Robust Independent Component Analysis via Minimum γ-Divergence Estimation

Pengwen Chen, Hung Hung, Osamu Komori, Su-Yun Huang, and Shinto Eguchi

***Abstract*—Independent component analysis (ICA) has been shown to be useful in many applications. However, most ICA methods are sensitive to data contamination. In this article we introduce a general minimum $\boldsymbol{U}$-divergence framework for ICA, which covers some standard ICA methods as special cases. Within the $\boldsymbol{U}$-family we further focus on the $\boldsymbol{\gamma}$-divergence due to its desirable property of super robustness for outliers, which gives the proposed method $\boldsymbol{\gamma}$-ICA. Statistical properties and technical conditions for recovery consistency of $\boldsymbol{\gamma}$-ICA are studied. In the limiting case, it improves the recovery condition of MLE-ICA known in the literature by giving necessary and sufficient condition. Since the parameter of interest in $\boldsymbol{\gamma}$-ICA is an orthogonal matrix, a geometrical algorithm based on gradient flows on special orthogonal group is introduced. Furthermore, a data-driven selection for the $\boldsymbol{\gamma}$ value, which is critical to the achievement of $\boldsymbol{\gamma}$-ICA, is developed. The performance, especially the robustness, of $\boldsymbol{\gamma}$-ICA is demonstrated through experimental studies using simulated data and image data.**

***Index Terms*—$\boldsymbol{\beta}$-divergence, $\boldsymbol{\gamma}$-divergence, geodesic, minimum divergence estimation, robust statistics, special orthogonal group.**

I. Introduction

CONSIDER the following generative model for independent component analysis (ICA)

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{\mu}, \tag{1}$$

where the elements of the non-Gaussian source vector $\boldsymbol{S} = (S_1, \ldots, S_p)^\top \in \mathbb{R}^p$ are mutually independent with zero mean, $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ is an unknown nonsingular mixing matrix, $\boldsymbol{X} \in \mathbb{R}^p$ is the observable signal, and $\boldsymbol{\mu} = E(\boldsymbol{X}) \in \mathbb{R}^p$ is a shift parameter. Let $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$ be the whitened data of $\boldsymbol{X}$, where $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{X})$. An equivalent expression of (1) is

$$\boldsymbol{Z} = \boldsymbol{A}^*\boldsymbol{S}, \tag{2}$$

Manuscript received October 03, 2012; revised December 18, 2012; accepted February 02, 2013. Date of publication February 13, 2013; date of current version July 15, 2013. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Shiro Ikeda.

P. Chen is with the Department of Applied Mathematics, National Chung Hsing University, Taichung 402, Taiwan (e-mail: pengwen@nchu.edu.tw).

H. Hung is with the Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei 10055, Taiwan (e-mail: hhung@ntu.edu.tw).

O. Komori is with the School of Statistical Thinking, Institute of Statistical Mathematics, Tachikawa 190-8562, Japan (e-mail: komori@ism.ac.jp).

S.-Y. Huang is with the Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan (e-mail: syhuang@stat.sinica.edu.tw).

S. Eguchi is with the Institute of Statistical Mathematics and Graduate University of Advanced Studies, Tachikawa 190-8562, Japan (e-mail: eguchi@ism.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/JSTSP.2013.2247024

where $\boldsymbol{A}^* = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}$ is the mixing matrix in $\boldsymbol{Z}$-scale. It is reported in literature that prewhitening the data usually makes the ICA inference procedure more stable [1]. In the rest of discussion, we will work on model (2) to estimating the mixing matrix $\boldsymbol{A}^*$ based on the prewhitened $\boldsymbol{Z}$. It is easy to transform $\boldsymbol{A}^*$ back to the original $\boldsymbol{X}$-scale via $\boldsymbol{A} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{A}^*$. Note that both $\boldsymbol{A}^*$ and $\boldsymbol{S}$ are unknown, and there exists the problem of non-identifiability [2]. This can be seen from the fact that $\boldsymbol{Z} = \boldsymbol{A}^*\boldsymbol{S} = (\boldsymbol{A}^*\boldsymbol{M})(\boldsymbol{M}^{-1}\boldsymbol{S})$ for any nonsingular diagonal matrix $\boldsymbol{M}$. To make $\boldsymbol{A}^*$ identifiable (up to permutation and sign ambiguities), we assume the following conditions for $\boldsymbol{S}$:

$$E(\boldsymbol{S}) = \boldsymbol{0} \quad \text{and} \quad \mathrm{cov}(\boldsymbol{S}) = \boldsymbol{I}_p, \tag{3}$$

where $\boldsymbol{I}_p \in \mathbb{R}^{p\times p}$ is the identity matrix. It then implies that $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^\top$ and

$$\boldsymbol{I}_p = \mathrm{cov}(\boldsymbol{Z}) = \boldsymbol{A}^*\,\mathrm{cov}(\boldsymbol{S})\,\boldsymbol{A}^{*\top} = \boldsymbol{A}^*\boldsymbol{A}^{*\top}, \tag{4}$$

which means that the mixing matrix $\boldsymbol{A}^*$ in $\boldsymbol{Z}$-scale is orthogonal. Let $\mathcal{O}_p$ be the space of orthogonal matrices in $\mathbb{R}^{p\times p}$. Note that, if $\boldsymbol{A}^* \in \mathcal{O}_p$ is a parameter of model (2), so is $-\boldsymbol{A}^* \in \mathcal{O}_p$. Thus, to fix one direction, we restrict $\boldsymbol{A}^* \in \mathcal{SO}_p$, where $\mathcal{SO}_p \subset \mathcal{O}_p$ consists of orthogonal matrices with determinant one. The set $\mathcal{SO}_p$ is called the special orthogonal group. The main purpose of ICA can thus be formulated as estimating the orthogonal $\boldsymbol{A}^* \in \mathcal{SO}_p$ based on the whitened data $\{\boldsymbol{z}_i\}_{i=1}^n$, the random copies of $\boldsymbol{Z}$, or equivalently, looking for a recovering matrix $\boldsymbol{W} \in \mathcal{SO}_p$ so that components in

$$\boldsymbol{Y} := \boldsymbol{W}^\top\boldsymbol{Z} = (\boldsymbol{w}_1^\top\boldsymbol{Z}, \ldots, \boldsymbol{w}_p^\top\boldsymbol{Z})^\top$$

have the maximum degree of independence, where $\boldsymbol{w}_j$ is the j-th column of $\boldsymbol{W}$. In the latter case, $\boldsymbol{W}$ provides an estimate of $\boldsymbol{A}^*$, and $\boldsymbol{Y}$ provides an estimate of $\boldsymbol{S}$.

We first briefly review some existing methods for ICA. One idea is to estimate $\boldsymbol{W}$ via *minimizing the mutual information*. Let $g_{\boldsymbol{Y}}$ be the joint probability density function of $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top$, and g_{Y_j} be the marginal probability density function of Y_j. The mutual information of random variables $(Y_1, \ldots, Y_p)$, denoted by $I(Y_1, \ldots, Y_p)$, is

$$I(Y_1, \ldots, Y_p) := \sum_{j=1}^{p} H(Y_j) - H(\boldsymbol{Y}), \tag{5}$$

where $H(\boldsymbol{Y}) = -\int g_{\boldsymbol{Y}} \ln g_{\boldsymbol{Y}}$ and $H(Y_j) = -\int g_{Y_j} \ln g_{Y_j}$ are the Shannon entropy. Ideally, if $\boldsymbol{W}$ is properly chosen so that $\boldsymbol{Y}$ has independent components, then $g_{\boldsymbol{Y}} = \prod_j g_{Y_j}$ and, hence,

$I(Y_1,\ldots,Y_p) = 0$. Thus, via minimizing $I(Y_1,\ldots,Y_p)$ with respect to $\boldsymbol{W}$, it leads to an estimate of $\boldsymbol{W}$. Another method is to estimate $\boldsymbol{W}$ via *maximizing the negentropy*, which is equivalent to minimizing the mutual information as described below. The negentropy of $\boldsymbol{Y}$ is defined to be

$$J(\boldsymbol{Y}) := H(\boldsymbol{Y}') - H(\boldsymbol{Y}), \tag{6}$$

where $\boldsymbol{Y}'$ is a Gaussian random vector having the same covariance matrix as $\boldsymbol{Y}$ [3]. Firstly, it can be deduced that

$$I(Y_1,\ldots,Y_p) = J(\boldsymbol{Y}) - \sum_{j=1}^{p} J(Y_j), \tag{7}$$

where the equality holds since $H(\boldsymbol{Y}') = \sum_{j=1}^{p} H(Y_j')$ due to $\mathrm{cov}(\boldsymbol{Y}') = \mathrm{cov}(\boldsymbol{Y}) = \boldsymbol{I}_p$. Moreover, as $\boldsymbol{Y} = \boldsymbol{W}^\top \boldsymbol{Z}$ with $\boldsymbol{W} \in \mathcal{SO}_p$, we have $J(\boldsymbol{Y}) = J(\boldsymbol{Z})$, which does not depend on $\boldsymbol{W}$. That is, the negentropy is invariant under orthogonal transformation. It concludes that minimizing $I(Y_1,\ldots,Y_p)$ is equivalent to maximizing the negentropy $\sum_{j=1}^{p} J(Y_j)$. The negentropy $J(Y_j)$, however, involves the unknown density g_{Y_j}. To avoid nonparametric estimation of g_{Y_j}, one can use the approximation [4] via a non-quadratic contrast function $G(\cdot)$,

$$J(Y_j) \approx J_G(Y_j) := [E\{G(Y_j)\} - E\{G(\nu)\}]^2, \tag{8}$$

where ν is a random variable having the standard normal distribution. Here J_G can be treated as a measure of non-Gaussianity, and minimizing the sample analogue of $J_G(Y_j)$ to search $\boldsymbol{W}$ corresponds to fast-ICA [5].

Another widely used estimation criterion for $\boldsymbol{W}$ is via *maximizing the likelihood*. Consider the model

$$g_{Y_j} = f_j, \quad j = 1,\ldots,p, \tag{9}$$

where f_j's are parametric density functions. Possible choices for f_j include $f_j(s) \propto \exp(-cs^4)$ for sub-Gaussian model, and $f_j(s) \propto 1/\cosh(cs)$ for super-Gaussian model. Define

$$f(\boldsymbol{W}^\top \boldsymbol{z}) = \prod_{j=1}^{p} f_j(\boldsymbol{w}_j^\top \boldsymbol{z}). \tag{10}$$

Then, under model (9) and when $\boldsymbol{W} \in \mathcal{SO}_p$ recovers all independent sources, the density function of $\boldsymbol{Z}$ takes the form

$$f_{\boldsymbol{Z}}(\boldsymbol{z};\boldsymbol{W}) = |\det(\boldsymbol{W})| \cdot \prod_{j=1}^{p} f_j(\boldsymbol{w}_j^\top \boldsymbol{z}) = f(\boldsymbol{W}^\top \boldsymbol{z}). \tag{11}$$

Let $\mathcal{D}_0(g,f) = \int g \ln(g/f)$ be the Kullback-Leibler divergence (KL-divergence). The MLE-ICA then searches $\boldsymbol{W}$ via

$$\underset{\boldsymbol{W}\in\mathcal{SO}_p}{\operatorname{argmin}}\, \mathcal{D}_0\left(g_{\boldsymbol{Z}}, f_{\boldsymbol{Z}}(\cdot;\boldsymbol{W})\right), \tag{12}$$

where $g_{\boldsymbol{Z}}$ is the true probability density function of $\boldsymbol{Z}$. The sample analogue is then obtained by replacing $g_{\boldsymbol{Z}}$ by $\widehat{g}_{\boldsymbol{Z}}$, the empirical distribution of $\{\boldsymbol{z}_i\}_{i=1}^n$.

