Natural Gradient Algorithm for Blind Separation of Overdetermined Mixture with Additive Noise

L.-Q. Zhang, A. Cichocki, *Member, IEEE,* and S. Amari, *Fellow, IEEE*

***Abstract*—In this letter, we study the natural gradient approach to blind separation of overdetermined mixtures. First we introduce a Lie group on the manifold of overdetermined mixtures, and endow a Riemannian metric on the manifold based on the property of the Lie group. Then we derive the natural gradient on the manifold using the isometry of the Riemannian metric. Using the natural gradient, we present a new learning algorithm based on the minimization of mutual information.**

I. Introduction

RECENTLY, blind separation of independent sources has become an increasingly important research area due to its rapidly growing applications in various fields, such as telecommunication systems, image enhancement and biomedical signal processing [1]–[11]. It has been shown that the natural gradient improves dramatically the learning efficiency in blind separation [1]–[8]. For the standard case where the number of sources is equal to the number of sensors, the natural gradient algorithm has been developed by Amari *et al.* [3], and independently as the relative gradient by Cardoso [7]. However, in most practical cases, the number of active source signals is unknown and changing over time. Therefore, in the general case the mixing matrix and demixing matrix are not square and not invertible. The blind separation of more sources than mixtures was discussed in [10] by using overcomplete representations. In this letter, we study blind separation of overdetermined mixtures, where the number of sensors is not less than the number of sources.

The main objective of this letter is to extend the idea of natural gradient to overdetermined mixtures, and apply the natural gradient to derive an efficient learning algorithm. It is a surprise that the optimal natural gradient algorithm, in the sense of minimizing the effect of noises on the output signals, is in the same form as the standard case where the number of sources is equal to the number of sensors. It is plausible to use overdetermined mixtures to improve upon blind source separation algorithms in extracting the signals of interest from mixtures.

II. Blind Separation of Overdetermined Mixture

Assume that the unknown source signals $\mathbf{s}(t) = (s_1(t), \cdots, s_n(t))^T$ are zero-mean processes and mutually statistically independent and $\mathbf{x}(t) = (x_1(t), \cdots, x_m(t))^T$ is an available sensor vector, which is a linear instantaneous mixture of sources by

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t) \tag{1}$$

where $\mathbf{A} \in \mathbf{R}^{m\times n}$ is an unknown mixing matrix of full rank, $\mathbf{n}(t)$ is the vector of additive white Gaussian noises. In this letter we consider the overdetermined case, $m \geq n$. The blind separation problem is to recover original signals $\mathbf{s}(t)$ from observations $\mathbf{x}(t)$ without prior knowledge on the source signals and the mixing matrix $\mathbf{A}$ except for independence of the source signals. The demixing model is a linear transformation in the form

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) \tag{2}$$

where $\mathbf{y}(t) = (y_1(t), \ldots, y_n(t))^T$ is an estimate of source signals $\mathbf{s}(t)$, $\mathbf{W} \in \mathbf{R}^{n\times m}$ is a demixing matrix to be determined. The general solution to the blind separation is to find a matrix $\mathbf{W}$ such that $\mathbf{WA} = \mathbf{\Lambda P}$, where $\mathbf{\Lambda} \in \mathbf{R}^{n\times n}$ is a nonsingular diagonal matrix and $\mathbf{P} \in \mathbf{R}^{n\times n}$ is a permutation.

III. Natural Gradient

In this section, we discuss some geometrical structures, such as the Lie group and the Riemannian metric, on the manifold of demixing matrices defined as $Gl(n,m) = \{\mathbf{W} \in \mathbf{R}^{n\times m} \mid \operatorname{rank}(\mathbf{W}) = \min(n,m)\}$. For $\mathbf{W} \in Gl(n,m)$, there exists an orthogonal matrix $\mathbf{Q} \in \mathbf{R}^{m\times m}$ such that

$$\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2]\mathbf{Q} \tag{3}$$

where $\mathbf{W}_1 \in \mathbf{R}^{n\times n}$ is nonsingular, and $\mathbf{W}_2 \in \mathbf{R}^{n\times(m-n)}$.

