi_i(\alpha_i = \kappa_i) = \left(\frac{1}{2} - \frac{1}{\kappa_i + 2}\right) E(y_i^4). \quad (26)$$

Therefore, the weight $(1/2) - (1/(\kappa_i + 2))$ is positive for super-Gaussian source ($\kappa_i > 0$) and negative otherwise ($\kappa_i < 0$). It is consistent with the previous ICA framework as shown in $\Omega_1 = \sum_i \pm E(y_i^4)$ of (6).

5) *Gram–Schmidt Orthonormalization*: Here, we analyze the cases where Ψ_i of (24) is maximized for each i in the Gram–Schmidt orthonormalization. It is called the deflation approach in the fast ICA. The Gram–Schmidt orthonormalization constrains each y_i to satisfy $E(y_i y_j) = \delta_{ij}$ for $j \leq i$. Consequently, it is proved that all the non-Gaussian sources are extracted in the descending order of a criterion of non-Gaussianity if each Ψ_i is globally maximized in the Gram–Schmidt orthonormalization. Rigorously, the Theorem 3 holds.

Theorem 3: We assume the linear ICA model, where $\mathbf{x} = \mathbf{A}\mathbf{s}$ in the real domain, $E(s_i) = 0$, $E(s_i s_j) = \delta_{ij}$, $\kappa_i > -2$ (no uniform Bernoulli source), and \mathbf{A} is invertible. Then, all the non-Gaussian sources are extracted in descending order of $\Upsilon(\kappa_i)$ if each Ψ_i of (24) is globally maximized in the Gram–Schmidt orthonormalization, where $\Upsilon(\kappa)$ is a criterion of non-Gaussianity given as

$$\Upsilon(\kappa) = \kappa - 2 \log \frac{\kappa + 2}{2}. \quad (27)$$

This theorem is proved by using the convexity of Φ_i of (25) with respect to $E(y_i^4 - 1)$. See Appendix F for the details of the proof. Note that $\Upsilon(\kappa_i)$ takes the minimum ($= 0$) only for $\kappa_i = 0$ (a Gaussian source). Therefore, $\Upsilon(\kappa_i)$ is regarded as a criterion of non-Gaussianity. This theorem guarantees that the maximization of the AIF in the Gram–Schmidt orthonormalization can find all the non-Gaussian sources. It can resolve the permutation ambiguity, because the sources are sorted by $\Upsilon(\kappa_i)$. In addition, it can reduce the time for estimating only the non-Gaussian sources by terminating the optimization once any Gaussian source is found. These advantages of

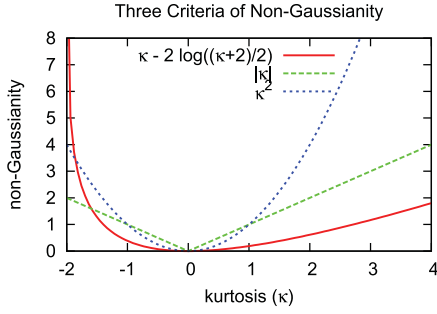


Fig. 1. Comparison of the three criteria of non-Gaussianity along the kurtosis κ [$\Upsilon(\kappa) = \kappa - 2 \log((\kappa + 2)/2)$, the absolute value $|\kappa|$, and the square κ^2].

the AIF in the Gram–Schmidt orthonormalization are verified experimentally in Section V.

C. Criteria of Non-Gaussianity

Section III-B5 shows that $\Upsilon(\kappa) = \kappa - 2 \log((\kappa + 2)/2)$ of (27) is regarded as a criterion of non-Gaussianity of a source. Here, we compare this criterion derived from the AIF with the usual two criteria shown in Section II-A. The first usual criterion is the absolute value of the kurtosis $|\kappa|$ derived from Ω_1 of (6), where \pm is removed. The second usual criterion is the square of the kurtosis κ^2 from Ω_3 of (8). Fig. 1 shows the three criteria along the kurtosis κ from -2 to 5 . They take the minimum 0 at $\kappa = 0$ (the Gaussian source). Moreover, they are larger as κ is further from 0 . It shows that all of them can be regarded as reasonable criteria of non-Gaussianity. However, there are some differences in detail. For example, as $|\kappa|$ is not differentiable at $\kappa = 0$, it causes the problem of switching from sub-Gaussian to super-Gaussian. On the other hand, κ^2 and the proposed criterion $\Upsilon(\kappa) = \kappa - 2 \log((\kappa + 2)/2)$ are always differentiable. Moreover, while $|\kappa|$ and κ^2 have small upper bounds in the sub-Gaussian area, $\Upsilon(\kappa)$ diverges to the infinity in both sub-Gaussian and super-Gaussian areas. In other words, $\Upsilon(\kappa)$ emphasizes the sub-Gaussian sources in comparison with the usual criteria.

D. Gaussianity Test by AIC

As shown in Section III-B5, if the currently estimated component is Gaussian, every succeeding component is Gaussian. By terminating the estimation process when we find the first Gaussian component, the process is expected to be much more efficient if there are only a few non-Gaussian sources. Here, a threshold for the Gaussianity test is proposed by utilizing the AIC [20] in the second-order polynomial feature space. Now, the Gaussianity of the currently extracted i th component is tested. It is assumed that y_i is constrained to be normalized and orthogonal to the preceding $i - 1$ components. Under these constraints, the log-likelihood of (17) is given as

$$\begin{aligned} \ell_i(w_{i1}, \dots, w_{iN}, \alpha_i) \\ = -\frac{M \log(\alpha_i + 2)}{2} + \left(\frac{1}{4} - \frac{1}{2(\alpha_i + 2)} \right) \sum_{t=1}^M (y_{it}^4 - 1) \end{aligned} \quad (28)$$

where M is the sample size and y_{it} is the value of y_i of the t th sample. It is derived in the similar way of the derivation

of (24) by keeping the constant factor $1/2$ and replacing the expectation with the summation over the samples. The constant factor $1/2$ is not neglected in order to keep the scale of the original log-likelihood. Then, the maximum of ℓ_i is given as the following form depending on only α_i :

$$\hat{\ell}_i(\hat{\alpha}_i) = \frac{M(\hat{\alpha}_i - 2 \log(\hat{\alpha}_i + 2))}{4} \quad (29)$$

where some constant terms are neglected and $\hat{\alpha}_i$ is the optimal value of α_i maximizing ℓ_i given as

$$\hat{\alpha}_i = \frac{\sum_{t=1}^M \hat{y}_{it}^4}{M} - 3. \quad (30)$$

\hat{y}_{it} is the optimal value of y_{it} . Though ℓ_i has $N + 1$ parameters, the degree of freedom is reduced by i , because (w_{i1}, \dots, w_{iN}) is constrained to be normalized and orthogonal to the preceding $i - 1$ sources. By maximizing the log-likelihood of the succeeding $N - i$ sources similarly, the maximum of the log-likelihood and the degree of freedom for all the remaining $N - i + 1$ components are given as $\sum_i \ell_i$ and $(N - i + 2)(N - i + 1)/2$, respectively. Consequently, the AIC in the estimation of the current i th component is given as

$$\text{AIC} = -\sum_{k=i}^N \hat{\ell}_k(\hat{\alpha}_k) + \frac{(N - i + 2)(N - i + 1)}{2}. \quad (31)$$

