NEW FAST FACTORIZATION METHOD FOR MULTIVARIATE OPTIMIZATION AND ITS REALIZATION AS ICA ALGORITHM

Toshinao Akuzawa

Lab. for Mathematical Neuroscience, RIKEN Brain Science Institute
2-1 Hirosawa, Wako, Saitama 351-0198, Japan
akuzawa@brain.riken.go.jp, http://www.mns.brain.riken.go.jp/~akuzawa/

ABSTRACT

A new framework for multivariate optimization by criteria invariant under componentwise scaling is constructed. These problems are naturally considered as problems on the coset $\mathbb{R}^{\times N} \backslash GL(N, \mathbb{R})$. We show that there is a duality between the optimization flow on this coset and the dynamics of quantum lattice with inner degrees of freedom. Then we propose a new algorithm for optimization problems on this coset named nested Newton's method, whose essence is the ignorance of n-body interactions with $n \geq 3$ if we explain it by using its analogy to quantum lattice. This method is readily applicable as a highly useful ICA algorithm, which is robust under gaussian noises, quite fast, and practical for large dimensional problems. The last feature comes from the fact that our method requires memory space of order N^2, whereas the conventional Newton's method for ICA (and the JADE) requires that of order N^4. The matlab code for this algorithm is available from our web-site.

1. INTRODUCTION

1.1. overview

Independent component analysis(ICA) in its conventional framework is nothing but a variant of principal component analysis(PCA) or factor analysis(FA). The concept of ICA, however, contains more than that. Proposed here are a new method for ICA, which is highly practical and is formulated quite differently from PCA and FA.

First, we will construct a fast and highly practical algorithm for optimization problems on $\mathbb{R}^{\times N} \backslash GL(N, \mathbb{R})$ by criteria invariant under componentwise scaling, which we call the nested Newton's method. The nested Newton's method is a method introduced in [1] which decomposes the flow of optimization on $\mathbb{R}^{\times N} \backslash GL(N, \mathbb{R})$ into dynamics of N-particles under two-body interactions. The interaction between two 'particles' is decided along the Newton's manner.

In this method thanks to the translation into two-body problems, the updating rule is decomposed into manipulations of 2×2 matrices. This method is quite new though it is in some sense related to the so-called natural gradient method or nonholonomic method.

Then we will show that an ICA algorithm without prewhitening (pre-WH) is constructed from the nested Newton's method, that is, factorized Newton's method to minimize sum of squared 4-th order cross cumulants between components (with some weight). Its advantages are as follows.

1. Robust under gaussian noises. It is not necessary to estimate the noise variance since this is pure fourth-order-cumulant-based method.
2. Globally stable even if the number of sources does not equal the number of observation channels.
3. Convergence is quite fast. Practical also in case where the number of observation channels are quite large.

It is extremely useful when factor analysis (FA) is unavairable. When $M > (2N + 1 - (8N + 1)^{1/2})/2$ where M and N are respectively the number of sources and observation channels, FA is unavailable. The main target of our method is problems for which this inequality holds.

1.2. notational conventions

We denote the sample mean by $\langle\rangle$ and the empirical cumulants by $\langle\rangle_c$. $GL(N, \mathbb{R})$ is the set of $N \times N$ nonsingular real matrices. $\mathbb{R}^{\times N} \backslash GL(N, \mathbb{R})$ is the quotient set of $GL(N, \mathbb{R})$ by the equivalence relation $\sim$, which is defined by

$$A \sim B \Leftrightarrow {}^{\exists}\text{nonsingular diagonal matrix } D \text{ and } A = DB \ .$$

