NATURAL GRADIENT USING SIMULTANEOUS PERTURBATION WITHOUT PROBABILITY DENSITIES FOR BLIND SOURCE SEPARATION

Yutaka Maeda, Takayuki Maruyama

3-3-35, Yamate-cho, Suita 564-8680
Faculty of Engineering, Department of Electrical Engineering, Kansai University
maedayut@kansai-u.ac.jp

ABSTRACT

When independent plural signals are mixed and the mixed plural signals are measured, blind signal separation technique is very interesting approach to separate these signals only based on the measured signals. It is a hot subject in the fields of engineering, for example, communication engineering, signal processing, image processing, analysis of organs inside a body and so on. In this paper, we propose recursive methods to obtain a separating matrix based on mutual information, via the simultaneous perturbation optimization method. The simultaneous perturbation method updates the separating matrix by using only two values of the mutual information. Then, probability densities of source signals, which are used in ordinary gradient methods, are not required. A simple example is shown to confirm a feasibility of the proposed methods.

1. INTRODUCTION

Blind Source Separation (BSS) is a technique that can extract original signals from their mixtures observed by the same number of sensors. BSS is an approach to separate these signals without knowing the mixing coefficients and the information of source signals as well.

BSS is realized on condition that the original signals are independent each other. Therefore, retrieved signals separated from the measured signals separated must be independent as well. Using a criterion to measure independence of the retrieved signals, we can construct a separating matrix and estimate the source signals[1, 2, 3].

When $\boldsymbol{a} \in \Re^n$ is a vector of independent source signals, and $\boldsymbol{A} \in \Re^{n\times n}$ is a mixing process matrix, we have a vector of observation signals $\boldsymbol{x} \in \Re^n$ which are assumed to be a linear transformation of a non-singular matrix $\boldsymbol{A}$. Then, we have

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{a} \tag{1}$$

On condition that mixing matrix $\boldsymbol{A}$ and source signals $\boldsymbol{a}$ are unknown, we use a criterion to measure independence of the retrieved signals $\boldsymbol{y} \in \Re^n$. Then we can construct a separating matrix $\boldsymbol{W} \in \Re^{n\times n}$, and estimate the original source signals $\boldsymbol{a}$. That is,

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} = \boldsymbol{W}\boldsymbol{A}\boldsymbol{a} \tag{2}$$

In BSS, we expect $\boldsymbol{W}$ to be an inverse of $\boldsymbol{A}$. Then, it is obvious that we can retrieve the original signals.

In other words, BSS is an optimization problem in which the criterion measuring the independence is an evaluation function and the separating matrix is a parameter.

Then, the gradient method is a simple approach. The method needs a gradient of an evaluation function; a criterion measuring the independence in this case.

However, the gradient type methods including the natural gradient method use probability densities of the original signals.

In this paper, we apply the simultaneous perturbation optimization method to BSS problem. By using the optimization method, we can update all parameters of the separating matrix based only on two values of the evaluation function. Then, we do not use the probability densities of the original signals.

2. EVALUATION FUNCTION

In this paper, the mutual information based on the information theory is applied to make an evaluation of independence of each separated signal. Let X and Y be two random variables, the mutual information implies a measure of the amount of information about Y contained in X.

This research was financially supported by the Kansai University Research Grants : Grant-in-Aid for Encouragement of Scientists, 2003 and High Technology Research Center of Kansai University.

The mutual information $I(\cdot)$ is defined as follows;

$$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X,Y) \\ &= \sum r_{ij} \log \frac{r_{ij}}{p_i q_j} \end{aligned} \quad (3)$$

Where, $H(\cdot)$ is called the entropy of two random variables $X (= x_1, x_2, , x_n), Y (= y_1, y_2, , y_m)$ or of the distribution $p_i = P(X = x_i), q_i = P(Y = y_j), r_{ij} = P(X = x_i, Y = y_j)$.

$$H(X) = \sum_i^n -p_i \log p_i \quad (4)$$

$$H(Y) = \sum_j^m -q_j \log q_j \quad (5)$$

$$H(X,Y) = \sum_i^n \sum_j^m -r_{ij} \log r_{ij} \quad (6)$$

If the random variables X, Y are independent each other, the mutual information $I(X,Y)$ equals zero.