On the other hand, if the current i th and the succeeding components are assumed to be Gaussian, the log-likelihood is given as $(N - i + 1)\hat{\ell}(0)$ with the degree of freedom of 0 . Therefore, the difference of the AIC between the two models is given as

$$\begin{aligned} \Delta \text{AIC} \\ = -\sum_{k=i}^N \hat{\ell}_k(\hat{\alpha}_k) + \frac{(N - i + 2)(N - i + 1)}{2} + (N - i + 1)\hat{\ell}(0) \\ = -\frac{M \sum_{k=i}^N \Upsilon(\hat{\alpha}_k)}{4} + \frac{(N - i + 2)(N - i + 1)}{2} \end{aligned} \quad (32)$$

where Υ is the criterion of non-Gaussianity defined in (27). As the model with a smaller AIC is preferable, the condition for preferring the Gaussian model is given as

$$-\frac{M \sum_{k=i}^N \Upsilon(\hat{\alpha}_k)}{4} + \frac{(N - i + 2)(N - i + 1)}{2} > 0. \quad (33)$$

Unfortunately, the summation of $\Upsilon(\hat{\alpha}_k)$ over the remaining components needs the estimation of all the components without termination. Considering that $\Upsilon(\hat{\alpha}_i)$ is the maximum over the remaining components, we propose the following simple condition:

$$\Upsilon(\hat{\alpha}_i) < \frac{2(N - i + 2)(N - i + 1)}{M} \quad (34)$$

which is employed as the termination condition in this paper. Though this approximation increases the possibility that a non-Gaussian source is decided to be Gaussian incorrectly, the experimental results in Section V show the usefulness of this condition.

Here, some properties of the proposed termination condition are discussed. As $\Upsilon(u) \simeq u^2/4$ holds for a small u value by

the Taylor series approximation, this condition is simplified as follows:

$$\hat{\alpha}_i^2 < \frac{8(N-i+2)(N-i+1)}{M}. \quad (35)$$

It depends on the number of unestimated components $(N+1-i)$ unlike a usual constant threshold. It seems plausible, because the fluctuation of the maximum over multiple Gaussian sources depends on the number of unestimated Gaussian sources. It is worth noting that this condition does not depend on any form of the source distributions, because it depends on only the Gaussian approximation in the second-order polynomial feature space. Moreover, it does not depend on any arbitrary parameters such as the significance level. It is also worth noting that this condition does not estimate the accurate number of non-Gaussian sources but avoids futile estimation according to the sample size. While the estimated number of non-Gaussian sources may be inaccurate if the sample size is limited, it is expected to be appropriate for the sample size. Though this condition is applicable to the usual ICA methods, such as fast ICA and JADE, by sorting all the extracted components in the descending order after the total estimation, it was not so useful in the numerical experiments (see Section V).

Note that the AIC is not theoretically appropriate if the true distribution cannot be represented by the parametric model. Takeuchi information criterion (TIC) is more appropriate in such cases [20]. Therefore, the TIC is expected to be more suitable for our proposed model, where $s_i s_j$ never obey the Gaussian distribution actually. However, the TIC is much more complicated than the AIC. In addition, the AIC is known to give a good approximation of TIC in many cases. Thus, we employed the AIC instead of the TIC.

IV. ALGORITHM

Here, we describe an algorithm for maximizing the AIF of (23) in the Gram–Schmidt orthonormalization. Though our previous works used the combination of the stochastic gradient algorithm and the fast ICA [9], we do not use the stochastic gradient but extend the deflation approach of the fast ICA using the kurtosis in this paper. If α_i is given appropriately, the maximization of Ψ_i is equivalent to the maximization or the minimization of $E(y_i^4)$ [see (26)]. Therefore, we can utilize the fast ICA using the kurtosis. The complete algorithm is described as follows.

- 1) *Initialization*: Prewhiten the observed signals $X = (x_{it})$ and initialize $i = 1$.
- 2) *Optimization of i th Component*:
 - a) *Initialization of Multiple Candidates*: Initialize L different candidates of (w_{i1}, \dots, w_{iN}) randomly.
 - b) *Fast ICA for Each Candidate*: Repeat the following steps until the solution converges or the number of iterations exceeds a given maximum for each candidate.
 - i) Calculate $y_{it} = \sum_k w_{ik} x_{kt}$.
 - ii) Update $w_{ik} := \frac{\sum_t x_{kt} y_{it}^3}{M} - 3w_{ik}$ for $k = 1, \dots, N$.

- iii) *Orthonormalize* (w_{i1}, \dots, w_{iN}) by $w_{ik} := w_{ik} - \sum_{j < i} (\sum_l w_{il} w_{jl}) w_{jk}$ and $w_{ik} := w_{ik} / \sqrt{\sum_l w_{il}^2}$.

- c) *Selection of the Best Estimation From the Candidates*: Select the estimation with the highest criterion of non-Gaussianity $\Upsilon(\hat{\alpha}_i) = \hat{\alpha}_i - 2 \log((\hat{\alpha}_i - 2)/2)$, where $\hat{\alpha}_i = \sum_t y_{it}^4 / M - 3$.

- 3) *Termination Decision*: If $\Upsilon(\hat{\alpha}_i) < (2(N+2-i)(N+1-i)/M)$ [see (34)], estimate the number of non-Gaussian sources K to be $i-1$ and terminate the algorithm. Otherwise, let i be $i+1$ (the next component) and return to Step 2 (optimization) if $i \leq N$.

It is almost equivalent to the deflation approach of the original fast ICA [5] except for the selection from the multiple candidates and the termination decision. The original fast ICA corresponds to a single candidate selection ($L=1$) without the termination decision (the algorithm is terminated only when it fails to converge). Here, we call this method “the ordering ICA,” because it extracts the independent components in a unique order. The ordering ICA essentially repeats the fast ICA L times. The time complexity of an iteration of the fast ICA for each component is $O(NM)$. Moreover, the fast ICA is guaranteed to converge quadratically [5]. Letting the averaged number of iterations until the convergence be R , the time complexity of the ordering ICA is $O(N^2MLR)$. It can be accelerated by employing an efficient matrix multiplication algorithm. Moreover, the appropriate termination may reduce the calculation time drastically if there are only a few non-Gaussian sources. In addition, the candidates can be processed in parallel. Its actual computational cost is investigated by the numerical experiments in Section V. It is also worth noting that the criterion of non-Gaussianity can be replaced with the other criteria such as the square of the kurtosis. In Section V, it was shown that the proposed criterion of non-Gaussianity was better than the square of kurtosis in the numerical experiments.

