ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9415059

ICA WITH ORTHOGONALITY CONSTRAINT: IDENTIFIABILITY AND A NEW EFFICIENT ALGORITHM

Ben Gabrielson, M. A. B. S. Akhonda*, Zois Boukouvalas**, Seung-Jun Kim*, and Tülay Adali**

*University of Maryland, Baltimore County, Baltimore, MD
**American University, Washington, D.C.

ABSTRACT

Given the prevalence of independent component analysis (ICA) for signal processing, many methods for improving the convergence properties of ICA have been introduced. The most utilized methods operate by iterative rotations over pre-whitened data, whereby limiting the space of estimated demixing matrices to those that are orthogonal. However, a proof of the identifiability conditions for orthogonal ICA methods has not yet been presented in the literature. In this paper, we derive the identifiability conditions, starting from the orthogonal ICA maximum likelihood cost function. We then review efficient optimization approaches for orthogonal ICA defined on the Lie group of orthogonal matrices. Afterwards, we derive a new efficient algorithm for orthogonal ICA, by defining a mapping onto a space of constrained matrices which we define as *hyper skew-symmetric*. Finally, we experimentally demonstrate the advantages of the new algorithm over the pre-existing Lie group methods.

Index Terms— Independent Component Analysis, Constrained Optimization, Lie Group Methods, Identifiability

1. INTRODUCTION

Independent component analysis (ICA) is a data-driven technique commonly used for blind source separation (BSS), as well as for studying the latent structure of datasets. ICA decomposes a dataset into latent *sources* according to the assumption that the sources are statistically independent. ICA has been successfully used in a diverse range of applications across the sciences [1–7].

ICA is typically performed on pre-whitened data. A key consequence of pre-whitening is that as both the pre-whitened data and the true sources are uncorrelated, then the true demixing matrix must be orthogonal (or nearly orthogonal for large number of samples T). This is exploited in many popular ICA algorithms to limit the solution space to orthogonal matrices, leading to considerably more efficient ICA algorithms [8, 9]. Despite the popularity of these *orthogonal* algorithms, in the literature there lacks a proof of how orthogonality constraint theoretically affects maximum likelihood estimation of ICA. Thus, in this paper we produce this proof: first writing the cost function of orthogonal ICA under maximum likelihood, then deriving the gradient, the Fisher Information matrix (FIM), and the orthogonal ICA identifiability conditions.

We then discuss existing methods for the orthogonal ICA, and focus on *symmetric* updates as these are preferable in ICA due to their robustness with respect to estimation error. We present several methods, all that can be considered as *Lie Group* methods, and compare their convergence properties.

This work was supported in part by NSF-CCF 1618551, NSF-NCS 1631838, and NIH R01 MH118695.

Optimization based on Lie theory is useful in that it both reduces the complexity of the problem, and defines a natural connection between unconstrained and constrained optimization. However, while Lie group methods have proven useful, it is also useful for optimization methods to have limits on the degree of change, e.g., to ensure stability for stochastic or second order updates. We thus derive a new orthogonal algorithm that allows this limiting behavior to occur naturally in mapping an unconstrained search direction to the constraint. This algorithm generalizes well across multiple types of data, and in general does not require any initial parameter tuning to achieve nearly the fastest convergence we observed (for fixed parameters chosen in the range of possible parameter values).

The paper is organized as follows. Section 2 outlines maximum likelihood for orthogonal ICA, after which the gradient, FIM, and identifiability conditions of orthogonal ICA are derived, and the existing Lie group methods are described. Section 3 presents a new orthogonal algorithm based on the derivation of a matrix type which we call *hyper skew-symmetric*. Section 4 compares the performance of these algorithms across real and simulated data, and Section 5 concludes with takeaways on the new algorithm.

2. ORTHOGONAL ICA: THEORY AND METHODS

2.1. ICA preliminaries

We consider the general ICA problem where the observed data is modeled as a random process. With $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]^\top \in \mathbb{R}^N$ denoting the N underlying sources at some sample index t, ICA assumes that sources are mixed by unknown invertible matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$, to produce observed mixtures $\mathbf{x}(t) = [x_1(t), \ldots, x_N(t)]^\top \in \mathbb{R}^N$. Here, $(.)^\top$ denotes the transpose. The ICA generative model is thus represented as:

$$\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t)\,, \qquad \text{or} \qquad \mathbf{X} = \mathbf{A}\,\mathbf{S} \tag{1}$$

across T observed samples of the random process. Here $\mathbf{X}, \mathbf{S} \in \mathbb{R}^{N\times T}$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top$ and $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_N]^\top$. Here $\mathbf{x}_n$ and $\mathbf{s}_n$ are column vectors of $\mathbf{X}$ and $\mathbf{S}$ respectively, for $n = 1, 2, \ldots, N$. ICA estimates a demixing matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$ that maximizes independence between the source estimates $\mathbf{Y} = \mathbf{W}\mathbf{X}$, with $\mathbf{Y} \in \mathbb{R}^{N\times T}$, and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^\top$.

