GEOMETRIC ALGORITHMS FOR THE NON-WHITENED ONE-UNIT LINEAR INDEPENDENT COMPONENT ANALYSIS PROBLEM

Hao Shen\textsuperscript{1}, Klaus Diepold\textsuperscript{1} and Knut Hüper\textsuperscript{2}

\textsuperscript{1}Institute for Data Processing, Technische Universität München, 80290 München, Germany.
Email: \{hao.shen, kldi\}@tum.de

\textsuperscript{2}Department of Mathematics, University of Würzburg, 97074 Würzburg, Germany.
Email: hueper@mathematik.uni-wuerzburg.de

ABSTRACT

In this paper, we study the problem of one-unit linear Independent Component Analysis (ICA) without whitening. The FastICA algorithm is arguably the most popular algorithm for solving the whitened one-unit linear ICA problem. Although a modified FastICA has been already proposed to solve the non-whitened one-unit linear ICA problem, there is unfortunately no known analysis regarding its effectiveness and efficiency. In this work, the non-whitened FastICA algorithm is revisited and analyzed in the framework of geometric optimization algorithms. In this paper, a conjugate gradient (CG) algorithm for the non-whitened one-unit linear ICA problem is developed as well. Local convergence properties of both algorithms are discussed. Finally, local convergence performance of the algorithms is investigated by several numerical experiments.

Index Terms— Independent component analysis (ICA), non-whitening, unit sphere, fixed point algorithm, conjugate gradient (CG) algorithm.

1. INTRODUCTION

Nowadays, linear Independent Component Analysis (ICA) has become a standard statistical tool for solving the linear Blind Source Separation (BSS) problem. It refers to the problem of recovering linearly mixed, statistically independent sources from only observed mixtures, without knowing information about either the source distributions or the mixing process. In the last two decades, various linear ICA algorithms have been developed. In general, there are two major categories of linear ICA algorithms, namely, the contrast-based ICA algorithm \cite{Hyvarinen}, which involves certain optimization procedures of a contrast function, which measures the statistical independence between extracted signals, and the tensorial ICA algorithm \cite{Amari}, which usually requires a joint diagonalisation procedure of a set of cumulant tensors.

Many linear ICA algorithms require a pre-process of the observed signals, i.e. the so-called whitening procedure, which is usually done by Principal Component Analysis (PCA), in order to reduce complexity of the problem and cope with uniqueness of source recovery \cite{Hyvarinen}. It has been shown in \cite{Comon}, however, that performance of linear ICA methods with whitening is limited due to statistical inefficiency in some applications. To avoid such problem, more recently, based on the concept of non-orthogonal joint diagonalization of a set of matrices, several tensorial linear ICA algorithms without whitening have been proposed, e.g., \cite{Lee, Cardoso}. Unfortunately, such tensorial algorithms have in general limited applications in high dimensional problems.

Arguably, the FastICA algorithm proposed in \cite{Hyvarinen} is the most popular algorithm for solving the whitened one-unit linear ICA problem. It enjoys promising properties of strong robustness, fast convergence and easy implementation. However, the classic FastICA requires the whitening approach as a compulsory pre-processing step. Although a modified FastICA algorithm without whitening has already been proposed in \cite{Wipf}, to our best knowledge, there is unfortunately no known analysis regarding its effectiveness and efficiency. Recently, geometric techniques have been successfully applied to analyze the classic FastICA and to develop efficient linear ICA algorithms \cite{Amari}. In this work, the non-whitened FastICA algorithm is revisited and analyzed in the framework of geometric optimization algorithms. A CG non-whitened one-unit linear ICA algorithm is developed as well.

This paper is organized as follows. Section 2 briefly introduces the non-whitened one-unit linear ICA problem and some basic concepts of geometric optimization algorithm. In Section 3, the non-whitened FastICA algorithm is analyzed, and a CG algorithm is developed. Finally in Section 4, local convergence performance of both algorithms is investigated by several numerical experiments.

