Natural Gradient Learning for Over- and Under-Complete Bases in ICA

Shun-ichi Amari
RIKEN Brain Science Institute, Wako-shi, Hirosawa, Saitama 351-01, Japan

Independent component analysis or blind source separation is a new technique of extracting independent signals from mixtures. It is applicable even when the number of independent sources is unknown and is larger or smaller than the number of observed mixture signals. This article extends the natural gradient learning algorithm to be applicable to these overcomplete and undercomplete cases. Here, the observed signals are assumed to be whitened by preprocessing, so that we use the natural Riemannian gradient in Stiefel manifolds.

1 Introduction

Let us consider m independent signals $s_1, \ldots, s_m$ summarized in a vector $\boldsymbol{s} = (s_1, \ldots, s_m)^T$, where T denotes the transposition. The m independent sources generate signals $\boldsymbol{s}(t)$ at discrete times $t = 1, 2, \ldots$. Let us assume that we can observe only their n linear mixtures, $\boldsymbol{x} = (x_1, \ldots, x_n)^T$,

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{1.1}$$

or in the component form,

$$x_i(t) = \sum_{b=1}^{m} A_{ib} s_b(t). \tag{1.2}$$

Given observed signals $\mathbf{x}(1), \ldots, \mathbf{x}(t)$, we would like to recover $\mathbf{s}(1), \ldots, \mathbf{s}(t)$ without knowing the mixing matrix $\mathbf{A}$ and probability distribution of $\mathbf{s}$. When $n = m$, the problem reduces to online estimation of $\mathbf{A}$ or its inverse, $\mathbf{W}$; there exists a lot of work on this subject (Jutten & Hérault, 1991; Bell & Sejnowski, 1995; Comon, 1994; Amari, Chen, & Cichocki, 1997; Cardoso & Laheld, 1996). The search space for $\mathbf{W}$ in this case of $n = m$ is the space of nonsingular matrices. The natural gradient learning algorithm (Amari, Cichocki, & Yang, 1996; Amari, 1998) is the true steepest descent method in the Riemannian parameter space of the nonsingular matrices. It is proved to be Fisher efficient in general, having the equivariant property. Therefore, it is desired to extend it to more general cases of $n \neq m$. This article reports on natural gradient learning in the cases of $n \neq m$.

In many cases, the number m of the sources is unknown. Lewicki and Sejnowski (1998a, 1998b) treated the overcomplete case where $n < m$, and proved that independent component analysis (ICA) provides a powerful new technique in the area of brain imaging and signal processing. In this case, the mixing matrix $\mathbf{A}$ is rectangular and is not invertible. The problem is split into two phases: estimation of $\mathbf{A}$ and estimation of $\mathbf{s}(t)$ based on the estimated $\hat{\mathbf{A}}$.

Let us denote the m columns of $\mathbf{A}$ by n-dimensional vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m$. Then,

$$\mathbf{x} = \sum_{b=1}^{m} s_b \mathbf{a}_b \tag{1.3}$$

is a representation of $\mathbf{x}$ in terms of sources s_b's. This is an overcomplete representation where $\{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ is the overcomplete basis (Chen, Donoho, & Saunders, 1996). This basis elucidates the mixing mechanism so that one may analyze the locations of the independent sources by using the estimated basis vectors. An algorithm for learning this type of basis was proposed by Lewicki and Sejnowski (1998a, 1998b). Another problem is to reconstruct $\mathbf{s}(t)$ by using an estimate $\hat{\mathbf{A}}$. Since $\hat{\mathbf{A}}$ is rectangular, it is not invertible and we do not have $\hat{\mathbf{A}}^{-1}$. One idea is to use the generalized inverse $\hat{\mathbf{A}}^{\dagger}$ and estimate $\mathbf{s}(t)$ by

$$\hat{\mathbf{s}}(t) = \hat{\mathbf{A}}^{\dagger}\mathbf{x}(t). \tag{1.4}$$

This gives the minimum square-norm solution of the ill-posed (underdetermined) equation,

$$\mathbf{x}(t) = \hat{\mathbf{A}}\mathbf{s}(t). \tag{1.5}$$

One interesting idea is to use the least L_1-norm solution corresponding to the Laplace prior on $\mathbf{s}$ (Chen, Donoho, & Saunders, 1996; Lewicki & Sejnowski, 1998a, 1998b). This gives a sparse solution (see also Girosi, 1998). Estimation of $\mathbf{A}$ or basis $\{\mathbf{a}, \ldots, \mathbf{a}_m\}$ is one important problem to understand hidden structures in observations $\mathbf{x}$. Recovery of $\mathbf{s}$ is another important problem, which is carried out based on a good estimate $\hat{\mathbf{A}}$. This article does not treat the latter interesting problem of recovering s, but focuses only on the natural gradient learning algorithm to estimate $\mathbf{A}$.