In this paper, we utilize the mutual information as an evaluation function.

3. SIMULTANEOUS PERTURBATION

The idea of the simultaneous perturbation was proposed by J.C.Spall as an extension of Kiefer-Wolfowitz stochastic approximation[4, 5]. J.Alespector et al. and G.Cauwenberghs also proposed the same idea[6, 7]. Independently, Y.Maeda introduced the same algorithm as a learning rule of neural networks[8]. Some applications of the simultaneous perturbation method in control problems are also reported[9, 10]. Now, we describe the simultaneous perturbation leaning rule used in this paper. Define a parameter vector $\boldsymbol{w}$ and a sign vector$\boldsymbol{s}$ as follows;

$$\mathbf{w}(t) = (w_1(t), w_2(t), \cdots, w_N(t))^T \quad (7)$$

$$\mathbf{s}(t) = (s_1(t), s_2(t), \cdots, s_N(t))^T \quad (8)$$

Where t denotes iteration, superscript T is transpose of a vector. Components $s_i(t)$of the sign vector $\boldsymbol{s}(t)$ are +1 or -1. The i-th component of the modifying vector for the parameter is defined as follows;

$$\Delta w_i(t) = \frac{J(\mathbf{w}(t) + c\mathbf{s}(t)) - J(\mathbf{w}(t))}{c} s_i(t) \quad (9)$$

Where c is a magnitude of the perturbation. The parameter is updated as follows;

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \alpha \Delta \mathbf{w}(t) \quad (10)$$

Where, α is a positive gain coefficient. Note that only two values of the objective function; $J(\boldsymbol{w}(t))$ and $J(\boldsymbol{w}(t) + c\boldsymbol{s}(t))$ are used to update the parameter. Any other information about the objective function does not includ in the algorithm. It is also known that the method is a stochastic gradient method.

4. SIMULTANEOUS PERTURBATION FOR BSS

4.1. Gradient type of simultaneous perturbation[11]

In this paper, we propose a recursive method to obtain the separating matrix based on the mutual information using the simultaneous perturbation optimization method. We apply the method to a recursive learning of the blind signal separation directly. The algorithm corresponding to an ordinary gradient method is as follows;

$$\mathbf{W}(t+1) = \mathbf{W}(t) - \alpha \Delta \mathbf{W}(t) \quad (11)$$

where, α is a positive gain coefficient to adjust a magnitude of a modifying quantity. The $i - j$ component of the matrix $\boldsymbol{\Delta} W$ is defined as follows;

$$\Delta W_{ij}(t) = \frac{J(\mathbf{W}(t) + c\mathbf{s}(t)) - J(\mathbf{W}(t))}{c s_{ij}(t)} \quad (12)$$

Then, unlike the sign vector described before, $\boldsymbol{s}$ is a sign matrix. $s_{ij}(t)$, which is an element of the sign matrix $\boldsymbol{s}(t)$, is +1 or -1, and has the following properties.

$$\mathbf{s}(t) = (s_{ij}(t)) \quad (13)$$

$$E(s_{ij}(t)) = 0 \quad (14)$$

$$E(s_{ij}(t1) \cdot s_{kl}(t2)) = \delta_{ik} \delta_{jl} \delta_{t1t2} \quad (15)$$

Where, $E(\cdot)$ denotes exception. δ is the Kronecker's delta. That is, $s_{ij}(t)$ has zero mean and is independent with respect to the other $s_{ij}(t)$ and time t. c is a magnitude of the perturbation. $J(\cdot)$ is an evaluation function to be minimized. In this, the evaluation is the mutual information of Eq.(3). That is,

$$J(\mathbf{W}(\mathbf{t})) = I(X, Y | \mathbf{W}(\mathbf{t})) \quad (16)$$

An important point is that this optimization method requires only two values of the mutual information $J(\cdot)$, even if number of the parameters; elements of the separating matrix, is large. Thus, in this method, we need only twice calculations of the evaluation function (the mutual information) to update for all parameters of $\boldsymbol{W}(t)$.