2. NESTED NEWTON'S METHOD

2.1. general framework

We assume that T samples of N-dimensional variables $\{(y(0))_{it} | 1 \leq i \leq N,\ 1 \leq t \leq T\}$ are available as observed data. The mean values have already subtracted from the data, that is, $\sum_{t=1}^{T}(y(0))_{it} = 0$ for all i. We often omit the last lower index denoting the sample number and denote the data as N-dimensional vectors. Let us set

$$y = By(0) \ , \tag{1}$$

where we assume that $B \in GL(N, \mathbb{R})$. We will study optimization problems with respect to B. First, we assume that there are M criteria $\{Q_k | \mathbb{R}^T \times \mathbb{R}^T \to \mathbb{R}, 1 \leq k \leq M\}$ which depends on two specific rows in y and become very close to zero at the optimal point. More explicitly, we want to find B with which $|Q_k(y_i, y_j)|$ become very small simultaneously for $1 \leq k \leq M$ and $1 \leq i \neq j \leq N$. Since $y(0)$ is given, the criteria $\{Q_k(y_i, y_j)\}$'s are completely determined by B. We also assume that Q_k is scale invariant, that is,

$$Q_k(ay_i, by_j) = Q_k(y_i, y_j) \tag{2}$$

for any $a, b \neq 0$. Then $\{Q_k(y_i, y_j)\}$'s become invariant under multiplication $B \to DB$ of

$$D \in \mathbb{R}^{\times N} \stackrel{\text{def}}{=} \{\text{diagonal matrix in } GL(N, \mathbb{R})\}$$

from the left-hand-side of B. Thus the optimal point has invariance under this multiplcation. The optimization of B on this setting is naturally considered as an optimization problem on the coset $\mathbb{R}^{\times N} \backslash GL(N, \mathbb{R})$.[2]

In this paper, we will construct the nested Newton's method to solve this kind of problems. The nested Newton's method is originally introduced in [1]. Here we will explain the nested Newton's method by using the analogy of the multiplicative update (3) of N-dimensional ICA optimization to quantum dynamics of N-particles under two-body interactions.

2.2. multiplicative updating

We consider an iterative multiplicative algorithm. Let us set

$$\mathfrak{m} \stackrel{\text{def}}{=} \left\{ \Delta \in N \times N \text{ matrices} \middle| \Delta_{ii} = 0 \quad (1 \leq i \leq N) \right\} .$$

We specify the flow of the sequence by $\{\Delta(s) \in \mathfrak{m}; s = 0, 1, 2, \cdots\}$, which describes the amount of individual steps:

$$y(0) = \text{observation} \ , \quad y(s+1) = (\exp \Delta(s))y(s) \ . \tag{3}$$

In other words, $y(n)$ depends on $\{\Delta(k)\}$'s as

$$y(n) = \mathrm{e}^{\Delta(n-1)}\mathrm{e}^{\Delta(n-2)} \cdots \mathrm{e}^{\Delta(0)} y(0) \ . \tag{4}$$

The meaning of the conditions $\Delta(k) \in \mathfrak{m}$ is intuitively understood as follows. As explained in the previous subsection an N-dimensional redundancy due to the scale invariance exists in the description of B. A natural way to suppress this redundancy is [2, 3] to attach N constraints, $\Delta(k)_{ii} = 0$ for $1 \leq i \leq N$, since the diagonal degrees of freedoms correspond to the componentwise scalings. Actually, we can identify $\Delta \in \mathfrak{m}$ with the tangent space of the coset $\mathbb{R}^{\times N} \backslash GL(N, \mathbb{R})$, which explains why we call this framework "optimization problems on $\mathbb{R}^{\times N} \backslash GL(N, \mathbb{R})$. This framework is also understood by using the concept of the nonholonomy. For details on these topics, see [2] and [1].

2.3. consideration on Newton-type algorithms

If we apply the conventional Newton's method on this setting, Δ_{ij} is determined by using all components $\{y_1, \cdots, y_N\}$. There it is necessary to deal with $N^2 \times N^2$ matrices which become quite gigantic when N is large.

Another choice is to decompose the optimization flow (3) as

$$\exp(\Delta) \to \mathrm{e}^{\Delta^{(N-1,N)}} \cdots \mathrm{e}^{\Delta^{(1,3)}} \mathrm{e}^{\Delta^{(1,2)}} \tag{5}$$

where we assume that only (k, l)-th and (l, k)-th elements are nonzero among the elements of $\Delta^{(k,l)} \in \mathfrak{m}$, we can strictly determine elements of $\{\Delta^{(k,l)}\}$'s only by using the two-body interaction. This is in its spirit similar to the Jacobi's method for diagonalization of hermitian matrices. It is obvious, however, that the number of steps becomes very large if we adopt this approach because a single step in (3) is decomposed into $N(N-1)$ steps. This is a serious demerit of this approach.