There exist other ICA procedures that are not covered in the above review. The joint approximate diagonalization of eigenmatrices (JADE) is a cumulant-based ICA method [6]. Instead of considering the modeling (9), approximation of the density function g_{Y_j} for MLE-ICA is proposed [7]. We also refer to [8] and the references therein for the ICA problem from an information geometry perspective and the corresponding learning algorithms.

As will become clear later that the above reviewed methods are related to *minimizing the KL-divergence*, which is not robust in the presence of outliers. Outliers, however, frequently appear in real data analysis, and a robust ICA procedure becomes urgent. For the purpose of robustness, instead of using KL-divergence, Minami and Eguchi [9] propose β-ICA by considering the *minimum β-divergence* estimation. On the other hand, the γ-divergence is shown to be super robust against data contamination [10]. We are therefore motivated to focus on *minimum γ-divergence* estimation to propose a robust ICA procedure, called γ-ICA. It is also important to investigate the consistency property of the proposed γ-ICA. Hyvärinen, Karhnen and Oja (page 206 in [11]) have provided a sufficient condition for the modeling (9) to ensure the validity of MLE-ICA when $\boldsymbol{W} \in \mathcal{SO}_p$, in the sense of being able to recover all independent components. Amari, Chen, and Cichocki [12] studied necessary and sufficient conditions for recovery consistency under a different constraint on $\boldsymbol{W}$, and this consistency result is further extended to the case of β-ICA [9]. In this work, we derive necessary and sufficient conditions regarding the modeling (9) for the recovery consistency of γ-ICA. In the limiting case $\gamma \to 0$, our necessary and sufficient condition improves the result of [11] (page 206) for MLE-ICA. To the best of our knowledge, this result is not explored in existing literature.

Some notations are defined here for reference. For any $\boldsymbol{M} \in \mathbb{R}^{p\times p}$, let $\boldsymbol{K}_p \in \mathbb{R}^{p^2\times p^2}$ be the commutation matrix such that $\mathrm{vec}(\boldsymbol{M}^\top) = \boldsymbol{K}_p \mathrm{vec}(\boldsymbol{M})$, where $\mathrm{vec}(\boldsymbol{M})$ stacks the columns of $\boldsymbol{M}$ into a long vector; $\boldsymbol{M} > 0$ (resp. < 0) means $\boldsymbol{M}$ is strictly positive (resp. negative) definite; and $\exp(\boldsymbol{M}) := \sum_{k=0}^{\infty} (1/k!)\boldsymbol{M}^k$ is the matrix exponential. Note that $\det(\exp(\boldsymbol{M})) = \exp(\mathrm{tr}(\boldsymbol{M}))$ for any nonsingular $\boldsymbol{M}$. For a lower triangular matrix $\boldsymbol{M}$ with 0 diagonals, $\mathrm{vecp}(\boldsymbol{M})$ stacks the nonzero elements of the columns of $\boldsymbol{M}$ into a vector with length $p(p-1)/2$. There exist matrices $\boldsymbol{P} \in \mathbb{R}^{p(p-1)/2\times p^2}$ and $\boldsymbol{Q} \in \mathbb{R}^{p^2\times p(p-1)/2}$ such that $\mathrm{vecp}(\boldsymbol{M}) = \boldsymbol{P}\mathrm{vec}(\boldsymbol{M})$ and $\mathrm{vec}(\boldsymbol{M}) = \boldsymbol{Q}\mathrm{vecp}(\boldsymbol{M})$. Each column vector of $\boldsymbol{Q}$ is of the form $(\boldsymbol{e}_i \otimes \boldsymbol{e}_j), i < j$, where $\boldsymbol{e}_i \in \mathbb{R}^p$ is a vector with a one in the i-th position and 0 elsewhere, and $\otimes$ is the Kronecker product. $\mathbf{1}_p$ is the p-vector of ones. For a function h, h' is the differential of h. Matrices and vectors are in bold letters.

The rest of this paper is organized below. A unified framework for ICA problem by minimum divergence estimation is introduced in Section II. A robust γ-ICA procedure is developed in Section III, wherein the related statistical properties are studied. A geometrical implementation algorithm for γ-ICA is illustrated in Section IV. In Section V, the issue of selecting the γ value is discussed. Numerical studies are conducted in Section VI to show the robustness of γ-ICA. The paper ends with a conclusion in Section VII. All the proofs are placed in Appendix.

II. Minimum $\boldsymbol{U}$-Divergence Estimation for ICA

The aim of ICA is understood to search a matrix $\boldsymbol{W} \in \mathcal{SO}_p$ so that the joint probability density function $g_{\boldsymbol{Y}}$ of $\boldsymbol{Y}$ is as close

to the marginal product $\prod_j g_{Y_j}$ as possible. It motivates estimating $\boldsymbol{W}$ by minimizing a divergence between $g_{\boldsymbol{Y}}$ and $\prod_j g_{Y_j}$. A general estimation scheme for $\boldsymbol{W}$ can then be formulated as the minimization problem

$$\min_{\boldsymbol{W}\in\mathcal{SO}_p} \mathcal{D}\left(g_{\boldsymbol{Y}}, \prod_j g_{Y_j}\right), \tag{13}$$

where $\mathcal{D}(\cdot,\cdot)$ denotes a divergence function. Starting from (13), different choices of $\mathcal{D}$ will lead to different estimation criteria for ICA. Here we will consider the class of U-divergence ([13], [14]) as described below.

The U-divergence is a general class of divergence functions. Consider a strictly convex function $U(t)$ defined on $\mathbb{R}$, or on some interval of $\mathbb{R}$ where $U(t)$ is well-defined. Let $\xi = (U')^{-1}$. The U-divergence is defined to be

$$\begin{aligned}\mathcal{D}_U(g,f) &= \int U(\xi(f)) - U(\xi(g)) - U'(\xi(g))\cdot\{\xi(f)-\xi(g)\} \\ &= \int U(\xi(f)) - U(\xi(g)) - g\cdot\{\xi(f)-\xi(g)\},\end{aligned}$$

which defines a mapping from $\mathcal{M}_U \times \mathcal{M}_U$ to $[0,\infty)$, where $\mathcal{M}_U = \left\{f : \int U(\xi(f)) < \infty,\ f \geq 0\right\}$. Define the U-cross entropy by $C_U(g,f) = -\int \xi(f)g + \int U(\xi(f))$ and the U-entropy by $H_U(g) = C_U(g,g)$. Then the U-divergence can be written as $\mathcal{D}_U(g,f) = C_U(g,f) - H_U(g) \geq 0$. In the subsequent sections, we introduce some special cases of U-divergence that correspond to different ICA methods.

A. KL-Divergence

By taking the (U,ξ) pair to be

$$U(t) = \exp(t), \qquad \xi(t) = \ln t,$$

the corresponding U-divergence is equivalent to the KL-divergence D_0. In this case, it can be deduced that

$$\mathcal{D}_0\left(g_{\boldsymbol{Y}}, \prod_j g_{Y_j}\right) = I(Y_1,\ldots,Y_p),$$

where $I(Y_1,\ldots,Y_p)$ is the mutual information defined in (5). As described in Section I that, up to a constant term, $I(Y_1,\ldots,Y_p) \propto -\sum_{j=1}^p J(Y_j) \approx -\sum_{j=1}^p J_G(Y_j)$, we conclude that the following criteria, minimum mutual information, maximum negentropy, and fast-ICA, are all special cases of (13). On the other hand, observe that

$$\mathcal{D}_0\left(g_{\boldsymbol{Y}}, \prod_j g_{Y_j}\right) = \mathcal{D}_0\left(g_{\boldsymbol{Z}}, |\det(\boldsymbol{W})| \prod_j g_{Y_j}(\boldsymbol{w}_j^\top \cdot)\right). \tag{14}$$

If we consider the model (9) and $\boldsymbol{W} \in \mathcal{SO}_p$, and if we estimate $g_{\boldsymbol{Z}}$ by $\widehat{g}_{\boldsymbol{Z}}$, minimizing (14) is equivalent to MLE-ICA in (12).

B. $\boldsymbol{\beta}$-Divergence

Define $\mathcal{M}_{\beta+1} := \left\{f : \int f^{\beta+1} < \infty,\ f \geq 0\right\}$ which is convex. Take the (U,ξ) pair to be

$$U(t) = \frac{(1+\beta t)^{(\beta+1)/\beta}}{(1+\beta)}, \quad \xi(t) = \frac{(t^\beta - 1)}{\beta}.$$

The resulting U-divergence defined on $\mathcal{M}_{\beta+1}\times\mathcal{M}_{\beta+1}$ is

$$\mathcal{B}_\beta(g,f) = \frac{1}{\beta}\int (g^\beta - f^\beta)g - \frac{1}{\beta+1}\int (g^{\beta+1} - f^{\beta+1}) \tag{15}$$

which is called β-divergence [9], or density power divergence [15]. Note that $\mathcal{B}_\beta(g,f) = 0$ if and only if $f = \lambda g$ for some $\lambda > 0$. In the limiting case $\lim_{\beta\to 0}\mathcal{B}_\beta = \mathcal{D}_0$, which gives the KL-divergence. Without considering the orthogonality constraint on $\boldsymbol{W}$, replacing $\mathcal{D}_0$ in (14) by $\mathcal{B}_\beta$ and using the model (9) give (up to a constant term) the quasi β-likelihood

$$|\det(\boldsymbol{W})|^\beta \left\{\int f^\beta(\boldsymbol{W}^\top \boldsymbol{z}) g_{\boldsymbol{Z}}(\boldsymbol{z}) d\boldsymbol{z} - c_\beta\right\}, \tag{16}$$

where $c_\beta = (\beta/(\beta+1))\prod_{j=1}^p \int f_j^{\beta+1}(s)ds$ is a constant, and $f(\cdot)$ is defined in (10). The β-ICA [9] searches $\boldsymbol{W}$ via maximizing the sample analogue of (16) by replacing $g_{\boldsymbol{Z}}$ with $\widehat{g}_{\boldsymbol{Z}}$.

C. $\boldsymbol{\gamma}$-Divergence

The γ-divergence can be obtained from β-divergence through a U-volume normalization as

$$\mathcal{D}_\gamma(g,f) := \mathcal{B}_\gamma\left(\alpha(g)\cdot g, \alpha(f)\cdot f\right),$$

where $\mathcal{B}_\gamma$ is defined the same way as (15) with the plug-in $\beta = \gamma$, and where $\alpha(f)$ is some normalizing constant. Here we adopt the volume-mass-one normalization

$$\int U\big(\xi(\alpha(f)\cdot f(x))\big)dx = 1.$$

It leads to $\alpha(f) = (\gamma+1)^{1/(\gamma+1)}\|f\|_{\gamma+1}^{-1}$, where $\|f\|_{\gamma+1} = \{\int f^{\gamma+1}(x)dx\}^{1/(\gamma+1)}$. Then, we have

$$\mathcal{D}_\gamma(g,f) = \frac{\gamma+1}{\gamma}\left\{1 - \int\left(\frac{f(x)}{\|f\|_{\gamma+1}}\right)^\gamma \frac{g(x)}{\|g\|_{\gamma+1}}dx\right\}. \tag{17}$$

It can be seen that γ-divergence is scale invariant. Moreover, $\mathcal{D}_\gamma(g,f) = 0$ if and only if $f = \lambda g$ for some $\lambda > 0$. The γ-divergence, indexed by a power parameter γ, is a generalization of KL-divergence. In the limiting case $\lim_{\gamma\to 0}\mathcal{D}_\gamma = \mathcal{D}_0$, it gives the KL-divergence.

