Blind Signal Separation: Mathematical Foundations of ICA, Sparse Component Analysis and Other Techniques

Shun-ichi Amari

RIKEN Brain Science Institute, Hirosawa 2-1, Wako-shi, Saitama 351-0198, Japan

ABSTRACT

The present paper shows mathematical foundations of ICA (independent component analysis) and related subjects of signal representations. Information geometry plays a basic role for elucidating the structure of the problem underlying signal representation and decomposition. The method of estimating function is used for the analysis of errors and stability for various ICA algorithms. The nonholonomic method is of particularly interest.

1. INTRODUCTION

There are abundant of signals which we should analyze in the real world. Observed signals are in many cases mixtures of various components, and we need to decompose these hidden components. Various techniques have been so far been proposed. The Fourier analysis and wavelets analyses are classical analytical techniques, and PCA is also a classical statistical technique. Independent component analysis (ICA) is a relatively new technique which has become popular for these ten years, and its applications are expanding.[1,2] The idea of ICA opened a way to fortify methods of signal processing, and new techniques are emerging in this field inspired by ICA, such as sparse component analysis,[3] non-negative matrix factorization[4] and others. It is important to understand mathematical structures of these techniques. The present paper intends to summarize their mathematical foundations, and to overview these new emerging techniques from the mathematical point of view, based mostly on information geometrical ideas[5] of the present author.

2. BASES AND DECOMPOSITION OF SIGNALS

Let us consider vector signals $\boldsymbol{x} = (x_1, \cdots, x_n)^T \in \boldsymbol{R}^n$ of n dimensions. Signals can be functions $x(t)$ of time, $x(u, v)$ of space, or $y = x(\boldsymbol{\xi})$ of input vectors. These functions are infinite-dimensional signals, but similar treatments are possible. We show various types of representations of signals.

1. Fixed bases
   Signals $\boldsymbol{x}$ are decomposed by using a fixed basis $\{\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m\}$, $\boldsymbol{x} = \sum s_i \boldsymbol{a}_i$.
   The Fourier basis, wavelet basis, and spline basis are well known examples.

2. Variable bases under stochastic criteria
   Given a set of observed signals $\boldsymbol{x}$'s, we search for the basis $\{\boldsymbol{a}_i\}$,

$$\boldsymbol{x} = \sum s_i \boldsymbol{a}_i, \tag{1}$$

   such that the decomposed signal $\boldsymbol{s} = (s_1, \cdots, s_m)^T$ has a specific property. Eq.(1) is represented by the matrix-vector notation as

$$\boldsymbol{x} = A\boldsymbol{s}, \tag{2}$$

   where A is an $n \times m$ matrix. Here, we assume $E[\boldsymbol{x}] = E[\boldsymbol{s}] = 0$.

Further author information: (Send correspondence to S.A.)
S.A.: E-mail: amari@brain.riken.jp, Telephone: +81 (0)48 467 9669

Independent Component Analyses, Wavelets, Unsupervised Smart Sensors, and Neural Networks III, edited by Harold H. Szu, Proceedings of SPIE Vol. 5818 (SPIE, Bellingham, WA, 2005) · 0277-786X/05/$15 · doi: 10.1117/12.607004

2-1. PCA (Principal Component Analysis):
PCA searches for an orthonormal basis $\{\boldsymbol{a}_i\}$, satisfying

$$\left\{\boldsymbol{a}_i^T \boldsymbol{a}_j = \delta_{ij}\right\}, \tag{3}$$

such that the decomposed signals s_i become uncorrelated,

$$E\left[s_i s_j\right] = 0. \tag{4}$$

When $\boldsymbol{x}$, and hence $\boldsymbol{s}$, are subject to Gaussian distributions, all the components s_i are independent. However, they are in general dependent in non-Gaussian cases. Orthonormal bases $\{\boldsymbol{a}_i\}$ are connected by orthogonal transformation ("rotation" of basis), so that PCA is a statistical technique of decorrelating components by rotations of a basis. The power of a signal is kept invariant,

$$\sum x_i^2 = \sum s_i^2 \tag{5}$$

by rotation.