A. Lie Group $Gl(n,m)$

The Lie group plays a crucial role in deriving natural gradient of the manifold $Gl(n,n)$. We introduce the Lie group structure on the manifold $Gl(n,m)$. It is easy to verify that $Gl(n,m)$ is a C^∞ manifold of dimension nm. The operations on the manifold $Gl(n,m)$ are defined as follows:

$$\mathbf{X} * \mathbf{Y} = [\mathbf{X}_1\mathbf{Y}_1, \mathbf{X}_1\mathbf{Y}_2 + \mathbf{X}_2]\mathbf{Q}, \tag{4}$$

$$\mathbf{X}^\dagger = \left[\mathbf{X}_1^{-1}, -\mathbf{X}_1^{-1}\mathbf{X}_2\right]\mathbf{Q} \tag{5}$$

where $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]\mathbf{Q}$ and $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2]\mathbf{Q}$ are in $Gl(n,m)$, $*$ is the multiplication operator of two matrices $\mathbf{X}$ and $\mathbf{Y}$, and $\dagger$ is the inverse operator on $Gl(n,m)$. The identity element in $Gl(n,m)$ is defined by $\mathbf{E} = [\mathbf{I}_n, \mathbf{0}]\mathbf{Q}$. It is easy to prove that both the multiplication and the inverse mappings are C^∞ mappings. The inverse operator satisfies $\mathbf{X}*\mathbf{X}^\dagger = \mathbf{X}^\dagger*\mathbf{X} = \mathbf{E}$. Therefore, the manifold $Gl(n,m)$ with the above operations forms a Lie Group.

Manuscript received April 3, 1999. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. G. Ramponi.

The authors are with the Brain-Style Information Systems Group, Riken Brain Science Institute, Saitama 351–0198, Japan.

Publisher Item Identifier S 1070-9908(99)08274-7.

B. Riemannian Metrics

The Lie Group has an important property that it admits an invariant Riemannian metric. Let $T_{\mathbf{W}}$ be the tangent space of $Gl(n,m)$, and $\mathbf{X}$, $\mathbf{Y} \in T_{\mathbf{W}}$ be the tangent vectors. We introduce an inner product on $T_{\mathbf{W}}$ with respect to $\mathbf{W}$ as $\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{W}}$. Since $Gl(n,m)$ is a Lie group, any $\mathbf{Z} \in Gl(n,m)$ defines an onto-mapping: $\mathbf{W} \rightarrow \mathbf{W} * \mathbf{Z}$. The multiplication transformation maps a tangent vector $\mathbf{X}$ at $\mathbf{W}$ to a tangent vector $\mathbf{X} * \mathbf{Z}$ at $\mathbf{W} * \mathbf{Z}$. Therefore we can define a Riemannian metric on $Gl(n,m)$, such that the right multiplication transformation is isometric, that is, it preserves the Riemannian metric on $Gl(n,m)$. Explicitly, we write it as follows:

$$\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{W}} = \langle \mathbf{X} * \mathbf{Z}, \mathbf{Y} * \mathbf{Z} \rangle_{\mathbf{W}*\mathbf{Z}}. \tag{6}$$

If we define the inner product at the identity $\mathbf{E}$ by $\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{E}} = \mathrm{tr}(\mathbf{X}\mathbf{Y}^T)$, then $\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{W}}$ is automatically induced by

$$\langle \mathbf{X}, \mathbf{Y} \rangle_{\mathbf{W}} = \langle \mathbf{X} * \mathbf{W}^{\dagger}, \mathbf{Y} * \mathbf{W}^{\dagger} \rangle_{\mathbf{E}}. \tag{7}$$