The maximum likelihood cost function for ICA is given by:

$$\mathcal{J}_{ICA}(\mathbf{W}) = \sum_{n=1}^{N} \log p_{S_n}(\mathbf{y}_n) + T \log\left(\det \mathbf{W}\right)\,, \tag{2}$$

where $p_{S_n}(\mathbf{y}_n)$ is the probability distribution function (PDF) of the nth underlying independent source. Maximization of the likelihood in (2) can be shown to be equivalent to the minimization of the mutual information among the source estimates.

When deriving ICA using the likelihood formulation for sources $\mathbf{S}$, one presumes a differentiable probability density $p_S(\mathbf{S})$, associated with the function $\mathbf{\Phi} : \mathbb{R}^{N\times T} \longrightarrow \mathbb{R}^{N\times T}$:

$$[\mathbf{\Phi}(\mathbf{S})]_{nt} = \frac{\partial \log p_S(\mathbf{S})}{\partial\, s_n(t)}; \qquad \begin{matrix} n = 1, 2, \ldots, N \\ t = 1, 2, \ldots, T \end{matrix}$$

This is called the *score function* for density model $p_S(\mathbf{S})$. The score function matrix $\mathbf{\Phi} \in \mathbb{R}^{N\times T}$ is expressed for true sources by $\mathbf{\Phi_S} = [\boldsymbol{\phi}_{\mathbf{s}_1}, \ldots, \boldsymbol{\phi}_{\mathbf{s}_N}]^\top$, and for source estimates by $\mathbf{\Phi_Y} = [\boldsymbol{\phi}_{\mathbf{y}_1}, \ldots, \boldsymbol{\phi}_{\mathbf{y}_N}]^\top$, for sources $n = 1, 2, \ldots, N$.

In our derivations, we assume that mixtures $\mathbf{x}_n$, and sources $\mathbf{s}_n$, are each standardized. Due to this, and the independence of sources, sources are both orthogonal with respect to themselves, and their respective score function components [10]. This is represented respectively by the identities, for source indices $1 \le m, n \le N$:

$$\mathrm{E}\{s_m(t)\, s_n(t)\} = \delta_{mn}\ , \qquad \mathrm{E}\{\phi_m(t)\, s_n(t)\} = \delta_{mn},$$

where $\delta_{mn} = 1$ if $m = n$, and $\delta_{mn} = 0$ otherwise.

In order to accurately estimate the demixing matrix, use of orthogonal ICA algorithms requires that the data $\mathbf{X}$ must be pre-whitened. The reasoning is that given standardized $\mathbf{X}$, it follows from the first identity that $\mathbf{R_s} = \mathrm{E}\{\mathbf{s}(t)\, \mathbf{s}^\top(t)\} = \mathbf{I}$, thus $\mathbf{R_x} = \mathrm{E}\{\mathbf{x}(t)\, \mathbf{x}^\top(t)\} = \mathbf{A}\mathbf{A}^\top$. Thus, $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$ is only guaranteed when $\mathbf{X}$ is pre-whitened. Note that $\mathbf{R_s}$ in practice will be approximated by a sample average $\hat{\mathbf{R}}_\mathbf{s}$ for finite T, hence $\hat{\mathbf{R}}_\mathbf{s} = \mathbf{I}$ is satisfied only as $T \to \infty$, provided that the random process is covariance-ergodic. However, given pre-whitened data with sufficient T such that $\hat{\mathbf{R}}_\mathbf{s} \approx \mathbf{I}$, then $\mathbf{A}\mathbf{A}^\top \approx \mathbf{I}$, thus justifying using an orthogonal algorithm to find the orthogonal solution nearest to $\mathbf{A}$.

2.2. Orthogonal ICA: cost function and gradient

We start by applying the constraint $\mathbf{W}\mathbf{W}^\top - \mathbf{I} = \mathbf{0} \in \mathbb{R}^{N\times N}$ to the unconstrained ICA maximum likelihood cost function: $\mathcal{J}_{ICA}(\mathbf{W}) = \sum_{n=1}^{N} \log p_{S_n}(\mathbf{y}_n) + T \log(\det\mathbf{W})$. To incorporate the constraint into the cost, we use the Lagrangian function [11] as applied to constrained matrices:

$$\mathcal{L}(\mathbf{W}, \mathbf{\Lambda}) = \mathcal{J}_{\mathrm{ICA}}(\mathbf{W}) - \frac{1}{2}\mathrm{tr}\left(\mathbf{\Lambda}\left(\mathbf{W}\mathbf{W}^\top - \mathbf{I}\right)\right)$$

where tr(.) is the trace operator, and $\mathbf{\Lambda}$ is the Lagrangian multiplier. Because $\mathbf{W}\mathbf{W}^\top$ is symmetric, the multiplier $\mathbf{\Lambda}$ corresponding to $\mathbf{W}\mathbf{W}^\top$ is also symmetric, thus $\mathbf{\Lambda} = \mathbf{\Lambda}^\top$.

