Intrinsic Newton's Method on Oblique Manifolds for Overdetermined Blind Source Separation

Martin Kleinsteuber and Hao Shen

***Abstract*— This paper studies the problem of Overdetermined Blind Source Separation (OdBSS), a challenging problem in signal processing. It aims to recover desired sources from outnumbered observations without knowing either the source distributions or the mixing process. It is well-known that performance of standard BSS algorithms, which usually utilize a whitening step as a pre-process to reduce the dimensionality of observations, might be seriously limited due to its blind trust on the data covariance matrix. In this paper, we develop and compare two locally quadratic OdBSS algorithms that forgo the dimensionality reduction step. In particular, our algorithms solve a problem of simultaneous diagonalization of a set of symmetric matrices. By exploiting the appropriate underlying manifold, namely the so-called oblique manifold, intrinsic Newton's method is developed to optimize two popular cost functions for the simultaneous diagonalization of symmetric matrices: the off-norm function and the log-likelihood function. Performance of the proposed algorithms is investigated and compared by several numerical experiments.**

I. INTRODUCTION

Linear Blind Source Separation (BSS) addresses the problem of recovering linearly mixed sources from only several observed mixtures without knowing either the source distributions or the mixing process. A popular assumption of the sources being *mutually statistically independent* leads to the concept of linear Independent Component Analysis (ICA), which has become a prominant statistical method for solving the linear BSS problem. A common linear BSS model, usually refered to as the *determined linear BSS model*, assumes that the number of sources is equal to the number of observations. In this work, we are interested in the problem of extracting a fewer number of signals from a number of observations, i.e. the problem of overdetermined linear ICA. Its applications can be found in image analysis and biomedical data analysis.

A widely-used linear ICA procedure consists of two steps [1]: (i) the observations are firstly whitened, usually by Principal Component Analysis (PCA), and (ii) a number of desired signals are extracted from the whitened observations according to mutual statistical independence. Step (i) reduces complexity of step (ii) and meanwhile copes with uniqueness of source extraction [2]. Such a procedure results in the so-called whitened linear ICA problem. It is well-known [3] that, in real applications, i.e. problems with only a finite number of samples and additional noises, performance of linear ICA methods with whitening is limited due to statistical inefficiency. Recent work in [4] shows that the so-called *oblique manifold* is the suitable setting for doing non-whitened linear ICA.

A popular category of ICA algorithms involve a joint diagonalization of a set of matrices, which are derived from certain statistics of the observations [2], [5]. Recently, several efficient simultaneous diagonalization based ICA algorithms have been developed in [6], [7], for the determined linear ICA problem, i.e. extracting all sources. Unfortunately, these works do not handle the current overdetermined situation. Moreover, we are aware of gradient descent algorithms on the non-compact Stiefel manifold for non-whitened overdetermined linear ICA without considering convergence properties of the proposed algorithms [8], [9]. In this work, we develop an intrinsic Newton's method for solving the OdBSS problem on the appropriate oblique manifold. The algorithms are locally quadratic convergent to a demixing matrix.

This paper is organized as follows. In Section II, we briefly introduce the overdetermined blind source separation problem. Section III presents some basic results about the oblique manifold, which are needed in our later analysis and development. Critical point analysis of two studied cost functions is provided in Section IV, followed by a formulation of intrinsic Newton's method on the oblique manifold. Finally in Section V, performance of the proposed algorithms is investigated by several numerical experiments.

II. OVERDETERMINED BSS PROBLEM

The mixing model of an instantaneous OdBSS problem is given as

$$w(t) = As(t) + n(t), \tag{1}$$

where $s(t) \in \mathbb{R}^n$ denotes the time series of n statistically independent signals, $A \in \mathbb{R}^{m\times n}$ with $m > n$ is the mixing matrix of full rank, $w(t) \in \mathbb{R}^m$ denotes m observed linear mixtures, and $n(t) \in \mathbb{R}^m$ are certain noises. We denote by $s_i(t) \in \mathbb{R}$ and $w_i(t) \in \mathbb{R}$ the i-th components of $s(t)$ and $w(t)$, respectively. By the construction of linear ICA, the source signals $s(t)$ are assumed to be *mutually statistically independent* and, without loss of generality, to have *zero mean* and *unit variance*, i.e.,

$$\mathbb{E}[s(t)] = 0, \quad \text{and} \quad \mathbb{E}[s(t)s(t)^\top] = I_n, \tag{2}$$

where $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix. The task of OdBSS is to extract k source signals with $k \leq n$, by finding a demixing matrix $X \in \mathbb{R}^{m\times k}$ based only on the observations

M. Kleinsteuber and H. Shen are with the Institute for Data Processing and CoTeSys at Technische Universität München, Germany. {Kleinsteuber, hao.shen}@tum.de.

