EXTENDED QUASI-NEWTON METHOD FOR THE ICA

Toshinao Akuzawa

Lab. for Information Synthesis, Brain Science Institute, RIKEN
2-1 Hirosawa, Wako, Saitama 351-0198, Japan

ABSTRACT

As extensions to the strict method developed in [1], variations of the Newton method for the independent component analysis(ICA) are proposed. Our method presented here is highly practical and simple. Concrete merits of our algorithm are as follows. i) Robust under gaussian noises. In the presence of strong gaussian noises it outperforms the existing methods like the JADE[2] and the FICA[3]. ii) By two deformations it becomes considerably stable globally. Although the first deformation is apparently unnatural, a justification is given based on the random matrix arguments. iii) Each step of the methods proposed here resolves itself into the determination of the inverse of 2×2 matrices or generalized inverse matrices of 3×2 matrices. There is no need to deal with gigantic matrices and little computational resources are required.

1. OVERVIEW

The ICA is considered as an optimization problem on the frame of the N-dimensional projective space. In [1] Akuzawa and Murata have proposed a multiplicative Newton algorithm for the ICA by regarding the frame as the coset $GL(1,\mathbb{R})^N\backslash GL(N,\mathbb{R})$. The method described in [1] is a pure-Newton method and the second-order-convergence is shown rigorously. In this paper we will propose new methods for the ICA starting from this pure-Newton method. First, we construct a cheap quasi-Newton method. Then the method is deformed by the introduction of a "stabilizing parameter" ξ and acquires a pretty good global convergence. Secondly, we propose an extended quasi-Newton method, which directly results in the construction of a highly practical algorithm. The extension introduced here may be useful for many overdetermined optimization problems other than the ICA. The algorithm thus constructed is extremely robust under gaussian noises even if the noises are correlated, which is a fatally important feature when we deal with data preprocessed by the PCA.

The framework is as follows. We denote by $<>_{\rm c}$ cumulants estimated from the observed data:

$$< \prod_i y_i^{m_i} >_{\rm c} = \left(\prod_k \frac{\partial^{m_k}}{\partial \xi_k^{m_k}} \right) \ln \left\langle (\exp(\sum_i y_i \xi_i) \right\rangle \Bigg|_{\xi_i=0, 1\le i\le N} , \tag{1.1}$$

where $<>$ is the sample average. In this paper we will mainly deal with the fourth order cumulants. The fourth order cumulants such as $< y_1^3 y_2 >_{\rm c}$, $< y_1^2 y_2 y_3 >_{\rm c}$, and $< y_1^2 y_2^2 >_{\rm c}$ are called, respectively, $(3,1)$-type, $(2,1,1)$-type, and $(2,2)$-type. The exploitation of the $(2,2)$-type cumulants is characteristic of our method. We assume that T samples of N-dimensional variables $\{y_{it}^{(0)} | 1\le i\le N,\ 1\le t\le T\}$ are available as observed data. The mean value have already drawn from the data, that is, $\sum_{t=1}^{T} y_{it} = 0$ for all i. Hereafter the last lower index denoting the sample number is omitted and the data are denoted as N-dimensional vectors. For example, $y^{(0)} = (y_1^{(0)}, y_2^{(0)}, \cdots, y_N^{(0)})^{\rm T}$, where we denote by the upper subscript T the transposition. It is presumed that N-dimensional mutually independent random variables $\{s_i | 1\le i\le N\}$ with zero means lie behind the observed data. We assume that two random variables $\{s_i\}$ and $\{y_i\}$ are related by

$$y_i = \sum_j A_{ij} s_j + \eta_i , \tag{1.2}$$

where $\{\eta_i\}$ constitutes an N-dimensional gaussain random variable with zero-mean and variance v_{ij}. The noises $\{\eta_i\}$'s are not assumed to be mutually independent. Note that the robustness under the gaussian noises is not acquired by methods which exhaust a half degrees of freedom in the prewhitening even if the noises are not correlated. This is one of the motivations for us to construct algorithms which do not require prewhitening.