Due to its super robustness for outliers, we adopt the γ-divergence to propose γ-ICA by replacing $\mathcal{D}_0$ in (14) with $\mathcal{D}_\gamma$. Similar to the derivation of (16), under model (9) and without considering the orthogonality constraint on $\boldsymbol{W}$, the objective function of γ-ICA being maximized is

$$|\det(\boldsymbol{W})|^{\gamma/(\gamma+1)} \left\{\int f^\gamma(\boldsymbol{W}^\top \boldsymbol{z}) g_{\boldsymbol{Z}}(\boldsymbol{z}) d\boldsymbol{z}\right\}, \tag{18}$$

which is different from (16), but is similar when c_β is small. This confirms the observation of [9] that setting $c_\beta = 0$ does not affect the performance of β-ICA. It should be emphasized that the quasi β-likelihood (16) is not guaranteed to be positive, and we found in our simulation studies that β-ICA maximizing (16) suffers the problem of numerical instability. On the other hand, the quasi γ-likelihood (18) is always positive for any γ value. Interestingly, γ-ICA and β-ICA are equivalent if we consider the orthogonality constraint. Obviously, when $\boldsymbol{W} \in \mathcal{SO}_p$, $|\det(\boldsymbol{W})| = 1$ and maximizing (16) is equivalent to maximizing (18). Note that the constraint $\boldsymbol{W} \in \mathcal{SO}_p$ is a consequence of

prewhitening, and it is reported in literature that prewhitening usually makes the ICA learning process more stable [1]. We are therefore motivated to consider γ-ICA with $\boldsymbol{W} \in \mathcal{SO}_p$ based on the prewhitened $\boldsymbol{Z}$. Detailed inference procedure and statistical properties of γ-ICA are investigated in next section.

III. The γ-ICA Inference Procedure

The considered ICA problem is a two-stage process consisting of prewhitening and estimation stages. Since our aim is to develop a robust ICA procedure, the robustness for both stages should be guaranteed. Here we utilize the γ-divergence to introduce a robust γ-prewhitening, followed by illustrating γ-ICA based on the γ-prewhitened data. In practice, the γ value for γ-divergence should be determined. We assume γ is given in this section, and leave its selection to Section V.

A. γ-Prewhitening

Although prewhitening is always possible by a straightforward standardization of $\boldsymbol{X}$, there exists the issue of robustness of such a whitening procedure. It is known that empirical moment estimates of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are not robust. In [1], the authors proposed a robust β-prewhitening procedure. In particular, let $\xi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ be the probability density function of p-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and let $\widehat{g}_{\boldsymbol{X}}$ be the empirical distribution of $\{\boldsymbol{x}_i\}_{i=1}^n$. With a given β, Mollah *et al.* [1] considered

$$(\widehat{\kappa}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = \underset{\kappa,\boldsymbol{\mu},\boldsymbol{\Sigma}}{\operatorname{argmin}} \mathcal{B}_\beta(\widehat{g}_{\boldsymbol{X}}, \kappa \cdot \xi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}), \tag{19}$$

and then suggested to use $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ for whitening the data, which is called β-prewhitening. Interestingly, $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ from (19) can also be derived from the minimum γ-divergence as

$$(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = \underset{\boldsymbol{\mu},\boldsymbol{\Sigma}}{\operatorname{argmin}} \mathcal{D}_\gamma(\widehat{g}_{\boldsymbol{X}}, \xi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}) \tag{20}$$

when $\gamma = \beta$. At the stationarity of (20), $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ will satisfy

$$\widehat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^n d_i^\gamma(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) \cdot \boldsymbol{x}_i}{\sum_{i=1}^n d_i^\gamma(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})} \quad \text{and}$$

$$\widehat{\boldsymbol{\Sigma}} = (1+\gamma) \cdot \frac{\sum_{i=1}^n d_i^\gamma(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) \cdot (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^\top}{\sum_{i=1}^n d_i^\gamma(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})},$$

where $d_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \exp\left\{-(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})/2\right\}$. The robustness property of $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ can be found in [1]. We call the prewhitening procedure

$$\boldsymbol{z}_i = \widehat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}), \quad i = 1, \ldots, n, \tag{21}$$

the γ-prewhitening. The whitened data $\{\boldsymbol{z}_i\}_{i=1}^n$ then enter the γ-ICA estimation procedure.

B. Estimation of γ-ICA

We are now in the position to develop our γ-ICA based on the γ-prewhitened data $\{\boldsymbol{z}_i\}_{i=1}^n$. As discussed in Section II-C, under the modeling (9), γ-ICA aims to estimate $\boldsymbol{W}$ via

$$\widehat{\boldsymbol{W}} = \underset{\boldsymbol{W} \in \mathcal{SO}_p}{\operatorname{argmin}} \mathcal{D}_\gamma(\widehat{g}_{\boldsymbol{Z}}, f_{\boldsymbol{Z}}(\cdot; \boldsymbol{W}))$$

where $f_{\boldsymbol{Z}}(\boldsymbol{z}; \boldsymbol{W})$ is defined in (11). Equivalently, paralleling to the derivation of (18) and using $\boldsymbol{W} \in \mathcal{SO}_p$, $\widehat{\boldsymbol{W}}$ can also be obtained via

$$\widehat{\boldsymbol{W}} = \underset{\boldsymbol{W} \in \mathcal{SO}_p}{\operatorname{argmax}} \mathcal{L}(\boldsymbol{W}) := \underset{\boldsymbol{W} \in \mathcal{SO}_p}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^n f^\gamma(\boldsymbol{W}^\top \boldsymbol{z}_i), \tag{22}$$

where $f(\cdot)$ is defined in (10). We remind the readers that $\mathcal{L}(\boldsymbol{W})$ is just the sample analogue of (18) by replacing $g_{\boldsymbol{Z}}$ with $\widehat{g}_{\boldsymbol{Z}}$, under the constraint $\boldsymbol{W} \in \mathcal{SO}_p$. With $\widehat{\boldsymbol{W}}$, the mixing matrix $\boldsymbol{A}$ is then estimated by $\widehat{\boldsymbol{A}} = \widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{W}}$. Let

$$\boldsymbol{\phi}(\boldsymbol{W}^\top \boldsymbol{z}) := \left[\phi_1(\boldsymbol{w}_1^\top \boldsymbol{z}), \ldots, \phi_p(\boldsymbol{w}_p^\top \boldsymbol{z})\right]^\top$$

with $\phi_j(y) = f_j'(y)/f_j(y)$. We have the following proposition.

Proposition 1: At the stationarity, $\widehat{\boldsymbol{W}}$ in (22) will satisfy

$$\frac{1}{n} \sum_{i=1}^n f^\gamma(\widehat{\boldsymbol{W}}^\top \boldsymbol{z}_i) \boldsymbol{\Phi}(\widehat{\boldsymbol{W}}, \boldsymbol{z}_i) = 0$$

with $\boldsymbol{\Phi}(\boldsymbol{W}, \boldsymbol{z}) = \boldsymbol{W}^\top \boldsymbol{z} \left[\boldsymbol{\phi}(\boldsymbol{W}^\top \boldsymbol{z})\right]^\top - \boldsymbol{\phi}(\boldsymbol{W}^\top \boldsymbol{z}) \left[\boldsymbol{W}^\top \boldsymbol{z}\right]^\top$.

From Proposition 1, it can be seen the robustness nature of γ-ICA: the stationary equation is a weighted sum with the weight function f^γ. When $\gamma > 0$, an outlier with extreme value will contribute less to the stationary equation. In the limiting case of $\gamma \to 0$, which corresponds to MLE-ICA, the weight f^γ becomes uniform and, hence, is not robust.

C. Consistency of γ-ICA

A critical step to the likelihood-based ICA method is the modeling (9) for g_{Y_j}, and it is important to investigate conditions of f_j under which ICA procedure is consistent. Here the ICA consistency means recovery consistency. An ICA procedure is said to be recovery consistent if it is able to recover all independent components, that is, the separation solutions are the (local) maximum of the objective function. A sufficient condition for the consistency of MLE-ICA can be found in [11] (page 206). Notably, the consistency of MLE-ICA does not rely on the correct specification of f_j, but only on the positivity of $E[\phi_j(S_j)S_j - \phi_j'(S_j)]$. This subsection investigates the recovery consistency of γ-ICA defined in (22). The main result is summarized below. We refer to the end of Section I for the definitions of $\boldsymbol{Q}$ and $\boldsymbol{K}_p$.

Theorem 1: Assume the ICA model (2) and the modeling (9). Assume the existence of $\Gamma = (0, \tau]$ for some $\tau > 0$ such that

(A) $E[f_j^\gamma(S_j)S_j] = 0, \forall\, \gamma \in \Gamma, j = 1, \ldots, p.$

Then, for $\gamma \in \Gamma$, the associated γ-ICA is recovery consistent if and only if $\boldsymbol{\Psi}_\gamma < 0$, where

$$\boldsymbol{\Psi}_\gamma = \boldsymbol{Q}^\top \left(\boldsymbol{I}_{p^2} - \boldsymbol{K}_p\right) \left\{\gamma \boldsymbol{\Psi}_{(1)} + \gamma^2 \boldsymbol{\Psi}_{(2)}\right\} \left(\boldsymbol{I}_{p^2} - \boldsymbol{K}_p\right) \boldsymbol{Q},$$

$\boldsymbol{\Psi}_{(1)} = \sum_{j=1}^{p}(\boldsymbol{e}_j\boldsymbol{e}_j^\top \otimes \boldsymbol{U}_j) - (\boldsymbol{D} \otimes \boldsymbol{I}_p)$, $\boldsymbol{U}_j = \mathrm{diag}(u_{j1}, \ldots, u_{jp})$ *with* $u_{jk} = E[f^\gamma(\boldsymbol{S})\phi_j'(S_j)S_k^2]$, $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_p)$ *with* $d_j = E[f^\gamma(\boldsymbol{S})\phi_j(S_j)S_j]$, *and* $\boldsymbol{\Psi}_{(2)} = E[f^\gamma(\boldsymbol{S})\{\boldsymbol{\phi}(\boldsymbol{S}) \otimes \boldsymbol{S}\}\{\boldsymbol{\phi}(\boldsymbol{S}) \otimes \boldsymbol{S}\}^\top]$.

Condition (A) of Theorem 1 can be treated as a weighted version of $E(S_j) = 0$. It is satisfied when S_j is symmetrically distributed about zero, and when the model probability density function f_j is an even function. We believe condition (A) is not restrictive and should be approximately valid in practice. Notice that $\boldsymbol{\Psi}_{(2)} > 0$. Thus, to ensure $\boldsymbol{\Psi}_\gamma < 0$, we must require that $\gamma\boldsymbol{\Psi}_{(1)} < 0$, and the effect of $\gamma^2\boldsymbol{\Psi}_{(2)} > 0$ can be exceeded by $\gamma\boldsymbol{\Psi}_{(1)} < 0$. Fortunately, due to the coefficient γ^2, when γ is small, the effect of $\gamma\boldsymbol{\Psi}_{(1)}$ will eventually outnumber the effect of $\gamma^2\boldsymbol{\Psi}_{(2)}$. In this situation, the negative definiteness of $\boldsymbol{\Psi}_\gamma$ mainly relies on the structure of $\boldsymbol{\Psi}_{(1)}$. Moreover, a direct calculation gives $\boldsymbol{Q}^\top(\boldsymbol{I}_{p^2} - \boldsymbol{K}_p)\boldsymbol{\Psi}_{(1)}(\boldsymbol{I}_{p^2} - \boldsymbol{K}_p)\boldsymbol{Q}$ to be a diagonal matrix with diagonal elements $\{(u_{jk} - d_j) + (u_{kj} - d_k) : j < k\}$. We thus have the following corollary.

Corollary 2: Assume the ICA model (2) and the modeling (9). Assume the existence of $\Gamma = (0, \tau]$ *for some* $\tau > 0$ *such that*

(A) $E[f_j^\gamma(S_j)S_j] = 0$, $\forall\, \gamma \in \Gamma$, $j = 1, \ldots, p$.