We solve for $\mathbf{\Lambda}$ such that $\frac{\partial \mathcal{L}(\mathbf{W},\mathbf{\Lambda})}{\partial \mathbf{W}} = \mathbf{0}$ at $\mathbf{W} = \mathbf{A}^{-1}$. First applying the derivative to the constraint term, we get $\frac{\partial\, \mathrm{tr}\left(\mathbf{\Lambda}\left(\mathbf{W}\mathbf{W}^\top - \mathbf{I}\right)\right)}{\partial \mathbf{W}} = 2\mathbf{\Lambda}\mathbf{W}$. With unconstrained ICA cost gradient, given by $\frac{\partial \mathcal{J}_{\mathrm{ICA}}(\mathbf{W})}{\partial \mathbf{W}} = -\mathbf{\Phi_Y}\mathbf{X}^\top + T\left(\mathbf{W}^{-1}\right)^\top$, we form the derivative of the Lagrangian, given by $\frac{\partial \mathcal{L}(\mathbf{W},\mathbf{\Lambda})}{\partial \mathbf{W}} = -\mathbf{\Phi_Y}\mathbf{X}^\top + T\left(\mathbf{W}^{-1}\right)^\top - \mathbf{\Lambda}\mathbf{W}$. At $\frac{\partial \mathcal{L}(\mathbf{W},\mathbf{\Lambda})}{\partial \mathbf{W}} = \mathbf{0}$, we have $\mathbf{W}\mathbf{A} = \mathbf{I}_N$, and $\mathbf{Y} = \mathbf{S}$. We thus set the Lagrangian equal to $\mathbf{0}$ to solve for $\mathbf{\Lambda}$, and obtain $\mathbf{\Lambda} = -\mathbf{\Phi_S}\mathbf{S}^\top + T\,\mathbf{I}_N$.

We thus plug in $\mathbf{\Lambda}$ into $\frac{\partial \mathcal{L}(\mathbf{W},\mathbf{\Lambda})}{\partial \mathbf{W}}$, and form the orthogonal ICA cost gradient, $\frac{\partial \mathcal{J}_{\mathrm{ICA}}^{\mathrm{ORTH}}(\mathbf{W})}{\partial \mathbf{W}}$. We can now compare this gradient (3) to the gradient of unconstrained ICA (4), derived from the maximum likelihood cost without the constraint:

$$\frac{\partial \mathcal{J}_{\mathrm{ICA}}^{\mathrm{ORTH}}(\mathbf{W})}{\partial \mathbf{W}} = -\mathbf{\Phi_Y}\mathbf{X}^\top + \mathbf{\Phi_S}\mathbf{S}^\top\mathbf{W} \tag{3}$$

$$\frac{\partial \mathcal{J}_{\mathrm{ICA}}(\mathbf{W})}{\partial \mathbf{W}} = -\mathbf{\Phi_Y}\mathbf{X}^\top + T\left(\mathbf{W}^{-1}\right)^\top \tag{4}$$

In comparing these, we first note that $\mathbf{W} = \left(\mathbf{W}^{-1}\right)^\top$ when $\mathbf{W}$ is orthogonal. Then due to our earlier identity $\mathrm{E}\{\phi_m(t)\, s_n(t)\} = \delta_{mn}$, we have that $\lim_{T\to\infty} \mathbf{\Phi_S}\mathbf{S}^\top = T\,\mathbf{I}$, at which point (3) and (4) are equivalent. Thus, given that $\mathbf{W}$ is orthogonal, the derivatives in (3) and (4) become asymptotically equivalent as T approaches infinity. In the next section, we consider the FIM of orthogonal ICA when dealing with finite samples.