We consider a sequence $y^{(0)}, y^{(1)}, y^{(2)}, \cdots$, which converges to the optimal point $y^{(\infty)}$ where each component becomes mutually independent. We specify the flow of this sequence by $N\times N$ matrices $\{\Delta^{(s)}; s=0,1,2,\cdots\}$, which describe the amount of individual steps:

$$y^{(s+1)} = (\exp \Delta^{(s)}) y^{(s)} . \tag{1.3}$$

akuzawa@brain.riken.go.jp, http://www.islab.brain.riken.go.jp/~akuzawa/

Since the ICA is an $N(N-1)$-dimensional optimization, an N-dimensional redundancy remains. To suppress the indefiniteness we attach constraints $\Delta_{ii}^{(s)} = 0$ for $1 \le i \le N$[1, 4]. These constraints are natural since the diagonal degrees of freedoms correspond to the componentwise scalings and the scalings are nothing to do with the independence[1]. The goal of this paper is the construction of sequences $\{\Delta^{(s)}\}$ which are robust under gaussian noises and converge rapidly.

In the following two sections we will deal with only one time step and show how to determine $\Delta^{(s)}$ from the data $\{y^{(s)}\}$. So the upper subscript $\{(s)\}$ will be omitted.

2. QUASI-NEWTON METHOD

2.1. normal quasi-Newton

A Newton algorithm on a p-dimensional space is interpreted in two ways:

1. the minimization of a cost function based on its second order expansion
2. the determination of a point where objective functions f_i for $1 \le i \le p$ vanish simultaneously based on their first order expansions.

Here we adopt the second interpretation since it can deal with problems of broader range. We choose as the objective functions

$$Q_{ij} = < y_i^3 y_j >_{\rm c} \ , \quad i \neq j \ , \quad 1 \le i, j \le N \ , \tag{2.1}$$

that is, we want to determine a set $\{y_i\}$ which satisfies $Q_{ij} = 0$ for all i and $j(\neq i)$. We denote by $O(\Delta^k)$ polynomials of matrix elements of Δ which does not contain terms with degrees less than k. Under the transformation $y \to \exp(\Delta)y$, Q_{ij} transforms as

$$Q_{ij} \to Q_{ij} + \Delta_{ji} < y_i^4 >_{\rm c} + 3\Delta_{ij} < y_i^2 y_j^2 >_{\rm c} + O(\Delta^2) \ , \tag{2.2}$$

where we have neglected cumulant terms which contain more than two distinct components such as $< y_1 y_2 y_3^2 >_{\rm c}$. This approximation is similar to that by Amari[5]. The amount of a step Δ is determined after the Newton manner. That is, Δ_{ij} and Δ_{ji} is determined by the condition

$$\begin{pmatrix} Q_{ij} \\ Q_{ji} \end{pmatrix} + \begin{pmatrix} < y_i^4 >_{\rm c} & 3 < y_i^2 y_j^2 >_{\rm c} \\ 3 < y_i^2 y_j^2 >_{\rm c} & < y_j^4 >_{\rm c} \end{pmatrix} \begin{pmatrix} \Delta_{ji} \\ \Delta_{ij} \end{pmatrix} = 0 \ , \tag{2.3}$$

or more explicitly,

$$\begin{pmatrix} \Delta_{ji} \\ \Delta_{ij} \end{pmatrix} = - \begin{pmatrix} < y_i^4 >_{\rm c} & 3 < y_i^2 y_j^2 >_{\rm c} \\ 3 < y_i^2 y_j^2 >_{\rm c} & < y_j^4 >_{\rm c} \end{pmatrix}^{-1} \begin{pmatrix} Q_{ij} \\ Q_{ji} \end{pmatrix} \ . \tag{2.4}$$

Note that conditions for $N(N-1)$ elements of Δ are divided into 2-dimensional linear equations thanks to the approximation. Thus the determination of Δ by this procedure is computationally much cheaper for a large N than the pure-Newton method and the JADE, which requires an eigenvalue decomposition of the $N^2 \times N^2$ cumulant matrix. The updating rule based on (2.4) works perfectly if we start from a point close to the optimal one. Unfortunately this method is not necessarily stable globally. In the next subsection we propose a deformation of the updating rule which makes the method considerable stable.

2.2. deformation and stability on earlier stages

Let us introduce a positive number ξ and define $S(\xi)$ as

$$S(\xi) = \begin{pmatrix} < y_i^4 >_{\rm c} & (3-\xi) < y_i^2 y_j^2 >_{\rm c} \\ (3-\xi) < y_i^2 y_j^2 >_{\rm c} & < y_j^4 >_{\rm c} \end{pmatrix} \ . \tag{2.5}$$