V. RESULTS

A. Methods of ICA

Here, we describe the settings of the following five methods of the ICA in the numerical experiments: the ordering ICA in Section IV (denoted by ordering-ICA); the ordering ICA using κ^2 instead of $\Upsilon(\kappa)$ (ordering-kurt). Note that the usage of κ^2 is completely equivalent to that of $|\kappa|$ in this method; the fast ICA using the kurtosis [5] (fast-kurt); the fast ICA robust to outliers using hyperbolic tangent function [5] (fast-tanh); and JADE [6]. In ordering-ICA and ordering-kurt (using multiple candidates), the convergence of the fast ICA for each candidate was determined by the Euclidean distance between the preceding estimation and the current one (smaller than 10^{-6}) and the maximum number of iterations was set to 100. The settings of these conditions did not have a strong effect on the results. In fast-kurt and fast-tanh (using a single candidate), the convergence determination was the same in ordering-ICA and ordering-kurt. However, a more complicated termination decision was used. The algorithm was terminated only if a component did not converge within

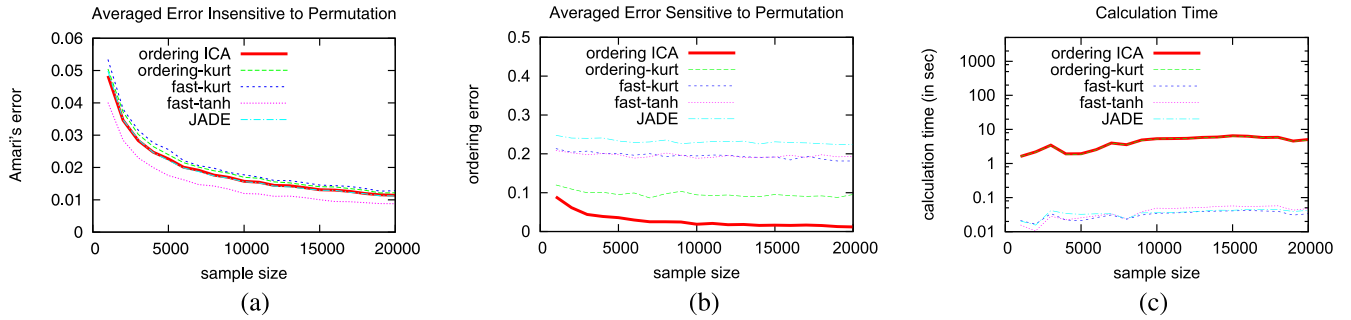


Fig. 2. Comparison of the five ICA methods in blind separation of artificial sources in the complete case ($N = K = 8$). The following five methods were applied: the ordering ICA (denoted by ordering-ICA), the ordering ICA using the square of kurtosis (ordering-kurt), fast ICA using kurtosis (fast-kurt), fast ICA using tanh (fast-tanh), and JADE. (a) E . (b) E^* . (c) Calculation time.

TABLE I
PARAMETERS AND STATISTICS OF GENERALIZED GAUSSIAN
DISTRIBUTIONS: RELATIVE RANKS IN $\Upsilon(\kappa)$ AND κ^2 ARE ALSO SHOWN

| ρ | κ | $\Upsilon(\kappa)$ | κ^2 |
|------------|----------|--------------------|------------|
| 2^{-1} | 22.2 | [1] 17.2 | [1] 492 |
| $2^{-0.5}$ | 7.84 | [2] 4.65 | [2] 61.4 |
| 2^{-0} | 3.00 | [3] 1.17 | [3] 9.00 |
| $2^{0.5}$ | 0.978 | [7] 0.182 | [6] 0.957 |
| $2^{1.5}$ | -0.520 | [8] 0.0821 | [8] 0.270 |
| 2^2 | -0.812 | [6] 0.229 | [7] 0.659 |
| $2^{2.5}$ | -0.980 | [5] 0.366 | [5] 0.959 |
| 2^3 | -1.08 | [4] 0.469 | [4] 1.16 |

five runs, each of which consists of 1000 iterations for a different random initialization. As the original fast ICA uses only a single candidate, the termination needs to be determined carefully and the settings of the conditions had a strong effect on the results. The simplicity of the termination decision is one advantage of the ordering ICA in comparison with the original fast ICA. In JADE, the convergence is determined when every Givens rotation is lower than a small threshold, which did not have a strong effect on the results. In addition, the number of candidates L in ordering-ICA and ordering-kurt was set to 100.

B. Generalized Gaussian Distribution

Here, the results on an artificial data set are shown. We employed the generalized Gaussian distributions as the sources, whose probability density function is given as

$$p(u) = \frac{\rho}{2\beta\Gamma(1/\rho)} \exp(-(|u|/\beta)^\rho) \quad (36)$$

where β and ρ are the parameters, and Γ is the gamma function. β was set so that the variance is 1. ρ is the parameter determining the kurtosis. The distribution is super-Gaussian for $\rho > 2$ and sub-Gaussian for $\rho < 2$. ρ was set to the eight different values. The parameter and the statistics of each generalized Gaussian distribution are shown in Table I, where the case for $\rho = 2^1$ is excluded, because it is Gaussian. The two ranks in the descending order of $\Upsilon(\kappa)$ and κ^2 are similar but not the same in the sixth and seventh places.

In the first experiment, we show the results on the separation of the eight sources without Gaussian noise, where the number of sources N is 8. The termination decision was not used. The mixing matrix \mathbf{A} was generated randomly, each element of which is given by a Gaussian distribution. The five methods in

Section V-A were applied to estimate the separating matrix \mathbf{W} . Two types of error on $\mathbf{Q} = (q_{ij}) = \mathbf{W}\mathbf{A}$ were employed for evaluating the results. The first type is the usual Amari's separating error [28]

$$E = \frac{1}{2N(N-1)} \sum_{i,j} \left(\frac{|q_{ij}|}{\max_k |q_{ik}|} + \frac{|q_{ij}|}{\max_k |q_{kj}|} \right) \quad (37)$$

which is insensitive to permutation. The second type is the difference between \mathbf{Q} and the identity matrix

$$E^* = \frac{1}{N^2} \sum_{i,j} |q_{i\iota(j)} - \delta_{ii(j)}| \quad (38)$$

where $\iota(j)$ is the index of the true source corresponding to the j th largest $\Upsilon(\kappa)$. It is sensitive to permutation, because it is small only if the estimated sources are sorted by $\Upsilon(\kappa)$. Fig. 2 shows the averaged results (E , E^* , and the calculation time) over 100 runs for various sample sizes. Fig. 2(a) shows that all the methods could separate the sources appropriately according to a given sample size. Though fast-tanh achieved the best results, ordering-ICA and the other methods were not so inferior. Fig. 2(b) shows that ordering-ICA could estimate the sorted sources without the permutation ambiguity. It is not so surprising but verifies that the proposed simple extension of the fast ICA using multiple candidates is useful. Fig. 2(c) shows that ordering-ICA and ordering-kurt were about $L = 100$ times as slow as the other methods, which is expected in Section IV. In summary, the ordering ICA is not so inferior to the other widely used methods of the ICA on the separation accuracy but is too slow in the complete case.