C. Natural Gradient

For a cost function $l(\mathbf{W})$ defined on the manifold $Gl(n,m)$, the natural gradient $\tilde{\nabla} l(\mathbf{W})$ is the steepest ascent direction of the cost function $l(\mathbf{W})$ as measured by the Riemannian metric on $Gl(n,m)$, which is the contravariant form of partial derivatives $\nabla l(\mathbf{W}) = (\frac{\partial l(\mathbf{W})}{\partial \mathbf{W}_{ij}})_{n\times m}$. The natural gradient of the function $l(\mathbf{W})$ is defined by [1]

$$\langle \mathbf{X}, \tilde{\nabla} l(\mathbf{W}) \rangle_{\mathbf{W}} = \langle \mathbf{X}, \nabla l(\mathbf{W}) \rangle_{\mathbf{E}} \tag{8}$$

for any $\mathbf{X} \in T_{\mathbf{W}}$. Comparing the both side of (8), we have

$$\tilde{\nabla} l(\mathbf{W}) = \nabla l(\mathbf{W})(\mathbf{W}^T\mathbf{W} + \mathbf{N}_{\mathbf{I}}) \tag{9}$$

where $\mathbf{N}_{\mathbf{I}} = \mathbf{Q}^T \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m-n} \end{bmatrix} \mathbf{Q} \in \mathbf{R}^{m\times m}$ is a block diagonal matrix, $\mathbf{I}_{m-n}$ is an $(m-n)\times(m-n)$ identity matrix. It is worthy noting that the natural gradient on the manifold $Gl(n,m)$ has an additional term compared with the one on the manifold $Gl(n,n)$. In the overdetermined case, the matrix $\mathbf{W}^T\mathbf{W}$ is singular, while $\mathbf{W}^T\mathbf{W} + \mathbf{N}_{\mathbf{I}}$ is a positive definite matrix for any $\mathbf{W} \in Gl(n,m)$. The property ensures that the natural gradient descent algorithm keeps the same kind of equilibria of the learning system as the ordinary gradient descent one.

Remark 1 It is easy to see that the $\mathbf{N}_{\mathbf{I}}$ is a projection matrix. The result indicates that the natural gradient for overdetermined mixtures is not unique, which depends on the orthogonal matrix $\mathbf{Q}$. The redundancy makes it possible to choose an optimal projection for learning algorithms.

IV. Learning Algorithm

Assume that $p(\mathbf{y}, \mathbf{W})$, $p_i(y_i, \mathbf{W})$ are the joint probability density function (pdf) of $\mathbf{y}$ and marginal pdf of y_i, $(i = 1, \cdots, m)$ respectively. Our target is to make the components of $\mathbf{y}$ as mutually independent as possible. To this end, we employ the Kullback–Leibler divergence as a risk function [3]

$$l(\mathbf{W}) = -H(\mathbf{y}, \mathbf{W}) + \sum_{i=1}^{n} H(y_i, \mathbf{W}) \tag{10}$$

where $H(\mathbf{y}, \mathbf{W}) = -\int p(\mathbf{y}, \mathbf{W}) \log p(\mathbf{y}, \mathbf{W})\, d\mathbf{y}$, $H(y_i, \mathbf{W}) = -\int p_i(y_i, \mathbf{W}) \log p_i(y_i, \mathbf{W}) dy_i$. The divergence $l(\mathbf{W})$ measures the mutual independence of the output signals $y_i(k)$. The output signals $\mathbf{y}$ are mutually independent if and only if $l(\mathbf{W}) = 0$. In order to develop an efficient on-line learning algorithm, we simplify (10) into the following cost function: $l(\mathbf{y}, \mathbf{W})$,

$$l(\mathbf{y}, \mathbf{W}) = -\log(|\det(\mathbf{W}\mathbf{E}^T)|) - \sum_{i=1}^{n} \log p_i(y_i(k), \mathbf{W}) \tag{11}$$

where $\mathbf{E}$ is the identity element of the Lie group $Gl(n,m)$, and $\det(\mathbf{W}\mathbf{E}^T)$ is the determinant of matrix $\mathbf{W}\mathbf{E}^T$. In the following discussion, we use the following decomposition:

$$\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2]\mathbf{Q}, \quad \mathbf{x} = \mathbf{Q}^T \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \tag{12}$$

where $\mathbf{W}_1 \in \mathbf{R}^{n\times n}$ and $\mathbf{x}_1 \in \mathbf{R}^n$. The ordinary gradient of $l(\mathbf{y}, \mathbf{W})$ with respect to $\mathbf{W}$ is given by.