Then we propose to use

$$\begin{pmatrix} \Delta_{ji} \\ \Delta_{ij} \end{pmatrix} = -S(\xi)^{-1} \begin{pmatrix} Q_{ij} \\ Q_{ji} \end{pmatrix} \tag{2.6}$$

as the updating rule instead of (2.4). The difference between (2.6) and (2.4) vanishes at the optimal point. Numerical experiments indicate that a positive ξ is crucially important for the global stability of the algorithm. This is an empirical fact. Let us illustrate the situation more theoretically. In this subsection we neglect the noise and assume that $y = As$. 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 exsemble the Laguerre orthogonal ensemble (LOE) as in the context of physics[6, 7]. Since there is no invariant measure on $GL(N, \mathbb{R})$, it is natural to adopt the LOE as a prior distribution. We can show, however, that

$$E_{\rm LOE}(S(\xi)) = a \begin{pmatrix} 3 & 3-\xi \\ 3-\xi & 3 \end{pmatrix} \tag{2.7}$$

where a is some real number. So if we choose $\xi = 0$, the average becomes singular. This explains the importance of nonzero ξ. In consideration of the fact that $< y_i^2 y_j^2 >_{\rm c}$ must vanish at the optimal point, it is understood that a negative value for ξ is not a good choice for the global stability. In the example below we adaptively control the value for ξ: we choose $\xi = 1$ in the beginning and when the updating width becomes small, we alter it to smaller value like $\xi = 0.3$. By the introduction of ξ, the algorithm becomes fairly stable. Further deformation is, however, in order, which is explained in the next section.

3. EXTENDED QUASI-NEWTON METHOD

In this section we extend our method furthermore. The method which will be proposed here is a new extension to the Newton method. The ICA is an optimization on an $N(N-1)$-dimensional space. This is essentially an overdetermined problem since the optimal point should be among the zero point of infinitely many functions. Any of these functions can be used as the objective function. In the algorithm above described, we have chosen the (3,1)-type fourth order cumulants. Here we introduce additional $(2,2)$-type objective functions,

$$R_{ij} = < y_i^2 y_j^2 >_{\mathrm{c}} \ . \tag{3.1}$$

It transforms under the action of e^{Δ} to y as

$$R_{ij} \to R_{ij} + 2\Delta_{ji} < y_i^3 y_j >_{\mathrm{c}} + 2\Delta_{ij} < y_i y_j^3 >_{\mathrm{c}} + O(\Delta^2) \ , \tag{3.2}$$

where we have omitted terms with more than two distinct components as in the previous case. We denote the kurtosis of each component of y by K_i:

$$K_i = < y_i^4 >_{\mathrm{c}} \ . \tag{3.3}$$

Let us combine (2.3) with (3.2) and consider the vector,

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = f + \begin{pmatrix} K_i & (3-\xi)R_{ij} \\ (3-\xi)R_{ij} & K_j \\ 2Q_{ij} & 2Q_{ji} \end{pmatrix} \begin{pmatrix} \Delta_{ji} \\ \Delta_{ij} \end{pmatrix} \ , \tag{3.4}$$

where

$$f = \begin{pmatrix} Q_{ij} \\ Q_{ji} \\ R_{ij} \end{pmatrix} \ . \tag{3.5}$$

At the optimal point, each element of f must vanish. Since $\alpha = 0$ is an overdetermined problem we choose two rows from it in the usual Newton approach to determine the updating width Δ. Here we try to utilize all of the three rows, that is, we choose Δ which minimizes the norm of α. We use the euclidean norm. For brevity's sake we introduce vectors r_1 and r_2 by

$$r_1 = \begin{pmatrix} K_i \\ (3-\xi)R_{ij} \\ 2Q_{ij} \end{pmatrix} \quad \text{and} \quad r_2 = \begin{pmatrix} (3-\xi)R_{ij} \\ K_j \\ 2Q_{ji} \end{pmatrix} \ . \tag{3.6}$$

Then Δ is determined by

$$\begin{pmatrix} \Delta_{ji} \\ \Delta_{ij} \end{pmatrix} = - \begin{pmatrix} (r_1, r_1) & (r_1, r_2) \\ (r_2, r_1) & (r_2, r_2) \end{pmatrix}^{-1} \begin{pmatrix} (r_1, f) \\ (r_2, f) \end{pmatrix} \ . \tag{3.7}$$

Otherwise, we can rephrase (3.7) by using

$$V = \begin{pmatrix} K_i & (3-\xi)R_{ij} \\ (3-\xi)R_{ij} & K_j \\ 2Q_{ij} & 2Q_{ji} \end{pmatrix} \tag{3.8}$$

as

$$\begin{pmatrix} \Delta_{ji} \\ \Delta_{ij} \end{pmatrix} = -(V^{\mathrm{T}} V)^{-1} V^{\mathrm{T}} f \ . \tag{3.9}$$

The updating rule (3.7) or (3.9) is the main result of this paper. The matrix on the right-hand-side of (3.9) is interpreted as a generalized inverse of V. Note that there might be other choices for the norm. This norm, however, works good.