In the second experiment, we use the undercomplete cases where the number of non-Gaussian sources ($K = 8$) is less than the total number of signals N . In other words, $N - K$ sources were given according to the Gaussian distribution. N was set to 20, 30, and 40. Here, the estimated sources in every method were sorted by $\Upsilon(\hat{\kappa})$ (where $\hat{\kappa}$ is the kurtosis of a separated source). Then, the number of non-Gaussian sources was estimated by applying the AIC-based Gaussianity test in Section III-D to the sorted sources. The estimated number of sources is denoted by K_{AIC} . As is expected, K_{AIC} of ordering-ICA and that of ordering-kurt were equal to the number of extracted sources in their optimization process. On the other hand, K_{AIC} of the fast ICA and JADE was often smaller than the number of extracted sources, because

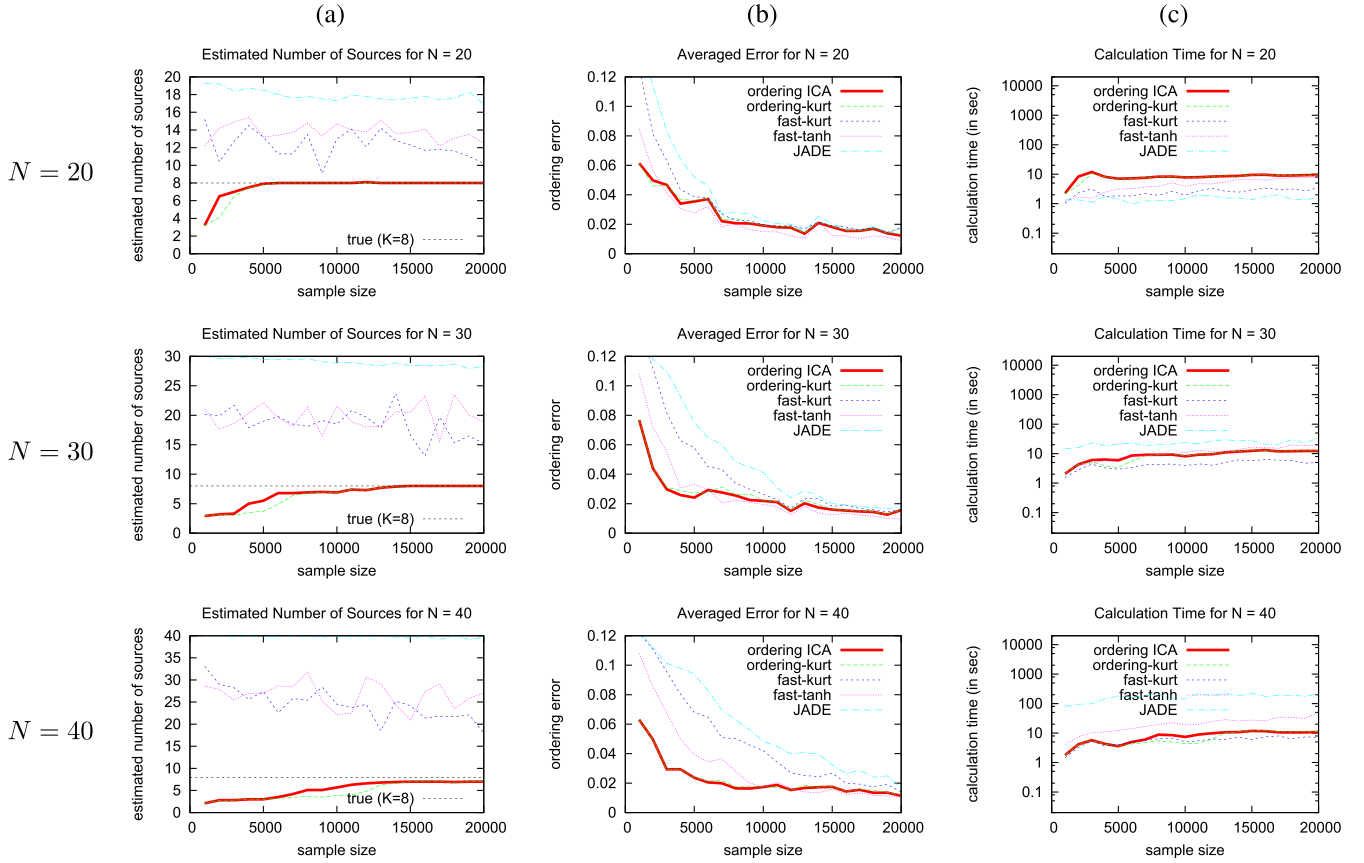


Fig. 3. Comparison of the five ICA methods in blind separation of artificial sources in the undercomplete case ($N > K$). Each row corresponds to $N = 20$, $N = 30$, and $N = 40$. Each column corresponds to (a) estimated number of non-Gaussian sources (K_{AIC}), (b) errors sensitive to permutation (E^*), and (c) calculation time.

their non-Gaussianity may be below the AIC-based threshold. We also used an extension of E^* of (38) as the separating error, which is defined as the difference between the nonsquare \mathbf{Q} and the corresponding nonsquare identity matrix because of $K < N$. Though K can be estimated as K_{AIC} , it was set to the true value 8 if $K_{AIC} > 8$, because the overestimation of K drastically increases the separating error. Fig. 3 shows the averaged results over 10 runs for $N = 20, 30$, and 40. Fig. 3(a) shows the results about K_{AIC} . Though ordering-ICA and ordering-kurt underestimated K for the small sample size, K_{AIC} approaches the true value $K = 8$ as the sample size becomes larger. In addition, ordering-ICA approached $K = 8$ slightly better than ordering-kurt for a small sample size. On the other hand, the other methods (the fast ICA and JADE) always overestimated K . Fig. 3(b) shows the results about the separating error E^* . Ordering-ICA and ordering-kurt outperformed the other methods for a small sample size. It is probably because the ordering ICA extracts only a few sources for such a small sample size and neglects the other sources which are not estimated accurately. All the methods achieved similar performance for a large sample size. Considering that it is hard to estimate the true number of sources for the fast ICA and JADE, it suggests that ordering-ICA and ordering-kurt achieve the best performance. Fig. 3(c) shows the calculation time. Ordering-ICA and ordering-kurt became relatively faster as the total number N became larger. It is surprising that ordering-ICA was faster than fast-tanh and was not so inferior

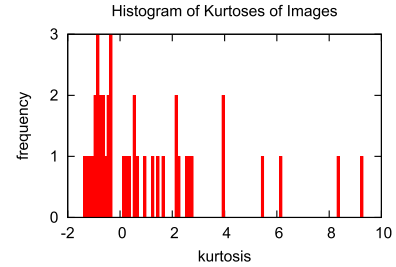


Fig. 4. Histogram of the kurtoses of the 44 original images: three images with the kurtoses beyond 10 are omitted here.

to fast-kurt for $N = 40$ while it needs to estimate all the $L = 100$ candidates. It is because the ordering ICA can terminate its optimization process when a Gaussian source is detected. On the other hand, the fast ICA is quite slow when all the sources are Gaussian. In summary, the ordering ICA can estimate the number of non-Gaussian sources appropriately in the undercomplete case and can separate them accurately for various sample sizes and various total numbers of signals. In addition, the ordering ICA is not slower than the other methods though it uses a large number of candidates.

