Multiplicative Updating Rule for Blind Separation Derived from the Method of Scoring

Howard Hua Yang
Department of Computer Science
Oregon Graduate Institute
PO Box 91000, Portland, OR 97291, USA
hyang@cse.ogi.edu

Abstract

For blind source separation, when the Fisher information matrix is used as the Riemannian metric tensor for the parameter space, the steepest descent algorithm to maximize the likelihood function in this Riemannian parameter space becomes the serial updating rule with equivariant property. This algorithm can be further simplified by using the asymptotic form of the Fisher information matrix around the equilibrium.

1 Introduction

The relative gradient was introduced by (Cardoso and Laheld, 1996) to design multiplicative updating algorithms with equivariant property for blind separation problems. The idea is to calculate differentials by using a relative increment instead of an absolute increment in the parameter space. This idea has been extended to compute the relative Hessian by (Pham, 1996).

For a matrix function $f = f(\boldsymbol{W})$, the relative gradient is defined by

$$\widehat{\nabla} f = \frac{\partial f}{\partial \boldsymbol{W}} \boldsymbol{W}^T. \tag{1}$$

From the differential of $f(\boldsymbol{W})$ based on the relative gradient, the following learning rule is given by (Cardoso and Laheld, 1996) to maximize the function f:

$$\frac{d\boldsymbol{W}}{dt} = \eta \widehat{\nabla} f \boldsymbol{W} = \eta \frac{\partial f}{\partial \boldsymbol{W}} \boldsymbol{W}^T \boldsymbol{W} \tag{2}$$

Also motivated by designing blind separation algorithms with equivariant property,

the natural gradient defined by

$$\widetilde{\nabla} f = \frac{\partial f}{\partial W} W^T W \tag{3}$$

was introduced in (Amari et al, 1996) which yields the same learning rule (2). The geometrical meaning of the natural gradient is given by (Amari, 1996). More details about the natural gradient can be found in (Yang and Amari, 1997) and (Amari, 1997).

The framework of the natural gradient learning was proposed by (Amari, 1997) . In this framework, the ordinary gradient descent learning algorithm in the Euclidean space is not optimal in minimizing a function defined in a Riemannian space. The ordinary gradient should be replaced by the natural gradient which is defined by operating the inverse of the metric tensor in the Riemannian space on the ordinary gradient. Let $\boldsymbol{w}$ denote a parameter vector. It is proved by (Amari, 1997) that if $C(\boldsymbol{w})$ is a loss function defined on a Riemannian space $\{\boldsymbol{w}\}$ with a metric tensor $\boldsymbol{G}$, the negative natural gradient of $C(\boldsymbol{w})$, namely, $-\boldsymbol{G}^{-1}\frac{\partial C}{\partial \boldsymbol{w}}$ is the steepest descent direction to decrease this function in the Riemannian space. Therefore, the steepest descent algorithm in this Riemannian space has the following form:

$$\frac{d\boldsymbol{w}}{dt} = -\eta \boldsymbol{G}^{-1}\frac{\partial C}{\partial \boldsymbol{w}}.$$

If the Fisher information matrix is used as the metric tensor for the Riemannian space and $C(\boldsymbol{w})$ is replaced by the negative log-likelihood function, the above learning rule becomes the method of scoring (Kay, 1993) which is the focus of this paper.

Both the relative gradient $\widehat{\nabla}$ and the natural gradient $\widetilde{\nabla}$ were proposed in order to design the multiplicative updating algorithms with the equivariant property. The former is due to a multiplicative increment in calculating differential while the latter is due to an increment based on a nonholonomic basis (Amari, 1997). Neither $\widehat{\nabla}$ nor $\widetilde{\nabla}$ depends on the data model. The Fisher information matrix is a special and important choice for the Riemannian metric tensor for statistical estimation problems. It depends on the data model. Operating the inverse of the Fisher information matrix on the ordinary gradient, we have another gradient operator. It is called a natural gradient induced by the Fisher information matrix.

In this paper, we show how to derive a multiplicative updating algorithm from the method of scoring. This approach is different from those based on the relative gradient and the natural gradient defined by (3).