4. PERFORMANCE

We have performed two numerical experiments to verify the performance of our method. For the updating rule we use (3.9). In the first experiment, the source signals are 6-dimensional sound data with 48000 samples. The source signals are mixed by a matrix choosen from the LOE. Although singular matrices have measure zero on the LOE, matrices with very small determinants may be choosen on this setting. After the mixture we add noises. For the noises we choose 6-dimensional mutually independent gaussian random variables, since the JADE and the FICA are not optimized for the correlated gaussian noises. Note that our method is little affected by the presence of the correlations among gaussian noises. In this experiment the noise level is low. The ratio of the standard deviation of the signals and noises are 8.61% on the average. We have iterated 50 times the demixing problem. As a whole, it can be said that the setting of this experiment is fairly tough. Indeed, the conventional simple methods do not work perfectly as in the usual cases. Of course our method also fails sometimes. The numerical results, however, shows the power of our method. The results are as follows(Fig.1). By our method time needed for one trial is 8.76 seconds, the crosstalk is 1.89%, and the maximum remaining crosstalk in one trial is 11.98% each on average. The median of the maximu remaining crosstalk is 6.23%. The results by JADE are 2.79 seconds, 4.40%, 26.88%, and 18.19%. By FICA the results are 4.26 seconds, 5.10%, 31.80%, and 20.08%. These results indicate our method is much stabler than the remaining two methods in this setting.

The next experiment is more noisy one with three signals. The noise level is 29.07%. The results are as follows.

	our method	JADE	FICA
time(sec.)	3.07	0.37	1.47
mean crosstalk(%)	8.51	20.62	21.52
mean max crosstalk(%)	12.7	31.93	33.62
max crosstalk median(%)	3.15	27.15	27.16

It also illustrates the power of our method. See also Fig.2 The results of the JADE and the FICA are quite similar and much worse than our method.

5. CONCLUDING REMARKS

In this paper we have constructed a deformed quasi-Newton method for the ICA. This method is simple and robust under gaussian noises. Moreover, its convergence is quite fast and considerably stable. It must also be noted that we can use our method for fairly high-dimensional problems since the computational quantity for this algorithm is of $O(N^2)$.

The robustness under gaussian noises is a result of two factors. First, since the objective functions are fourth order cumulants they are not affected by the gaussian noises. The second factor is related to the prewhitening process. We do not need to prewhiten the data since the quasi-Newton flow constructed in this paper can move toward any direction in the $N(N-1)$-dimensional space. In the presence of noises, the prewhitening 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 prewhitening is not preferable.

The stability of our method is also the consequence of two ideas: the deformation described in the subsection 2.2 and the addition of the $(2,2)$-type cumulants to the objective functions. The latter is put into shape by the introduction of the generalized inverse technique to the Newton method. Of course, it is fair to mention that our method is slower than the JADE and the FICA in noiseless lower dimensional cases for now.

Let us examine other possibilities. If the dimension of the observations are much greater than the number of signals, the factor analysis can be used to estimate the distribution of the noises and there might be alternative simple methods to obtain results comparable with ours. We can, however, not estimate the variance of the noises accurately by the factor analysis if the number of the source signals are greater than the half of the channel number of the observations. So the method developed in this paper is especially useful in such occasions.

Acknowledgments
The author would like to express my gratitude to Noboru Murata and Shun-ichi Amari for invaluable discussions and comments.

6. REFERENCES

[1] T.Akuzawa and N.Murata, Multiplicative Nonholonomic/Newton -like Algorithm, *preprint* *(available from http://www.islab.brain.riken.go.jp/˜akuzawa/)* (1999).

[2] J.-F. Cardoso and A.Souloumiac, Blind beamforming for non Gaussian signals, in *IEE Proceedings-F*, vol. 140, 362–370 (1993).

[3] J. Hurri, H. Gävert, J. Sälelä, and A. Hyvärinen, FastICA package for MATLAB (1998), http://www.cis.hut.fi/projects/ica/fastica/.

[4] S. Amari, T.-P. Chen, and A. Cichocki, Non-holonomic Constraints in Learning Algorithms for Blind Source Separation, *preprint* (1997).