ISBN 978-963-311-370-7

$w(t)$ via the demixing model

$$y(t) = X^\top w(t), \tag{3}$$

where $y(t) \in \mathbb{R}^k$ denotes k extracted source signals.

Under certain conditions, the OdBSS problem can be solved effectively by only using second-order statistics [5]. The first scenario studied in this work assumes that the sources $s(t)$ are nonstationary, namely, covariance of $s(t)$, and thus consequently, $w(t)$ as well, is time-varying. A simple approach to separate nonstationary sources [10] is to simultaneously diagonalize a set of covariance matrices of $w(t)$ in different time periods, which are symmetric positive semi-definite. A more general approach of using second-order statistics is to simultaneously diagonalize a set of time-lagged covariance matrices

$$R(\tau) := \mathbb{E}\left[w(t)w(t+\tau)^\top\right] \tag{4}$$

where $\tau > 0$ is a time lag. Note that $R(\tau)$ is symmetric, but not necessarily positive definite, which makes this approach unsuitable for optimizing the second cost function discussed in this paper.

Here, we are interested in solving the following problem. Given a set of symmetric matrices $\{C_i\}_{i=1}^N \subset \mathbb{R}^{m\times m}$, constructed as second-order statistics of the observations $w(t)$, the task is to find a matrix $X \in \mathbb{R}^{m\times k}$ being of full rank, such that

$$X^\top C_i X, \qquad \text{for all } i = 1, \ldots, N, \tag{5}$$

are simultaneously diagonalized, or approximately simultaneously diagonalized subject to certain diagonality measure.

It is well-known [2], [11] that, if an $X^* \in \mathbb{R}^{m\times k}$ extracts k desired sources, the demixing matrix X^* can only be identified up to arbitrary column-wise scaling and permutation of columns, i.e., any X^*DP, where D is a $k\times k$ invertible diagonal matrix and P a $k\times k$ permutation matrix, extracts the same k sources. In the noise-free case, i.e. $n(t) = 0$, an X^* is referred to as an *exact joint diagonalizer*. It is known that the ambiguity due to column-wise scaling can be eliminated by a pre-whitening process. For a non-whitened approach, an appropriate setting to handle the column-wise scaling is given by the oblique manifold [12],

$$\mathcal{O}(m,k) := \left\{X \in \mathbb{R}^{m\times k} \,|\, \operatorname{ddiag}(X^\top X) = I_k, \operatorname{rk} X = k\right\}, \tag{6}$$

where rk is the rank, and $\operatorname{ddiag}(Z)$ forms a diagonal matrix, whose diagonal entries are those of Z.

Straightforwardly, we adapt two popular cost functions of measuring diagonality of matrices, namely, the off-norm function [13] and the log-likelihood based cost function [10], to the present overdetermined scenario,

$$\begin{aligned} f_1 \colon \mathcal{O}(m,k) &\to \mathbb{R}, \\ X &\mapsto \frac{1}{4}\sum_{i=1}^N \left\| \operatorname{off}(X^\top C_i X)\right\|_{\mathrm{F}}^2, \end{aligned} \tag{7}$$

where $\operatorname{off}(Z) = Z - \operatorname{ddiag}(Z)$ is a matrix by setting the diagonal entries of Z to zero and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm, and

$$\begin{aligned} f_2 \colon \mathcal{O}(m,k) &\to \mathbb{R}, \\ X &\mapsto \frac{1}{2}\sum_{i=1}^N \log \frac{\det \operatorname{ddiag}\left(X^\top C_i X\right)}{\det\left(X^\top C_i X\right)}. \end{aligned} \tag{8}$$

It is important to notice that the off-norm function (7) is *column-wise scale invariant* with respect to the matrix X, only if the OdBSS problem is noiseless. In other words, solutions of OdBSS problems with additional noises, provided by minimizing the off-norm function, might be sensitive to certain properties of the noises, e.g. type of distributions and magnitudes.