2 Fisher Information Matrix For Blind Separation

Consider a linear mixing system:

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$$

where $\boldsymbol{A} \in \Re^{n\times n}$, $\boldsymbol{x} = (x_1, \cdots, x_n)^T$ and $\boldsymbol{s} = (s_1, \cdots, s_n)^T$. Assume that sources are independent with a factorized joint pdf:

$$r(\boldsymbol{s}) = \prod_{i=1}^{n} r(s_i).$$

The likelihood function is

$$p(\boldsymbol{x}; \boldsymbol{A}) = \frac{r(\boldsymbol{A}^{-1}\boldsymbol{x})}{|\boldsymbol{A}|}$$

where $|\boldsymbol{A}| = |det(\boldsymbol{A})|$. Let $\boldsymbol{W} = \boldsymbol{A}^{-1}$ and $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ (a demixing system), then we have the log-likelihood function

$$L(\boldsymbol{W}) = \sum_{i=1}^{n} \log r_i(y_i) + \log |\boldsymbol{W}|.$$

It is easy to obtain

$$\frac{\partial L}{\partial w_{ij}} = \frac{r_i'(y_i)}{r_i(y_i)} x_j + \boldsymbol{W}_{ij}^{-T} \tag{4}$$

where $\boldsymbol{W}_{ij}^{-T}$ is the (i,j) entry in $\boldsymbol{W}^{-T} = (\boldsymbol{W}^{-1})^T$. Writing (4) in a matrix form, we have

$$\frac{\partial L}{\partial \boldsymbol{W}} = \boldsymbol{W}^{-T} - \Phi(\boldsymbol{y})\boldsymbol{x}^T = (\boldsymbol{I} - \Phi(\boldsymbol{y})\boldsymbol{y}^T)\boldsymbol{W}^{-T} = \boldsymbol{F}(\boldsymbol{y})\boldsymbol{W}^{-T} \tag{5}$$

where $\Phi(\boldsymbol{y}) = (\phi_1(y_1), \cdots, \phi_n(y_n))^T$, $\phi_i(y_i) = -\frac{r_i'(y_i)}{r_i(y_i)}$ and $\boldsymbol{F}(\boldsymbol{y}) = \boldsymbol{I} - \Phi(\boldsymbol{y})\boldsymbol{y}^T$.

The maximum likelihood algorithm based on the ordinary gradient $\frac{\partial L}{\partial \boldsymbol{W}}$ is

$$\frac{d\boldsymbol{W}}{dt} = \eta(\boldsymbol{I} - \Phi(\boldsymbol{y})\boldsymbol{y}^T)\boldsymbol{W}^{-T} = \eta \boldsymbol{F}(\boldsymbol{y})\boldsymbol{W}^{-T}$$

which has the high computational complexity due to the matrix inverse $\boldsymbol{W}^{-1}$. The maximum likelihood algorithm based on the natural gradient of matrix functions is

$$\frac{d\boldsymbol{W}}{dt} = \eta \widetilde{\nabla} L = \eta(\boldsymbol{I} - \Phi(\boldsymbol{y})\boldsymbol{y}^T)\boldsymbol{W}. \tag{6}$$

The same algorithm is obtained from $\frac{d\boldsymbol{W}}{dt} = \eta \widehat{\nabla} L \boldsymbol{W}$ by using the relative gradient. An apparent reason for using this algorithm is to avoid the matrix inverse $\boldsymbol{W}^{-1}$. Another good reason for using it is due to the fact that the matrix $\boldsymbol{W}$ driven by (6) never becomes singular if the initial matrix $\boldsymbol{W}$ is not singular. This is proved by (Yang and Amari, 1997). In fact, this property holds for any learning rule of the following type:

$$\frac{d\boldsymbol{W}}{dt} = \boldsymbol{H}(\boldsymbol{y})\boldsymbol{W}. \tag{7}$$

Let $< \boldsymbol{U}, \boldsymbol{V} >= \mathrm{Tr}(\boldsymbol{U}^T\boldsymbol{V})$ denote the inner product of $\boldsymbol{U}$ and $\boldsymbol{V} \in \Re^{n\times n}$. When $\boldsymbol{W}(t)$ is driven by the equation (7), we have
$\frac{d|\boldsymbol{W}|}{dt} =< \frac{\partial |\boldsymbol{W}|}{\partial \boldsymbol{W}}, \frac{d\boldsymbol{W}}{dt} >=< |\boldsymbol{W}|(\boldsymbol{W}^{-1})^T, \frac{d\boldsymbol{W}}{dt} >$
$= \mathrm{Tr}(|\boldsymbol{W}|\boldsymbol{W}^{-1}\boldsymbol{H}(\boldsymbol{y})\boldsymbol{W}) = \mathrm{Tr}(\boldsymbol{H}(\boldsymbol{y}))|\boldsymbol{W}|.$

Therefore,

$$|\boldsymbol{W}(t)| = |\boldsymbol{W}(0)| \exp\{\int_0^t \mathrm{Tr}(\boldsymbol{H}(\boldsymbol{y}(\tau)))d\tau\} \tag{8}$$

which is non-singular when the initial matrix $\boldsymbol{W}(0)$ is non-singular.