Another situation is the undercomplete case where $m < n$ and one wants to extract p independent signals from mixtures of an unknown number $m < n$ of original signals. Cichocki, Thawonmas, and Amari (1997) proposed a method of sequential extraction. We give the natural gradient learning algorithm in this case too.

2 Orthogonal Matrices and Stiefel Manifolds

It is a useful technique to whiten $\mathbf{x}$ by preprocessing (Cardoso & Laheld, 1996). We assume that observed vector $\boldsymbol{x}$ has already been whitened by preprocessing so that the covariances of x_i and x_j are 0. This does not imply that x_i and x_j are independent. Principal component analysis can be used for this preprocessing. This gives

$$E\left[\mathbf{x}\mathbf{x}^T\right] = \mathbf{I}_n, \tag{2.1}$$

where $\mathbf{I}_n$ denotes the $n \times n$ unit matrix and E denotes the expectation. Since the scales of the source signals are unidentifiable, we assume that source signals $\mathbf{s}$ are normalized,

$$E\left[\mathbf{s}\mathbf{s}^T\right] = \mathbf{I}_m, \tag{2.2}$$

without loss of generality.

By substituting equation 1.1 in 2.1, we have

$$E\left[\mathbf{A}\mathbf{s}\mathbf{s}^T\mathbf{A}^T\right] = \mathbf{A}\mathbf{I}_m\mathbf{A}^T = \mathbf{A}\mathbf{A}^T = \mathbf{I}_n. \tag{2.3}$$

In the overcomplete case where $n < m$, this implies that n row vectors of $\mathbf{A}$ are mutually orthogonal m-dimensional unit vectors. Let $S_{m,n}$ be the set of all such matrices. This set forms a manifold known as a Stiefel manifold. When $n = m$, such a matrix is an orthogonal matrix, and $S_{m,n}$ reduces to the orthogonal group O_n. The search space of matrices $\mathbf{A}$ in the overcomplete case is, hence, the Stiefel manifold $S_{m,n}$. Algebraically, it is represented by the quotient set

$$S_{m,n} = O_m / O_{m-n}. \tag{2.4}$$

Since O_n is a Lie group, we can introduce the Riemannian metric in it in the same manner as we did in the case of the set $Gl(n)$ of all the nonsingular matrices (Yang & Amari, 1997; Amari, 1998). Since $S_{m,n}$ is the quotient space of two orthogonal groups, the natural Riemannian structure is given to $S_{m,n}$. (See Edelman, Arias, & Smith, 1998, for the explicit form of the metric and mathematical details of derivation.)

In the undercomplete case, prewhitening may eliminate the redundant components from $\mathbf{x}$, so that the observed signals span only m dimensions in the larger n-dimensional space of observed signals $\mathbf{x}$. In this case, $\mathbf{A}$ can be regarded as an orthogonal matrix, mapping m-dimensional $\mathbf{s}$ to a m-dimensional subspace of $\mathbf{x}$. However, it often happens because of noise that $\mathbf{x}$'s span the whole n dimensions, where n is not equal to the number m of the source signals, which we do not know. In such a case, we try to extract p independent signals ($p \leq n$) by

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}, \tag{2.5}$$

where $\mathbf{W}$ is an $p \times n$ matrix. When $\mathbf{W}$ is chosen adequately, $\mathbf{y}$ gives p components of $\mathbf{s}$. The recovered signals by an unmixing matrix $\mathbf{W}$ can be written as

$$\mathbf{y} = \mathbf{WAs}. \tag{2.6}$$

Therefore, p signals among m sources are extracted when $\mathbf{WA}$ is an $p \times m$ matrix whose p rows are different and each has only one nonzero entry with value 1 or -1. This shows that

$$\begin{aligned} \mathbf{I}_p &= E\left[\mathbf{y}\mathbf{y}^T\right] \\ &= \mathbf{W}E\left[\mathbf{x}\mathbf{x}^T\right]\mathbf{W}^T \\ &= \mathbf{W}\mathbf{W}^T. \end{aligned} \tag{2.7}$$

Hence, p rows of $\boldsymbol{W}$ are mutually orthogonal n-dimensional unit vectors. The set of all such matrices $\boldsymbol{W}$ is the Stiefel manifold $S_{n,p}$.