C. Various Data Sets

Here, we focus on the estimation of the number of non-Gaussian sources. We used the following three data sets.

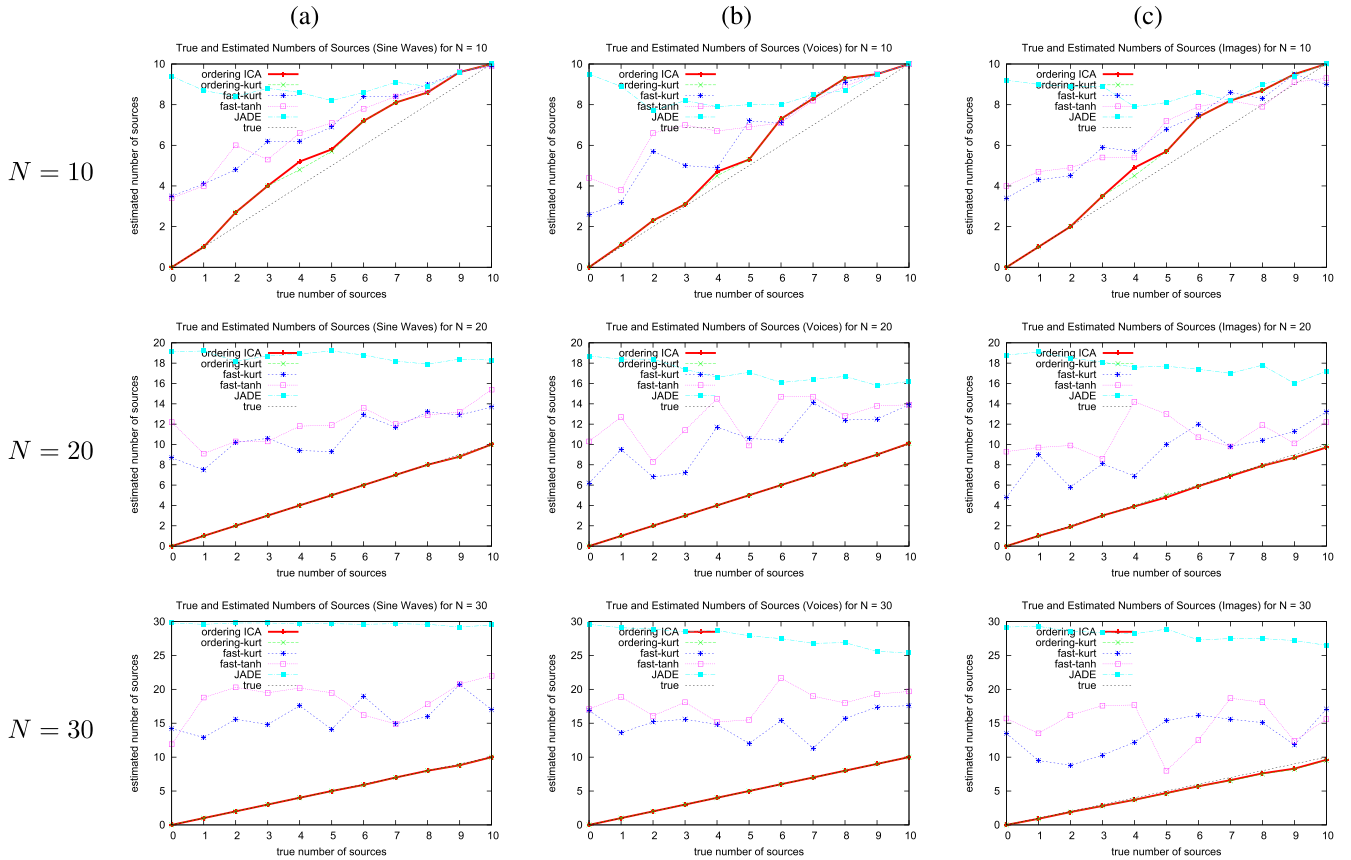


Fig. 5. Comparison of the ICA methods for estimating the number of non-Gaussian sources in various source separation problems (sine waves, voices, and images) for $N = 10$, $N = 20$, and $N = 30$. Each black dotted line corresponds to the true number of non-Gaussian sources, which was set from 0 to 10. (a) Sine waves. (b) Voices. (c) Images.

1) *Sine Waves*: Several sine waves were generated randomly by $\sin(100\theta_1 t + 2\pi\theta_2)$, where θ_1 and θ_2 were given according to the uniform distribution over $[0, 1)$. t ranged over $(0, 10]$ at intervals of 0.001. Therefore, the sample size M was 10000. These sine waves have an identical negative (sub-Gaussian) kurtosis whose theoretical value is -1.5 .

2) *Voice Separation*: The original data set consists of 160 voices of the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. The sequence of 20000 sampling points was extracted from the head for each voice (namely, $M = 20000$). Their kurtoses were always positive (super-Gaussian) around from 0.1 to 120 (their mean was 14.2).

3) *Image Separation*: The original data set consists of 44 images in the USC-SIPI image database (Volume 3; Miscellaneous). They were transformed into grayscale images of 256×256 pixels. Each pixel corresponds to a sample. M was set to $256 \times 256 = 65536$ (the total number of pixels). The distributions of the source images are quite diverse. Fig. 4 shows the histogram of the kurtoses of the 44 images. This histogram shows that there are various source distributions: 25 of the kurtoses are positive (super-Gaussian) and 19 of them are negative (sub-Gaussian).

The experiment with the following common setting was carried out for each data set. We set the number of non-Gaussian sources K from 0 to 10. K sources were selected randomly from the data set in each run of the experiment. The total number of signals N was set to 10, 20, and 30.

In other words, $N - K$ Gaussian sources (noises) were given. The $N \times N$ mixing matrix \mathbf{A} was randomly initialized in each run. Fig. 5 shows the averaged results of the estimated numbers of sources (K_{AIC}) over 10 runs for each data set. There seems to be no distinct difference among the three data sets. The accurate estimation is achieved if K_{AIC} is on the black dotted line. It shows that the ordering ICA could estimate K accurately especially for $K \ll N$ (strongly undercomplete). Even if K is near to N , the estimation was roughly appropriate. On the other hand, the fast ICA overestimated K with large fluctuations. JADE always estimated $K_{\text{AIC}} \simeq N$. In summary, the ordering ICA can estimate the number of non-Gaussian sources appropriately even if the sources consists of actual signals with various distributions. It is surprising that the proposed Gaussianity test is effective without any parameter setting for such a variety of sources and data sets.