[5] S. Amari, Superefficiency in Blind Source Separation, *IEEE Transactions on Signal Processing* **47**, 936–944 (1999).

[6] K. Slevin and T. Nagao, New Random Matrix Theory of Scattering in Mesoscopic Systems, *Phys.Rev.Lett.* **70**, 635–638 (1993).

[7] T.Akuzawa and M.Wadati, Non-Hermitian Random Matrices and Integrable Quantum Hamiltonians, *J.Phys.Soc.Jpn* **65**, 1583–1688 (1996).

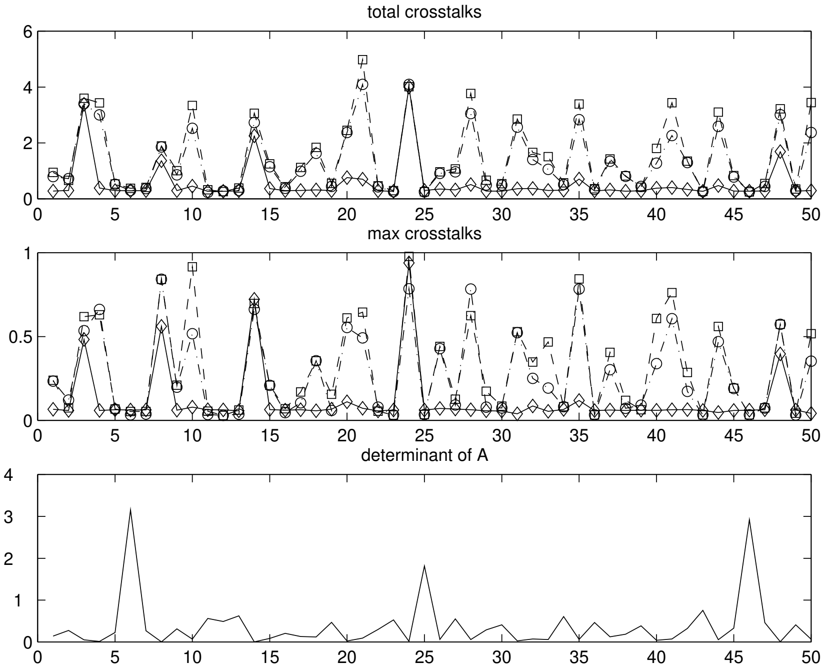

Figure 1: The results of six signal demixing problems. The horizontal axis denotes the trial number. The solid lines with diamonds, broken lines with squares, and dotted lines with circles denote, respectively, the results of our method, the JADE, and the FICA.

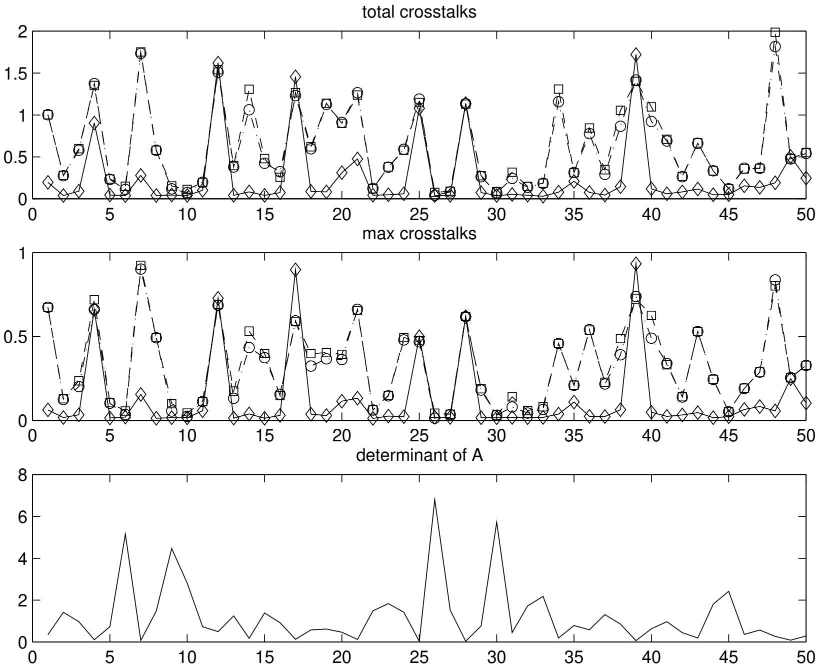

Figure 2: The results of three signal demixing problems. Stronger noises are present than in the previous example.