3 Minimizing Cost Function

Let us first consider a candidate $\mathbf{A}$ of the mixing matrix in the overcomplete case and put

$$\mathbf{y} = \mathbf{A}^T\mathbf{x}. \tag{3.1}$$

Since the true $\mathbf{A}$ satisfies equation 2.3, we have

$$\mathbf{x} = \mathbf{Ay}, \tag{3.2}$$

so that $\mathbf{y}$ is an estimate of original $\mathbf{s}$. However, there are infinitely many $\mathbf{y}$ satisfying equation 3.2 and equation 3.1 does not give original $\mathbf{s}$ even when $\boldsymbol{A}$ is the true mixing matrix. We do not touch on the problem of extracting $\mathbf{s}$ by the technique of sparse representation (see Lewicki & Sejnowski, 1998a, 1998b). We focus only on the problem of estimation of $\mathbf{A}$.

Let us consider the probability density function $p(\mathbf{y}, \mathbf{A})$ of $\mathbf{y}$ determined by $\mathbf{A} \in S_{m,n}$. Here, $\mathbf{A}$ is not a random variable but a parameter to specify a distribution of $\mathbf{y}$. The probability density $p(\mathbf{y}, \mathbf{A})$ is degenerate in the sense that nonzero probabilities are concentrated on the n-dimensional subspace determined by $\mathbf{A}$.

Our target is to make the components of $\mathbf{y}$ as independent as possible. To this end, let us choose an adequate independent distribution of $\mathbf{y}$,

$$q(\mathbf{y}) = \prod_{a=1}^{m} q_a(y_a). \tag{3.3}$$

One idea is to define a cost function to be minimized by the Kullback divergence between two distributions $p(\mathbf{y}, \mathbf{A})$ and $q(\mathbf{y})$,

$$\begin{aligned} C(\mathbf{A}) &= KL\left[p(\mathbf{y}, \mathbf{A}) : q(\mathbf{y})\right] \\ &= \int p(\mathbf{y}, \mathbf{A}) \log \frac{p(\mathbf{y}, \mathbf{A})}{q(\mathbf{y})} d\mathbf{y}. \end{aligned} \tag{3.4}$$

This shows how far the current $p(\mathbf{y}, A)$ is from the prescribed independent distribution $q(\mathbf{y})$ and is minimized when $\mathbf{y} = \mathbf{A}^T\mathbf{x}$ are independent under a certain condition (Amari et al., 1997). Note that $p(\mathbf{y}, \mathbf{A})$ is singular, but $C(\mathbf{A})$ has a finite value, whereas $KL[q(\mathbf{y}) : p(\mathbf{y}, \mathbf{A})]$ diverges. The entropy term

$$-H = \int p(\boldsymbol{y}, \boldsymbol{A}) \log p(\boldsymbol{y}, \boldsymbol{A}) d\boldsymbol{y} \tag{3.5}$$

does not depend on $\boldsymbol{A}$ because $\log |\boldsymbol{A}\boldsymbol{A}^T| = \log |\boldsymbol{I}_n| = 0$. Hence, this is equivalent to the following cost function,

$$C(\boldsymbol{A}) = -E\left[\sum_{a=1}^{m} \log q_a(y_a)\right] - c, \tag{3.6}$$

where c is the entropy of $\boldsymbol{y}$. Such a cost function has been derived by various considerations (Amari et al., 1997; Bell & Sejnowski, 1995; and many others). We apply the stochastic gradient descent method to obtain a learning algorithm.

In the underdetermined case, we also use the cost function

$$C(\boldsymbol{W}) = -E\left[\sum \log q_a(y_a)\right], \tag{3.7}$$

where $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$.