2.3. Orthogonal ICA: Fisher Information Matrix

We now consider the elementwise FIM:

$$[\mathbf{F}(\mathbf{W})]_{m_2 n_2}^{m_1 n_1} = E\left\{\left(\frac{\partial \mathcal{J}_{\mathrm{ICA}}^{\mathrm{ORTH}}(\mathbf{W})}{\partial \mathbf{W}_{m_1 n_1}}\right)\left(\frac{\partial \mathcal{J}_{\mathrm{ICA}}^{\mathrm{ORTH}}(\mathbf{W})}{\partial \mathbf{W}_{m_2 n_2}}\right)^\top\right\}, \tag{5}$$

which is given as the covariance of the elementwise orthogonal ICA gradient $\frac{\partial \mathcal{J}_{\mathrm{ICA}}^{\mathrm{ORTH}}(\mathbf{W})}{\partial \mathbf{W}_{mn}}$. It is easy to show that the elementwise gradient of (3) (defined as $\boldsymbol{\Delta_{mn}}$) is given by:

$$\boldsymbol{\Delta_{mn}} = -\boldsymbol{\phi}_{\mathbf{y}_m}^\top \mathbf{x}_n + \sum_{k=1}^{N} [\mathbf{W}]_{kn}\, \boldsymbol{\phi}_{\mathbf{s}_m}^\top \mathbf{s}_k \tag{6}$$

As these are scalars, they are equal to their own transpose, which we use in forming the interior of the FIM expectation. For simplicity, we denote $a_i \triangleq \boldsymbol{\phi}_{\mathbf{y}_{m_i}}^\top \mathbf{x}_{n_i}$, and $b_i \triangleq \sum_{k=1}^{N} [\mathbf{W}]_{kn_i}\, \boldsymbol{\phi}_{\mathbf{s}_{n_i}}^\top \mathbf{s}_k$. Thus using this shorthand notation, the interior of the FIM is given by: $\boldsymbol{\Delta_{mn}}\boldsymbol{\Delta_{mn}}^\top = a_1 a_2^\top + b_1 b_2^\top - b_2 a_1^\top - b_1 a_2^\top$.

For identifiability, we can study what values the interior of the expectation (5) takes at the optimal solution, $\mathbf{W}\mathbf{A} = \mathbf{I}$. To simplify this quantity, due to the equivariance of the maximum likelihood estimator, we can evaluate this quantity at $\mathbf{A} = \mathbf{I}$, $\mathbf{W} = \mathbf{I}$, at which $\mathbf{Y} = \mathbf{X} = \mathbf{S}$, and $\mathbf{\Phi_Y} = \mathbf{\Phi_S}$.

To evaluate the interior of the expectation, it is easy to show that for the 4 terms in the sum:

$$a_1 a_2^\top\big|_{\substack{\mathbf{A}=\mathbf{I}\\ \mathbf{W}=\mathbf{I}}} = b_1 b_2^\top\big|_{\substack{\mathbf{A}=\mathbf{I}\\ \mathbf{W}=\mathbf{I}}} = b_2 a_1^\top\big|_{\substack{\mathbf{A}=\mathbf{I}\\ \mathbf{W}=\mathbf{I}}} = b_1 a_2^\top\big|_{\substack{\mathbf{A}=\mathbf{I}\\ \mathbf{W}=\mathbf{I}}} \tag{7}$$

Given this result, the 4 terms in the interior of the FIM summed together cancel out completely. Therefore,

$$\text{at } \mathbf{W} = \mathbf{A} = \mathbf{I}, \quad [\mathbf{F}(\mathbf{W})]_{m_2 n_2}^{m_1 n_1} = 0 \tag{8}$$

Thus at the true solution, the orthogonal ICA FIM is always singular. Given this result, it is important to understand what can lead to a singular FIM. This can occur when there is a true singularity, e.g., for ICA, when the data consists of two or more Gaussian sources. However, a singular FIM can also occur when the parameters do not cover the entire space of interest. This is the case for orthogonal ICA: from section 2.1, when $\mathbf{R_s} = \mathbf{I}$, then $\mathbf{A}\mathbf{A}^\top = \mathbf{I}$, but $\hat{\mathbf{R}}_\mathbf{s} = \mathbf{I}$ is satisfied only in the limit as $T \to \infty$. For finite samples, or when $\mathbf{X}$ is not whitened, $\mathbf{A}\mathbf{A}^\top \neq \mathbf{I}$, and the FIM will be singular.

As shown in section 2.2, the estimated gradient of orthogonal ICA asymptotically reaches the true gradient of unconstrained ICA as $T \to \infty$. As the FIM is the covariance of the gradient, this asymptotic equivalence also extends to the FIM of unconstrained and orthogonal ICA, thus the FIM of orthogonal ICA asymptotically reaches the true FIM of unconstrained ICA as $T \to \infty$. With the identifiability conditions defined on the FIM, this means that in the limit, given pre-whitened $\mathbf{X}$, orthogonal ICA has the same identifiability conditions as unconstrained ICA. Therefore, with orthogonal ICA, sources can be sufficiently estimated even for finite T, with this estimate further improved for larger T (reflective of improved statistical power, in addition to the true $\mathbf{A}$ being closer to orthogonal).

Having shown that orthogonal ICA asymptotically achieves the identifiabilty conditions of unconstrained ICA, in the next section we introduce the methods used to perform orthogonal ICA.