(B) *For all pairs* (j, k), $j \neq k$

$$E[f^\gamma(\boldsymbol{S})\{\phi_j(S_j)S_j - \phi_j'(S_j)S_k^2\}] + E[f^\gamma(\boldsymbol{S})\{\phi_k(S_k)S_k - \phi_k'(S_k)S_j^2\}] > 0, \quad \forall\, \gamma \in \Gamma.$$

Then, for $\gamma \in \Gamma$ *small enough, the associated* γ*-ICA is recovery consistent.*

To understand the meaning of condition (B), we first consider an implication of Corollary 2 in the limiting case $\gamma \to 0$, which corresponds to MLE-ICA. In this case, condition (A) becomes $E(S_j) = 0$, which is automatically true by (3). Moreover, since $E(S_j^2) = 1$, condition (B) becomes

$$E[\phi_j(S_j)S_j - \phi_j'(S_j)] + E[\phi_k(S_k)S_k - \phi_k'(S_k)] > 0 \quad \text{for all pairs } (j, k), j \neq k. \tag{23}$$

A sufficient condition to ensure the validity of (23) is

$$E[\phi_j(S_j)S_j - \phi_j'(S_j)] > 0 \qquad \forall j, \tag{24}$$

which is also the condition given in Theorem 9.1 of [11] (page 206) for the consistency of MLE-ICA. We should note that (23) is a weaker condition than (24). In fact, from the proof of Theorem 1, (23) is also a necessary condition. One implication of (23) is that, we can have at most one f_j to be wrongly specified or at most one Gaussian component involved, and MLE-ICA is still able to recover all independent components. See [16] for more explications. This can also be intuitively understood that once we have determined $p - 1$ directions in $\mathbb{R}^p$, the last direction is automatically determined. However, this fact cannot be observed from (24). We note that condition (23) is also explored to be the stability condition of the equivariant adaptive separation via independence (EASI) algorithm [17], and of Amari's gradient algorithm [18] for the ICA problem. We summarize the result for MLE-ICA below.

Corollary 3: Assume the ICA model (2) and the modeling (9). Then, MLE-ICA is recovery consistent if and only if $E[\phi_j(S_j)S_j - \phi_j'(S_j)] + E[\phi_k(S_k)S_k - \phi_k'(S_k)] > 0$ *for all pairs* (j, k), $j \neq k$.

Turning to the case of γ-ICA, condition (B) of Corollary 2 is the weighted version of (23) with the weight function f^γ. However, one should notice that the validity of γ-ICA has nothing to do with that of MLE-ICA, since there is no direct relationship between condition (B) and its limiting case (23). For example, even if (23) is violated (i.e., MLE-ICA fails), with a proper choice of γ, it is still possible that condition (B) holds and, hence, the recovery consistency of γ-ICA can be guaranteed. Finally, we remind the readers that the recovery consistency discussed in this section should be understood locally at the separation solution (see Remark 5). Moreover, the developed conditions for recovery consistency is with respect to the objective function of γ-ICA in (22) itself, but not for any specific learning algorithm. A gradient algorithm constrained on $\mathcal{SO}_p$ for γ-ICA is introduced in Section IV.

Remark 4: By Theorem 1, a valid γ*-ICA must correspond to* $\boldsymbol{\Psi}_\gamma < 0$*, i.e., the maximum eigenvalue of* $\boldsymbol{\Psi}_\gamma$*, denoted by* $\lambda_{\max}(\boldsymbol{\Psi}_\gamma)$*, must be negative. This suggests a rule of thumb to pick a* Γ*-interval for* γ*. Let* $\widehat{\boldsymbol{\Psi}}_\gamma$ *be the empirical estimator of* $\boldsymbol{\Psi}_\gamma$ *based on the estimated source* $\widehat{\boldsymbol{s}}_i := \widehat{\boldsymbol{W}}^\top \boldsymbol{z}_i$*. The plot* $\{(\gamma, \lambda_{\max}(\widehat{\boldsymbol{\Psi}}_\gamma))\}$ *then provides a guide to determine* Γ*, over which* $\lambda_{\max}(\widehat{\boldsymbol{\Psi}}_\gamma)$ *should be far away below zero. With the* Γ*-interval, a further selection procedure (see Section V) can be applied to select an optimal* γ *value. It is confirmed in our numerical study in Section VI that the range for* $\lambda_{\max}(\widehat{\boldsymbol{\Psi}}_\gamma) < 0$ *is quite wide, and the suggested rule does provide adequate choice of* Γ*. It also implies that the choice of* τ *in Corollary 2 is not critical, as* γ *is allowed to vary in a wide range. It is the condition (B) that plays the most important role to ensure the consistency of* γ*-ICA.*

Remark 5: Let $\mathcal{W}$ *be the set of local maximizers of* $E[\mathcal{L}(\boldsymbol{W})]$ *from (22). If* $\boldsymbol{\Psi}_\gamma < 0$*, we have shown in the proof of Theorem 1 that* $\boldsymbol{A}^* \in \mathcal{W}$*. Generally,* $\mathcal{W}$ *contains more than one element. Consider the simple case of (9) with* $f_j = f_0$ *for a common* f_0*. In this situation, the same argument for Theorem 1 shows that any column permutation of* $\boldsymbol{A}^*$ *is also an element of* $\mathcal{W}$*. See [17] for further discussion on this issue. On the other hand, under regularity conditions of* f_0*, it can be shown that* $\sup_{\boldsymbol{W} \in \mathcal{SO}_p} |\mathcal{L}(\boldsymbol{W}) - E[\mathcal{L}(\boldsymbol{W})]| = o_p(1)$*. Consequently,* $\widehat{\boldsymbol{W}}$ *in (22) is proven to be statistically consistent in the sense that* $P(\widehat{\boldsymbol{W}} \in \mathcal{W})$ *goes to unity as* $n \to \infty$.

IV. Gradient Method for γ-ICA on $\mathcal{SO}_p$

In this section, we introduce an algorithm for estimating $\boldsymbol{W}$ constrained to the special orthogonal group $\mathcal{SO}_p$, which is a Lie group and is endowed with a manifold structure.[1] The Lie group $\mathcal{SO}_p$, which is a path-connected subgroup of $\mathcal{O}_p$, consists of all orthogonal matrices in $\mathbb{R}^{p\times p}$ with determinant one.[2] Recall $\mathcal{L}$ in (22) being the objective function of γ-ICA. A desirable algorithm is to generate an increasing sequence $\{\mathcal{L}(\boldsymbol{W}_k)\}_{k=1}^\infty$ with $\boldsymbol{W}_k \in \mathcal{SO}_p$, such that $\{\boldsymbol{W}_k\}_{k=1}^\infty$ converges to a local

[1] $\mathcal{G}$ is a Lie group if the group operations $\mathcal{G} \times \mathcal{G} \to \mathcal{G}$ defined by $(x, y) \to xy$ and $\mathcal{G} \to \mathcal{G}$ defined by $x \to x^{-1}$ are both C^∞ mappings [19].

[2] The reason to consider $\mathcal{SO}_p$ is that $\mathcal{O}_p$ is not connected. When the desired orthogonal matrix $\boldsymbol{W}$ has determinant -1, our algorithm in fact searches for $\boldsymbol{\Pi}\boldsymbol{W} \in \mathcal{SO}_p$ for some permutation matrix $\boldsymbol{\Pi}$ with $\det(\boldsymbol{\Pi}) = -1$.

maximizer $\boldsymbol{W}^*$ of $\mathcal{L}$. Various approaches can be used to generate such a sequence $\{\boldsymbol{W}_k\}_{k=1}^{\infty}$ in $\mathcal{SO}_p$, for instance, geodesic flows and quasi-geodesic flows [20]. Here we focus on geodesic flows on $\mathcal{SO}_p$. In particular, starting with the current $\boldsymbol{W}_k$, the update $\boldsymbol{W}_{k+1}$ is selected from one geodesic path of $\boldsymbol{W}_k$ along the steepest ascent direction such that $\mathcal{L}(\boldsymbol{W}_{k+1}) > \mathcal{L}(\boldsymbol{W}_k)$. In fact, this approach has been applied to the general Stiefel manifold [20]. Below we briefly review the idea and then introduce our implementation algorithm for γ-ICA. We note that the proposed algorithm is also applicable to MLE-ICA by using the corresponding objective function.

Let $T_{\boldsymbol{W}}\mathcal{SO}_p$ be the tangent space of $\mathcal{SO}_p$ at $\boldsymbol{W}$. Consider a smooth path $\boldsymbol{W}(t)$ on $\mathcal{SO}_p$ with $\boldsymbol{W}(0) = \boldsymbol{W}$. Differentiating $\boldsymbol{W}(t)^\top \boldsymbol{W}(t) = \boldsymbol{I}_p$ yields the tangent space at $\boldsymbol{W}$

$$T_{\boldsymbol{W}}\mathcal{SO}_p = \left\{ \boldsymbol{W}\boldsymbol{V} : \boldsymbol{V} \in \mathbb{R}^{p\times p}, \boldsymbol{V}^\top = -\boldsymbol{V} \right\}.$$

Clearly, $T_{\boldsymbol{I}_p}\mathcal{SO}_p$ is the set of all skew-symmetric matrices. Each geodesic path starting from $\boldsymbol{I}_p$ has an intimate relation with the matrix exponential function. In fact, $\exp(\boldsymbol{V}) \in \mathcal{SO}_p$ if and only if $\boldsymbol{V}$ is skew-symmetric (see [19, page 148]; Proposition 9.2.5 in [21]). Moreover, for any $\boldsymbol{M} \in \mathcal{SO}_p$, there exists (not unique) a skew-symmetric $\boldsymbol{V}$ such that $\boldsymbol{M} = \exp(\boldsymbol{V})$. If we consider the Killing metric [20],

$$g_{\boldsymbol{W}}(\boldsymbol{M}_1, \boldsymbol{M}_2) := \operatorname{tr}(\boldsymbol{M}_1^\top \boldsymbol{M}_2), \quad \boldsymbol{M}_1, \boldsymbol{M}_2 \in T_{\boldsymbol{W}}\mathcal{SO}_p,$$

the geodesic path starting from $\boldsymbol{I}_p$ in the direction $\boldsymbol{V}$ is

$$\{\exp(t\boldsymbol{V}) : t \in \mathbb{R}\}. \tag{25}$$

Since the Lie group is homogeneous, we can compute the gradient and geodesic at $\boldsymbol{W}_k \in \mathcal{SO}_p$ by pulling them back to the identity $\boldsymbol{I}_p$ and then transform back to $\boldsymbol{W}_k$. In the implementation algorithm, to ensure all the iterations lying on the manifold $\mathcal{SO}_p$, we update $\boldsymbol{W}_{k+1}$ through

$$\boldsymbol{W}_{k+1} := \boldsymbol{W}_k \exp(t_k \boldsymbol{V}_k), \tag{26}$$

where the skew-symmetric matrix $\boldsymbol{V}_k$ and the step size t_k are chosen properly to meet the ascending condition $\mathcal{L}(\boldsymbol{W}_{k+1}) > \mathcal{L}(\boldsymbol{W}_k)$. Since, from (25), $\exp(t_k\boldsymbol{V}_k)$ lies on the geodesic path of $\boldsymbol{I}_p$, then $\boldsymbol{W}_{k+1} = \boldsymbol{W}_k \exp(t_k\boldsymbol{V}_k)$ must lie on the geodesic path of $\boldsymbol{W}_k$. Moreover, since $\det(\boldsymbol{W}_{k+1}) = \det(\boldsymbol{W}_k)\exp(0) = 1$ by $\operatorname{tr}(\boldsymbol{V}_k) = 0$, the sequence in (26) satisfies $\boldsymbol{W}_k \in \mathcal{SO}_p$ for all k. The determination of the gradient direction $\boldsymbol{V}_k$ and the step size t_k is discussed below.