4 Gradient and Natural Gradient

The gradient of

$$l(\boldsymbol{y}, \boldsymbol{A}) = -\sum \log q_a(y_a) \tag{4.1}$$

is calculated easily by

$$dl = \varphi(\boldsymbol{y})^T d\boldsymbol{y} = \varphi(\boldsymbol{y})^T d\boldsymbol{A}^T \boldsymbol{x}, \tag{4.2}$$

where $\varphi(\boldsymbol{y})$ is a vector composed of $\varphi_a(y_a)$, $\varphi(\boldsymbol{y}) = [\varphi_1(y_1), \ldots, \varphi_n(y_n)]^T$,

$$\varphi_i(y_i) = -\frac{d}{dy_i} \log q_i(y_i), \tag{4.3}$$

and

$$d\boldsymbol{y} = d\mathbf{A}^T \boldsymbol{x} \tag{4.4}$$

is used. We then have the ordinary gradient

$$\nabla l = \left(\frac{\partial l}{\partial A_{ib}} \right) = \boldsymbol{x} \varphi(\boldsymbol{y})^T = \boldsymbol{A} \boldsymbol{y} \varphi(\boldsymbol{y})^T. \tag{4.5}$$

Since $\boldsymbol{A}$ belongs to the Stiefel manifold, the steepest descent direction of the cost function C is given by the natural gradient $\tilde{\nabla} l$, which takes the Riemannian structure of the parameter space. When we know the explicit form of $p(\boldsymbol{y}, \boldsymbol{A})$, we can use the Fisher information matrix to define a Riemannian metric in this manifold. However, we do not know the probability density functions of the source signals in the case of blind source separation. In such cases, we cannot calculate the Fisher information. However, when the parameter space has a Lie group structure, we can introduce an invariant Riemannian metric, as has been done in the case of $n = m$ (Amari et al., 1996). Note that the Fisher information metric is also Lie group invariant.

In the present case, an invariant metric is derived from the Lie group structure of the two orthogonal groups into account. Edelman et al. (1998) showed an explicit form of the natural gradient in a general Stiefel manifold. In the present case, it is given by

$$\begin{aligned} \tilde{\nabla} l &= \nabla l - \boldsymbol{A} \left(\nabla l \right)^T \boldsymbol{A} \\ &= \boldsymbol{A} \left\{ \boldsymbol{y} \varphi(\boldsymbol{y})^T - \varphi(\boldsymbol{y}) \boldsymbol{y}^T \boldsymbol{A}^T \boldsymbol{A} \right\}. \end{aligned} \tag{4.6}$$

Therefore, the increment $\Delta \boldsymbol{A}_t = \boldsymbol{A}_{t+1} - \boldsymbol{A}_t$ by natural gradient learning is given by

$$\Delta \boldsymbol{A}_t = \eta_t \boldsymbol{A}_t \left\{ \varphi(\boldsymbol{y}_t) \boldsymbol{y}_t^T \boldsymbol{A}_t^T \boldsymbol{A}_t - \boldsymbol{y}_t \varphi(\boldsymbol{y}_t)^T \right\}, \tag{4.7}$$

where η is a learning constant. Since

$$\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}_n \tag{4.8}$$

should hold throughout the learning processes, $\Delta \boldsymbol{A}$ should satisfy

$$\Delta \boldsymbol{A}\boldsymbol{A}^T + \boldsymbol{A}\Delta \boldsymbol{A}^T = 0. \tag{4.9}$$

Equation 4.7 satisfies this constraint.

In the underdetermined case,

$$dl(\boldsymbol{y}) = \varphi(\boldsymbol{y})^T d\boldsymbol{W}\boldsymbol{x}. \tag{4.10}$$

Hence, the gradient is

$$\nabla l = \varphi(\boldsymbol{y})\boldsymbol{x}^T. \tag{4.11}$$

The natural Riemannian gradient in a Stiefel manifold is $\tilde{\nabla} l = \nabla l - \boldsymbol{W}\{\nabla l\}^T \boldsymbol{W}$. We use their result and apply it to our case. Then the natural gradient is given by

$$\tilde{\nabla} l = \varphi(\boldsymbol{y})\boldsymbol{x}^T - \boldsymbol{y}\varphi(\boldsymbol{y})^T \boldsymbol{W}. \tag{4.12}$$