2.4. Methods for orthogonal ICA: Lie Group Methods

Orthogonal methods operate via *rotations*, called special orthogonal matrices (denoted by SO(N)). The SO(N) manifold forms a Lie group corresponding to a vector space called a Lie algebra, which for SO(N) are the skew-symmetric matrices ($\mathbf{D} + \mathbf{D}^\top = \mathbf{0}$). The Lie group O($N$) is fully characterized by the connection between the Lie group and Lie algebra, via a mapping called the *exponential map*. This reduces the number of parameters from N^2 to $\frac{N(N-1)}{2}$, considerably reducing the complexity of the optimization problem. These orthogonal methods thus operate by first calculating an unconstrained search direction, then mapping the direction to the nearest skew-symmetric matrix, mapping the skew-symmetric matrix to the method's corresponding rotation matrix $\mathbf{R}$, and then updating $\mathbf{W}$ by the rotation $\mathbf{W} \rightarrow \mathbf{RW}$ [9, 12].

One method of calculating $\mathbf{R}$ is via the matrix exponential, often called the Geodesic Flow method [9,13]. However, calculation of the matrix exponential is in general only an approximation. In contrast, the Cayley Transform calculates this without approximation by using matrix inversion, and is popular for SO(N) optimization [11]. As these methods can have significant computational complexity, another method produces an approximation of the rotation by use of only matrix multiplication [14]. This technique, *Infinitesimal skew-symmetrix rotation*, or INF-SSM, converts the skew-symmetric matrix into an infinitesimal rotation by a chosen scaling factor $\frac{1}{\beta}$, then scales the rotation back by raising the infinitesimal rotation to the power β. Given stepsize α and skew-symmetric matrix $\mathbf{D}$, these three techniques have the update rotation matrix $\mathbf{R}$ of the form:

$$\mathbf{R}_{\text{GEODESIC FLOW}}(\alpha, \mathbf{D}) = \exp(\alpha \mathbf{D}) \tag{9}$$

$$\mathbf{R}_{\text{CAYLEY TRANSFORM}}(\alpha, \mathbf{D}) = (\mathbf{I} - \frac{\alpha}{2}\mathbf{D})^{-1}(\mathbf{I} + \frac{\alpha}{2}\mathbf{D}) \tag{10}$$

$$\mathbf{R}_{\text{INF SSM}}(\alpha, \mathbf{D}) = (\mathbf{I} + \frac{1}{\beta}\alpha \mathbf{D})^{\beta} \tag{11}$$

These techniques are useful in operating within a vector space, where linear operators are defined and the norm of the original unconstrained direction is proportional to the degree of the orthogonal rotation (e.g., multiplying $\mathbf{D}$ by some scalar μ is equivalent to raising the corresponding rotation $\mathbf{R}$ to the exponent μ). However, in optimization it may be useful to naturally limit the norm of a search direction, e.g., for cases where stability is not guaranteed. In the next section, we define a *hyper skew-symmetric* matrix, and use it to create a new algorithm for orthogonal ICA.

3. IGLOO: DERIVATION FOR A HYPER SKEW-SYMMETRIC MATRIX

Given an orthogonal matrix $\mathbf{W}$, we would like to find update $\mathbf{W} + \mathbf{DW}$, such that this updated matrix is still orthogonal. To see what this entails, we can look at what is required to make $\mathbf{W} + \mathbf{DW}$ orthogonal:

$$(\mathbf{W} + \mathbf{DW})(\mathbf{W} + \mathbf{DW})^\top = \mathbf{I} + \mathbf{D} + \mathbf{D}^\top + \mathbf{D}\mathbf{D}^\top = \mathbf{I} \tag{12}$$

This expression shows that in order for $\mathbf{W} + \mathbf{DW}$ to remain orthogonal, we require that matrix $\mathbf{D}$ satisfy the constraint $\mathbf{D} + \mathbf{D}^\top + \mathbf{D}\mathbf{D}^\top = \mathbf{0}$. Given the similarity to the skew symmetric matrices, we call these *hyper skew-symmetric* matrices. Our goal is to develop a mapping of any square matrix $\mathbf{D}_p$ into this constraint $\mathbf{D}$. There are some observations that emerge out of this constraint:

$$\mathbf{I} = \mathbf{I} + \mathbf{D} + \mathbf{D}^\top + \mathbf{D}\mathbf{D}^\top = (\mathbf{I} + \mathbf{D})(\mathbf{I} + \mathbf{D})^\top \tag{13}$$