4.2. Natural Gradient type of simultaneous perturbation

It is known that the natural gradient method is more efficient than the ordinary gradient method. From the same point of view, we can propose the natural gradient method using the simultaneous perturbation optimization method.

The form of the natural gradient is shown as follows;

$$\Delta \boldsymbol{W}(t) = \frac{\partial J(\mathbf{W}(t))}{\partial \boldsymbol{W}} \boldsymbol{W}(t)^T \boldsymbol{W}(t) \tag{17}$$

Now, we propose the following algorithm using the simultaneous perturbation method, paying attention to the perturbation added to the separation matrix.

$$\Delta W_{ij}(t) = \frac{J(\mathbf{W}(t) + c\mathbf{s}(t)\boldsymbol{W}^T(t)\boldsymbol{W}(t)) - J(\mathbf{W}(t))}{c s_{ij}(t)} \tag{18}$$

We investigate a quantity of the right hand side of the above equation. For convenience, we consider the following 2×2 matrix of a mixing matrix and a separating matrix and if not necessary, we omit the time index t.

$$\boldsymbol{W} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \tag{19}$$

We expand $J(\boldsymbol{W} + c\mathbf{s}\boldsymbol{W}^T\boldsymbol{W})$ at $\boldsymbol{W}$. Then, there exists $\boldsymbol{W}_s$ such that

$$\begin{aligned} J\left(\mathbf{W} + c\mathbf{s}\mathbf{W}^T\mathbf{W}\right) &= J(\mathbf{W}) + c\left(D_1 s_{11} + D_2 s_{12}\right)\frac{\partial J}{\partial W_{11}} \\ &+ c\left(D_2 s_{11} + D_3 s_{12}\right)\frac{\partial J}{\partial W_{12}} \\ &+ c\left(D_1 s_{21} + D_2 s_{22}\right)\frac{\partial J}{\partial W_{21}} \\ &+ c\left(D_2 s_{21} + D_3 s_{22}\right)\frac{\partial J}{\partial W_{22}} \\ &+ \frac{1}{2}c^2\left\{\left(D_1 s_{11} + D_2 s_{12}\right)^2 \frac{\partial^2 J(\mathbf{W}_S)}{\partial W_{11}^2}\right. \\ &+ \left(D_1 s_{11} + D_2 s_{12}\right)\left(D_2 s_{11} + D_3 s_{12}\right)\frac{\partial^2 J(\mathbf{W}_S)}{\partial W_{11}\partial W_{12}} \\ &+ \left(D_1 s_{11} + D_2 s_{12}\right)\left(D_1 s_{21} + D_2 s_{22}\right)\frac{\partial^2 J(\mathbf{W}_S)}{\partial W_{11}\partial W_{21}} \\ &+ \left(D_1 s_{11} + D_2 s_{12}\right)\left(D_2 s_{21} + D_3 s_{22}\right)\frac{\partial^2 J(\mathbf{W}_S)}{\partial W_{11}\partial W_{22}} \\ &\left. + \cdots \right\} \end{aligned}$$