2.4. similarity to quantum lattice

The nested Newton's method which we propose here provides a good prescription to these disadvantages of conventional methods. Let us consider discretized time evolutions on quantum lattice on the following setting:

- There are N fixed particles on lattice.
- Each particle has inner degree of freedom. The inner degree of freedom for each particle is represented by T-dimensional vector.
- Only two-body interactions between particles are present.

The time evolution in this system is described as

$$\phi(t) = \mathrm{e}^{\mathrm{i}H(t-1)}\mathrm{e}^{\mathrm{i}H(t-2)} \cdots \mathrm{e}^{\mathrm{i}H(0)} \phi(0) \ , \tag{6}$$

where the hamiltonian $H(t)$ for each t is an $N \times N$ hermitian matrix. Since there is no n-body interactions with $n \geq 3$, the matrix element $H_{ij}(t)$ for a pair (i, j) depends only on the inner degrees of freedom of i-th and j-th particles. Since the wave function ϕ is the direct sum of S-dimensional vectors which represent the inner degrees of freedom:

$$\phi(t) \in \{(\mathbb{C}^T)^N\} \cong \mathbb{C}^T \times \mathbb{C}^T \times \cdots \times \mathbb{C}^T ,$$

we can express ϕ by an $N \times T$ complex matrix.

Each off-diagonal elements of H, say H_{ij}, is considered as an interaction between ϕ_i and ϕ_j. On many of quantum lattice systems, $H_{ij}(t)$ is dependent only on $\phi_i(t)$ and $\phi_j(t)$ for a specific pair (i,j). (Of course, three-body interactions are not exceptional.)

On the other hand, the ICA flow is described as

$$y(n) = \mathrm{e}^{\Delta(n-1)}\mathrm{e}^{\Delta(n-2)} \cdots \mathrm{e}^{\Delta(0)} y(0) ,$$

where $y(k)$ is interpreted as a direct sum,

$$y(k) \in \{\mathbb{R}^{N\times T}\} \cong \mathbb{R}^T \times \mathbb{R}^T \times \cdots \times \mathbb{R}^T .$$

Now, it is easily understood that there is a similarity between the quantum lattice and optimization problems on $(R^{\times})^N \backslash GL(N,\mathbb{R})$.

ICA	quantum analogy
N-channel observation	N particles (on lattice)
T samples	T dimensional representation of inner degree of freedom

2.5. optimization flow as dynamics under two-body interactions

We can construct a natural and cheap updating rule by using the analogy between the updating rules (4) and (6). That is exactly what we call the nested Newton's method. The essence of the nested Newton's method are as follows:

1. We do not decompose the problem as (5). Thus, we can avoid explosions of the step size.
2. We ignore n-body interactions with $n \geq 3$, which are, in fact, not important. Then, components of Δ in (3) is determined directly by using only two-body interactions. Note that this is quite usual in the gradient descent literatures. We, however, adopt the Newton's method to determine the interaction between two components.
3. Thanks to this translation into two-body problems, the updating rule is decomposed into manipulations of 2×2 matrices. Thus the nested method is practical for large N. Furthermore, its convergence is quite fast since this is a quasi-Newton's method.

The concrete procedures are explained in the following. Note that this method is quite new though it is in some sense related to the so-called natural gradient method or nonholonomic method.

2.6. concrete procedure of nested Newton's method

Let us return to the optimization problem given at (2.1). When $M > 3$, it is impossible to determine simultaneous zeros of all $\{Q_k(y_i,y_j)\}$ in general. How can we get good solutions to these over-determined optimization problems? Our choice is to translate the problem to the minimization of squared sum

$$F = \sum_{i,j,k} w_{ijk} Q_k(y_i,y_j)^2 \tag{7}$$

of criteria $\{Q_k(y_i,y_j) | 1 \leq k \leq M\}$ with weights $\{w_{ijk} \geq 0\}$. We define M-dimensional vectors $\{\boldsymbol{f}^{(ij)}\}$ by