To compute the gradient and geodesic at $\boldsymbol{W}_k$ by pulling them back to $\boldsymbol{I}_p$, define

$$\mathcal{F}_{\boldsymbol{W}_k}(\boldsymbol{W}) := \mathcal{L}(\boldsymbol{W}_k\boldsymbol{W}). \tag{27}$$

We then determine $\boldsymbol{W}_{k+1} = \boldsymbol{W}_k \exp(t_k\boldsymbol{V}_k)$ from one geodesic at $\boldsymbol{I}_p$ in the direction of the projected gradient of $\mathcal{F}_{\boldsymbol{W}_k}$. Specifically, to ensure the ascending condition, we choose each skew-symmetric $\boldsymbol{V}_k$ to be $\nabla_{/\!/}\mathcal{F}_{\boldsymbol{W}_k}$, the projected gradient of $\mathcal{F}_{\boldsymbol{W}_k}$ at $\boldsymbol{I}_p$, defined to be

$$\begin{aligned}\nabla_{/\!/}\mathcal{F}_{\boldsymbol{W}_k} &:= \operatorname*{argmin}_{\boldsymbol{V}\in T_{\boldsymbol{I}_p}\mathcal{SO}_p} \|\nabla\mathcal{F}_{\boldsymbol{W}_k} - \boldsymbol{V}\| = \frac{\nabla\mathcal{F}_{\boldsymbol{W}_k} - \nabla\mathcal{F}_{\boldsymbol{W}_k}^\top}{2} \\ &= \frac{\gamma}{2n}\sum_{i=1}^{n} f^\gamma(\boldsymbol{W}_k^\top \boldsymbol{z}_i)\boldsymbol{\Phi}(\boldsymbol{W}_k, \boldsymbol{z}_i),\end{aligned} \tag{28}$$

where $\nabla\mathcal{F}_{\boldsymbol{W}_k} := \partial\mathcal{F}_{\boldsymbol{W}_k}/\partial\boldsymbol{W}|_{\boldsymbol{W}=\boldsymbol{I}_p}$ and $\boldsymbol{\Phi}$ is defined in Proposition 1. This particular choice of $\boldsymbol{V}_k$ ensures the existence of the step size t_k for the ascending condition. Note that in the case of $\mathcal{SO}_p$ imposed with the Killing metric, the projected gradient coincides with the natural gradient introduced by [22]. See also Fact 5 in [20] for further details.

As to the selection of the step size t_k at each iteration k with $\boldsymbol{W}_k$ and $\boldsymbol{V}_k = \nabla_{/\!/}\mathcal{F}_{\boldsymbol{W}_k}$, we propose to select t_k such that $\boldsymbol{W}_k\exp(t_k\boldsymbol{V}_k)$ is the "first improved rotation". In particular, we consider $t_k = \alpha\rho^{\ell_k}$ for some $\alpha > 0$ and $0 < \rho < 1$, where ℓ_k is a nonnegative integer. To proceed, we search ℓ_k such that $\mathcal{L}\big(\boldsymbol{W}_k\exp(\alpha\rho^{\ell_k}\boldsymbol{V}_k)\big) > \mathcal{L}\big(\boldsymbol{W}_k\big)$, where $\boldsymbol{V}_k = \nabla_{/\!/}\mathcal{F}_{\boldsymbol{W}_k}$, and then update $\boldsymbol{W}_{k+1} = \boldsymbol{W}_k\exp(\alpha\rho^{\ell_k}\boldsymbol{V}_k)$. In our implementation, $\alpha = 1$ and $\rho = 0.5$ are used. For the convergence issue, one can instead consider the Armijo rule for t_k (given in (29)). Our experiments show that the above "first improved rotation" rule works quite well. Lastly, in the implementation, to save the storage for $\boldsymbol{W}_k$, we "rotate $\boldsymbol{\mathcal{Z}}$ directly" instead of manipulating $\boldsymbol{W}$, where $\boldsymbol{\mathcal{Z}}$ is the $p \times n$ data matrix whose columns are $\boldsymbol{z}_i$, $i = 1, \ldots, n$. That is, we use the update $\boldsymbol{\mathcal{Z}}_k = \boldsymbol{W}_k^\top\boldsymbol{\mathcal{Z}}$. To retrieve the matrix $\boldsymbol{W}$, we simply do a matrix right division of the final $\boldsymbol{\mathcal{Z}}$ and the initial $\boldsymbol{\mathcal{Z}}$. The algorithm for γ-ICA based on gradient ascend on $\mathcal{SO}_p$ is summarized below.

1) Initialization: $\alpha = 1$, $\rho = 0.5$, whitened data $\boldsymbol{\mathcal{Z}}_1 = \boldsymbol{\mathcal{Z}}$.
2) For each iteration $k = 1, 2, 3, \ldots$,
   (i) Compute the skew-symmetric matrix $\boldsymbol{V}_k$ in (28).
   (ii) For $\ell_k = 0, 1, 2, \ldots$, if $\mathcal{F}_{\boldsymbol{W}_k}(\exp(\alpha\rho^{\ell_k}\boldsymbol{V}_k)) > \mathcal{F}_{\boldsymbol{W}_k}(\boldsymbol{I}_p)$, then break the loop.
   (iii) Update $\boldsymbol{\mathcal{Z}}_{k+1} = \exp(\alpha\rho^{\ell_k}\boldsymbol{V}_k)^\top\boldsymbol{\mathcal{Z}}_k$. If the convergence criterion is not met, go back to (i).
3) Output $\widehat{\boldsymbol{W}} = \left(\boldsymbol{\mathcal{Z}}_1\boldsymbol{\mathcal{Z}}_1^\top\right)^{-1}\boldsymbol{\mathcal{Z}}_1\boldsymbol{\mathcal{Z}}_k^\top$.

Finally, we mention the convergence issue. The statement is similar to Proposition 1.2.1 of [23].

Theorem 6: Let $\mathcal{L}$ be continuously differentiable on $\mathcal{SO}_p$, and $\mathcal{F}$ be defined in (27). Let $\{\boldsymbol{W}_k \in \mathcal{SO}_p\}$ be a sequence generated by $\boldsymbol{W}_{k+1} = \boldsymbol{W}_k\exp(t_k\boldsymbol{V}_k)$, where $\boldsymbol{V}_k$ is a projected gradient related (see (30) below) and t_k is a properly chosen step size by the Armijo rule: reduce the step size $t_k = \alpha\rho^{\ell_k}$, $\ell_k = 0, 1, 2, \ldots$, until the inequality holds for the first nonnegative ℓ_k,

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}_{k+1}) - \mathcal{L}(\boldsymbol{W}_k) &= \mathcal{F}_{\boldsymbol{W}_k}(\exp(t_k\boldsymbol{V}_k)) - \mathcal{F}_{\boldsymbol{W}_k}(\boldsymbol{I}_p) \\ &\geq \eta\, t_k \operatorname{tr}\left(\nabla_{/\!/}\mathcal{F}_{\boldsymbol{W}_k}^\top\boldsymbol{V}_k\right),\end{aligned} \tag{29}$$

where $0 < \eta < 1$ is a constant. Then, every limit point $\boldsymbol{W}^*$ of $\{\boldsymbol{W}_k \in \mathcal{SO}_p\}$ is a stationary point, i.e., $\operatorname{tr}(\nabla\mathcal{F}_{\boldsymbol{W}^*}^\top\boldsymbol{V}) = 0$ for all $\boldsymbol{V} \in T_{\boldsymbol{W}^*}\mathcal{SO}_p$, or equivalently, $\nabla_{/\!/}\mathcal{F}_{\boldsymbol{W}^*} = 0$.

The statement that $\boldsymbol{V}_k$ is a projected gradient related corresponds to the condition

$$\limsup_{k\to\infty} \operatorname{tr}\left(\nabla_{//}\mathcal{F}_{\boldsymbol{W}_k}^{\top}\boldsymbol{V}_k\right) > 0. \tag{30}$$

This condition is true when $\boldsymbol{V}_k$ is the projected gradient $\nabla_{//}\mathcal{F}_{\boldsymbol{W}_k}$ itself or some natural gradient $\boldsymbol{M}^{-1}\nabla_{//}\mathcal{F}_{\boldsymbol{W}_k}$ (Theorem 1 in [22]), where $\boldsymbol{M}$ is a Riemannian metric tensor, which is positive definite.

V. Selection of γ

The estimation process of γ-ICA consists of two steps: γ-prewhitening and the geometry-based estimation for $\boldsymbol{W}$, in which the values of γ are essential to have robust estimators. Hence, we carefully select the value of γ based on the adaptive selection procedures proposed by [24] and [1]. We first introduce a general idea and then apply the idea to the selection of γ in both γ-prewhitening and γ-ICA. Define the measurement of generalization performance as

$$C_{\gamma_0}(\gamma) = E[\mathcal{D}_{\gamma_0}(g, f_{\widehat{\boldsymbol{\theta}}_\gamma})],$$

where g is the underlying true joint probability density function of the data, $f_{\boldsymbol{\theta}}$ is the considered model for fitting, $\widehat{\boldsymbol{\theta}}_\gamma := \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{D}_\gamma(\widehat{g}, f_{\boldsymbol{\theta}})$ is the minimum γ-divergence estimator of $\boldsymbol{\theta}$, and $\widehat{g}$ is the empirical estimate of g. The γ_0 is called the anchor parameter and is fixed at $\gamma_0 = 1$ throughout this paper. This value is empirically shown to be insensitive to the resultant estimators [24]. Let $\widehat{C}_{\gamma_0}(\gamma)$ be the sample analogue of $C_{\gamma_0}(\gamma)$. We propose to select the value of γ over a predefined set Γ through

$$\widehat{\gamma} = \operatorname*{argmin}_{\gamma\in\Gamma} \widehat{C}_{\gamma_0}(\gamma).$$

For γ-prewhitening, $g = g_{\boldsymbol{X}}$ and $f_{\boldsymbol{\theta}} = \xi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ with $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For γ-ICA, $g = g_{\boldsymbol{Z}}$ and $f_{\boldsymbol{\theta}} = f_{\boldsymbol{Z}}(\cdot;\boldsymbol{W})$ with $\boldsymbol{\theta} = \boldsymbol{W}$.

The above selection criterion requires the estimation of $C_{\gamma_0}(\gamma)$. To avoid overfitting, we apply the K-fold cross-validation. Let $\mathcal{T}$ be the whole data, and let K partitions of $\mathcal{T}$ be $\mathcal{T}_1, \ldots, \mathcal{T}_K$, that is, $\mathcal{T}_i \bigcap \mathcal{T}_j = \emptyset$ if $i \neq j$ and $\mathcal{T} = \cup_{i=1}^K \mathcal{T}_i$. The whole selection procedure is summarized below.

1) For $k = 1, \ldots, K$,
   (i) Obtain $\widehat{\boldsymbol{\theta}}_\gamma^{(-k)} = \operatorname{argmin}_{\boldsymbol{\theta}} \mathcal{D}_\gamma(\widehat{g}^{(-k)}, f_{\boldsymbol{\theta}})$ for $\gamma \in \Gamma$, where $\widehat{g}^{(-k)}$ is the empirical estimate of g based on $\mathcal{T} \setminus \mathcal{T}_k$.
   (ii) Compute $\mathcal{D}_{\gamma_0}(\widehat{g}^{(k)}, f_{\widehat{\boldsymbol{\theta}}_\gamma^{(-k)}})$, where $\widehat{g}^{(k)}$ is the empirical estimate of g based on $\mathcal{T}_k$.
2) Obtain $\widehat{C}_{\gamma_0}(\gamma) = (1/K)\sum_{k=1}^K \mathcal{D}_{\gamma_0}(\widehat{g}^{(k)}, f_{\widehat{\boldsymbol{\theta}}_\gamma^{(-k)}})$, and obtain $\widehat{\gamma} = \operatorname{argmin}_{\gamma\in\Gamma} \widehat{C}_{\gamma_0}(\gamma)$.