The matrix function $\boldsymbol{F}(\boldsymbol{y})$ is also called an estimating function. At the equilibrium of the system (6), it satisfies the zero condition $E[\boldsymbol{F}(\boldsymbol{y})] = \boldsymbol{0}$, i.e.,

$$E[\phi_i(y_i)y_j] = \delta_{ij} \tag{9}$$

where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise.

To calculate the Fisher information matrix, we need a vector form of the equation (5). Let Vec$(\cdot)$ denote an operator on a matrix which cascades the columns of the

matrix from the left to the right and forms a column vector. This operator has the following property:

$$\mathrm{Vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^T \otimes \boldsymbol{A})\mathrm{Vec}(\boldsymbol{B}) \tag{10}$$

where $\otimes$ denotes the Kronecker product. Applying this property, we first rewrite (5) as

$$\frac{\partial L}{\partial \mathrm{Vec}(\boldsymbol{W})} = \mathrm{Vec}(\frac{\partial L}{\partial \boldsymbol{W}}) = (\boldsymbol{W}^{-1} \otimes \boldsymbol{I})\mathrm{Vec}(\boldsymbol{F}(\boldsymbol{y})), \tag{11}$$

and then obtain the Fisher information matrix

$$\begin{aligned} \boldsymbol{G} &= E[\frac{\partial L}{\partial \mathrm{Vec}(\boldsymbol{W})}(\frac{\partial L}{\partial \mathrm{Vec}(\boldsymbol{W})})^T] \\ &= (\boldsymbol{W}^{-1} \otimes \boldsymbol{I})E[\mathrm{Vec}(\boldsymbol{F}(\boldsymbol{y}))\mathrm{Vec}^T(\boldsymbol{F}(\boldsymbol{y}))](\boldsymbol{W}^{-T} \otimes \boldsymbol{I}). \end{aligned} \tag{12}$$

The inverse of $\boldsymbol{G}$ is

$$\boldsymbol{G}^{-1} = (\boldsymbol{W}^T \otimes \boldsymbol{I})\boldsymbol{D}^{-1}(\boldsymbol{W} \otimes \boldsymbol{I}) \tag{13}$$

where $\boldsymbol{D} = E[\mathrm{Vec}(\boldsymbol{F}(\boldsymbol{y}))\mathrm{Vec}^T(\boldsymbol{F}(\boldsymbol{y}))]$.

3 Natural Gradient Induced By Fisher Information Matrix

Define a Riemannian space

$$\mathcal{V} = \{\mathrm{Vec}(\boldsymbol{W});\ \ \boldsymbol{W} \in Gl(n)\}$$

in which the Fisher information matrix $\boldsymbol{G}$ is used as its metric. Here, $Gl(n)$ is the space of all the $n \times n$ invertible matrices.

Let $C(\boldsymbol{W})$ be a matrix function to be minimized. It is shown by (Amari, 1997) that the steepest descent direction in the Riemannian space $\mathcal{V}$ is $-\boldsymbol{G}^{-1}\frac{\partial C}{\partial \mathrm{Vec}(\boldsymbol{W})}$.

Let us define the natural gradient in $\mathcal{V}$ by

$$\overline{\nabla} C(\boldsymbol{W}) = (\boldsymbol{W}^T \otimes \boldsymbol{I})\boldsymbol{D}^{-1}(\boldsymbol{W} \otimes \boldsymbol{I})\frac{\partial C}{\partial \mathrm{Vec}(\boldsymbol{W})} \tag{14}$$

which is called the natural gradient induced by the Fisher information matrix. The time complexity of computing the natural gradient in the space $\mathcal{V}$ is high since inverting the matrix $\boldsymbol{D}$ of $n^2 \times n^2$ is needed.