$$\boldsymbol{f}^{(ij)} = (Q_1(y_i,y_j), Q_2(y_i,y_j), \cdots, Q_M(y_i,y_j))^{\mathrm{T}}$$

and $M \times M$ positive definite diagonal matrices $\{S^{(ij)}\}$ by $S^{(ij)}_{kl} = \delta_{kl} w_{ijk}$. Then we decompose the single criterion F to the sum of smaller squared sums:

$$F = \sum_{i<j} F^{(ij)} ,$$

where we have set

$$F^{(ij)} \stackrel{\mathrm{def}}{=} {\boldsymbol{f}^{(ij)}}^{\mathrm{T}} S^{(ij)} \boldsymbol{f}^{(ij)} . \tag{8}$$

Let us focus on a specific pair i,j $(i<j)$. We will examine the variation of a part of criteria $F^{(ij)}$ under

$$y \to \exp(\Delta) y \quad \text{where } \Delta \in \mathfrak{m} . \tag{9}$$

and neglect n-body interactions with $n \geq 3$. First, the variation of $Q_k(y_i,y_j)$ up to second order terms is expressed by using only Δ_{ij} and Δ_{ji} as

$$\begin{aligned} Q_k(y_i,y_j) \to & Q_k(y_i,y_j) + (Q^{(1)}_k(y_i,y_j))^{\mathrm{T}} \vec{\Delta} \\ & + \frac{1}{2} \vec{\Delta}^{\mathrm{T}} Q^{(2)}_k(y_i,y_j) \vec{\Delta} , \end{aligned} \tag{10}$$

where $\vec{\Delta} \stackrel{\mathrm{def}}{=} (\Delta_{ij},\ \Delta_{ji})^{\mathrm{T}}$ and 2×1 and 2×2 matrices $Q^{(1)}_k(y_i,y_j)$ and $Q^{(2)}_k(y_i,y_j)$ are determined from given data. It leads to the expression of the variation of $F^{(ij)}$:

$$\begin{aligned} F^{(ij)} \to & F^{(ij)} + 2\vec{\Delta}^{\mathrm{T}} {V^{(ij)}}^{\mathrm{T}} S^{(ij)} \boldsymbol{f}^{(ij)} \\ & + \vec{\Delta}^{\mathrm{T}} ({V^{(ij)}}^{\mathrm{T}} S^{(ij)} V^{(ij)}) \vec{\Delta} + \vec{\Delta}^{\mathrm{T}} r^{(ij)} \vec{\Delta} \end{aligned} \tag{11}$$

where $V^{(ij)}$ is an $M \times 2$ matrix given by

$$V^{(ij)} = \left[Q^{(1)}_1(y_i,y_j),\ Q^{(1)}_2(y_i,y_j),\ \cdots,\ Q^{(1)}_M(y_i,y_j)\right]^{\mathrm{T}}$$

and $r^{(ij)}$ is a 2×2 matrix determined by

$$r^{(ij)} = \sum_{k=1}^{M} (S^{(ij)} \boldsymbol{f}^{(ij)})_k Q^{(2)}_k(y_i,y_j) . \tag{12}$$

Note that elements of Δ other than Δ_{ij} and Δ_{ji} do not appear in (11) thanks to the two-body interaction principle. Inversely speaking, Δ_{ij} and Δ_{ji} do not appear in the variation of $F^{(pq)}$ up to second order if $(i,j) \neq (p,q)$. Thus we can determine Δ_{ij} and Δ_{ji} readily from the variation of $F^{(ij)}$. That is, $(\Delta_{ij}, \Delta_{ji})$ which minimizes the decomposition of the criteria up to second order is determined by

$$\vec{\Delta} = -\left({V^{(ij)}}^{\mathrm{T}} S^{(ij)} V^{(ij)} + r^{(ij)} \right)^{-1} {V^{(ij)}}^{\mathrm{T}} S^{(ij)} \boldsymbol{f}^{(ij)} \, . \tag{13}$$

By repeating this procedure for all $1 \leq i < j \leq N$, an updating width Δ is completely determined. Note that (13) is 2-dimensional and there is no need to deal with large matrices.