2. PRELIMINARIES

We consider the general noiseless instantaneous linear ICA model,

\[ w = As, \tag{1} \]

where \( s \in \mathbb{R}^m \) is an \( m \)-dimensional random variable representing \( m \) source signals, \( A \in \mathbb{R}^{m \times m} \) is the mixing matrix of full rank and \( w \in \mathbb{R}^m \) denotes \( m \) observed linear mixtures. We denote by \( s_i \in \mathbb{R} \) and \( w_i \in \mathbb{R} \) the \( i \)-th components of \( s \) and \( w \), respectively. The source signals \( s \) are assumed to be statistically independent and, without loss of generality, to have zero mean and unit variance, i.e.,

\[ E[s] = 0, \quad \text{and} \quad E[ss^\top] = I_m, \tag{2} \]

where \( I_m \in \mathbb{R}^{m \times m} \) is the identity matrix.

The task of the linear ICA problem (1) is to recover the source signals \( s \) by estimating the mixing matrix \( A \) or its inverse \( A^{-1} \) based only on the observations \( w \) via the demixing model

\[ y = X^Tw, \tag{3} \]

where \( X \in \mathbb{R}^{m \times m} \) is the demixing matrix, an estimation of \( A^{-1} \) and \( y \in \mathbb{R}^m \) denotes the corresponding extracted source signals. We refer to such a problem as the \textit{parallel linear ICA problem}. In some applications, one might prefer to extract only one desired source,
rather than all sources. Let $X = [x_1, \ldots, x_m] \in \mathbb{R}^{m \times m}$. Then a single source is recovered by simply the following
\[ y_i = x_i^T w, \] (4)
which is here referred as the one-unit linear ICA problem.

According to theorem 11 in [1], a correct demixing matrix $X^* \in \mathbb{R}^{m \times m}$ can only be identified up to an arbitrary order and scaling, i.e., $X^*$ is the inverse of $A^*$ up to an $m \times m$ permutation matrix $P$ and an $m \times m$ diagonal matrix $D$
\[ X^* = A^{-T} DP. \] (5)
Then, a correct demixing vector $x_i^* \in \mathbb{R}^m$, which can, without loss of generality, be assumed to extract the $i$-th source signal, is identified up to an arbitrary scaling $d_i \in \mathbb{R} \setminus \{0\}$, i.e.,
\[ x_i^* = d_i (A^{-T} e_i), \] (6)
where $e_i$ denotes the $i$-th standard basis vector of $\mathbb{R}^m$.

Let $\mathbb{R}^*_m := \mathbb{R}^m \setminus \{0\}$ be the Euclidean space $\mathbb{R}^m$ with the origin removed. The original non-whitened FastICA algorithm can be considered as iterating the following map
\[ \Phi: \mathbb{R}^*_m \to \mathbb{R}^*_m, \]
\[ x \mapsto C^{-1} \mathbb{E} \left[ G' \left( \frac{x^T w}{\|x\|^2} \right) w \right] - \mathbb{E} \left[ G'' \left( \frac{x^T w}{\|x\|^2} \right) \right] x \phi(\xi), \] (7)
where $C := \mathbb{E} [ww^T] = AA^T$ is the covariance matrix of the observations $w$, the term $\phi(\xi)$ normalizes an estimated signal $x_w$ and, for $G': \mathbb{R} \to \mathbb{R}$ and $G''': \mathbb{R} \to \mathbb{R}$ are the first and second derivatives of a smooth nonlinear function $G': \mathbb{R} \to \mathbb{R}$, respectively. The function $G$ is chosen according to the specific application and in general to satisfy the following assumption. We refer to theorem 1 in [7] for further discussions.

**Assumption 1** The nonlinear function $G: \mathbb{R} \to \mathbb{R}$ is smooth, even and chosen such that the following inequality holds true for all sources in the linear ICA problem (1), i.e., for all $i = 1, \ldots, m$
\[ \mathbb{E} \left[ G''(s_i) \right] - \mathbb{E} \left[ G'(s_i) s_i \right] \neq 0. \] (8)

It is easily seen that by assumption 1, the map (7) is scale invariant, i.e.,
\[ \Phi(\xi) = \Phi(\lambda \xi), \] (9)
for any $\lambda \in \mathbb{R} \setminus \{0\}$. Therefore, the map $\Phi$ as defined in (7) induces a map on the $(m-1)$-dimensional real projective space $\mathbb{RP}^{m-1}$, which can only be identified up to an arbitrary order and scaling.