2-2. ICA (Independent Component Analysis):
There are cases where observed signals $\boldsymbol{x}$ are linear mixtures of independent signals $\boldsymbol{s}$,

$$\boldsymbol{x} = A\boldsymbol{s} = \sum s_i \boldsymbol{a}_i. \tag{6}$$

In such a case, $\{\boldsymbol{a}_i\}$ are not necessarily orthogonal. ICA is a new framework searching for a linear basis $\{\boldsymbol{a}_i\}$ such that s_i become not only uncorrelated but independent. When observations $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t$ are iid subject to a Gaussian distribution, the basis $\{\boldsymbol{a}_i\}$ is unidentifiable, having infinitely many solutions. Hence, we need to use the non-Gaussian structure such as higher-order cumulants, temporal structure of signals such as temporal correlations or variations of amplitudes over time, etc.
ICA provided a new perspective with signal processing, and stimulated emergence of new ideas and techniques. The present article focuses on the mathematical structure underlying ICA.

2-3. Adaptive filter:
Given a stationary temporal signal $x(t)$, one may decompose it into

$$x(t) = \int h(t-\tau) s(\tau) d\tau \tag{7}$$

such that $s(\tau)$ is a white sequence. When the signal is a vector,

$$\boldsymbol{x}(t) = \int H(t-\tau)\boldsymbol{s}(\tau) d\tau, \tag{8}$$

where $H(t)$ is the transfer function matrix. The problem is very similar to the standard ICA, and some techniques are common.

3. Variable bases under non-stochastic criteria

3-1. Sparse Component Analysis (SCA):
Given $\boldsymbol{x}$, one may search for a representation

$$\boldsymbol{x} = \sum s_i \boldsymbol{a}_i \tag{9}$$

such that many of s_i are zero or nearly zero. In other words, we search for a basis $\{\boldsymbol{a}_i\}$ such that non-zero components of $\boldsymbol{s}$ are "sparse". This is called sparse component analysis.

3-2. Sparse representation under a fixed basis:
Let us consider an overcomplete basis $\{\boldsymbol{a}_i\}$. In such a case, there are infinitely many ways of representations $\boldsymbol{s}$

$$\boldsymbol{x} = \sum s_i \boldsymbol{a}_i, \tag{10}$$

because $\{\boldsymbol{a}_i\}$ are linearly dependent. Given $\boldsymbol{x}$, one searches for the representation $\boldsymbol{s}$ in which the number of non-zero components is minimized.

3-3. Non-negative Matrix Factorization (NMF):
There are cases where $\boldsymbol{x}$ is a linear mixture of non-negative signals. The case of visual signals is a good example. In such a case, we have techniques of determining the basis $\{\boldsymbol{a}_i\}$ from many observed $\boldsymbol{x}$'s.

4. Multilayer perceptrons:
Let us consider a multilayer perceptron, which receives input signal $\boldsymbol{u}$ and emits scalar output y. When it includes h hidden units, the input-output relation is given by

$$y = \sum_{i=1}^{h} v_i \varphi\left(\boldsymbol{w}_i \cdot \boldsymbol{u}\right), \tag{11}$$

where φ is the sigmoidal function (activation function), $\boldsymbol{w}_i$ is the weight vector of ith hidden unit, and v_i is a weight from ith unit to the linear output unit. Writing the above as

$$y = x(\boldsymbol{u}) = \sum v_i a_i(\boldsymbol{u}) \tag{12}$$

where $a_i(\boldsymbol{u}) = \varphi\left(\boldsymbol{w}_i \cdot \boldsymbol{u}\right)$, we see that $\{a_i(\boldsymbol{u})\}$ forms a basis in the function space of $\boldsymbol{u}$. Here $\{a_i(\boldsymbol{u})\}$ is a variable basis including adjustable parameters $\boldsymbol{w}_i$. It is known that this type of variable basis representations has merits over the fixed basis representation.

3. MATHEMATICAL STRUCTURE OF ICA

3.1. Statistical formulation

Let us consider a simplest case, where $\boldsymbol{x}_t (t = 1, 2, \cdots)$ are given by

$$\boldsymbol{x}_t = A\boldsymbol{s}_t, \quad t = 1, 2, \cdots \tag{13}$$

and A is an $n \times n$ nonsingular matrix. We further assume that the components of $\boldsymbol{s}_t$ are stochastically independent. Their distributions are unknown, but do not depend on t. The problem is to estimate A and recover $\boldsymbol{s}_t$ from $\{\boldsymbol{x}_t\}$.