III. Geometry of the Oblique Manifold

In order to formulate an intrinsic Newton's method on the oblique manifold $\mathcal{O}(m,k)$, we need to introduce the Riemannian gradient and the Riemannian Hessian on $\mathcal{O}(m,k)$. We endow $\mathcal{O}(m,k)$ with the Riemannian metric inherited from $\mathbb{R}^{m\times k}$ by the inner product

$$g \colon \mathbb{R}^{m\times k} \times \mathbb{R}^{m\times k} \to \mathbb{R}, \quad g(A,B) := \operatorname{tr}(A^\top B). \tag{9}$$

It will be useful for understanding the upcoming formulas, if we recall the fact that $\mathcal{O}(m,k)$ is an open and dense Riemannian submanifold of the well-understood k-times product of the $(m-1)$-sphere with the usual Euclidean metric

$$\overline{\mathcal{O}(m,k)} = \underbrace{S^{m-1} \times \cdots \times S^{m-1}}_{k-\text{times}} =: (S^{m-1})^k. \tag{10}$$

Here, $\overline{\mathcal{O}}(m,k)$ denotes the closure of $\mathcal{O}(m,k)$. It follows, that

$$\dim \mathcal{O}(m,k) = k \dim S^{m-1} = k(m-1) \tag{11}$$

and, the tangent spaces and the geodesics for $\mathcal{O}(m,k)$ and $(S^{m-1})^k$ coincide. In other words, a geodesic on $\mathcal{O}(m,k)$ is exactly the connected component of a geodesic on $(S^{m-1})^k$ restricted to $\mathcal{O}(m,k)$. Concretely, the tangent space at some $X \in \mathcal{O}(m,k)$ is given by

$$T_X\mathcal{O}(m,k) = \{\Xi \in \mathbb{R}^{m\times k} \mid \operatorname{ddiag}(X^\top \Xi) = 0\}, \tag{12}$$

and the normal space by

$$N_X\mathcal{O}(m,k) = \{X\Gamma \mid \Gamma \in \mathbb{R}^{k\times k} \text{ is diagonal}\}. \tag{13}$$

Lemma 1: The orthogonal projection onto the tangent space $T_X\mathcal{O}(m,k)$ at $X \in \mathcal{O}(m,k)$ is given by

$$\begin{aligned} \operatorname{pr}_X \colon \mathbb{R}^{m\times k} &\to T_X\mathcal{O}(m,k), \\ \operatorname{pr}_X(A) &:= A - X\operatorname{ddiag}(X^\top A). \end{aligned} \tag{14}$$

Proof: We first show that for $\operatorname{pr}_X(A) \in T_X\mathcal{O}(m,k)$,

$$\begin{aligned} &\operatorname{ddiag}(X^\top \operatorname{pr}_X(A)) \\ =\ &\operatorname{ddiag}(X^\top(A - X\operatorname{ddiag}(X^\top A)) \\ =\ &\operatorname{ddiag}(X^\top A) - \operatorname{ddiag}(X^\top X(\operatorname{ddiag} X^\top A)) \\ =\ &\operatorname{ddiag}(X^\top A) - \operatorname{ddiag}(X^\top X)(\operatorname{ddiag} X^\top A) \\ =\ &0, \end{aligned} \tag{15}$$

because $\operatorname{ddiag}(X^\top X) = I_k$. For orthogonality of the projection, let $\Xi \in T_X\mathcal{O}(m,k)$, i.e. $\operatorname{ddiag}(\Xi^\top X) = 0$. We compute

$$\begin{aligned}\operatorname{tr}(\Xi^\top \operatorname{pr}_X(A)) &= \operatorname{tr}(\Xi^\top A) - \operatorname{tr}(\Xi^\top X \operatorname{ddiag}(X^\top A))\\ &= \operatorname{tr}(\Xi^\top A),\end{aligned} \tag{16}$$

following the fact that $\operatorname{ddiag}(\Xi^\top X) = 0$ implies $\operatorname{ddiag}(\Xi^\top X\Gamma) = 0$ for any diagonal matrix Γ, thus, $\operatorname{tr}(\Xi^\top X \operatorname{ddiag}(X^\top A)) = 0$. The result follows. ■