$$\frac{dl(\mathbf{y}, \mathbf{W})}{d\mathbf{W}_1} = -\mathbf{W}_1^{-T} + \boldsymbol{\varphi}(\mathbf{y})\mathbf{x}_1^T, \quad \frac{dl(\mathbf{y}, \mathbf{W})}{d\mathbf{W}_2} = \boldsymbol{\varphi}(\mathbf{y})\mathbf{x}_2^T \tag{13}$$

where $\boldsymbol{\varphi}(\mathbf{y})$ is a vector of nonlinear activation functions $\varphi_i(y_i) = -\frac{d \log p_i(y_i)}{dy_i} = -\frac{p_i'(y_i)}{p_i(y_i)}$. Therefore, the natural gradient learning algorithm on $Gl(n,m)$ can be implemented as follows:

$$\Delta\mathbf{W} = \eta((\mathbf{I} - \boldsymbol{\varphi}(\mathbf{y})\mathbf{y}^T)\mathbf{W} - \boldsymbol{\varphi}(\mathbf{y})\mathbf{x}^T\mathbf{N}_{\mathbf{I}}). \tag{14}$$

V. Optimization of Learning Algorithm

The demixing model projects the sensor signals into $\mathbf{R}^n$, and the projection depends on the matrix $\mathbf{N}_{\mathbf{I}}$. In this section we consider the optimization of such projection. Decompose the mixing matrix in the following form:

$$\mathbf{A} = \mathbf{Q}^T \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{0} \end{bmatrix} \tag{15}$$

where matrix $\mathbf{Q}$ is an orthogonal matrix, $\mathbf{A}_1 \in \mathbf{R}^{n\times n}$ is a nonsingular matrix. The mixing model transforms the source signal into a hyperplane $\mathcal{S} = \{\mathbf{x} \mid \mathbf{x} = \mathbf{Q}^T[\begin{smallmatrix}\mathbf{x}_1 \\ \mathbf{0}\end{smallmatrix}], \mathbf{x}_1 \in \mathbf{R}^n\}$. The orthogonal complement of $\mathcal{S}$ is denoted by $\mathcal{S}^{\perp}$. The question here is that which projection matrix $\mathbf{N}_{\mathbf{I}}$ is the best for learning algorithms in the sense of minimizing the influence of noises. To this end, we assume that the noises are Gaussian with the covariance matrix $E(\mathbf{n}\mathbf{n}^T) = \sigma^2\mathbf{I}$, and are independent of sources. We decompose the demixing matrix using the same orthogonal matrix $\mathbf{Q}$ in (15) as $\mathbf{W} = [\mathbf{W}_1, \mathbf{W}_2]\mathbf{Q}$, and we have

$$\mathbf{y} = \mathbf{W}\mathbf{A}\mathbf{s} + \mathbf{W}\mathbf{n} = \mathbf{W}_1\mathbf{A}_1\mathbf{s} + \mathbf{W}\mathbf{n}. \tag{16}$$

This means that $\mathbf{W}_2$ does not contribute to the main term. To minimize the effect of noises on the output $\mathbf{y}$, We introduce the following cost functional

$$F(\mathbf{W}_2) = E(\|\mathbf{W}\mathbf{n}\|_2^2) = E(\mathbf{n}^T\mathbf{W}^T\mathbf{W}\mathbf{n}). \tag{17}$$

On the other hand, we decompose the vector of noises into the following form $\mathbf{n} = \mathbf{Q}^T\mathbf{v} = \mathbf{Q}^T[\mathbf{v}_1, \mathbf{v}_2]$. It is easy to derive