VI. CONCLUSION

In this paper, we constructed a new objective function of the ICA named the AIF by applying the Gaussian approximation to the second-order polynomial feature space of sources. Various mathematical analysis guarantees that the AIF is an appropriate objective function of the ICA. In addition, we proposed a simple Gaussianity test in the feature space. We also constructed a new method named the ordering ICA by extending the fast ICA simply. It can estimate the number of non-Gaussian sources accurately and efficiently in

undercomplete problems. Experimental results verified that the ordering ICA is useful in various data sets, including both artificial and real problems. Our proposed method is promising for other practical applications, such as the analysis of fMRI data (needing to estimate the number of sources) and the multiple sound source tracking (needing to fix the permutation). The essential advantage of our approach based on the second-order polynomial feature space is that it has both simplicity and flexibility. Its simplicity leads to the simple Gaussianity test, and its flexibility leads to the applicability to various sources. We are planning to utilize the simplicity further by using other statistical techniques, such as a Bayesian approach. In addition, we are planning to improve our method in various ways, such as the postprocessing by the hyperbolic function instead of the kurtosis, the extension to the complex domain, and the utilization of the TIC.

Note again that our proposed model is not consistent with the true model. Nevertheless, the important properties for an objective function of the ICA are guaranteed mathematically and experimentally. We are planning to investigate further the appropriateness of our proposed model.

APPENDIX A PROOF OF (16)

First, \mathbf{W} is triangularized to $\mathbf{W} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$, where $\mathbf{\Lambda} = (\lambda_{ij})$ is an upper triangular matrix. Note that $|\mathbf{W}|$ is equal to $\prod_i \lambda_{ii}$. Second, the following equation is utilized:

$$\mathbf{E}(\mathbf{A} \otimes \mathbf{A})\mathbf{D}\mathbf{E}(\mathbf{B} \otimes \mathbf{B})\mathbf{D} = \mathbf{E}(\mathbf{A} \otimes \mathbf{A})(\mathbf{B} \otimes \mathbf{B})\mathbf{D} \quad (39)$$

where \mathbf{E} and \mathbf{D} are the elimination and duplication matrices defined in Section III-A. \mathbf{A} and \mathbf{B} are any $N \times N$ matrices. Equation (39) is proved by expanding each element on both sides. Then, the left-hand side of (16) is rewritten as

$$\begin{aligned} & |\mathbf{E}(\mathbf{W} \otimes \mathbf{W})\mathbf{D}| \\ &= |\mathbf{E}(\mathbf{P} \otimes \mathbf{P})\mathbf{D}||\mathbf{E}(\mathbf{P}^{-1} \otimes \mathbf{P}^{-1})\mathbf{D}||\mathbf{E}(\mathbf{\Lambda} \otimes \mathbf{\Lambda})\mathbf{D}| \\ &= |\mathbf{E}(\mathbf{\Lambda} \otimes \mathbf{\Lambda})\mathbf{D}| = \prod_{i,j \geq i} \lambda_{ii} \lambda_{jj}. \end{aligned} \quad (40)$$

As each λ_{ii} is multiplied $N+1$ times in $\prod_{i,j \geq i} \lambda_{ii} \lambda_{jj}$, the following equation holds:

$$\prod_{i,j \geq i} \lambda_{ii} \lambda_{jj} = \left(\prod_i \lambda_{ii} \right)^{N+1} = |\mathbf{W}|^{N+1}. \quad (41)$$

□

APPENDIX B PROOF OF THEOREM 1

First, \mathbf{x} is assumed to be prewhitened without loss of generality, because \mathbf{A} is invertible. Then, (22) is rewritten as

$$\begin{aligned} \Phi(\mathbf{\Sigma}) = & - \sum_i \log(3\sigma_{ii}^2 - 2\sigma_{ii} + 1) - \sum_{i,j \neq i} \frac{\sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2}{2} \\ & + (N+1) \log |\mathbf{\Sigma}| \end{aligned} \quad (42)$$

where $\mathbf{\Sigma} = (\sigma_{ij})$ is the covariance matrix of \mathbf{y} . We also utilize $\mathbf{\Sigma} = \mathbf{W}\mathbf{W}^T$, because \mathbf{x} is whitened. Equation (42) is rewritten

further as

$$\begin{aligned} & \Phi(\sigma_{11}, \dots, \sigma_{NN}, \tau_1, \dots, \tau_N) \\ &= - \sum_i \log(3\sigma_{ii}^2 - 2\sigma_{ii} + 1) - \frac{(\sum_i \tau_i)^2 + 2 \sum_i \tau_i^2}{2} \\ & \quad + \frac{3 \sum_i \sigma_{ii}^2}{2} + (N+1) \sum_i \log \tau_i \end{aligned} \quad (43)$$

where τ_i is the i th eigenvalue of $\mathbf{\Sigma}$. We also utilize $\sum_i \tau_i = \sum_i \sigma_{ii}$ and $\sum_i \tau_i^2 = \sum_{i,j} \sigma_{ij}^2$. The KKT condition shows that the following equation holds at the maximum of Φ of (43) with respect to every τ_i under the constraint $\sum_i \tau_i = \sum_i \sigma_{ii}$:

$$-2\tau_i + \frac{N+1}{\tau_i} = \sum_i \tau_i + \lambda \quad (44)$$

where λ is a Lagrange multiplier. Because the left-hand side of (44) does not depend on i and $-2\tau_i + (N+1/\tau_i)$ is monotonically decreasing, every τ_i value must be the same as the τ value. Therefore, $\mathbf{\Sigma}$ should be a diagonal matrix where $\sigma_{ij} = \tau \delta_{ij}$. Consequently, (43) is given as

$$\Phi(\tau) = - \log(3\tau^2 - 2\tau + 1) - \frac{(N-1)\tau^2}{2} + (N+1) \log \tau \quad (45)$$

where the constant factor N is neglected. It is shown that $\tau = 1$ is a solution maximizing $\Phi(\tau)$ because of $(\partial\Phi(\tau)/\partial\tau)|_{\tau=1} = 0$. Moreover, it is shown that $(\partial^2\Phi(\tau)/\partial\tau^2)$ is always negative. Thus, $\hat{\tau} = 1$ is the unique solution. □

APPENDIX C NECESSARY CONDITION ON THEOREM 2

Here, it is shown that the KKT condition of the constrained optimization in Theorem 2 is satisfied when $\mathbf{W} = \mathbf{A}^{-1}$. The following equation always holds:

$$E(y_i^2 y_j^2) = \sum_k b_{ik}^2 b_{jk}^2 \kappa_k + \sum_k b_{ik}^2 \sum_k b_{jk}^2 + 2 \left(\sum_k b_{ik} b_{jk} \right)^2 \quad (46)$$

where $\mathbf{B} = (b_{ij})$ is defined as $\mathbf{B} = \mathbf{W}\mathbf{A}$, and κ_k is the kurtosis of s_k . By $\sum_k b_{ik}^2 = 1$ for every i [derived from $E(y_i^2) = 1$], (22) is equivalent to

$$\begin{aligned} \Phi(\mathbf{B}) = & - \sum_i \left(\log \left(\sum_k b_{ik}^4 \kappa_k + 2 \right) - \frac{\sum_k b_{ik}^4 \kappa_k}{2} \right) \\ & - \frac{\sum_k (\sum_i b_{ik}^2)^2 \kappa_k}{2} - \sum_{i,j} \left(\sum_k b_{ik} b_{jk} \right)^2 \\ & + 2(N+1) \log |\mathbf{B}| \end{aligned} \quad (47)$$

where it is utilized that $\log |\mathbf{A}|$ is a constant. By using that $b_{ij} = \delta_{ij}$ is the solution of the ICA, the KKT condition is given as $(\partial\mathcal{L}(b_{ij}, \lambda_i)/\partial b_{ij})|_{b_{ij}=\delta_{ij}} = 0$ where

$$\mathcal{L} = \Phi(\mathbf{B}) + \sum_l \lambda_l \left(\sum_k b_{lk}^2 - 1 \right) \quad (48)$$

and λ_i is the i th Lagrange multiplier. It is rewritten as

$$\frac{\partial \mathcal{L}(b_{ij}, \lambda_i)}{\partial b_{ij}} \Big|_{b_{ij}=\delta_{ij}} = \left(-\frac{4\kappa_j}{\kappa_i + 2} + 2N - 2 + 2\lambda_i \right) \delta_{ij} = 0. \quad (49)$$