Where,

$$D_1 = W_{11}^2 + W_{21}^2 \tag{20}$$

$$D_2 = W_{11}W_{12} + W_{21}W_{22} \tag{21}$$

$$D_3 = W_{12}^2 + W_{22}^2 \tag{22}$$

Therefore, we have

$$\begin{aligned} \Delta W_t^{ij} &= s_{ij}\left(D_1 s_{11} + D_2 s_{12}\right)\frac{\partial J}{\partial W_{11}} \\ &+ s_{ij}\left(D_2 s_{11} + D_3 s_{12}\right)\frac{\partial J}{\partial W_{12}} \\ &+ s_{ij}\left(D_1 s_{21} + D_2 s_{22}\right)\frac{\partial J}{\partial W_{21}} \\ &+ s_{ij}\left(D_2 s_{21} + D_3 s_{22}\right)\frac{\partial J}{\partial W_{22}} \\ &+ \frac{1}{2}c s_{ij}\left\{\left(D_1 s_{11} + D_2 s_{12}\right)^2 \frac{\partial^2 J(\mathbf{W}_S)}{\partial W_{11}^2}\right. \\ &+ \left(D_1 s_{11} + D_2 s_{12}\right)\left(D_2 s_{11} + D_3 s_{12}\right)\frac{\partial^2 J(\mathbf{W}_S)}{\partial W_{11}\partial W_{12}} \\ &+ \left(D_1 s_{11} + D_2 s_{12}\right)\left(D_1 s_{21} + D_2 s_{22}\right)\frac{\partial^2 J(\mathbf{W}_S)}{\partial W_{11}\partial W_{21}} \\ &+ \left(D_1 s_{11} + D_2 s_{12}\right)\left(D_2 s_{21} + D_3 s_{22}\right)\frac{\partial^2 J(\mathbf{W}_S)}{\partial W_{11}\partial W_{22}} \\ &\left. + \cdots \right\} \end{aligned} \tag{23}$$

Using the conditions (13), (14) and (15) on the sign matrix,

$$E(\Delta W^{11}) = D_1 \frac{\partial J}{\partial W_{11}} + D_2 \frac{\partial J}{\partial W_{12}} \tag{24}$$

The right hand side of the above equation is equal to the 1-1 element of Eq.(17) for the 2×2 separation matrix. We can obtain the same relation for the other elements.

This means that the algorithm described here is a natural gradient method in the sense of expectation. In other words, this method is a kind of stochastic natural gradient method.

Note that the method uses only two values of the evaluation function, that is, the mutual information. This method does not require probability densities of the original signal as well.

4.3. Hybrid type of simultaneous perturbation

As we mentioned in the subsection of Gradient type of simultaneous perturbation, we can estimate the following quantity using the simultaneous perturbation method.

$$\frac{\partial J(\mathbf{W}(t))}{\partial \mathbf{W}} \tag{25}$$

Then, using $\boldsymbol{W}^T(t)\boldsymbol{W}(t)$, we can estimate the natural gradient, similar to the natural gradient method. Therefore, we can propose the following algorithm;

$$\mathbf{W}(t+1) = \mathbf{W}(t) - \alpha \boldsymbol{\Delta W}(t)\boldsymbol{W}^T(t)\boldsymbol{W}(t) \tag{26}$$