Now, let us recall the great circle μ_x of S^{m-1} at $x \in S^{m-1}$ for a given tangent direction $\xi \in T_x S^{m-1}$, defined as follows

$$\begin{aligned}&\mu_{x,\xi}\colon \mathbb{R} \to S^{m-1},\\ &\mu_{x,\xi}(t) := \begin{cases} x, & \|\xi\| = 0;\\ x\cos t\|\xi\| + \xi\frac{\sin t\|\xi\|}{\|\xi\|}, & \text{otherwise.}\end{cases}\end{aligned} \tag{17}$$

Clearly, $\mu_{x,\xi}(0) = x$ and $\dot{\mu}_{x,\xi}(0) = \xi$. Geodesics of $\mathcal{O}(m,k)$ are given as follows.

Lemma 2: Geodesics $\gamma_{X,\Xi}\colon \mathbb{R} \to \mathcal{O}(m,k)$ through $X = [x_1, \ldots, x_k] \in \mathcal{O}(m,k)$ in direction $\Xi = [\xi_1, \ldots, \xi_k] \in T_X\mathcal{O}(m,k)$ on $(S^{m-1})^k$ and hence, by restriction, on $\mathcal{O}(m,k)$ are given by

$$\gamma_{X,\Xi}(t) = [\mu_{x_1,\xi_1}(t), \ldots, \mu_{x_k,\xi_k}(t)]. \tag{18}$$

Proof: Since $x_i^\top x_i = 1$ and $\xi_i^\top x_i = 0$ for $i = 1, \ldots k$, we have $\dot{\gamma}_{X,\Xi}(0) = \Xi$. Moreover, it can be shown that $\ddot{\gamma}_{X,\Xi}(0)$ lies in the normal space of X since

$$\ddot{\gamma}_{X,\Xi}(0) = X \operatorname{diag}(-\|\xi_1\|^2, \ldots, -\|\xi_k\|^2), \tag{19}$$

hence the result follows. ■

IV. INTRINSIC NEWTON'S METHOD

In this section, we firstly provide a critical point analysis of the two cost functions defined in (7) and (8), followed by development of an intrinsic Newton's method for optimizing both functions.

We compute the first derivative of f_1 as defined in (7) at $X \in \mathcal{O}(m,k)$ in direction $\Xi \in T_X\mathcal{O}(m,k)$ as

$$\mathrm{D}\, f_1(X)\Xi = \sum_{i=1}^{N} \operatorname{tr}\left(\Xi^\top C_i X \operatorname{off}(X^\top C_i X)\right). \tag{20}$$

Let $X^* \in \mathcal{O}(m,k)$ be an exact joint diagonalizer, obviously

$$\mathrm{D}\, f_1(X^*)\Xi = 0, \tag{21}$$

i.e., any exact diagonalizer of the simultaneous diagonalization problem (5) is a critical point of f_1. Taking the second derivative of f_1 at $X \in \mathcal{O}(m,k)$ in direction $\Xi \in T_X\mathcal{O}(m,k)$ results in

$$\begin{aligned}&\mathrm{D}^2 f_1(X)(\Xi,\Xi) = \tfrac{\mathrm{d}^2}{\mathrm{d}\,t^2}(f_1 \circ \gamma_{X,\Xi})(t)\Big|_{t=0}\\ &= \sum_{i=1}^{N} \operatorname{tr}\left(\Xi^\top C_i \Xi \operatorname{off}\left(X^\top C_i X\right)\right)\\ &\quad - \operatorname{tr}\left(\operatorname{ddiag}(\Xi^\top \Xi) X^\top C_i X \operatorname{off}\left(X^\top C_i X\right)\right)\\ &\quad + \operatorname{tr}\left(\Xi^\top C_i X \left(\operatorname{off}\left(X^\top C_i \Xi\right) + \operatorname{off}\left(\Xi^\top C_i X\right)\right)\right),\end{aligned} \tag{22}$$