It is satisfied for

$$\hat{\lambda}_i = -N + 1 + \frac{2\kappa_i}{\kappa_i + 2}. \quad (50)$$

□

APPENDIX D

SUFFICIENT CONDITION ON THEOREM 2

Here, the convergence of the solution of the ICA in the constrained optimization in Theorem 2 is proved by using the bordered Hessian matrix. The $N^2 \times N^2$ Hessian matrix $\mathbf{H} = (h_{ij,pq})$ of \mathcal{L} of (48) at $b_{ij} = \delta_{ij}$ ($b_{pq} = \delta_{pq}$) and $\lambda_i = \hat{\lambda}_i$ of (50) is given as

$$\begin{aligned} & h_{ij,pq} \\ &= \frac{\partial^2 \mathcal{L}(\mathbf{B}, \lambda_i)}{\partial b_{ij} \partial b_{pq}} \Big|_{b_{ij}=\delta_{ij}, b_{pq}=\delta_{pq}, \lambda_i=\hat{\lambda}_i} \\ &= \begin{cases} -4N - 12 + \frac{16\kappa_i}{(\kappa_i + 2)^2} - \frac{8\kappa_i}{\kappa_i + 2} & (i = j = p = q) \\ -2N - 6 - 2\kappa_j + \frac{4\kappa_i}{\kappa_i + 2} & (i \neq j, i = p, j = q) \\ -2N - 6 & (i \neq j, i = q, j = p) \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (51)$$

Using the first derivative of each constraint $\sum_l b_{kl}^2 - 1$, an $N \times N^2$ matrix $\mathbf{G} = (g_{k,pq})$ is defined as

$$g_{k,pq} = \frac{\partial (\sum_l b_{kl}^2 - 1)}{\partial b_{pq}} \Big|_{b_{kq}=\delta_{kq}} = -2\delta_{kp}\delta_{kq}. \quad (52)$$

Consequently, the bordered Hessian matrix $\tilde{\mathbf{H}}$ is given as

$$\tilde{\mathbf{H}} = \begin{pmatrix} \mathbf{0} & \mathbf{G} \\ \mathbf{G}^T & \mathbf{H} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & -2\mathbf{I} & \mathbf{0} \\ -2\mathbf{I} & \mathbf{\Lambda}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Lambda}_2 \end{pmatrix}. \quad (53)$$

Here, \mathbf{I} is the $N \times N$ identity matrix. $\mathbf{0}$ is a matrix of zeros. $\mathbf{\Lambda}_1$ is a diagonal matrix, each diagonal element of which is $h_{ii,ii}$. $\mathbf{\Lambda}_2$ is a block diagonal matrix given as

$$\mathbf{\Lambda}_2 = \begin{pmatrix} \mathbf{J}^{12} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}^{13} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{J}^{(N-1)N} \end{pmatrix}, \quad (54)$$

each diagonal block of which [denoted by \mathbf{J}^{ij} ($i \neq j$)] is given as

$$\mathbf{J}^{ij} = \begin{pmatrix} h_{ij,ij} & h_{ij,ji} \\ h_{ji,ij} & h_{ji,ji} \end{pmatrix} \quad (55)$$

where $h_{ij,pq}$ is given in (51). Then, Lemma 4 holds.

Lemma 4: $\Delta(r)$ is defined as the determinant of the square submatrix consisting of the first r rows and the first r columns of $\tilde{\mathbf{H}}$. If $\kappa_i > -2$ holds for every i and $\kappa_i \neq 0$ holds

except at most one i , $(-1)^{N+k} \Delta(2N+k) > 0$ holds for $0 \leq k \leq N^2 - N$.

Proof: Here, we show only the outline of the proof. The rigorous mathematical induction is described in [7]. For $k = 0$, $\Delta(2N)$ is given as

$$\Delta(2N) = \begin{vmatrix} \mathbf{0} & -2\mathbf{I} \\ -2\mathbf{I} & \mathbf{\Lambda}_1 \end{vmatrix}. \quad (56)$$

This matrix is transformed into an upper triangular matrix whose diagonal elements are -2 by interchanging the rows suitably at N times. Therefore, this matrix is transformed into a matrix whose diagonal elements are the same and constant (-2) . Therefore, the following inequality holds:

$$(-1)^N \Delta(2N) = (-1)^N (-1)^N (-2)^{2N} > 0. \quad (57)$$

For $k > 0$, the following two inequalities holds:

$$h_{ij,ij} < -2N - 6 + 4 + 4 = -2N + 2 \leq 0 \quad (58)$$

because of $\kappa_i > -2$ and

$$\begin{aligned} |\mathbf{J}^{ij}| &= \frac{4(N-1)((\kappa_i + 2)\kappa_j^2 + (\kappa_j + 2)\kappa_i^2)}{(\kappa_i + 2)(\kappa_j + 2)} \\ &\quad + \frac{4(\kappa_i\kappa_j + 2\kappa_i + 2\kappa_j)^2}{(\kappa_i + 2)(\kappa_j + 2)} > 0 \end{aligned} \quad (59)$$

because at least one of κ_i and κ_j is not equal to 0. They mean that $\Delta(2N+k)$ is negative if k is odd and positive if k is even. Therefore, $\Delta(2N+k)(-1)^{N+k} > 0$ holds for any $k \geq 0$. □

The constrained optimization theory guarantees that a critical point of an objective function under some constraints is a local maximum if $(-1)^{N+k} \Delta(2N+k) > 0$ holds for every $k \geq 1$, which is called the bordered determinantal criterion (see [27]). This criterion is satisfied by Lemma 4. □

APPENDIX E

PROOF OF (24)