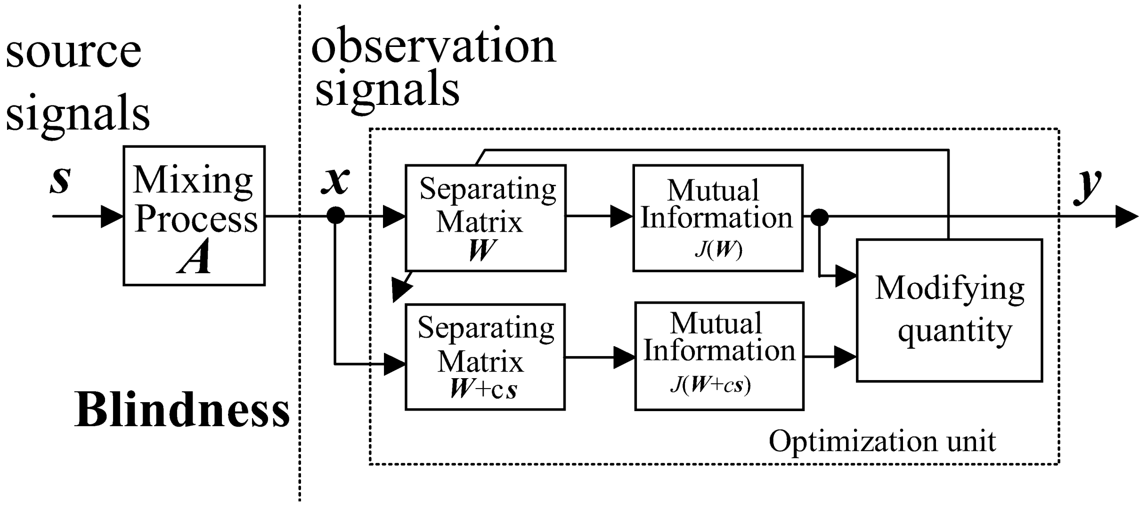

Fig. 1. Flow of signal

where, α is a positive gain coefficient to adjust a magnitude of a modifying quantity. The $i-j$ component of the matrix $\boldsymbol{\Delta} W$ is defined in (12).

Only the gradient of Eq.(25) is estimated by the simultaneous perturbation method in the algorithm.

Total flowchart of the algorithm is shown in Fig.1. First, we set up a mixing matrix $\boldsymbol{A}$ and source signals a. These are blind data. From these two data, observed signals $\boldsymbol{x}$ is calculated.

Next, without perturbation $c\boldsymbol{s}$, we input $\boldsymbol{x}$ into the separating matrix $\boldsymbol{W}$, and obtain the separated signals $\boldsymbol{y}$. Then, we calculate the mutual information $J(\boldsymbol{W})$ by Eq.(3). On the other hand, simultaneously, we add perturbation $c\boldsymbol{s}$ to all parameters of $\boldsymbol{W}$. Then, we have a value of the mutual information $J(\boldsymbol{W}+c\boldsymbol{s})$.

Using two values of the mutual information, we can obtain modifying quantities for all elements of the separating matrix. We update all parameters of the separating matrix and repeat this procedure.

Note that this method can calculate modifying qualities, by using only two values of the mutual information without perturbation $J(\boldsymbol{W})$ and with perturbation $J(\boldsymbol{W}+c\boldsymbol{s})$. This procedure does not include any information of the original signals as mentioned before. Moreover, the procedure itself is very simple and easy. This is suitable for a hardware implementation of the algorithm.

5. EXPERIMENT

In order to confirm a feasibility and usefulness, we handle a basic experiment for speech signals. The experiment is for a simple instantaneous mixture of two signals.

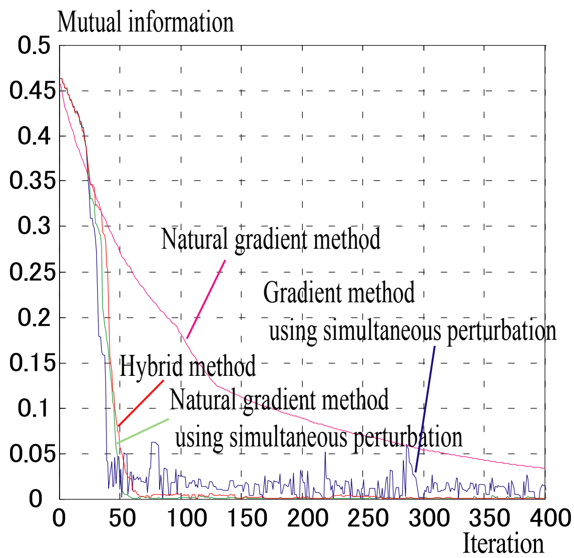

Fig. 2. Change of the mutual information.

Now, two speech signals are mixed using the following mixing matrix. Then, only instantaneous mixture is carried out on computer.