The result of this shows that any hyper skew-symmetric matrix $\mathbf{D}$, plus an identity matrix, always equals some orthogonal matrix. While it is not immediately clear how to map a matrix $\mathbf{D}_p$ onto the space of hyper skew-symmetric matrices $\mathbf{D}$, we can map $\mathbf{D}_p$ onto its image in the space of orthogonal matrices, O$(\mathbf{D}_p)$:

$$\text{O}(\mathbf{D}_p) = \mathbf{D}_p \left(\mathbf{D}_p^\top \mathbf{D}_p\right)^{-\frac{1}{2}} \tag{14}$$

Thus, one possible way to map a matrix $\mathbf{D}_p$ into a hyper skew-symmetric matrix, is to orthogonalize $\mathbf{D}_p$, and then subtract an identity matrix. This is given by $\mathbf{D} = \text{O}(\mathbf{D}_p) - \mathbf{I}$. While this is one possible mapping, it may not be the optimal mapping: we desire the optimal image of $\mathbf{D}_p$ onto the space of hyper skew-symmetric matrices. This ambiguity is realized by arbitrary rotations $\mathbf{W}_r$: $\mathbf{I} = (\mathbf{I} + \mathbf{D})\mathbf{W}_r\mathbf{W}_r^\top(\mathbf{I} + \mathbf{D})^\top$.

Here we have the subscript on arbitrary orthogonal matrix $\mathbf{W}_r$, to distinguish it from our ICA parameter $\mathbf{W}$. Given this ambiguity, we can represent any mapping into the space of hyper skew-symmetric matrices by the following:

$$\mathbf{D} = \text{O}(\mathbf{D}_p)\mathbf{W}_r^\top - \mathbf{I} \tag{15}$$

This shows the ambiguity of possible mappings by existence of arbitrary rotation $\mathbf{W}_r$. To find $\mathbf{W}_r$ that gives the optimal mapping, we seek $\mathbf{W}_r$ that minimizes the distance between $\mathbf{D}$ and $\mathbf{D}_p$:

$$\min_{\mathbf{W}_r} \text{dist}\left(\text{O}(\mathbf{D}_p)\mathbf{W}_r^\top - \mathbf{I} \quad , \quad \mathbf{D}_p\right)$$

To isolate $\mathbf{W}_r$ in this expression, we can add $\mathbf{I}$ to both quantities, and then rotate by an ortogonal matrix O$(\mathbf{D}_p)^\top$, and the problem will remain the same: $\min_{\mathbf{W}_r} \text{dist}\left(\mathbf{W}_r^\top \quad , \quad \text{O}(\mathbf{D}_p)^\top(\mathbf{D}_p + \mathbf{I})\right)$.

Now because we know that $\mathbf{W}_r^\top$ is an orthogonal matrix, $\mathbf{W}_r^\top$ can thus be obtained by mapping O$(\mathbf{D}_p)^\top(\mathbf{D}_p + \mathbf{I})$ to the nearest orthogonal matrix, using (14):

$$\mathbf{W}_r^\top = \text{O}\left(\text{O}(\mathbf{D}_p)^\top(\mathbf{D}_p + \mathbf{I})\right) = \text{O}(\mathbf{D}_p)^\top \text{O}(\mathbf{D}_p + \mathbf{I}) \tag{16}$$

Having found optimal $\mathbf{W}_r^\top$, we incorporate it into (15) to get the optimal hyper skew-symmetric matrix:

$$\mathbf{D} = \text{O}(\mathbf{D}_p + \mathbf{I}) - \mathbf{I} \tag{17}$$

This technique can be applied to any search direction $\mathbf{U}$, by representing $\mathbf{U}$ as $\mathbf{U}\mathbf{W}^\top\mathbf{W}$, then mapping $\mathbf{D}_p = \mathbf{U}\mathbf{W}^\top$ to the nearest hyper skew-symmetric matrix via (17).

It is easy to show that this technique achieves a natural limiting behavior on the rotation when the norm of $\mathbf{D}_p$ is large. In applying a stepsize α to $\mathbf{D}_p$, as $\alpha \rightarrow \infty$, the rotation asymptotically approaches a fixed rotation:

$$\lim_{\alpha \rightarrow \infty} \text{O}(\alpha \mathbf{D}_p + \mathbf{I}) - \mathbf{I} = \text{O}(\mathbf{D}_p) - \mathbf{I} \tag{18}$$

Clearly this applies to both α and the norm of $\mathbf{D}_p$. While this limiting behavior has benefits for optimization, the limit to the rotation may be too restrictive. Therefore, we can extend it via introducing a *stretch* parameter β: allowing the extension on the limit by raising the rotation matrix to the power β. This generalization of the hyper skew-symmetric mapping is what we introduce as IGLOO for ICA: Independence by Geometrically Limited Orthogonal Optimizer. Like methods (9) (10) (11), we give the imposed rotation on $\mathbf{W}$, given direction $\mathbf{U}$, stepsize α, and stretch parameter β:

$$\mathbf{R}_{\text{IGLOO}}(\alpha, \mathbf{U}) = \text{O}(\alpha \mathbf{U}\mathbf{W}^\top + \mathbf{I})^{\beta} \tag{19}$$