Under the orthonormality constraint of $\mathbf{y} = \mathbf{W}\mathbf{x}$ where $E(y_i) = 0$ and $E(y_i y_j) = \delta_{ij}$, \mathbf{x} is prewhitened without loss of generality and \mathbf{W} is constrained to be orthonormal. Therefore, $|\mathbf{W}|$ is a constant of 1. In addition, $\sum_i y_i^2 = \sum_i x_i^2$ always holds so that $\sum_{i,j} y_i^2 y_j^2$ is a constant. Thus, (18) is rewritten as

$$\Psi(\mathbf{W}, \boldsymbol{\alpha}) = \sum_i \left(-\log(\alpha_i + 2) + \left(\frac{1}{2} - \frac{1}{\alpha_i + 2} \right) E(y_i^4 - 1) \right) \quad (60)$$

where constant terms are neglected. □

APPENDIX F

PROOF OF THEOREM 3

Here, we show the outline of the proof. See [8] for the more rigorous proof. We employ Φ_i (free of α_i) instead of Ψ_i as the objective function and $\mathbf{B} = (b_{ij}) = \mathbf{W}\mathbf{A}$ instead of \mathbf{W} as the parameters. \mathbf{B} is orthonormal under the orthonormality constraint. Then, Φ_i of (25) is rewritten as

$$\Phi_i(b_{i1}, \dots, b_{iN}) = -\log(Z + 2) + \frac{Z}{2} \quad (61)$$

where

$$Z = \sum_k b_{ik}^4 \kappa_k. \quad (62)$$

Note that $Z > -2$ holds because of $\kappa_k > -2$. Now, Lemmas 5 and 6 hold.

Lemma 5: Φ_i is globally maximized if and only if Z is globally maximized or minimized.

Lemma 6: We assume that there is at least one $\kappa_k \neq 0$. Then, Z is globally maximized or minimized if and only if $b_{1p} = \pm 1$ and $b_{1q} = 0$ for $q \neq p$, where p corresponds to the largest or the smallest value of κ_p .

Lemma 5 is easily proved, because $\Phi_i(Z) = -\log(Z + 2) + Z/2$ is a convex function with respect to Z if $Z > -2$. Lemma 6 is easily proved by $\sum_k b_{1k}^2 = 1$. These lemmas guarantee that the p th source with the largest $\Phi_1(Z = \kappa_p)$ is extracted as the first component. They also guarantee that the q th source with the second largest $\Phi_2(Z = \kappa_q)$ is extracted under the orthonormality constraint. Therefore, the mathematical induction guarantees that all the non-Gaussian sources are extracted in the descending order of $\Upsilon(\kappa_k)$ if each Φ_i is globally maximized in the Gram–Schmidt orthonormalization. \square

REFERENCES

- [1] A. Cichocki and S.-I. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. Hoboken, NJ, USA: Wiley, 2002.
- [2] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. San Diego, CA, USA: Academic, 2010.
- [3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Hoboken, NJ, USA: Wiley, 2001.
- [4] T.-W. Lee, M. Girolami, and T. J. Sejnowski, “Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and supergaussian sources,” *Neural Comput.*, vol. 11, no. 2, pp. 417–441, 1999.
- [5] 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.
- [6] 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.
- [7] Y. Matsuda and K. Yamaguchi, “Adaptive objective function of ICA by Gaussian approximation in second-order polynomial feature space,” in *Proc. IJCNN*, Vancouver, BC, Canada, Jul. 2016, pp. 2382–2389.
- [8] Y. Matsuda and K. Yamaguchi, “Gram-schmidt orthonormalization to the adaptive ICA function for fixing the permutation ambiguity,” in *Proc. Int. Conf. Neural Inf. Process.*, 2016, pp. 152–159.
- [9] Y. Matsuda and K. Yamaguchi, “Efficient optimization of the adaptive ICA function with estimating the number of non-Gaussian sources,” in *Proc. Int. Conf. Latent Variable Anal. Signal Separat.*, 2017, pp. 469–478.
- [10] 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.
- [11] E. Oja and Z. Yuan, “The FastICA algorithm revisited: Convergence analysis,” *IEEE Trans. Neural Netw.*, vol. 17, no. 6, pp. 1370–1381, Nov. 2006.
- [12] V. Zarzoso and A. K. Nandi, “Blind separation of independent sources for virtually any source probability density function,” *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2419–2432, Sep. 1999.
- [13] J.-F. Cardoso, “High-order contrasts for independent component analysis,” *Neural Comput.*, vol. 11, no. 1, pp. 157–192, Jan. 1999.
- [14] P. Comon, “Independent component analysis, a new concept?” *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.
- [15] V. Zarzoso and P. Comon, “Robust independent component analysis by iterative maximization of the kurtosis contrast with algebraic optimal step size,” *IEEE Trans. Neural Netw.*, vol. 21, no. 2, pp. 248–261, Feb. 2010.
- [16] V. Zarzoso, P. Comon, and R. Phlypo, “A contrast function for independent component analysis without permutation ambiguity,” *IEEE Trans. Neural Netw.*, vol. 21, no. 5, pp. 863–868, May 2010.
- [17] H. Sawada, R. Mukai, S. Araki, and S. Makino, “A robust and precise method for solving the permutation problem of frequency-domain blind source separation,” *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.
- [18] T. Amishima, A. Okamura, S. Morita, and T. Kirimoto, “Permutation method for ICA separated source signal blocks in time domain,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 2, pp. 899–904, Apr. 2010.
- [19] Y.-O. Li, T. Adahi, and V. D. Calhoun, “Estimating the number of independent components for functional magnetic resonance imaging data,” *Human Brain Mapping*, vol. 28, no. 11, pp. 1251–1266, 2007. [Online]. Available: <http://dx.doi.org/10.1002/hbm.20359>
- [20] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. New York, NY, USA: Springer-Verlag, 2002.
- [21] Y. Matsuda and K. Yamaguchi, “An adaptive threshold in joint approximate diagonalization by assuming exponentially distributed errors,” *Neurocomputing*, vol. 74, no. 11, pp. 1994–2001, 2011.
- [22] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. New York, NY, USA: Cambridge Univ. Press, 2004.
- [23] F. R. Bach and M. I. Jordan, “Kernel independent component analysis,” *J. Mach. Learn. Res.*, vol. 3, pp. 1–48, Jan. 2002.
- [24] Z. Koldovský, P. Tichavský, 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.
- [25] H. Attias, “Independent factor analysis,” *Neural Comput.*, vol. 11, no. 4, pp. 803–851, 1999.
- [26] K. Chan, T. Lee, and T. Sejnowski, “Variational Bayesian learning of ICA with missing data,” *Neural Comput.*, vol. 15, no. 8, pp. 1991–2011, Aug. 2003.
- [27] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus With Applications in Statistics and Econometrics*. Hoboken, NJ, USA: Wiley, 1999.
- [28] S. Amari and A. Cichocki, “A new learning algorithm for blind signal separation,” in *Advances in Neural Information Processing Systems 8*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds. Cambridge, MA, USA: MIT Press, 1996, pp. 757–763.



Yoshitatsu Matsuda (M'08) received the Ph.D. degree from The University of Tokyo, Tokyo, Japan, in 2002.

He is currently a Research Fellow with The University of Tokyo. His current research interests include independent component analysis, massive data analysis, and self-organizing neural networks.



Kazunori Yamaguchi received the B.S., M.S., and D.Sc. degrees in information science from The University of Tokyo, Tokyo, Japan, in 1979, 1981, and 1985, respectively.

He is currently a Professor with the University of Tokyo. His current research interests include data models and data analysis.