$E(\mathbf{v}\mathbf{v}^T) = \mathbf{Q}E(\mathbf{n}\mathbf{n}^T)\mathbf{Q}^T = \sigma^2\mathbf{I}$. Then the cost functional (17) can be rewritten as

$$F(\mathbf{W}_2) = E\big(\mathbf{v}_1^T\mathbf{W}_1^T\mathbf{W}_1\mathbf{v}_1\big) + \sigma^2\,\mathrm{tr}\big(\mathbf{W}_2^T\mathbf{W}_2\big). \tag{18}$$

The minimal solution of the cost functional $F(\mathbf{W}_2)$ is $\mathbf{W}_2 = \mathbf{0}$. This means that the transform $\mathbf{y} = \mathbf{W}\mathbf{x}$ should be orthogonal to the normal space $\mathcal{S}^{\perp}$, that is, for any $\mathbf{z} \in \mathcal{S}^{\perp}$, $\mathbf{y} = \mathbf{W}\mathbf{z} = \mathbf{0}$. In this case our natural learning algorithm (14) is simplified to the standard form

$$\Delta\mathbf{W} = \eta(\mathbf{I} - \varphi(\mathbf{y})\mathbf{y}^T)\mathbf{W} \tag{19}$$

where $\mathbf{W}$ is, in general, a nonsquare $n \times m$ demixing matrix $(n \leq m)$. It should be noted that the learning algorithm (19) has been proposed for the special case when the number of sensors is exactly equal to the number of sources [3], [8]. It is apparent that the learning algorithm (19) is of the equivariance property in the sense of Lie multiplication. It has been proved the natural gradient improves the learning efficiency in blind separation [1]. Here, we present a rigorous geometric interpretation why the algorithm (19) works efficiently for overdetermined mixtures. Using the theory of information geometry, we can also analyze the effect of noises on the performance of the learning algorithm. Due to the limited space, the problem is left for discussion in future work.

VI. Conclusion

In this letter, we discuss some geometrical structures on the manifold of the nonsquare demixing matrices and derive the natural gradient on the manifold. Using the natural gradient, we present a novel learning algorithm for blind separation of overdetermined mixtures. The learning algorithm works efficiently in blind separation. The detailed derivation of the natural gradient algorithm and computer simulations will be given in future work.

References

[1] S. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, pp. 251–276, 1998.

[2] S. Amari and A. Cichocki, "Adaptive blind signal processing—Neural network approaches," *Proc. IEEE*, vol. 86, pp. 2026–2048, 1998,.

[3] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," in *Advances in Neural Information Processing Systems 8 (NIPS*95)*, G. Tesauro, D. S. Touretzky, and T. K. Leen, Eds. Cambridge, MA: MIT Press, 1996, pp. 757–763.

[4] A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, pp. 1129–1159, 1995.

[5] J. Cardoso and S. Amari, "Maximum likelihood source separation: Equivalance and adaptivity," in *Proc. SYSID'97*, pp. 1063–1968.

[6] J.-F. Cardoso, "Blind signal separation: Statistical principles," in *Proc. IEEE*, vol. 86. pp. 2009–2025, 1998.

[7] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing*, vol. 44, pp. 3017–3030, Dec. 1996.

[8] A. Cichocki and R. Unbehauen, "Robust neural networks with on-line learning for blind identification and blind separation of sources," *IEEE Trans. Circuits Syst.*, vol. 43, pp. 894–906, 1996.

[9] S. Douglas, A. Cichocki, and S. Amari, "Multichannel blind separation and deconvolution of sources with arbitrary distributions," in *Proc. IEEE Workshop on Neural Networks for Signal Processing*, Sept. 1997, pp. 436–445.

[10] T. W. Lee, M. S. Lewicki, and M. Girolami, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Processing Lett.*, vol. 6, pp. 87–90, Apr. 1999.

[11] L. Zhang, S. Amari, and A. Cichocki, "Natural gradient approach to blind separation of over- and undercomplete mixtures," in *Proc. Independent Component Analysis and Signal Separation*, Aussois, France, 1999, pp. 455–460.