where $\gamma_{X,\Xi}$ is the geodesic on $\mathcal{O}(m,k)$ as defined in (18). It is easy to see that the first two terms on the right-hand side in (22) evaluated at a joint diagonalizer $X^* \in \mathcal{O}(m,k)$ are equal to zero. Let $\Xi = [\xi_1, \ldots, \xi_k] \in T_{X^*}\mathcal{O}(m,k)$. Then we evaluate the third term at X^* as

$$\begin{aligned}&\operatorname{tr}\left(\Xi^\top C_i X^* \left(\operatorname{off}\left(X^{*\top} C_i \Xi + \Xi^\top C_i X^*\right)\right)\right)\\ &= \sum_{p,q=1}^{k} \sum_{i=1}^{N} \xi_p^\top C_i x_q^* x_q^{*\top} C_i \xi_p + \xi_p^\top C_i x_q^* x_p^{*\top} C_i \xi_q.\end{aligned} \tag{23}$$

A direct computation shows that the second summation on the right-hand side of (23) is equal to zero as well. By the construction of noiseless OdBSS, it can be shown that the second-order statistics C_i's are not of full rank. We then conclude the following result.

Lemma 3: Any exact joint diagonalizer $X^* \in \mathcal{O}(m,k)$ of the noiseless OdBSS problem as defined in (5) is a critical point of the off-norm function f_1, defined in (7). The Hessian of f_1 at X^* is *positive semidefinite*.

Remark 1: Obviously, to ensure local quadratic convergence for a Newton's method, the Hessian has to be nondegenerated. It is reasonable to assume the nondegeneracy in the generic case with additional noises.

In what follows, we derive a critical point analysis of the log-likelihood based cost function f_2, defined in (8), in the same manner as for f_1. The first derivative of f_2 at $X \in \mathcal{O}(m,k)$ in direction $\Xi \in T_X\mathcal{O}(m,k)$ is computed by

$$\begin{aligned}\mathrm{D} f_2(X)\Xi = \sum_{i=1}^{N} \operatorname{tr}\Big(\Xi^\top C_i X \Big(\left(\operatorname{ddiag}(X^\top C_i X)\right)^{-1}\\ - (X^\top C_i X)^{-1}\Big)\Big).\end{aligned} \tag{24}$$

It can be shown that an exact joint diagonalizer $X^* \in \mathcal{O}(m,k)$ is a critical point of f_2. A tedious computation leads to the the second derivative of f_2 at $X \in \mathcal{O}(m,k)$ in direction $\Xi \in T_X\mathcal{O}(m,k)$ as follows

$$\begin{aligned}&\mathrm{D}^2 f_2(X)(\Xi,\Xi) = \tfrac{\mathrm{d}^2}{\mathrm{d}\,t^2}(f_2 \circ \gamma_{X,\Xi})(t)\Big|_{t=0}\\ &= \sum_{i=1}^{N} \operatorname{tr}\Big(\Big(\left(\operatorname{ddiag}(X^\top C_i X)\right)^{-1} - \left(X^\top C_i X\right)^{-1}\Big)\cdot\\ &\quad \cdot \Big(\Xi^\top C_i \Xi - \operatorname{ddiag}(\Xi^\top \Xi) X^\top C_i X\Big)\Big)\\ &\quad + \operatorname{tr}\Big(\Xi^\top C_i X \Big(\left(X^\top C_i X\right)^{-1}\left(\Xi^\top C_i X + X^\top C_i \Xi\right)\cdot\\ &\quad \cdot \left(X^\top C_i X\right)^{-1} - \left(\operatorname{ddiag}(X^\top C_i X)\right)^{-1}\cdot\\ &\quad \cdot \operatorname{ddiag}\left(\Xi^\top C_i X + X^\top C_i \Xi\right)\left(\operatorname{ddiag}\left(X^\top C_i X\right)\right)^{-1}\Big)\Big).\end{aligned} \tag{25}$$