3. NESTED METHOD AS ICA ALGORITHM WITHOUT PREWHITENING

3.1. fast method without 2nd order statistics

We assume that N mutually independent random variables $\{s_i | 1 \leq i \leq N\}$ with zero means lie behind the observed data and that two random variables $\{s_i\}$ and $\{y(0)_i\}$ are related by $y(0)_i = \sum_j A_{ij} s_j + \eta_i$, where $\{\eta_i\}$ constitutes an N-dimensional gaussain random variable with zero-mean and variance v_{ij}. The goal of independent component analysis(ICA) is estimation of A or its inverse.

Here we construct a nested method for ICA by slightly deforming the nested method explained thus far. First, we set

$$Q_k(y_i, y_j) = \langle y_i^k y_j^{4-k} \rangle_{\mathrm{c}} \tag{14}$$

and

$$w_{ijk} = \frac{1}{|\langle y_i^4 \rangle_{\mathrm{c}}|^{k/2} |\langle y_j^4 \rangle_{\mathrm{c}}|^{2-k/2}} \tag{15}$$

for $k = 1, 2, 3$, which leads to the slight violation of the invariance under scaling and at the same time leads to the stability. On this setting we use only fourth order cumulants for the criteria of independence. Thus, the algorithm becomes robust under gaussian noises. Note that for this method, there is no need to estimate the noise variance. Many practical ICA algorithms use the 2nd order statistics as a part of criteria. To apply these algorithms to noisy cases, we have to estimate the noise variance, which is usually accomplished by Factor analysis(FA). It is, however, known that FA is available only if the inequality

$$M \leq \frac{1}{2} \left(2N + 1 - \sqrt{8N + 1} \right) \tag{16}$$

holds where M and N are repspectively the number of signals and observation channels. Thus in case where (16) does not hold, ICA algorithms based on FA is not available. Our method works, however, perfectly also in that case. This is a major advantage of our method.

3.2. deformation to increase global stability

Here we deform the updating rule to increase the global stability of the algorithm. We will use the notations,

$$K_i = Q_3(y_i, y_i) \, , \quad Q_{ij} = Q_3(y_i, y_j) \, , \quad R_{ij} = Q_2(y_i, y_j) \, .$$

Then, the gradient $V^{(ij)}$ of $\boldsymbol{f}^{(ij)}$ is expressed as

$$V^{(ij)} = \begin{pmatrix} K_j & 3R_{ij} \\ 2Q_{ji} & 2Q_{ij} \\ 3R_{ij} & K_i \end{pmatrix} \, . \tag{17}$$

Now we introduce a nontrivial deformation of $V^{(ij)}$,

$$V^{(ij)} \to V^{(ij)}_{(\epsilon)} = \begin{pmatrix} K_j & (3-\epsilon)R_{ij} \\ 2Q_{ji} & 2Q_{ij} \\ (3-\epsilon)R_{ij} & K_i \end{pmatrix} \, , \tag{18}$$

where ϵ is some small positive number. Then the updating rule becomes

$$\vec{\Delta} = -\left({V^{(ij)}_{(\epsilon)}}^{\mathrm{T}} S^{(ij)} V^{(ij)}_{(\epsilon)} + r^{(ij)} \right)^{-1} {V^{(ij)}_{(\epsilon)}}^{\mathrm{T}} S^{(ij)} \boldsymbol{f}^{(ij)} \, . \tag{19}$$

We will adopt the modified updating rule (19). Note that the difference between these two updating rules (13) and (19) vanishes around the optimal point since there it is expected that $R_{ij} = 0$. This deformation, however, increases significantly the global stability.

We will illustrate this situation from an $N = 2$ case. Let us introduce functions $\boldsymbol{f}^{(ij)}_{(\epsilon)}$ and assume that the gradient of $\boldsymbol{f}^{(ij)}_{(\epsilon)}$ is $V^{(ij)}_{(\epsilon)}$. Now we define the symmetric part $\Delta^{\mathrm{S}} = (\Delta_{ij} + \Delta_{ji})/2$ and the anti-symmetric part $\Delta^{\mathrm{A}} = (\Delta_{ij} - \Delta_{ji})/2$ of of Δ. Then the variation of ${\boldsymbol{f}_{(\epsilon)}}^{(ij)}$ is written as