Using the natural gradient in $\mathcal{V}$ to maximize the likelihood function $L(\boldsymbol{W})$ or the method of scoring, from (11) and (14) we have the following learning rule

$$\mathrm{Vec}(\frac{d\boldsymbol{W}}{dt}) = \eta(\boldsymbol{W}^T \otimes \boldsymbol{I})\boldsymbol{D}^{-1}\mathrm{Vec}(\boldsymbol{F}(\boldsymbol{y})) \tag{15}$$

We shall prove that the above learning rule has the equivariant property.

Denote Vec^{-1} the inverse of the operator Vec. Let matrices $\boldsymbol{B}$ and $\boldsymbol{A}$ be of $n^2 \times n^2$ and $n \times n$, respectively. Denote $\boldsymbol{B}(i,\cdot)$ the i-th row of $\boldsymbol{B}$ and $\boldsymbol{B}_i = \mathrm{Vec}^{-1}(\boldsymbol{B}(i,\cdot))$, $i = 1, \cdots, n^2$. Define an operator $\boldsymbol{B}\star$ as a mapping from $\Re^{n\times n}$ to $\Re^{n\times n}$:

$$\boldsymbol{B} \star \boldsymbol{A} = \begin{bmatrix} < \boldsymbol{B}_1, \boldsymbol{A} > & \cdots & < \boldsymbol{B}_{n^2-n+1}, \boldsymbol{A} > \\ \cdots & \cdots & \cdots \\ < \boldsymbol{B}_n, \boldsymbol{A} > & \cdots & < \boldsymbol{B}_{n^2}, \boldsymbol{A} > \end{bmatrix}$$

where $< \cdot, \cdot >$ is the inner product in $\Re^{n\times n}$. With the operation $\star$, we have

$$\boldsymbol{B}\mathrm{Vec}(\boldsymbol{A}) = \begin{bmatrix} < \boldsymbol{B}_1, \boldsymbol{A} > \\ \vdots \\ < \boldsymbol{B}_{n^2}, \boldsymbol{A} > \end{bmatrix} = \mathrm{Vec}(\mathrm{Vec}^{-1}(\begin{bmatrix} < \boldsymbol{B}_1, \boldsymbol{A} > \\ \vdots \\ < \boldsymbol{B}_{n^2}, \boldsymbol{A} > \end{bmatrix})) = \mathrm{Vec}(\boldsymbol{B} \star \boldsymbol{A}),$$

i.e.,

$$\boldsymbol{B}\text{Vec}(\boldsymbol{A}) = \text{Vec}(\boldsymbol{B} \star \boldsymbol{A}).$$

Applying the above relation, we first rewrite the equation (15) as

$$\text{Vec}(\frac{d\boldsymbol{W}}{dt}) = \eta(\boldsymbol{W}^T \otimes \boldsymbol{I})\text{Vec}(\boldsymbol{D}^{-1} \star \boldsymbol{F}(\boldsymbol{y})),$$

then applying (10) to the above equation we obtain

$$\frac{d\boldsymbol{W}}{dt} = \eta(\boldsymbol{D}^{-1} \star \boldsymbol{F}(\boldsymbol{y}))\boldsymbol{W}. \tag{16}$$

Theorem 1 *For the blind separation problem, the maximum likelihood algorithm based on the natural gradient induced by the Fisher information matrix or the method of scoring has the form (16) which is a multiplicative updating rule with the equivariant property.*

To implement the algorithm (16), we estimate $\boldsymbol{D}$ by sample average. Let $f_{ij}(\boldsymbol{y})$ be the (i,j) entry in $\boldsymbol{F}(\boldsymbol{y})$. A general form for the entries in $\boldsymbol{D}$ is

$$d_{ij,kl} = E[f_{ij}(\boldsymbol{y})f_{kl}(\boldsymbol{y})]$$

which depends on the source pdfs $r_i(s_i)$. When the source pdfs are unknown, in practice we choose $r_i(s_i)$ as our prior assumptions about the source pdfs. To simplify the algorithm (16), we replace $\boldsymbol{D}$ by its asymptotic form at the solution points $\boldsymbol{a} = (c_1 s_{\sigma(1)}, \cdots, c_n s_{\sigma(n)})^T$ where $(\sigma(1), \cdots, \sigma(n))$ is a permutation of $(1, \cdots, n)$.