The first term on the right-hand side from above can be shown to be equal to zero at $X^* \in \mathcal{O}(m,k)$. Let us denote

$$\begin{aligned}\Lambda &:= \operatorname{diag}(\lambda_1, \ldots, \lambda_k)\\ &= \left(X^{*\top} C_i X^*\right)^{-1} = \left(\operatorname{ddiag}(X^{*\top} C_i X^*)\right)^{-1},\end{aligned} \tag{26}$$

with $\lambda_j > 0$ for all $j = 1, \ldots, k$. Then the second term in (25) evaluated at X^* is computed as

$$\begin{aligned} &\operatorname{tr}\left(\Xi^\top C_i X^* \Lambda \operatorname{off}\left(X^{*\top} C_i \Xi + \Xi^\top C_i X^*\right) \Lambda\right) \\ &= \sum_{p,q=1}^{k} \sum_{i=1}^{N} \lambda_p \lambda_q \xi_p^\top C_i x_q^* x_q^{*\top} C_i \xi_p \\ &\quad + \sum_{p,q=1}^{k} \sum_{i=1}^{N} \lambda_p \lambda_q \xi_p^\top C_i x_q^* x_p^{*\top} C_i \xi_q. \end{aligned} \tag{27}$$

By the same argument, the second summation on the right-hand side of (27) is equal to zero. Following the fact that $\lambda_j > 0$ for all $j = 1, \ldots, k$, we conclude a similar result as Lemma 3 as follows

Lemma 4: Any exact joint diagonalizer $X^* \in \mathcal{O}(m,k)$ of the noiseless OdBSS problem as defined in (5) is a critical point of the off-norm function f_2, defined in (8). The Hessian of f_2 at X^* is *positive semidefinite*.

In the rest of the section, we develop an intrinsic Newton's method on the oblique manifold for minimizing the cost functions f_1 and f_2. Intrinsic Newton's method on smooth manifold, *cf.* algorithm 4.3 in [14], can be adapted to our current setting as follows.

Algorithm 1: Intrinsic Newton's OdBSS method

Step 1: Given an initial guess $X_0 \in \mathcal{O}(m,k)$ and set $i = 0$.

Step 2: Compute the Riemannian Newton direction $\Xi_i \in T_{X_i}\mathcal{O}(m,k)$ by solving the linear system $\mathcal{H}_{f_t}(X_i)(\Xi_i) = -\nabla_{f_t}(X_i)$.

Step 3: Move from X_i in direction Ξ_i, and update $X_{i+1} \leftarrow \gamma_{X_i,\Xi_i}(1)$.

Step 4: If $\|X_{i+1} - X_i\|_F$ is small enough, stop. Otherwise, set $i = i + 1$, and go to Step 2.

Here, $t \in \{1,2\}$, $\nabla_{f_t}(X_i)$ and $\mathcal{H}_{f_t}(X_i)$ are the Riemannian gradient and the Riemannian Hessian operator, respectively.

Following the first derivative of f_1 as computed in (20), the Riemannian gradient of f_1 at $X \in \mathcal{O}(m,k)$ with respect to the Riemannian metric (9) can be computed as

$$\nabla_{f_1}(X) = \operatorname{pr}_X\left(\sum_{i=1}^{N} C_i X \operatorname{off}(X^\top C_i X)\right). \tag{28}$$

Let $X = [x_1, \ldots, x_k] \in \mathcal{O}(m,k)$ and $\Xi = [\xi_1, \ldots, \xi_k] \in T_X\mathcal{O}(m,k)$. By polarization, the Riemannian Hessian operator $\mathcal{H}_{f_1}(X)\colon T_X\mathcal{O}(m,k) \to T_X\mathcal{O}(m,k)$ is given as

$$\mathcal{H}_{f_1}(X)(\Xi) = \operatorname{pr}_X([\psi_1(\Xi), \ldots, \psi_k(\Xi)]) \tag{29}$$

where

$$\begin{aligned} \psi_p(\Xi) = &\sum_{q\neq p}^{k} \sum_{i=1}^{N} \left(C_i x_q x_q^\top C_i - \left(x_p^\top C_i x_q\right)^2 I_m\right) \xi_p \\ &+ \sum_{q\neq p}^{k} \sum_{i=1}^{N} \left(x_p^\top C_i x_q C_i + C_i x_q x_p^\top C_i\right) \xi_q. \end{aligned} \tag{30}$$