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0.7 \\ 0.8 & 1 \end{pmatrix} \tag{27}$$

Initial separating matrix is an identity matrix. The perturbation c and the gain coefficient α are 5×10^{-2}

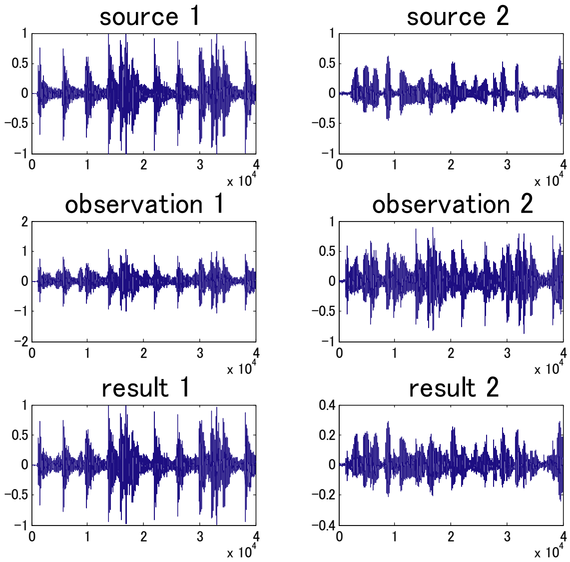

Fig. 3. Actual Data.

and 5×10^{-2}, respectively. These values are determined empirically.

Fig.2 shows a change of the mutual information for three proposed methods. Then, values of the mutual information for three algorithms decrease very quickly. We can confirm a feasibility of these methods.

Fig.3 shows real data; original signals, measured signals and separated signals after 500 iterations. The separated signals are very close to the original ones.

It seems that these methods of simultaneous perturbation have equal capability to the ordinary natural gradient method.

6. CONCLUSION

We propose the natural gradient methods using the simultaneous perturbation. Details of these methods for the blind source separation were described. The probability densities of source signals are not required in the algorithm.

This method is also useful for blind signal deconvolution and learning of neural networks. We are investigating detailed performance of these types of methods for some applications.

7. REFERENCES

[1] S.Amari, J.F.Cardoso : Blind Source Separation, IEEE Trans. on Signal Processing, Vol.45, No.11, pp.2692-2700(1997)

[2] S.Amari : Natural Gradient Works Efficiently in Learning, Neural Computation, Vol.10, pp.251-276(1998)

[3] S.Amari : Superefficiency in Blind Source Separation, IEEE Trans. on Signal Processing, Vol.47, No.4, pp.936-944(1999)

[4] J.C.Spall : A stochastic approximation technique for generating maximum likelihood parameter estimates, Proc. of the 1987 American Control Conference, pp.1161-1167(1987)

[5] J.C.Spall : Multivariate stochastic approximation using a simultaneous perturbation gradient approximation, IEEE trans. on Automatic Control, Vol.37, No.3, pp.332-341(1992)

[6] J.Alespector, R.Meir, B.Yuhas, A.Jayakumar and D.Lippe : A parallel gradient descent method for learning in analog VLSI neural networks, in S.J.Hanson, J.D.Cowan and C.Lee(eds.), Advances in neural information processing systems 5, Morgan Kaufmann Publisher, pp.836-844(1993)

[7] G.Cauwenberghs : A fast stochastic error-descent algorithm for supervised learning and optimization, in S.J.Hanson, J.D.Cowan and C.Lee(eds.), Advances in neural information processing systems 5, Morgan Kaufmann Publisher, pp.244-251(1993)

[8] Y.Maeda, H. Hirano, Y. Kanata : A Learning Rule of Neural Networks via Simultaneous Perturbation and Its Hardware Implementation, Neural Networks, Vol.8, No.2, pp.251-259(1995)

[9] Y.Maeda, R.J.P. de Figueiredo : Learning rules for neuro-controller via simultaneous perturbation, IEEE trans. on Neural Networks, Vol.8, No.5, pp.1119-1130(1997)

[10] Y.Maeda, A.Nakazawa and Y.Kanata : Hardware Implementation of a Pulse Density Neural Network Using Simultaneous Perturbation Learning Rule, Analog Integrated Circuits and Signal Processing, 18, pp.1-10(1999)

[11] T.Akiyama, Y.Maeda : Blind Signal Separation Using Simultaneous Perturbation Optimization Method, Trans. IEE of Japan, Vol.121-C, No.4, pp.782-788(2001)(in Japanese)