The learning rule is

$$\nabla \boldsymbol{W}_t = -\eta_t \tilde{\nabla} l = -\eta_t \left\{ \varphi(\boldsymbol{y}_t)\boldsymbol{x}_t^T - \boldsymbol{y}_t \varphi(\boldsymbol{y}_t)^T \boldsymbol{W}_t \right\}. \tag{4.13}$$

When $n = m$, $\boldsymbol{A}$ or $\boldsymbol{W}$ is orthogonal, and our result reduces to the known formula (Cardoso & Laheld, 1996) of the natural gradient in the space of orthogonal matrices,

$$\tilde{\nabla} l = \left\{ \varphi(\boldsymbol{y})\boldsymbol{y}^T - \boldsymbol{y}\varphi(\boldsymbol{y})^T \right\} W. \tag{4.14}$$

This is the natural gradient in the prewhitened case where the parameter space is the set of orthogonal matrices.

When $n = m$ and no prewhitening preprocessing takes place, the natural gradient is given by

$$\tilde{\nabla} l = \left(\boldsymbol{I} - \varphi(\boldsymbol{y})\boldsymbol{y}^T \right) \boldsymbol{W} \tag{4.15}$$

(Amari, Cichocki & Yang, 1996; Amari, 1998; Yang & Amari, 1997). When prewhitening takes place, the set of $\boldsymbol{W}$ (or $\boldsymbol{A}$) reduces from the general linear group to the orthogonal group. In the orthogonal group, $\Delta \boldsymbol{X} = \Delta \boldsymbol{W}\boldsymbol{W}^T$ is skew symmetric so that $\tilde{\nabla} l \boldsymbol{W}^T$ is skew symmetric. The natural gradient automatically satisfies this condition. This is the reason that the natural gradient in the Lie group of orthogonal matrices takes the skew-symmetric form of equation 4.14.

We may consider the natural gradient without prewhitening. In this case, a general $\boldsymbol{A}$ can be decomposed into

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V} \tag{4.16}$$

by the singular value decomposition, where $\boldsymbol{\Lambda}$ is a diagonal matrix. We may derive the natural gradient in the general nonprewhitened case by considering this decomposition of matrices.

Acknowledgments

The idea for this article emerged from discussions with Terry Sejnowski when I was at Newton Institute at Cambridge University. I thank Dr. Sejnowski and Newton Institute for giving me this opportunity.

References

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, *10*, 251–276.

Amari, S., Chen, T.-P., & Cichocki, A. (1997). Stability analysis of adaptive blind source separation. *Neural Networks*, *10*, 1345–1351.

Amari, S., Cichocki, A., & Yang, H. (1996). A new learning algorithm for blind signal separation. In D. S. Touretzky, C. M. Mozer, & M. E. Hasselmo (Eds.), Advances in neural information processing systems, 8 (pp. 757–763). Cambridge, MA: MIT Press.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*, 1129–1159.

Cardoso, J. F., & Laheld, B. (1996). Equivariant adaptive source separation. *IEEE Trans. on Signal Processing*, *44*, 3017–3030.

Chen, S., Donoho, D. L., & Saunders, M. A. (1996). *Atomic decomposition by basis pursuit* (Tech. Rep.). Stanford: Stanford University.

Cichocki, A., Thawonmas, R., & Amari, S. (1997). Sequential blind signal extraction in order specified by stochastic properties. *Electronics Letters*, *33*, 64–65.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, *36*, 287–314.

Edelman, A., Arias, T., & Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal of Matrix Analysis and Applications*, *20*, 303–353.

Girosi, F. (1998). An equivalence between sparse approximation and support vector machines. *Neural Computation*, *10*, 1455–1480.

Jutten, C., & Herault, J. (1991). Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, *24*, 1–20.

Lewicki, M. S., & Sejnowski, T. (1998a). Learning nonlinear overcomplete representations for efficient coding. In M. Kearns, M. Jordan, & S. Solla (Eds.), *Advances in neural information processing systems*, *10* (pp. 556–562). Cambridge, MA: MIT Press.

Lewicki, M. S., & Sejnowski, T. (1998b). Learning overcomplete representations. Unpublished manuscript, Salk Institute.

Yang, H. H., & Amari, S. (1997). Adaptive online learning algorithms for blind separation: Maximum entropy and minimal mutual information. *Neural Computation*, *9*, 1457–1482.

Received January 28, 1998; accepted December 7, 1998.