Regarding the structure of the asymptotic $\boldsymbol{D}$, we have the following theorem:

Theorem 2 *Assume that the pdfs of the sources s_i are even functions.*

Then at the solution point $\boldsymbol{a} = (c_1 s_{\sigma(1)}, \cdots, c_n s_{\sigma(n)})^T$, $\boldsymbol{D}$ is a diagonal matrix and its n^2 diagonal entries have two forms, namely,

$$E[f_{ij}(\boldsymbol{a})f_{ij}(\boldsymbol{a})] = \mu_i\lambda_j, \quad \textit{for } i \neq j \textit{ and}$$
$$E[(f_{ii}(\boldsymbol{a}))^2] = \nu_i$$

where $\mu_i = E[\phi_i^2(a_i)]$, $\lambda_i = E[a_i^2]$ and $\nu_i = E[\phi_i^2(a_i)a_i^2] - 1$. More concisely, we have

$$\boldsymbol{D} = \textit{diag}(\textit{Vec}(\boldsymbol{H})) \tag{17}$$

where

$$\boldsymbol{H} = (\mu_i\lambda_j)_{n\times n} - \textit{diag}(\mu_1\lambda_1, \cdots, \mu_n\lambda_n) + \textit{diag}(\nu_1, \cdots, \nu_n)$$

The proof of Theorem 2 is given in Appendix 1.

Let $\boldsymbol{H} = (h_{ij})_{n\times n}$. Since all μ_i, λ_i, and ν_i are positive, and so are all h_{ij}. We define

$$\frac{1}{\boldsymbol{H}} = (\frac{1}{h_{ij}})_{n\times n}.$$

Then from (17), we have

$$\boldsymbol{D}^{-1} = \text{diag}(\text{Vec}(\frac{1}{\boldsymbol{H}})).$$

The results in Theorem 2 enable us to simplify the algorithm (16) to obtain a low complexity learning rule. Since $\boldsymbol{D}^{-1}$ is a diagonal matrix, for any $n \times n$ matrix $\boldsymbol{A}$ we have

$$\boldsymbol{D}^{-1}\text{Vec}(\boldsymbol{A}) = \text{Vec}(\frac{1}{\boldsymbol{H}} \odot \boldsymbol{A}) \tag{18}$$

where $\odot$ denotes the componentwise multiplication of two matrices of the same dimension. Applying (18) to the learning rule (15), we obtain the following learning rule

$$\text{Vec}(\frac{d\boldsymbol{W}}{dt}) = \eta(\boldsymbol{W}^T \otimes \boldsymbol{I})\text{Vec}(\frac{1}{\boldsymbol{H}} \odot \boldsymbol{F}(\boldsymbol{y})).$$

Again, applying (10) to the above equation we have the following learning rule

$$\frac{d\boldsymbol{W}}{dt} = \eta(\frac{1}{\boldsymbol{H}} \odot \boldsymbol{F}(\boldsymbol{y}))\boldsymbol{W}. \tag{19}$$

Like the learning rule (16), the algorithm (19) is also multiplicative; but unlike (16), there is no need to inverse the $n^2 \times n^2$ matrix in (19). The computation of $\frac{1}{\boldsymbol{H}}$ is straightforward by computing the reciprocals of the entries in $\boldsymbol{H}$.

$(\mu_i, \lambda_i, \nu_i)$ are $3n$ unknowns in $\boldsymbol{G}$. Let us impose the following constraint

$$\nu_i = \mu_i \lambda_i. \tag{20}$$

Under this constraint, the number of unknowns in $\boldsymbol{G}$ is $2n$, and $\boldsymbol{D}$ can be written as

$$\boldsymbol{D} = \boldsymbol{D}_\lambda \otimes \boldsymbol{D}_\mu \tag{21}$$

where $\boldsymbol{D}_\lambda = \text{diag}(\lambda_1, \cdots, \lambda_n)$ and $\boldsymbol{D}_\mu = \text{diag}(\mu_1, \cdots, \mu_n)$.

From (14), using (21) we have the natural gradient descent rule in the Riemannian space $\mathcal{V}$

$$\frac{d\text{Vec}(\boldsymbol{W})}{dt} = -\eta(\boldsymbol{W}^T \boldsymbol{D}_\lambda^{-1} \boldsymbol{W} \otimes \boldsymbol{D}_\mu^{-1}) \frac{\partial C}{\partial \text{Vec}(\boldsymbol{W})}. \tag{22}$$

Applying the property (10), we rewrite the above equation in a matrix form

$$\frac{d\boldsymbol{W}}{dt} = -\eta \boldsymbol{D}_\mu^{-1} \frac{\partial C}{\partial \boldsymbol{W}} \boldsymbol{W}^T \boldsymbol{D}_\lambda^{-1} \boldsymbol{W}. \tag{23}$$

Since μ_i and λ_i are unknown, $\boldsymbol{D}_\mu$ and $\boldsymbol{D}_\lambda$ are replaced by the identity matrix in practice. Therefore, the algorithm (2) is an approximation of the algorithm (23).