$$\boldsymbol{f}^{(ij)}_{(\epsilon)} \to \boldsymbol{f}^{(ij)}_{(\epsilon)} + v^{\mathrm{S}}_{(\epsilon)} \Delta^{\mathrm{S}} + v^{\mathrm{A}}_{(\epsilon)} \Delta^{\mathrm{A}} \tag{20}$$

where we have set

$$v^{\mathrm{S}}_{(\epsilon)}(t) = \left. \begin{pmatrix} K_j + (3-\epsilon)R_{ij} \\ 2Q_{ji} + 2Q_{ij} \\ (3-\epsilon)R_{ij} + K_i \end{pmatrix} \right|_{y=y(t)} \tag{21}$$

and

$$v^{\mathrm{A}}_{(\epsilon)}(t) = \left. \begin{pmatrix} K_j - (3-\epsilon)R_{ij} \\ 2Q_{ji} - 2Q_{ij} \\ (3-\epsilon)R_{ij} - K_i \end{pmatrix} \right|_{y=y(t)} \tag{22}$$

and omitted (t). Let us consider the initial stage. Our assumption is that $y(0) = As + \eta$. At the initial point, we do not know anything about the transfer matrix A. Suppose that every A_{ij} is a gaussian random variable with an identical variance σ and zero-mean. We call this ensemble the Laguerre orthogonal ensemble (LOE) as in the context of physics[4, 5]. Since there is no convergent invariant measure on the space of $N \times N$ real matrices, it is natural to discuss the global stability by assuming that the mixing matrix is distributed from the LOE. First, we can show that

$$\mathrm{E}_{\mathrm{LOE}}(v^{\mathrm{A}}_{(\epsilon)}(0)) = q\epsilon \begin{pmatrix} 1 & 0 & -1 \end{pmatrix}^{\mathrm{T}} \overset{\mathrm{def}}{=} \bar{v}^{\mathrm{A}} , \tag{23}$$

where $\mathrm{E}_{\mathrm{LOE}}()$ is the expectation under LOE and q is a constant independent of ϵ. Then, since

$$\mathrm{E}_{\mathrm{LOE}}(v^{\mathrm{A}}_{(0)}(0)) = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}^{\mathrm{T}} , \tag{24}$$

is always satisfied, the original updating rule (13) is instable along the Δ^{A} direction at the initial stage. In this way, it is understood that non-zero ϵ is desirable. When the algorithm stops successfully, it is expected that $|R_{ij}|$, $|Q_{ij}|$, and $|Q_{ji}|$ become very small and the main contributions come from $K_i \overset{\mathrm{def}}{=} K^{(\infty)}_i$ and $K_j \overset{\mathrm{def}}{=} K^{(\infty)}_j$. Thus it is reasonable to expect that

$$v^{\mathrm{A}}_{(\epsilon)} \to v^{\mathrm{A}}_{(\epsilon)}(\infty) = \begin{pmatrix} K^{(\infty)}_j & 0 & -K^{(\infty)}_i \end{pmatrix}^{\mathrm{T}} \tag{25}$$

around the optimal point. Generally, the Newton's method becomes instable when $|v^{\mathrm{A}}_{(\epsilon)}(t)|$ is very small during the updating. If $v^{\mathrm{A}}_{(\epsilon)}(\infty)$ is located at the opposite side of $\bar{v}^{\mathrm{A}}_{(\epsilon)}$ from the origin, it is likely that the flow enters the dangerous regions where $|v^{\mathrm{A}}_{(\epsilon)}(t)| \approx 0$. Let us assume that ϵ is positive.

1. In case $K^{(\infty)}_i K^{(\infty)}_j > 0$ is satisfied, it is shown that the sign of q is the same as that of $K^{(\infty)}_i$ (and of $K^{(\infty)}_j$). Since $v^{\mathrm{A}}_{(\epsilon)}(\infty)$ is written as

   $$\begin{pmatrix} |K^{(\infty)}_j| & 0 & -|K^{(\infty)}_i| \end{pmatrix}^{\mathrm{T}}$$

   in this case, we can easily show that the inequality

   $$\left(v^{\mathrm{A}}_{(\epsilon)}(\infty)\right)^{\mathrm{T}} \bar{v}^{\mathrm{A}}_{(\epsilon)} > 0$$

   holds, which means that the angle between $v^{\mathrm{A}}_{(\epsilon)}(\infty)$ and $\bar{v}^{\mathrm{A}}_{(\epsilon)}$ is acute.