The $(m-1)$-dimensional real projective space $\mathbb{RP}^{m-1}$ can be defined as
\[ \mathbb{RP}^{m-1} := \left\{ [x] \mid x \in \mathbb{R}^m \setminus \{0\} \right\}, \] (10)
with the equivalence class of $x \in \mathbb{R}^m$ defined by $[x] := \lambda x$, where $\lambda \in \mathbb{R} \setminus \{0\}$.

The non-whitened FastICA algorithm can then be considered as the following map on $\mathbb{RP}^{m-1}$
\[ \Phi: \mathbb{RP}^{m-1} \to \mathbb{RP}^{m-1}, \quad [x] \mapsto [\Phi(x)]. \] (11)

Nevertheless, in practice, for the sake of simplicity, the unit sphere $S^{m-1} := \{ x \in \mathbb{R}^m \mid \|x\| = 1 \}$ is usually used instead of $\mathbb{RP}^{m-1}$, due to the local diffeomorphism between $\mathbb{RP}^{m-1}$ and $S^{m-1}$. Therefore, all solutions of the non-whitened one-unit linear ICA problem (4) can be identified as:

**Corollary 1** Let $A \in \mathbb{R}^{m \times m}$ be the mixing matrix in the linear ICA problem (1). Then all correct solutions for the one-unit linear ICA problem (4) form the set
\[ \Theta := \left\{ e \in \mathbb{R}^m \mid x = A^{-T} e_i \right\} \subset S^{m-1}. \] (12)

Thus, the non-whitened FastICA can also be considered as a map on $S^{m-1}$
\[ \tilde{\Phi}: S^{m-1} \to S^{m-1}, \quad x \mapsto \Phi(x), \] (13)
which induces the same map $\Phi$ as the map $\tilde{\Phi}$. Moreover, inspired by the construction of $\Phi$ as defined in (7), we propose the following cost function for the non-whitened one-unit linear ICA problem
\[ f: S^{m-1} \to \mathbb{R}, \quad x \mapsto \mathbb{E} \left[ G \left( \frac{x^T w}{\|x\|^2} \right) \right]. \] (14)

Before we continue, let us review some basic concepts of geometric optimization algorithm on $S^{m-1}$. It is well-known that the unit sphere $S^{m-1}$ is an $(m-1)$ dimensional smooth submanifold of the embedding space $\mathbb{R}^m$. The tangent space of $S^{m-1}$ at $x \in S^{m-1}$ is
\[ T_x S^{m-1} = \left\{ \xi \in \mathbb{R}^m \mid x^T \xi = 0 \right\}. \] (15)

To develop Riemannian algorithms on $S^{m-1}$, it requires the concept of parallel transport of tangent vectors along a geodesic. Given the Riemannian metric on $S^{m-1}$ induced by the Euclidean metric of the embedding space $\mathbb{R}^m$, i.e.,
\[ \langle \xi_1, \xi_2 \rangle := \xi_1^T \xi_2, \] (16)
with $\xi_1, \xi_2 \in T_x S^{m-1}$, a great circle $\mu_x$ of $S^{m-1}$ at $x \in S^{m-1}$ for a given tangent direction $\xi \in T_x S^{m-1}$ is defined as follows
\[ \mu_x : \mathbb{R} \to S^{m-1}, \quad \mu_x(t) := x \cos t + \xi \sin t, \quad \|\xi\| = 0; \] (17)
\[ \|\xi\| \neq 0, \quad \text{otherwise}. \]
Clearly, $\mu_x(0) = x$ and $\mu_x(0) = \xi$. The parallel transport $\tau$ of a tangent vector $\xi \in T_x S^{m-1}$ along the great circle $\mu_x$ is given by
\[ \tau(\xi) := x^T \xi \xi^T \xi \sin t \|\xi\| + \xi (1 - \cos t \|\xi\|) . \] (18)

3. GEOMETRIC NON-WHITENED ONE-UNIT LINEAR ICA ALGORITHMS

In this section, we investigate two geometric optimization algorithms for the non-whitened one-unit linear ICA problem, namely, the non-whitened FastICA algorithm and a conjugate gradient algorithm.