Let us denote

$$Y^{(i)} := \operatorname{diag}\left(y_{11}^{(i)}, \ldots, y_{kk}^{(i)}\right) = \left(\operatorname{ddiag}(X^\top C_i X)\right)^{-1} \tag{31}$$

and

$$Z^{(i)} := \left(z_{pq}^{(i)}\right) = \left(X^\top C_i X\right)^{-1}. \tag{32}$$

We compute the Riemannian gradient of f_2 at $X \in \mathcal{O}(m,k)$ with respect to the Riemannian metric (9) as

$$\nabla_{f_2}(X) = \operatorname{pr}_X\left(\sum_{i=1}^{N} C_i X \left(Y^{(i)} - Z^{(i)}\right)\right), \tag{33}$$

and by polarization, the Riemannian Hessian operator $\mathcal{H}_{f_2}(X)\colon T_X\mathcal{O}(m,k) \to T_X\mathcal{O}(m,k)$ as

$$\mathcal{H}_{f_2}(X)(\Xi) = \operatorname{pr}_X\left([\phi_1(\Xi), \ldots, \phi_k(\Xi)]\right) \tag{34}$$

where

$$\begin{aligned} \phi_p(\Xi) = &\sum_{i=1}^{N} \left(y_{pp}^{(i)} C_i - 2\left(y_{pp}^{(i)}\right)^2 C_i x_p x_p^\top C_i \right. \\ &\left. + \left(1 - y_{pp}^{(i)}\left(x_p^\top C_i x_p\right)\right) I_m\right) \xi_p \\ &+ \sum_{q=1}^{k} \left(\sum_{r,s=1}^{k} \sum_{i=1}^{N} \frac{1}{2}\left(z_{sq}^{(i)} z_{rp}^{(i)} + z_{sr}^{(i)} z_{qp}^{(i)}\right) \cdot \right. \\ &\cdot C_i x_s x_r^\top C_i + \frac{1}{2}\left(z_{sp}^{(i)} z_{rq}^{(i)} + z_{sr}^{(i)} z_{pq}^{(i)}\right) \cdot \\ &\left. \cdot C_i x_r x_s^\top C_i - \sum_{i=1}^{N}\left(z_{qp}^{(i)} C_i\right)\right) \xi_q, \end{aligned} \tag{35}$$

respectively.

V. Numerical Experiments

The task of our experiment is to approximately diagonalize a set of symmetric positive definite matrices $\{\widetilde{C}_i\}_{i=1}^N$ jointly, constructed by

$$\widetilde{C}_i = A \Lambda_i A^\top + \varepsilon E_i, \qquad i = 1, \ldots, N, \tag{36}$$

where $A \in \mathbb{R}^{m\times k}$ is a randomly picked matrix in $\mathcal{O}(m,k)$, diagonal entries of Λ_i are drawn from a uniform distribution on the interval $(9, 11)$, $E_i \in \mathbb{R}^{m\times m}$ is the symmetric part of an $m\times m$ matrix, whose entries are generated from a uniform distribution on the unit interval $(-0.5, 0.5)$, representing additive noise, and $\varepsilon \in \mathbb{R}$ is the noise level. We set $m = 5$, $k = 3$, $N = 20$, and $\varepsilon = 0.01$.

The convergence of algorithms is measured by the distance of the accumulation point $X^* \in \mathcal{O}(m,k)$ to the current iterate $X_k \in \mathcal{O}(m,k)$, i.e., by $\|X_k - X^*\|_{\mathrm{F}}$. According to Fig. 1 and Fig. 2, it is clear that both our proposed algorithms converge locally quadratically fast to a joint diagonalizer.

Acknowledgments

This work has been supported in parts by the German DFG funded Cluster of Excellence CoTeSys - Cognition for Technical Systems.

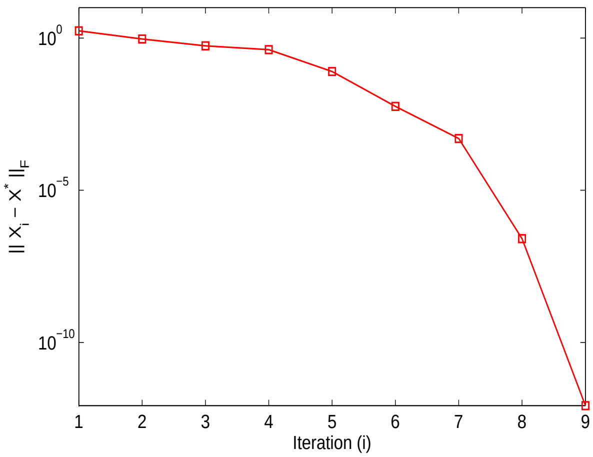

Fig. 1. Local quadratic convergence: intrinsic Newton's method minimizing the off-norm function.