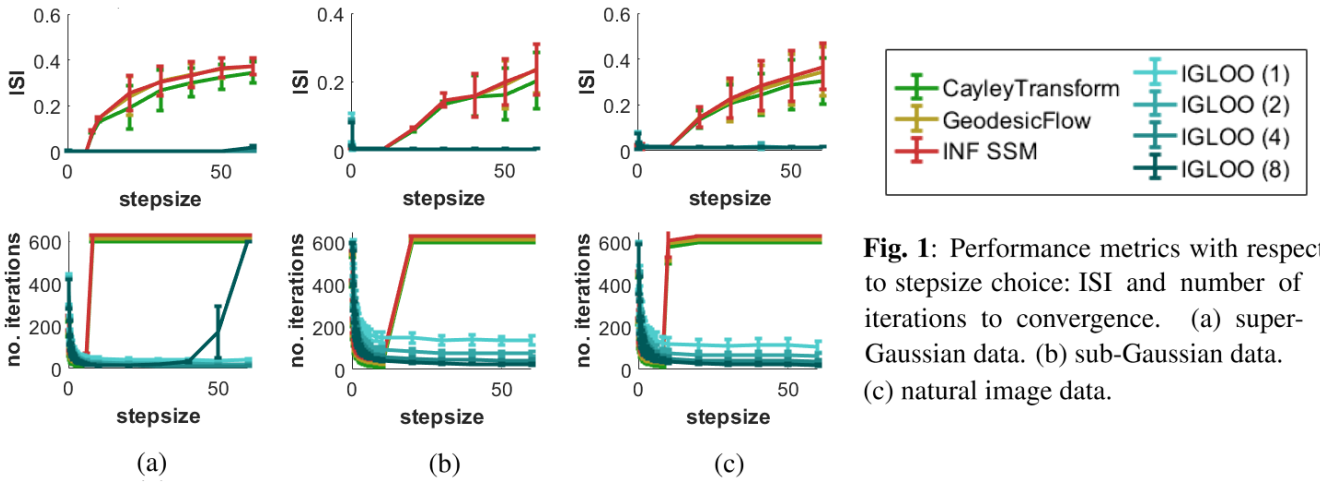

Fig. 1: Performance metrics with respect to stepsize choice: ISI and number of iterations to convergence. (a) super-Gaussian data. (b) sub-Gaussian data. (c) natural image data.

While this can be applied to any search direction $\mathbf{U}$, below we give the pseudocode for the update rule of IGLOO applied specifically to the relative gradient algorithm for ICA [6, 15] :

IGLOO applied to the relative gradient:

1. source estimates: $\mathbf{Y} = \mathbf{WX}$
2. unconstrained direction: $\mathbf{U} = (\mathbf{I} - \mathbf{\Phi_Y}\mathbf{Y}^\top)\mathbf{W}$
3. IGLOO rotation: $\mathbf{R}(\alpha, \mathbf{U}) = \mathrm{O}(\alpha\mathbf{U}\mathbf{W}^\top + \mathbf{I})^\beta$
4. $\mathbf{W}$ update: $\mathbf{W} = \mathbf{RW}$

4. EXPERIMENTAL RESULTS AND CONCLUSIONS

We demonstrate the performance of these methods over simulated and real data. Intersymbol interference (ISI) is the standard metric for measuring ICA performance when the true mixing $\mathbf{A}$ is known, such as for simulated studies [16]. Smaller ISI values reflect a superior demixing performance.

We apply these orthogonal techniques on the relative gradient algorithm for ICA [6,15]. As the ICA cost function is costly to compute, and because optimal stepsize varies considerably between unconstrained and orthogonal ICA, line search is not economical for orthogonal ICA when algorithmic efficiency is a priority. Thus a fixed stepsize can be implemented, and thus we explore algorithm performance across different stepsize choices. For a range of stepsizes, we simulate 200 different mixed datasets (either all super-Gaussian or sub-Gaussian sources, 10 sources of each 80000 samples), and report the average ISI and number of iterations to convergence for these methods. We also note that the number of iterations was nearly exactly proportional to wall time performance, due to the methods' calculation of rotation mappings taking a negligible proportion of the total computational cost per each update. We also experimented over real satellite image data [17], where we likewise mix the data over 200 different mixing matrices. Fig. 1 compares performance across the methods. For INF-SSM, we fixed β to be 2^{20}, and observed that varying β did not significantly affect the performance. For IGLOO, we included 4 choices of β to show how choice of β affects performance.