3.1. The Non-Whitened FastICA Algorithm

In this subsection, local convergence properties of the whitened FastICA algorithm are briefly studied in the framework of a scalar shifted fixed point algorithm, i.e., interpreting the whitened FastICA algorithm, considered as $\Phi$ in (16), as a scalar shifted fixed point algorithm. We refer to [8] for more details.

By denoting the first term of Eq. (7) as
\[ K: S^{m-1} \to \mathbb{R}, \quad K(x) := C^{-1} \mathbb{E} \left[ G' \left( \frac{x^T w}{\|x\|^2} \right) \right], \] (19)
and constructing a base map as
\[ \Psi: S^{m-1} \to S^{m-1}, \quad x \mapsto \frac{K(x)}{\|K(x)\|}. \] (20)
the algorithmic map \( \Phi \) as defined in (16) is just a scalar shifted version of \( \Psi \) as in (23) with the scalar shift being simply

\[
\sigma: S^{m-1} \to \mathbb{R}, \quad \sigma(x) := \frac{1}{\varphi(x)} \mathbb{E} \left[ \frac{\partial^2 f}{\partial x^i \partial x^j} \right].
\]

(24)

Given an \( x^* \in \Theta \), assumed to extract the \( i \)-th source, a direct computation leads to

\[
\Phi(x^*) = \varphi(x^*) \left( \mathbb{E} \left[ G' (s_i) s_i \right] - \mathbb{E} \left[ G'' (s_i) \right] \right)x^*.
\]

(25)

Following Assumption 1, it is easily seen

**Lemma 1** Any \( [x^*] \in \mathbb{R}^m \) with \( x^* \in \Theta \) is a fixed point of the non-whitened FastICA considered as a map \( \Phi \), defined in (13).

Computing the first derivative of \( K \) at \( x^* \in \Theta \) in tangent direction \( \xi \in T_{x^*} S^{m-1} \) gives

\[
D K(x) \xi \big|_{x=x^*} = \frac{1}{\varphi(x^*)} \mathbb{E} \left[ G'' (s_i) \right] \xi.
\]

(26)

Then, by theorem 4.2 in [8], we conclude

**Theorem 1** The non-whitened FastICA algorithm considered as the map \( \Phi \), defined in (13), is locally quadratically convergent to a correct demixing \( [x^*] \in \mathbb{R}^m \) for any \( x^* \in \Theta \).

**Remark 1** There is an obvious drawback of the non-whitened FastICA algorithm. The algorithm requires to compute the inverse of the covariance matrix \( C \), which is in general computationally equivalent to a whitening approach via PCA, and turns to be unstable when \( C \) is ill-conditioned. Nevertheless, the expression \( \Phi(x) \) as defined in (7) can be computed more efficiently and stably by solving the following linear system for \( h \in \mathbb{R}^m \)

\[
C \cdot h = \mathbb{E} \left[ G' \left( \frac{x \cdot w}{\varphi(x)} \right) w \right] - \mathbb{E} \left[ G'' \left( \frac{x \cdot w}{\varphi(x)} \right) \right] \xi_{x^*}.
\]

(27)

However, such an alternative approach requires to solve the linear system (27) at each iteration, which is certainly not always feasible for a high dimensional problem.

### 3.2. A Conjugate Gradient Algorithm

In this subsection, we develop a conjugate gradient algorithm, which avoids the difficulties as described in Remark 1, for solving the non-whitened one-unit linear ICA problem, i.e., optimizing the cost function \( f \) as defined in (17). Without loss of generality, we assume to minimize the function \( f \).