Eventually, we have two optimal values of γ: $\widehat{\gamma}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ for γ-prewhitening, and $\widehat{\gamma}_{\boldsymbol{W}}$ for γ-ICA.

VI. Numerical Experiments

We conduct two numerical studies to demonstrate the robustness of γ-ICA procedure. In the first study, the data is generated with known distributions. In the second study, we use transformations of Lena images to form mixed images.

A. Simulated Data

We independently generate two sources S_j, $j = 1, 2$, from a non-Gaussian distribution with sample size $n = 150 + n_1$. The observable $\boldsymbol{X}$ is then given by $\boldsymbol{X} = \boldsymbol{AS}$, where

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 1 & 0.5 \end{bmatrix}.$$

Among the n observations, we add to each of the last n_1 observations a random noise $\boldsymbol{e}$. The data thus contains 150 uncontaminated i.i.d. observations from the ICA model, $\boldsymbol{X} = \boldsymbol{AS}$, and n_1 contaminated i.i.d. observations from $\boldsymbol{X} = \boldsymbol{AS} + \boldsymbol{e}$, where $\boldsymbol{e} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{I}_2)$ with $\boldsymbol{\mu} = (0, 5)^\top$ and $\sigma = 5$. We consider two situations for $\boldsymbol{S} = (S_1, S_2)$:

(i) Uniform source: Each S_j, $j = 1, 2$, is generated from $\operatorname{Uniform}(-3, 3)$.
(ii) Student-t source: Each S_j, $j = 1, 2$, is generated from t-distribution with 3 degrees of freedom.

For uniform source, we use sub-Gaussian model $f_j(s) \propto \exp(-0.15s^4)$, while it is super-Gaussian model $f_j(s) \propto 1/\cosh(1.5s)$ for the case of t source, so that the variance under f_j is close to unity. To determine the γ value for γ-prewhitening, the selection criterion in Section V with $\gamma_0 = 1$ and $K = 5$ is considered. For comparison, we also use the same γ-prewhitened data to implement MLE-ICA (using the geometrical algorithm introduced in Section IV), fast-ICA (using the code available at *www.cis.hut.fi/projects/ica/fastica/*), and JADE (using the code available at *perso.telecom-paristech.fr/~cardoso/Algo/Jade/jadeR.m*), and use the original data $\boldsymbol{X}$ to implement β-ICA [9]. To evaluate the performance of each method, we modify from the performance index of [25] by a rescaling and by replacing the 2-norm with 1-norm and define the performance index

$$\pi = \frac{0.5}{p(p-1)} \sum_i \left\{ \left(\frac{\sum_k |\pi_{ik}|}{\max_j\{|\pi_{ij}|\}} - 1 \right) + \left(\frac{\sum_k |\pi_{ki}|}{\max_j\{|\pi_{ji}|\}} - 1 \right) \right\}$$

with π_{ij} being the (i, j)-th element of $\boldsymbol{\Pi} = \boldsymbol{A}^* \widehat{\boldsymbol{W}}^\top$. Note that $0 \leq \pi \leq 1$. We will expect $\boldsymbol{\Pi}$ to be a permutation matrix when the method performs well. In this situation, the value of π should be very close to 0, and attains 0 if $\boldsymbol{\Pi}$ is indeed a permutation matrix. Simulation results with 100 replications are reported in Fig. 1.

For the case of no outliers ($n_1 = 0$), all methods perform well as expected. When data is contaminated ($n_1 = 30$), it is detected that the performance of γ-prewhitening followed by γ-ICA is not heavily affected by the presence of outliers, while MLE-ICA, fast-ICA, and JADE are not able to recover the latent sources. Comparing with β-ICA, γ-ICA does have a better performance. Obviously, γ-ICA is applicable for a wider range of γ values, while β-ICA tends to perform worse at small β values. This is an appealing property for γ-ICA since in practice, γ should also be determined from the data. A wider range for γ then implies that γ-ICA is more reliable. One can see that the performance of γ-ICA becomes worse when γ is small. This is reasonable since in the limiting case $\gamma \to 0$, γ-ICA reduces to the non-robust MLE-ICA. We note that both γ-prewhitening

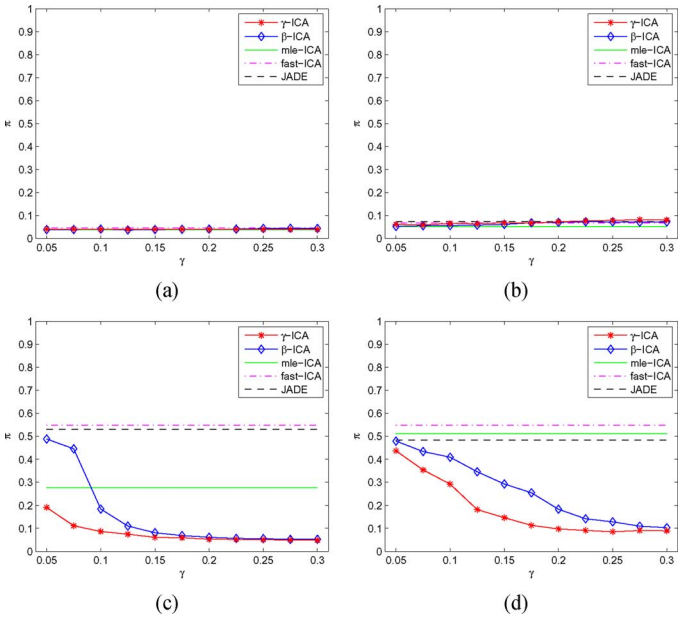

Fig. 1. The medians of the performance index π under different settings. (a) Uniform Source ($\mathrm{n}_1 = 0$). (b) t Source ($\mathrm{n}_1 = 0$). (c) Uniform Source ($\mathrm{n}_1 = 30$). (d) t Source ($\mathrm{n}_1 = 30$).

and γ-ICA are critical. This can be seen from the poor performance of MLE-ICA, fast-ICA, and JADE in the presence of outliers, even they use the same γ-prewhitened data as the input. Indeed, γ-prewhitening only ensures that we shift and rotate the data in a robust manner, while the outliers will still enter into the subsequent estimation process and, hence, produce non-robust results.

B. Lena Image

We use the Lena picture (512×512 pixels) to evaluate the performance of γ-ICA. We construct four types of Lena as the latent independent sources $\boldsymbol{S}$ as shown in Fig. 2. We randomly generate the mixing matrix to be $\boldsymbol{A} = \mathbf{1}_4\mathbf{1}_4^T + \boldsymbol{C}$, where the elements of $\boldsymbol{C} \in \mathbb{R}^{4\times 4}$ are independently generated from $\mathrm{Uniform}(-0.3, 0.3)$. The observed mixed pictures are also placed in Fig. 2, wherein about 30% of the pixels are added with random noise generated from $N(20, 50^2)$ for contamination. The aim of this data analysis is to recover the original Lena pictures based on the observed contaminated mixed pictures. In this analysis, the pixels are treated as the random sample, each with dimension 4. We randomly select 1000 pixels to estimate the demixing matrix, and then apply it to reconstruct the whole source pictures. We conduct two scenarios to evaluate the robustness of each method:

1) Using the mixed image $\boldsymbol{X}$ as the input (see Fig. 2).
2) Using the filtered image $\boldsymbol{X}^*$ as the input (see Fig. 2).

The filtering process in Scenario-2 replaces the mixed pixel value by the median of the pixel values over its neighborhood. In both scenarios, the estimated demixing matrix is applied to the mixed images $\boldsymbol{X}$ to recover $\boldsymbol{S}$. We apply γ-ICA, MLE-ICA, and fast-ICA, all with the sub-Gaussian modeling, to the same γ-prewhitened data for fair comparisons. The plot $\{(\gamma, \lambda_{\max}(\widehat{\boldsymbol{\Psi}}_\gamma))\}$ introduced in Remark 4 is placed in Fig. 3, which suggests that $\Gamma = (0, 1]$ is a good candidate for

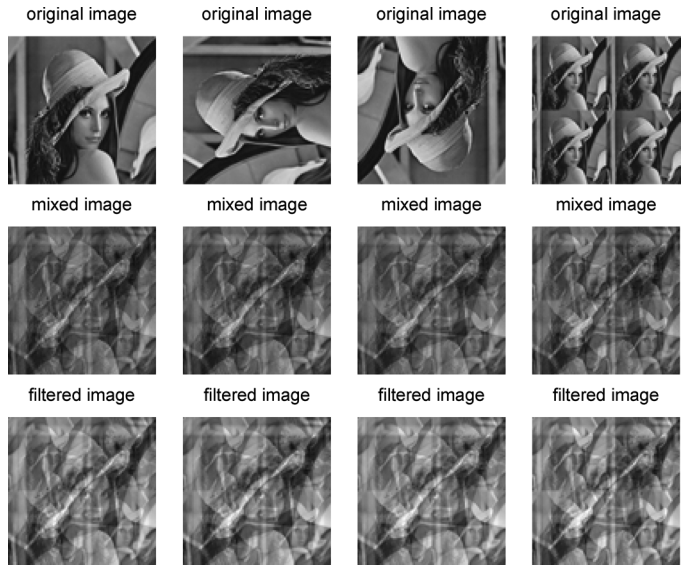

Fig. 2. Four images of Lena (the first row), the mixed images with contamination (the second row), and the filtered images (the third row).

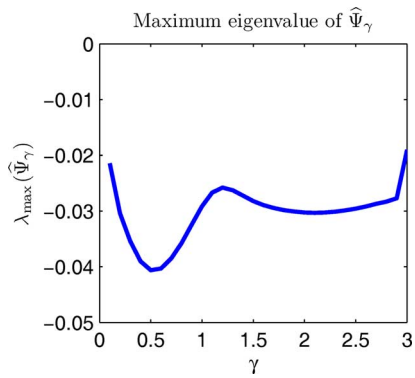

Fig. 3. The maximum eigenvalue of $\widehat{\boldsymbol{\Psi}}_\gamma$ at different γ values.

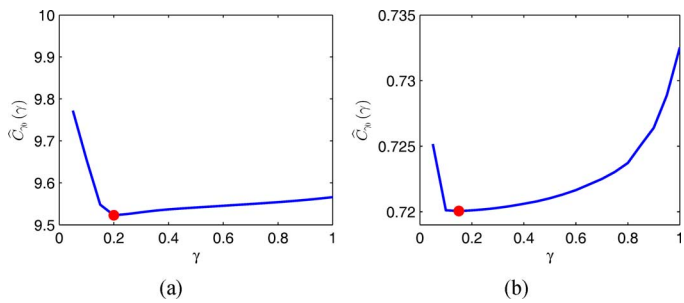

Fig. 4. The cross-validation estimates $\widehat{C}_{\gamma_0}(\gamma)$ with $\gamma_0 = 1$ for (a) γ-prewhitening and (b) γ-ICA. The dot indicates the minimum value.

possible γ values. We then apply the cross-validation method in Section V to determine the optimal $\gamma \in \Gamma$. The estimated values of $\widehat{C}_{\gamma_0}(\gamma)$ are plotted in Fig. 4, from which we select $\widehat{\gamma}_{\boldsymbol{\mu},\boldsymbol{\Sigma}} = 0.2$ for γ-prewhitening and $\widehat{\gamma}_{\boldsymbol{W}} = 0.15$ for γ-ICA. The recovered pictures are placed in Figs. 5–7.