Taking $C = -L(\boldsymbol{W})$ as the negative likelihood function and applying the expression (5), we have the following maximum likelihood algorithm based on the natural gradient in $\mathcal{V}$:

$$\frac{d\boldsymbol{W}}{dt} = \eta \boldsymbol{D}_\mu^{-1} (\boldsymbol{I} - \Phi(\boldsymbol{y})\boldsymbol{y}^T) \boldsymbol{D}_\lambda^{-1} \boldsymbol{W}. \tag{24}$$

Again, replacing $\boldsymbol{D}_\mu$ and $\boldsymbol{D}_\lambda$ by the identity matrix we obtain the maximum likelihood algorithm (6) based on the relative gradient or natural gradient of matrix functions.

In the context of the blind separation, the source pdfs are unknown. The prior assumption $r_i(s_i)$ used to define the functions $\phi_i(y_i)$ may not match the true pdfs of the sources. However, the algorithm (24) is generally robust to the mismatch between the true pdfs and the pdfs employed by the algorithm if the mismatch is not too large. See (Cardoso, 1997) and (Pham, 1996) for example.

4 Conclusion

In the context of blind separation, when the Fisher information matrix is used as the Riemannian metric tensor for the parameter space, maximizing the likelihood function in this Riemannian space based on the steepest descent method is the method of scoring. This method yields a multiplicative updating rule with the equivariant property. It is further simplified by using the asymptotic form of the Fisher information matrix around the equilibrium.

5 Appendix

Appendix 1 *Proof of Theorem 2:*

By definition $f_{ij}(\boldsymbol{y}) = \delta_{ij} - \phi_i(y_i)y_j$. At the equilibrium $\boldsymbol{a} = (c_1 s_{\sigma(1)}, \cdots, c_n s_{\sigma(n)})^T$, we have $E[\phi_i(a_i)a_j] = 0$ for $i \neq j$ and $E[\phi_i(a_i)a_i] = 1$. So $E[f_{ij}(\boldsymbol{a})] = 0$. Since the source pdfs are even functions, we have $E[a_i] = 0$ and $E[\phi_i(a_i)] = 0$. Applying these equalities , it is not difficult to verify that

$$E[f_{ij}(\boldsymbol{a})f_{kl}(\boldsymbol{a})] = 0, \quad \text{for } (i,j) \neq (k,l). \tag{25}$$

So, $\boldsymbol{D}$ is a diagonal matrix and

$$E[f_{ii}(\boldsymbol{a})f_{ii}(\boldsymbol{a})] = E[(1 - \phi_i(a_i)a_i)^2] = E[\phi_i^2(a_i)a_i^2] - 1,$$

$$E[f_{ij}(\boldsymbol{a})f_{ij}(\boldsymbol{a})] = E[\phi_i^2(a_i)a_j^2] = \mu_i\lambda_j$$

for $i \neq j$.
Q.E.D.

References

[1] S. Amari. Natural gradient works efficiently in learning. *Accepted by Neural Computation*, 1997.

[2] S. Amari. Neural learning in structured parameter spaces – natural Riemannian gradient. In *Advances in Neural Information Processing Systems, 9, ed. M. C. Mozer, M. I. Jordan and T. Petsche, The MIT Press: Cambridge, MA.*, pages 127–133, 1997.

[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, eds. David S. Touretzky, Michael C. Mozer and Michael E. Hasselmo, MIT Press: Cambridge, MA.*, pages 757–763, 1996.

[4] J.-F. Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, April 1997.

[5] J.-F. Cardoso and B. Laheld. Equivariant adaptive source separation. *IEEE Trans. on Signal Processing*, 44(12):3017–3030, December 1996.

[6] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory.* PTR Prentice Hall, Englewood Cliffs, 1993.

[7] D. T. Pham. Blind separation of instantaneous mixture of sources via an ica. *IEEE Trans. on Signal Processing*, 44(11):2768–2779, November 1996.

[8] H. H. Yang and S. Amari. Adaptive on-line learning algorithms for blind separation: Maximum entropy and minimum mutual information. *Neural Computation*, 9(7):1457–1482, 1997.