Firstly, we give a critical point analysis of \( f \). By the chain rule, the first derivative of \( f \) at \( x \in S^{m-1} \) in direction \( \xi \in T_x S^{m-1} \) is computed by

\[
D f(x) \xi = \mathbb{E} \left[ G' \left( \frac{x \cdot w}{\varphi(x)} \right) \left( \frac{x \cdot w}{\varphi(x)} - \frac{(x^T w)(\xi^T C x)}{\varphi(x)^2} \right) \right].
\]

(28)

Let \( x^* \in \Theta \), we compute

\[
D f(x) \xi \big|_{x=x^*} = 0,
\]

(29)

for \( \xi \in T_{x^*} S^{m-1} \), i.e., any correct demixing vector \( x^* \) is a critical point of the function \( f \).

Now, let us assume that \( x^* \in \Theta \) extracts the \( i \)-th source and let the mixing matrix \( A = [a_1, \ldots, a_m] \in \mathbb{R}^m \times m \). By tedious computations, the second derivative of \( f \) at \( x^* \in \Theta \) in direction \( \xi \in T_{x^*} S^{m-1} \) is calculated as follows

\[
D^2 f(x)(\xi, \xi) \big|_{x=x^*} = \frac{\xi^T (C-a_i a_i^T) \xi}{x^T C x^*} \left( \mathbb{E} \left[ G'' \left( \frac{x^T w}{\varphi(x)} \right) \right] - \mathbb{E} \left[ G' \left( \frac{x^T w}{\varphi(x)} \right) \right] x^T \right) \cdot \left( \mathbb{E} \left[ G'' \left( \frac{x^T w}{\varphi(x)} \right) \right] - \mathbb{E} \left[ G' \left( \frac{x^T w}{\varphi(x)} \right) \right] x^T \right) \xi.
\]

(30)

By exploring the relation between \( x^* \) and \( A \), as in Corollary 1, and exploiting the structure of \( \xi \in T_{x^*} S^{m-1} \), it can be shown that the following symmetric bilinear form

\[
\mathcal{F}: T_{x^*} S^{m-1} \times T_{x^*} S^{m-1} \to \mathbb{R},
\]

\[
(\xi, \xi) \mapsto \xi^T \left( C - a_i a_i^T \right) \xi
\]

(31)

is positive definite. Therefore, recalling Assumption 1, i.e., \( \alpha(x^*) \neq 0 \), we conclude

**Corollary 2** Any \( x^* \in \Theta \) is a non-degenerated critical point of the cost function \( f \) as defined in (17).

**Remark 2** It is known that the classic FastICA algorithm can be interpreted as either an approximate Newton algorithm or a scalar shifted fixed point algorithm [8]. The analysis in Section 3.1 has shown that the non-whitened FastICA algorithm is indeed a scalar shifted fixed point algorithm as well. Although an explicit structure of the Hessian of \( f \) at a desired critical point \( x^* \in \Theta \) is already given in (30), it is, however, hardly possible to estimate the term \( a_i \) in (30). In other words, the strategy of Hessian approximation, which leads to an interpretation of the classic FastICA as an approximate Newton algorithm, does not apply to the non-whitened scenario.

In the rest of this section, we quickly develop a conjugate gradient algorithm for minimizing the cost function \( f \). A general scheme of a conjugate gradient algorithm on a Riemannian manifold (algorithm 5.2 in [9]) can be easily adapted to the current scenario, i.e. on \( S^{m-1} \), as follows.

Firstly, by recalling the first derivative of \( f \) as in (28), the Riemannian gradient of \( f \) is computed as

\[
\nabla f(x) = \pi(x) \mathbb{E} \left[ G' \left( \frac{x \cdot w}{\varphi(x)} \right) \left( \frac{x \cdot w}{\varphi(x)} - \frac{(x^T w)(\xi^T C x)}{\varphi(x)^2} \right) \right],
\]

(32)

where \( \pi(x) := I_m - xx^T \) is the orthogonal projection operator onto the tangent space \( T_{x^*} S^{m-1} \). Then, a conjugate gradient algorithm for minimizing the function \( f \) as defined in (17) is as follows