2. When $K^{(\infty)}_i K^{(\infty)}_j < 0$, $v^{\mathrm{A}}_{(\epsilon)}$ around the optimal point is written as

   $$v^{\mathrm{A}}_{(\epsilon)}(\infty) = \pm \begin{pmatrix} |K^{(\infty)}_j| & 0 & |K^{(\infty)}_i| \end{pmatrix}^{\mathrm{T}} .$$

   If $|K^{(\infty)}_i| = |K^{(\infty)}_j|$, the angle between $v^{\mathrm{A}}_{(\epsilon)}(\infty)$ and $\bar{v}^{\mathrm{A}}_{(\epsilon)}$ is maximized when $|K^{(\infty)}_i|$ or $|K^{(\infty)}_j|$ becomes very small and the upper bound is $3\pi/4$.

Combining these, the angle between $v^{\mathrm{A}}_{(\epsilon)}(\infty)$ and $\bar{v}^{\mathrm{A}}_{(\epsilon)}$ is at most $3\pi/4$ for positive ϵ. Thus it is realized that the possibility to enter the dangerous regions is reduced greatly by a positive ϵ.

4. PERFORMANCE

We have performed numerical experiments to show the power of our method as an ICA algorithm. For reference, we have also analyzed the same data by FICA[6] and JADE[7]. The source signals are 6-dimensional sound data with 48000 samples, which are mixed by a matrix chosen from the LOE. This setting should make the experiences more convincing. Although singular matrices have measure zero on the LOE, matrices with very large condition numbers may be choosen on this setting. After the mixture we add noises. For the noises we choose 6-dimensional mutually independent gaussian random variables. (Actually, the nested method is also robust under the correlated gaussian noises.) We will measure the performances by the following criteria. First, let us set $G = BA(C^S)^{1/2}$, where $C^S = \mathrm{diag}(\langle s_1^2\rangle, \cdots, \langle s_N^2\rangle)$. We also introduce G_M by $G_M = \mathrm{diag}(\max_k |G_{1k}|, \max_k |G_{2k}|, \cdots, \max_k |G_{Nk}|)$. Then the matrix elements $\{(GG_M^{-1})_{ij}\}$ for $j \neq \mathrm{argmax}_k|G_{ik}|$ are a natural measure of crosstalks for an estimated demixing matrix B. In this paper we measure the signal-to-noise ratio (SNR) by $-10\log_{10}\{\mathrm{median}(\mathfrak{B})\}$ (db).

[No noise case]

	CT(mean)	max CT(med.)	time(med.,sec.)
nested Newton	0.0069	0.0342	6.51
JADE	0.0052	0.0110	2.88
FICA	0.0061	0.0180	3.58

Here, JADE are better than our method. For the total CT our method is better than FICA. It is natural that methods with pre-WH outperforms our method provided an prior knowledge that ther is no noise. Nevertheless, our method is competitive also in this case.

[Small noise (37.3518 db) case]

	CT(mean)	max CT(med.)	time(med.,sec.)
nested Newton	0.0074	0.0341	6.93
JADE	0.0085	0.0118	3.01
FICA	0.0097	0.0209	3.63

For the max CT, JADE is the best. For the total CT, our method is the best.

[Middle noise case (The SNR is 16.9171 db.)]

	CT(mean)	max CT(med.)	time(med.,sec.)
nested Newton	0.0247	0.0359	7.38
JADE	0.0505	0.2268	3.13
FICA	0.0525	0.2684	3.91

Our method outperforms the other methods in this region. As a whole, it is demonstrated that our algorithm is extremely useful and practical in cases where there might be some noises.