It can be seen that γ-ICA is the best performer under both scenarios, while MLE-ICA and fast-ICA cannot recover the source images well when data is contaminated. It also demonstrates the applicability of the proposed γ-selection procedure. We detect that MLE-ICA and fast-ICA perform better when using filtered images $\boldsymbol{X}^*$, but can still not reconstruct images as good as

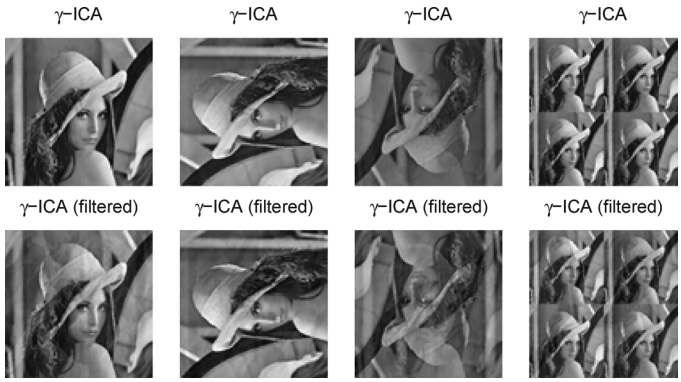

Fig. 5. Recovered Lena images from γ-ICA based on the mixed images (the first row) and the filtered images (the second row).

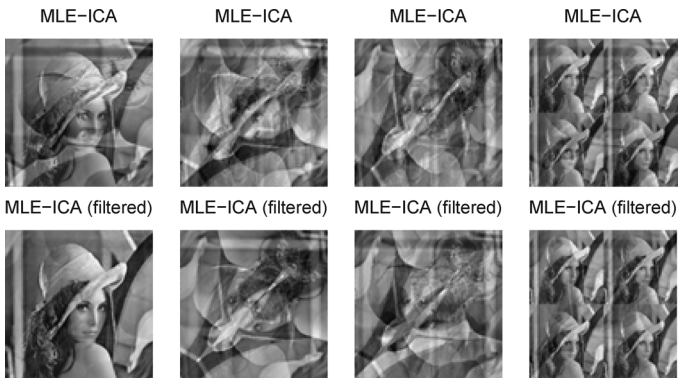

Fig. 6. Recovered Lena images from MLE-ICA based on the mixed images (the first row) and the filtered images (the second row).

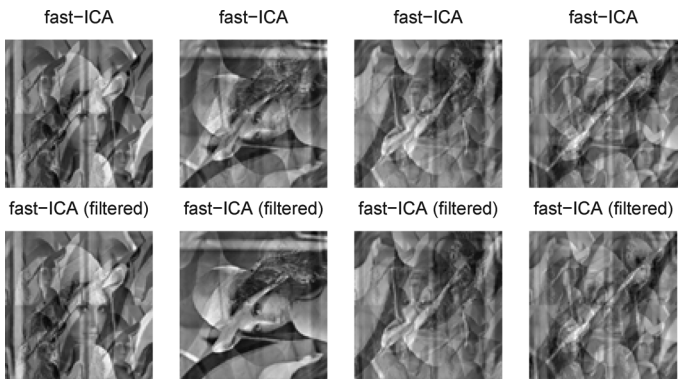

Fig. 7. Recovered Lena images from fast-ICA based on the mixed images (the first row) and the filtered images (the second row).

γ-ICA does. Notably, γ-ICA has a reverse performance, where the best reconstructed images are estimated from the mixed images instead of the filtered ones. Reasonably, it is still possible to lose useful information during the filtering process. For instance, a pixel without being contaminated will still be replaced with a median value during the filtering process. γ-ICA, however, is able to work on the mixed data $\boldsymbol{X}$ that possesses all the information available, and then weights each pixel according to its observed value to achieve robustness. Hence, a better performance for γ-ICA based on the mixed images is reasonably expected.

VII. Conclusions

In this paper, we introduce a unified framework for the ICA problem by means of minimum U-divergence estimation. For the sake of robustness, we further focus on γ-divergence to propose γ-ICA. Statistical properties are rigorously investigated. A geometrical algorithm based on gradient flows on $\mathcal{SO}_p$ is introduced to implement γ-ICA. The performance of γ-ICA is evaluated through synthetic and real data examples. Notably, the proposed γ-ICA procedure is equivalent to β-prewhitening [1] pluses β-ICA [9]. However, the performance of the combination of γ-prewhitening and γ-ICA has not been clarified so far. See [1], wherein the authors apply fast-ICA after β-prewhitening. One aim of this paper is to emphasize the importance of the combination. Simulation studies also demonstrate the superiority of γ-ICA over β-ICA.

There are still many important issues that are not covered by this work. For example, we only consider full ICA problem, i.e., simultaneous extraction of all p independent components, which is unpractical when p is large. It is of interest to extend our current γ-ICA to partial γ-ICA. In this work, data have to be prewhitened before entering the γ-ICA procedure. Prewhitening can be very unstable especially when p is large. How to avoid such a difficulty is an interesting and challenging issue. One approach is to follow the idea of [9] to consider γ-ICA under the original data $\boldsymbol{X}$ directly. Though the idea is simple, there are many issues needed to be investigated, such as the study of stability condition and the problem of non-identifiability. Tensor data analysis is now becoming popular and attracts the attention of many researchers. Many statistical methods include ICA have been extended to deal with such a data structure by means of multilinear algebra techniques. Extension of γ-ICA to a multilinear setting to adapt to tensor data is also of great interest for the future study.

Appendix
Proofs of Theorems

For any symmetric matrix $\boldsymbol{M} \in \mathbb{R}^{p\times p}$, $\mathrm{vech}(\boldsymbol{M})$ takes the unique elements of the columns of $\boldsymbol{M}$ as a vector with length $p(p+1)/2$. There exist matrices $\boldsymbol{H} \in \mathbb{R}^{p(p+1)/2\times p^2}$ and $\boldsymbol{G} \in \mathbb{R}^{p^2\times p(p+1)/2}$ [26] such that $\mathrm{vech}(\boldsymbol{M}) = \boldsymbol{H}\mathrm{vec}(\boldsymbol{M})$ and $\mathrm{vec}(\boldsymbol{M}) = \boldsymbol{G}\mathrm{vech}(\boldsymbol{M})$. Moreover, $\boldsymbol{HG} = \boldsymbol{I}_{p(p+1)/2}$ and $\boldsymbol{GH} = (1/2)(\boldsymbol{I}_{p^2} + \boldsymbol{K}_p)$.

Proof of Proposition 1: Since the objective function $\mathcal{L}(\boldsymbol{W})$ is defined on $\mathcal{SO}_p$, by [27, equation (2.53)], the natural gradient of $\mathcal{L}(\boldsymbol{W})$ with respect to $\boldsymbol{W}$ on $\mathcal{SO}_p$ is

$$\frac{\partial \mathcal{L}(\boldsymbol{W})}{\partial \boldsymbol{W}} - \boldsymbol{W}\left(\frac{\partial \mathcal{L}(\boldsymbol{W})}{\partial \boldsymbol{W}}\right)^{\top}\boldsymbol{W} = (\gamma \boldsymbol{W}) \cdot \frac{1}{n}\sum_{i=1}^{n} f^{\gamma}(\boldsymbol{W}^{\top}\boldsymbol{z}_i)\boldsymbol{\Phi}(\boldsymbol{W}, \boldsymbol{z}_i). \quad (31)$$

The proof is completed by equating (31) to $\boldsymbol{0}$. ■

Proof of Theorem 1: By $\boldsymbol{Z} = \boldsymbol{A}^{*}\boldsymbol{S}$, the population objective function of γ-ICA in (22) can be expressed as

$$E[f^{\gamma}(\boldsymbol{W}^{\top}\boldsymbol{Z})] = E[f^{\gamma}(\boldsymbol{W}^{\top}\boldsymbol{A}^{*}\boldsymbol{S})] := E[f^{\gamma}(\boldsymbol{B}^{\top}\boldsymbol{S})],$$

where $\boldsymbol{B} := \boldsymbol{A}^{*\top}\boldsymbol{W} \in \mathcal{SO}_p$. Considering the orthogonality of $\boldsymbol{B}$ gives the objective function

$$\ell(\boldsymbol{B}, \boldsymbol{\Lambda}) = E[f^{\gamma}(\boldsymbol{B}^{\top}\boldsymbol{S})] - \frac{1}{2}\mathrm{tr}\{\boldsymbol{\Lambda}(\boldsymbol{B}^{\top}\boldsymbol{B} - \boldsymbol{I}_p)\}, \quad (32)$$

where $\boldsymbol{\Lambda}$ is a symmetric matrix containing the Lagrange multipliers. Let $\boldsymbol{\lambda} = \mathrm{vech}(\boldsymbol{\Lambda})$ and $\mathrm{vec}(\boldsymbol{\Lambda}) = \boldsymbol{G}\boldsymbol{\lambda}$. To see if γ-ICA is able to recover $\boldsymbol{S}$, we first show that $\boldsymbol{B} = \boldsymbol{I}_p$ (which implies $\boldsymbol{W} = \boldsymbol{A}^*$) attains the stationarity of (32) for some symmetric $\boldsymbol{\Lambda}$. The gradient $\dot{\ell}_{\boldsymbol{B}}(\boldsymbol{B}, \boldsymbol{\Lambda}) := \partial \ell(\boldsymbol{B}, \boldsymbol{\Lambda})/\partial \mathrm{vec}(\boldsymbol{B})$ is

$$\dot{\ell}_{\boldsymbol{B}}(\boldsymbol{B}, \boldsymbol{\Lambda}) = \gamma E[f^{\gamma}(\boldsymbol{B}^{\top}\boldsymbol{S})\mathrm{vec}(\boldsymbol{S}\boldsymbol{\phi}^{\top}(\boldsymbol{B}^{\top}\boldsymbol{S}))] - \mathrm{vec}(\boldsymbol{B}\boldsymbol{\Lambda}). \tag{33}$$

Solving $\dot{\ell}_{\boldsymbol{B}}(\boldsymbol{B}, \boldsymbol{\Lambda}) = 0$ gives the solution of $\boldsymbol{\lambda}$ to be

$$\boldsymbol{\lambda}_{\boldsymbol{B}} = \gamma \boldsymbol{H} E[f^{\gamma}(\boldsymbol{B}^{\top}\boldsymbol{S})\mathrm{vec}\{(\boldsymbol{B}^{\top}\boldsymbol{S})\boldsymbol{\phi}^{\top}(\boldsymbol{B}^{\top}\boldsymbol{S})\}],$$

or equivalently, gives $\mathrm{vec}(\boldsymbol{\Lambda}_{\boldsymbol{B}}) := \boldsymbol{G}\boldsymbol{\lambda}_{\boldsymbol{B}}$ to be

$$\frac{\gamma}{2}(\boldsymbol{I}_{p^2} + \boldsymbol{K}_p) E[f^{\gamma}(\boldsymbol{B}^{\top}\boldsymbol{S})\mathrm{vec}\{(\boldsymbol{B}^{\top}\boldsymbol{S})\boldsymbol{\phi}^{\top}(\boldsymbol{B}^{\top}\boldsymbol{S})\}]$$

by $\boldsymbol{G}\boldsymbol{H} = (1/2)(\boldsymbol{I}_{p^2} + \boldsymbol{K}_p)$. Substituting $\boldsymbol{\Lambda}_{\boldsymbol{B}}$ into (33), we have

$$\dot{\ell}_{\boldsymbol{B}}(\boldsymbol{B}, \boldsymbol{\Lambda}_{\boldsymbol{B}}) = \frac{\gamma}{2}(\boldsymbol{I}_p \otimes \boldsymbol{B})(\boldsymbol{I}_{p^2} - \boldsymbol{K}_p) \times E[f^{\gamma}(\boldsymbol{B}^{\top}\boldsymbol{S})\mathrm{vec}\{(\boldsymbol{B}^{\top}\boldsymbol{S})\boldsymbol{\phi}^{\top}(\boldsymbol{B}^{\top}\boldsymbol{S})\}].$$

By condition (A) and the independence of $\boldsymbol{S}$, it is deduced that $\dot{\ell}_{\boldsymbol{B}}(\boldsymbol{B}, \boldsymbol{\Lambda}_{\boldsymbol{B}})\big|_{\boldsymbol{B}=\boldsymbol{I}_p} = 0$, i.e., $\boldsymbol{B} = \boldsymbol{I}_p$ attains the stationarity.