Optimal performance of IGLOO was nearly identical to that of the Lie group methods (e.g., for super-Gaussian data, 12.5 iters for Lie group methods with $\alpha = 3.5$, vs. 11.5 iters for IGLOO with $\alpha = 10$, $\beta = 4$). However, IGLOO demonstrated the ability to converge even when the stepsize was very high; in fact IGLOO performed nearly optimally for both arbitrarily chosen higher stepsizes and higher value of β. However, IGLOO is not completely immune to when both the stepsize and β may be too high (see IGLOO (β = 8) in Fig.1 (a)). Despite this, this shows IGLOO has a larger range on the direction norm where the algorithm can still converge.

The primary utility in IGLOO's mapping is that when the search direction norm is not economical to control, IGLOO can appropriately limit the corresponding rotation, even in general where parameters α and β are not optimally chosen. Furthermore, these parameters can generally be chosen with arbitrary large values, and IGLOO will perform nearly optimally with respect to parameter choice. As the Lie group methods have a much smaller region of stepsize where the algorithms are able to converge, this presents a considerable advantage of IGLOO for general use in an ICA setting.

We should note that our derivation of IGLOO and the hyper skew-symmetric matrix are not necessarily limited to ICA; both are also applicable to other problems where optimization is done over orthogonal matrices. IGLOO could also be useful in these other applications, especially in situations where this stability property is even more useful than it is for ICA optimization.

5. CONCLUSIONS

In this paper, we provide a comprehensive analysis of orthogonal ICA starting with the maximum likelihood cost function, proving that orthogonal ICA asymptotically has the same identifiability conditions as unconstrained ICA.

After describing the commonly used natural ways of optimizing over orthogonal matrices, we introduce a new algorithm for orthogonal ICA optimization by deriving a matrix we call *hyper skew-symmetric*. We demonstrate that the new algorithm IGLOO is distinct from the Lie group methods by its ability to naturally limit the orthogonal rotation imposed by the search direction, all with minimal use of hyperparameter tuning to obtain optimal performance. These results present IGLOO as an efficient and highly generalizable approach to orthogonal ICA.

6. REFERENCES

[1] Aapo Hyvärinen and Erkki Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[2] Scott Makeig, Anthony J Bell, Tzyy-Ping Jung, and Terrence Sejnowski, "Independent component analysis of electroencephalographic data," in *Advances in neural information processing systems*, 1996, pp. 145–151.

[3] Ganesh R Naik and Dinesh K Kumar, "An overview of independent component analysis and its applications," *Informatica*, 2011.

[4] Tülay Adali, Matthew Anderson, and Geng-Shen Fu, "Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 18–33, 2014.

[5] Tülay Adali, Yuri Levin-Schwartz, and Vince D Calhoun, "Multimodal data fusion using source separation: Two effective models based on ICA and IVA and their properties," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1478–1493, 2015.

[6] J-F Cardoso and Beate H Laheld, "Equivariant adaptive source separation," *IEEE Transactions on signal processing*, vol. 44, no. 12, pp. 3017–3030, 1996.

[7] J-F Cardoso, "Blind signal separation: statistical principles," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009–2025, 1998.

[8] Alper T Erdogan, "On the convergence of ica algorithms with symmetric orthogonalization," *IEEE Transactions on Signal Processing*, vol. 57, no. 6, pp. 2209–2221, 2009.

[9] Mark D Plumbley, "Geometry and manifolds for independent component analysis," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. IEEE, 2007, vol. 4, pp. IV–1397.

[10] Pierre Comon and Christian Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*, Academic press, 2010.

[11] Zaiwen Wen and Wotao Yin, "A feasible method for optimization with orthogonality constraints," *Mathematical Programming*, vol. 142, no. 1-2, pp. 397–434, 2013.

[12] Mark D Plumbley, "Geometrical methods for non-negative ica: Manifolds, lie groups and toral subalgebras," *Neurocomputing*, vol. 67, pp. 161–197, 2005.

[13] Yasunori Nishimori, "Learning algorithm for independent component analysis by geodesic flows on orthogonal group," in *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)*. IEEE, 1999, vol. 2, pp. 933–938.

[14] Arthur D Snider and Marcus McWaters, "Infinitesimal rotations," *American Journal of Physics*, vol. 48, no. 3, pp. 250–251, 1980.

[15] Shun-ichi Amari, Andrzej Cichocki, and Howard Hua Yang, "A new learning algorithm for blind signal separation," in *Advances in neural information processing systems*, 1996.

[16] Pauliina Ilmonen, Klaus Nordhausen, Hannu Oja, and Esa Ollila, "On asymptotics of ica estimators and their performance indices," *arXiv preprint arXiv:1212.3953*, 2012.

[17] Paul Bourke, *Google Earth Fractals*, 2010 (accessed September, 2020).