5. REMARKS

We have constructed a new method called the nested Newton's method. A close connection between cosets $\mathbb{R}^{\times}{}^{N}\backslash GL(N,\mathbb{R})$ and $U(1)^N\backslash U(N)$ is known, where the former coset is said to be paracomplex manifold and the latter is known as a complex manifold. Since dynamics on the latter coset is nothing but the quantum dynamics, it is reasonable to call dynamics on the latter coset para-quantum dynamics. Thus the nested method is naturally understood as para-quantum dynamical method for optimization. This explains why it is natural to use the analogy the quantum dynamics in order to construct a useful algorithm for the multivariate optimization. Characteristics of various methods is listed in Table 1 and 2.

The nested method can be applied to various optimization problems on $GL(N,\mathbb{R})$ provided the optimum is scale invariant. Given the scale invariance, the optimization is formulated as a flow on the coset $(\mathbb{R}^{\times})^N\backslash GL(N,\mathbb{R})$. This coset structure is fatally important for the factorization of the problem to a collection of 2-dimensional problems. Indeed, it is easily understood that on the general linear groups, we can not factorized the procedures in the same way.

Table 1. time evolution and interaction

	generator of time evolution	continu -ous limit	typical interaction
quantum lattice	i × hermitian matrix ($\mathfrak{u}(N)$)	○	2-body
gradient flow on $GL(N,\mathbb{R})$	$\Delta \in \mathbb{R}^{N\times N}$	○	2-body
gradient flow on $\mathfrak{M}$	$\Delta \in \mathbb{R}^{N\times N}$ $\Delta_{kk} = 0$	○	2-body
Newton's method on $\mathfrak{M}$	$\Delta \in \mathbb{R}^{N\times N}$ $\Delta_{kk} = 0$	×	N-body (expensive)
nested New-ton's method	$\Delta \in \mathbb{R}^{N\times N}$ $\Delta_{kk} = 0$	×	2-body (cheap)

Here, we have set $\mathfrak{M} = (\mathbb{R}^{\times})^N\backslash GL(N,\mathbb{R})$.

As we have shown, nested Newton's method is useful as ICA algorithm robust under gaussian noises. Remember that we do not prewhiten the data. In the presence of noises, the pre-WH results in 'the overwhitening' since at the optimal point the off-diagonal elements of the covariance matrix do not necessarily vanish. This explains partly the reason that the pre-WH is not preferable. In this paper we use, for the criteria of the independence, solely the fourth order cumulants, which are obviously not sensitive to the gaussian noises. As a result, our method is robust under strong gaussian noises. Other advantages are as follows. **(1)** It works well also in cases where the number of the signal and the number of the channel of the observation differ. **(2)** Thanks to the factorization, the memory space and computational quantity required for each step becomes of order N^2. **(3)** The algorithm stops after fairly small iterations because of its quasi-Newton nature. **(4)** It is considerably stable globally by the deformation in §3.2. It is thanks to all these features that our method becomes highly practical.

Table 2. advantage as an optimization method

	learning rate tuning	required step #	conver -gence	each step
gradient descent	required	large	linear	cheap
Newton's method	not required	small	2nd order	expensive
nested Newton's method	not required	rather small	almost 2nd or-der	cheap

6. REFERENCES

[1] T.Akuzawa, "Extended quasi-newton method for the ica," in *Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation*, 2000, pp. 521–525.

[2] T.Akuzawa and N.Murata, "Multiplicative nonholonomic/newton -like algorithm," *Chaos, Solitons & Fractals*, vol. 12, pp. 785–793, 2001.

[3] S.Amari, T.-P. Chen, and A. Cichocki, "Non-holonomic constraints in learning algorithms for blind source separation," *preprint*, 1997.

[4] K.Slevin and T.Nagao, "New random matrix theory of scattering in mesoscopic systems," *Phys.Rev.Lett.*, vol. 70, pp. 635–638, 1993.

[5] T.Akuzawa and M.Wadati, "Non-hermitian random matrices and integrable quantum hamiltonians," *J.Phys.Soc.Jpn*, vol. 65, pp. 1583–1688, 1996.

[6] A.Hyvärinen, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, pp. 1483–1492, 1997.

[7] J-F. Cardoso and A.Souloumiac, "Blind beamforming for non gaussian signals," in *IEE Proceedings-F*, 1993, vol. 140, pp. 362–370.