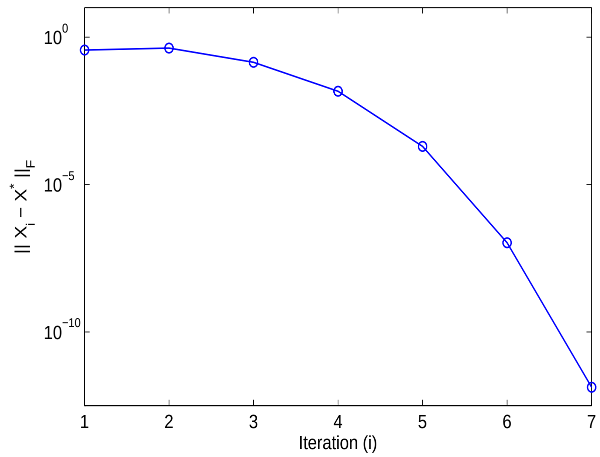

Fig. 2. Local quadratic convergence: intrinsic Newton's method minimizing the log-likelihooh function.

References

[1] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley, New York, 2001.

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

[3] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proceedings of the IEEE, Special Issue on Blind Identification and Estimation*, vol. 86, no. 10, pp. 2009–2025, 1998.

[4] P.-A. Absil and K. A. Gallivan, "Joint diagonalization on the oblique manifold for independent component analysis," in *Proceedings of the* 31^{st} *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006)*, Toulouse, France, 2006, pp. V945–V948.

[5] A. Belouchrani, K. A. Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique based on second order statistics," *IEEE Transactions on Signal Processing*, vol. 45, no. 2, pp. 434–444, 1997.

[6] B. Afsari, "Simple LU and QR based non-orthogonal matrix joint diagonalization," in *Lecture Notes in Computer Science, Proceedings of the* 6^{th} *International Conference on Independent Component Analysis and Blind Source Separation (ICA 2006)*, Berlin/Heidelberg, 2006, vol. 3889, pp. 1–7, Springer-Verlag.

[7] H. Shen and K. Hüper, "Block Jacobi-type methods for non-orthogonal joint diagonalisation," in *Proceedings of the* 34^{th} *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009)*, Taipei, Republic of China (Taiwan), 2009, pp. 3285–3288.

[8] S.-I. Amari, "Natural gradient learning for over- and under-complete bases in ICA," *Neural Computation*, vol. 11, no. 8, pp. 1875–1883, 1999.

[9] L. Q. Zhang, A. Cichocki, and S. I. Amari, "Natural gradient algorithm for blind separation of overdetermined mixture with additive noise," *IEEE Signal Processing Letters*, vol. 6, no. 11, pp. 293–295, 1999.

[10] D. T. Pham, "Joint approximate diagonalization of positive definite Hermitian matrices," *SIAM J. on Matrix Anal. and Appl*, vol. 22, pp. 1136–1152, 2001.

[11] F. J. Theis and W. Nakamura, "Quadratic independent component analysis," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E87-A, no. 9, pp. 2355–2363, 2004.

[12] B. Afsari, "Sensitivity analysis for the problem of matrix joint diagonalization," *SIAM Journal of Matrix Analysis and Application*, vol. 30, pp. 1148–1171, 2008.

[13] M. Joho and K. Rahbar, "Joint diagonalization of correlation matrices by using Newton methods with application to blind signal separation," in *Proceedings of the* 2^{nd} *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM 2002)*, Rosslyn, VA USA, 2002, pp. 403–407.

[14] S. T. Smith, "Optimization techniques on Riemannian manifolds," in *Hamiltonian and Gradient Flows, Algorithms and Control*, A. Bloch, Ed., Fields Institute Communications, pp. 113–136. American Math. Soc., Providence, 1994.