Secondly, we will give condition so that $\boldsymbol{B} = \boldsymbol{I}_p$ indeed attains the maximum value and, hence, the recovery consistency of γ-ICA is guaranteed. Using $\boldsymbol{\Lambda}_{\boldsymbol{B}} = \gamma \boldsymbol{D}$ at $\boldsymbol{B} = \boldsymbol{I}_p$, the Hessian matrix of (32) with respect to $\mathrm{vec}(\boldsymbol{B})$ evaluated at $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_{\boldsymbol{B}}$ and $\boldsymbol{B} = \boldsymbol{I}_p$ is calculated to be

$$\left[\frac{\partial^2 \ell(\boldsymbol{B}, \boldsymbol{\Lambda})}{\partial \mathrm{vec}(\boldsymbol{B}) \partial \mathrm{vec}(\boldsymbol{B})^{\top}}\bigg|_{\boldsymbol{B}=\boldsymbol{I}_p, \boldsymbol{\Lambda}=\gamma\boldsymbol{D}}\right] = \gamma \boldsymbol{\Psi}_{(1)} + \gamma^2 \boldsymbol{\Psi}_{(2)}. \tag{34}$$

Note also that each $\boldsymbol{B} \in \mathcal{SO}_p$ can be expressed as $\boldsymbol{B}_{\boldsymbol{\theta}} = \exp(\boldsymbol{C} - \boldsymbol{C}^{\top})$, where $\boldsymbol{C}$ is a lower triangular matrix with zero diagonal and $\boldsymbol{\theta} := \boldsymbol{P}\mathrm{vec}(\boldsymbol{C})$. By chain rule and the fact that $\partial \mathrm{vec}\{\exp(\boldsymbol{M})\}/\partial \mathrm{vec}(\boldsymbol{M})^{\top}\big|_{\boldsymbol{M}=\boldsymbol{0}} = \boldsymbol{I}_{p^2}$, the tangent vector of $\mathrm{vec}(\boldsymbol{B}_{\boldsymbol{\theta}})$ at $\boldsymbol{\theta} = \boldsymbol{0}$ (which corresponds to $\boldsymbol{B}_{\boldsymbol{\theta}} = \boldsymbol{I}_p$) must lie in the span of $(\boldsymbol{I}_{p^2} - \boldsymbol{K}_p)\boldsymbol{Q}$. Thus, together with (34), $\boldsymbol{B} = \boldsymbol{I}_p$ attains the maximum value of $E[f^{\gamma}(\boldsymbol{B}^{\top}\boldsymbol{S})]$ over $\mathcal{SO}_p$ if and only if $\boldsymbol{\Psi}_{\gamma} < 0$. ■

Proof of Theorem 6: Similar to the proof of Proposition 1.2.1 of [23], the theorem will be proved by a contradiction. Suppose that $\boldsymbol{W}^*$ is a limit point of $\boldsymbol{W}_k$ with $\nabla_{/\!/}\mathcal{F}_{\boldsymbol{W}^*} \neq 0$. Since $\mathcal{F}$ is continuous on the compact set $\mathcal{SO}_p$, we have $\mathcal{F}_{\boldsymbol{W}^*}(\boldsymbol{I}_p) < \infty$ and $\mathcal{F}_{\boldsymbol{W}_{k+1}}(\boldsymbol{I}_p) - \mathcal{F}_{\boldsymbol{W}_k}(\boldsymbol{I}_p) \to 0$. According to the Armijo rule, we have $\eta\, t_k \mathrm{tr}(\nabla_{/\!/}\mathcal{F}_{\boldsymbol{W}_k}^{\top}\boldsymbol{V}_k) \to 0$. Since $\boldsymbol{V}_k$ is the projected gradient related, a subsequence of $\{t_k\}$ converges to 0. Then, for this subsequence the Armijo rule fails with step size t_k/ρ, i.e.,

$$\eta\, t_k \rho^{-1} \mathrm{tr}\left(\nabla_{/\!/}\mathcal{F}_{\boldsymbol{W}_k}^{\top}\boldsymbol{V}_k\right) > \mathcal{F}_{\boldsymbol{W}_k}(\exp(t_k \boldsymbol{V}_k \rho^{-1})) - \mathcal{F}_{\boldsymbol{W}_k}(\boldsymbol{I}_p), \tag{35}$$

where the right hand side in fact equals to

$$t_k \rho^{-1}\, \mathrm{tr}\left(\left[\exp(\widetilde{t}_k \boldsymbol{V}_k)\, \nabla \mathcal{F}_{\boldsymbol{W}_k}(\exp(\widetilde{t}_k \boldsymbol{V}_k \rho^{-1}))\right]^{\top} \boldsymbol{V}_k\right), \tag{36}$$

for some $\widetilde{t}_k \in (0, t_k)$ by the Mean Value Theorem. Since the set of the tangent vectors $\{\boldsymbol{V}_k\}$ is bounded, taking a further subsequence $\boldsymbol{V}'_k$, we have $\boldsymbol{V}'_k \to \bar{\boldsymbol{V}}$. Taking limits $k \to \infty$ with $t_k \to 0$ on (35)–(36), we have

$$\eta\, \mathrm{tr}(\nabla_{/\!/}\mathcal{F}_{\boldsymbol{W}^*}^{\top}\bar{\boldsymbol{V}}) > \mathrm{tr}(\nabla \mathcal{F}_{\boldsymbol{W}^*}^{\top}\bar{\boldsymbol{V}}) = \mathrm{tr}(\nabla_{/\!/}\mathcal{F}_{\boldsymbol{W}^*}^{\top}\bar{\boldsymbol{V}}),$$

where the equality comes from $\bar{\boldsymbol{V}} \in T_{\boldsymbol{W}^*}\mathcal{SO}_p$. Since $\nabla_{/\!/}\mathcal{F}_{\boldsymbol{W}^*}^{\top}\bar{\boldsymbol{V}} > 0$, then the above inequality contradicts to the assumption $0 < \eta < 1$. ■

ACKNOWLEDGMENT

This work is initiated during the visit of H. Hung and S.-Y. Huang to The Institute of Statistical Mathematics hosted by S. Eguchi. The authors thank J.-R. Liu in Institute of Statistical Science, Academia Sinica for preparing figures.

REFERENCES

[1] M. N. H. Mollah, S. Eguchi, and M. Minami, "Robust prewhitening for ICA by minimizing β-divergence and its application to FastICA," *Neural Process. Lett.*, vol. 25, no. 2, pp. 91–110, 2007.

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

[3] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4, pp. 411–430, 2000.

[4] A. Hyvärinen, "New approximations of differential entropy for independent component analysis and projection pursuit," in *Proc. Conf. Adv. Neural Inf. Process. Syst. 10*, Cambridge, MA, USA, 1998, pp. 273–279.

[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 and Signal Process.*, vol. 140, pp. 362–370, 1993.

[7] F. Harroy and J. L. Lacoume, "Maximum likelihood estimators and Cramer-Rao bounds in source separation," *Signal Process.*, vol. 55, no. 2, pp. 167–177, 1996.

[8] S. Amari and J. Cardoso, "Blind source separation-semiparametric statistical approach," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2692–2700, Nov. 1997.

[9] M. Mihoko and S. Eguchi, "Robust blind source separation by β-divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1859–1886, 2002.

[10] H. Fujisawa and S. Eguchi, "Robust parameter estimation with a small bias against heavy contamination," *J. Multivariate Anal.*, vol. 99, no. 9, pp. 2053–2081, 2008.

[11] A. Hyvärinen, J. Karhnen, and E. Oja, *Independent Component Analysis*. New York, NY, USA: Wiley Inter-Science, 2001.

[12] S. I. Amari, T. P. Chen, and A. Cichocki, "Stability analysis of learning algorithms for blind source separation," *Neural Netw.*, vol. 10, no. 8, pp. 1345–1351, 1997.

[13] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi, "Information geometry of U-boost and Bregman divergence," *Neural Comput.*, vol. 16, no. 7, pp. 1437–1481, 2004.

[14] S. Eguchi, *Information Divergence Geometry and the Application to Statistical Machine Learning*. Berlin, Germany: Springer, 2009, ch. 13, pp. 309–332.

[15] A. Basu, I. R. Harris, N. L. Hjort, and M. Jones, "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, vol. 85, no. 3, pp. 549–559, 1998.

[16] J. F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

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

[18] S. A. Cruces-Alvarez, A. Cichocki, and S. I. Amari, "On a new blind signal extraction algorithm: Different criteria and stability analysis," *IEEE Signal Process. Lett.*, vol. 9, no. 8, pp. 233–236, Aug. 2002.

[19] W. M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*. New York, NY, USA: Academic, 1986.

[20] Y. Nishimori and S. Akaho, "Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold," *Neurocomput.*, vol. 67, pp. 106–135, 2005.

[21] J. E. Marsden and S. T. Ratiu, *Introduction to Mechanics and Symmetry: A Basic Exposition of Classical Mechanical Systems*. New York, NY, USA: Springer, 1998.
[22] S. I. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, 1998.
[23] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA, USA: Athena Scientific, 2003.
[24] M. Minami and S. Eguchi, "Adaptive selection for minimum β-divergence method," in *Proc. ICA-03 Conf.*, Nara, Japan, 2003, pp. 475–480.
[25] S. D. Parmar and B. Unhelkar, "Performance analysis of ICA algorithms against multiple-sources interference in biomedical systems," *Int. J. Recent Trends Eng.*, vol. 2, pp. 19–21, 2009.
[26] J. R. Magnus and H. Neudecker, "The commutation matrix: Some properties and applications," *Ann. Statist.*, vol. 7, no. 2, pp. 381–394, 1979.
[27] A. Edelman, T. A. Arias, and S. T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal. Applicat.*, vol. 20, no. 2, pp. 303–353, 1998.

Pengwen Chen received his Ph.D degree from the University of Florida in 2007. Currently, he is an Assistant Professor at the Department of Applied Mathematics in National Chung Hsing University. His main interest is in image processing and analysis, in particular, point-set matching problems.

Hung Hung received his B.S. degree in management, M.S. degree in engineering, and Ph.D. degree in mathematics from National Taiwan University, Taiwan, in 2002, 2004, and 2009, respectively. He is currently an Assistant Professor at the Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taiwan. His research interests include statistical methods for dimension reduction analysis, tensor learning, and survival analysis.

Osamu Komori received his Ph.D degree at The Graduate University for Advanced Studies in 2010. He is a project Assistant Professor at School of Statistical Thinking in The Institute of Statistical Mathematics. His research field includes bioinformatics, machine learning and linguistics.

Su-Yun Huang received her B.S. and M.S. degrees from the Department of Mathematics, National Taiwan University, in 1983 and 1985, respectively, and the Ph.D. degree from the Department of Statistics, Purdue University in 1990. She is currently a Research Fellow in the Institute of Statistical Science, Academia Sinica, Taiwan. Her research interests are mainly on mathematical statistics and statistical machine learning.

Shinto Eguchi received his Ph.D. from Hiroshima University, Japan, in 1984. He is currently a Professor at the Institute of Statistical Mathematics and Graduate University of Advanced Studies, Japan. His research interest is primarily statistics, including statistical machine learning, bioinformatics, and information geometry.