**Algorithm 1** A conjugate gradient non-whitened one-unit linear ICA algorithm

**Step 1:** Given an initial guess \( x^{(0)} \in S^{m-1} \) and set \( i = 0 \).

**Step 2:** Set \( i = i + 1 \), let \( x^{(i)} = x^{(i-1)} \), and compute

\[
\xi^{(i)} = \xi^{(i)} - \nabla f(x^{(i)}).
\]

**Step 3:** For \( j = 1, \ldots, m - 1 \):

(i) Update \( x^{(i)} \leftarrow \mu_{x^{(i)}} (\lambda^*) \), where

\[
\lambda^* = \arg \min_{\lambda} f(\mu_{x^{(i)}} (\lambda));
\]

(ii) Compute \( \xi^{(j+1)} = -\nabla f(x^{(i)}) \);

(iii) Update \( \xi^{(j+1)} = \xi^{(j+1)} + \gamma \, \eta(\xi^{(j)}) \), where \( \gamma \) is chosen such that \( \eta(\xi^{(j)}) \) and \( \xi^{(j+1)} \) conjugate with respect to the Hessian of \( f \) at \( x^{(i)} \).

**Step 4:** If \( ||x^{(i+1)} - x^{(i)}|| \) is small enough, stop. Otherwise, go to Step 2.
For updating the direction parameter $x$ in Step 3-(iii), we confine ourselves to a formula proposed recently in [10], which is based on a one-dimensional Newton step, i.e.

$$
\lambda^* = -\frac{\partial f_{\mu_x(i)}}{\partial x} |_{x=0}. \tag{33}
$$

For updating the direction parameter $\gamma$ in Step 3-(iii), we confine ourselves to a formula proposed recently in [10]

$$
\gamma_{KH} = \frac{(\zeta(j+1))\top(\zeta(j+1)-\gamma(j))}{(\zeta(j))\top(\zeta(j))}. \tag{34}
$$

Finally, following the arguments in [10] and theorem 5.3 in [9], we conclude the following theorem without proof due to the page limit. We refer to a forthcoming paper by the authors for a proof.

**Theorem 2** The CG non-whitened one-unit linear ICA algorithm (Algorithm 1) with step size selection (33) and direction update (34) is locally $(m-1)$-step quadratically convergent to a correct demixing $x^* \in \Theta$.

### 4. NUMERICAL EXPERIMENTS

It is important to know that the local quadratic convergence properties of the non-whitened FastIICA algorithm shown in Theorem 1 are proven under the population setting. In the finite sample setting, however, such a theoretical convergence rate can hardly be observed, and the actual convergence rate might reduce significantly. On the other hand, since the development of the CG algorithm does not depend on the population setting, its actual convergence rate as stated in Theorem 2 is expected to hold true. The task of our experiment is to separate three mutually statistically independent signals with uniform distributions. The convergence of algorithms is measured by the distance of the accumulation point $x^* \in S^{m-1}$ to the $k$-th iterate $x(k) \in S^{m-1}$, i.e. by $\|x(k) - x^*\|$.

For the non-whitened FastIICA algorithm, we run four tests in accordance with an increasing sample size, $n = 10^4, 10^5, 10^6,$ and $10^7$. It can be seen from Figure 1 that, the algorithm appears to converge only linearly fast. A closer observation shows that, when the number of samples $n$ increases, the non-whitened FastIICA tends to converge faster. Suggested by this asymptotic behavior when the number of samples increases towards infinity, we claim that the non-whitened FastIICA algorithm will converge locally quadratically fast to a correct source separation under the population setting.

For investigating our CG ICA algorithm, we compare it with several popular direction updates, such as Hestenes-Stiefel, Polak-Ribiére, Fletcher-Reeves and Dai-Yuan. We refer to [10] and references therein for the corresponding adapted formulas on manifolds. The number of samples is equal to $10^7$. It is obvious from Figure 2 that only the CG algorithm with the direction update (34) converges locally $(m-1)$-step quadratically fast.

### 5. REFERENCES