The probability density function of $\boldsymbol{s}$ is factorized as

$$p(\boldsymbol{s}) = r_1\left(s_1\right) r_2\left(s_2\right) \cdots r_n\left(s_n\right). \tag{14}$$

Since $\boldsymbol{x}$ is derived from $\boldsymbol{s}$, its probability density function is given by

$$p(\boldsymbol{x}; A, \boldsymbol{r}) = |W| r(W\boldsymbol{x}), \tag{15}$$

where $W = A^{-1}$ and

$$r(\boldsymbol{y}) = r_1\left(y_1\right) \cdots r_n\left(y_n\right). \tag{16}$$

The statistical model (15) of $\boldsymbol{x}$ includes two parameters. One is A or its inverse W which we want to know. The other is n functions $r_1, \cdots, r_n$. Given t observations $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t$, we can estimate $\hat{W}$, and we can recover $\boldsymbol{s}_t$ by

$$\boldsymbol{y}_t = \hat{W}\boldsymbol{x}_t. \tag{17}$$

This is a simple statistical problem when we know r, but when r is unknown, the statiscal model includes unknown parameters of function degrees of freedom. Such a model is called a semiparametric model, which is difficult to solve in general.

3.2. Cost functions

In order to recover $\boldsymbol{s}$, we use a matrix W, and put

$$\boldsymbol{y} = W\boldsymbol{x}. \tag{18}$$

If the components of $\boldsymbol{y}$ are independently distributed, $W = A^{-1}$ (more precisely a rescaled and permuted version of A^{-1}). In this case, $\boldsymbol{y}$ gives the original signal $\boldsymbol{s}$ (except for permutation and rescaling of components). Therefore, if we have a function $l(\boldsymbol{y})$, whose expectation

$$L(W) = E\left[l(\boldsymbol{y})\right] \tag{19}$$

is a measure of independence among the components of $\boldsymbol{y}$, we may use this as a cost function, and a gradient descent learning algorithm follows,

$$\Delta W = -\eta \frac{\partial l(\boldsymbol{y})}{\partial W}, \tag{20}$$

where η is a learning constant.

There is a number of candidates of such cost functions. One is the type derived from the probability density functions. Assume that we know $r_1, \cdots, r_n$. Then, the maximum likelihood estimator (mle)[6] is the one that maximizes the log likelihood function. Hence, its negative is

$$\begin{aligned} l(\boldsymbol{y}) &= -\log p(\boldsymbol{x}, W, r) \\ &= -\log|W| - \sum \log r_i\left(y_i\right). \end{aligned} \tag{21}$$

This is a cost function to be minimized, because mle maximizes the log likelihood. When we do not know r_i, choose arbitrary density functions $q_i\left(y_i\right)$, and define

$$l(\boldsymbol{y}) = -\log|W| - \sum \log q_i\left(y_i\right). \tag{22}$$

This also works as a cost function,[7] and the true W is at its critical point. There is an information theoretic interpretation on this function. But there is no guarantee that the true W minimizes it, so that one should choose q_i carefully. The stability analysis is required for this purpose.

Another idea is based on cumulants. The central limit theorem shows that, given n independent signals s_i, the distribution of their linear combination or mixing

$$x = \sum w_i s_i \tag{23}$$

converges to a Gaussian distribution as n tends to infinity. The Gaussian distribution has no higher-order cumulants (higher than the second). In general, the absolute values of higher order cumulants decrease by mixing independent signals.

Hence, one may use the cumulant, such as

$$\sum \mathrm{Cum}\left[y_i^4\right], \quad \sum \mathrm{Cum}\left[y_i, y_j, y_k, y_l\right], \text{ etc.} \tag{24}$$

as a cost function,[8] together with some constraints on the magnitude of W. Here, cum denotes the cumulant function, for example

$$\begin{aligned} \mathrm{Cum}\left[y_i, y_j, y_k, y_l\right] &= E\left[y_i y_j y_k y_l\right] \\ &\quad - E\left[y_i y_j\right] E\left[y_k y_l\right] - E\left[y_i y_k\right] E\left[y_j y_l\right] - E\left[y_i y_l\right] E\left[y_j y_k\right]. \end{aligned} \tag{25}$$

In any case, for a function $f(\boldsymbol{y})$ of $\boldsymbol{y}$, we have its gradient or the total differential df with respect to W by

$$df(\boldsymbol{y}) = df(W\boldsymbol{x}) = f'(\boldsymbol{y}) dW \boldsymbol{x} = f'(\boldsymbol{y}) dW W^{-1} \boldsymbol{y}. \tag{26}$$

The gradient $\nabla f(\boldsymbol{y}) = \partial f(\boldsymbol{y})/\partial W$ is easily calculated from this. This is a useful formula leading to various learning algorithms.

3.3. Natural gradient algorithm[9]

The gradient of a function $\nabla l(W)$ represents the steepest direction in which $f(W)$ changes, provided the space of parameters W is an orthonormal Cartesian coordinate system in a Euclidean space. However, here is the space of all non-singular matrices W, called the general linear group $Gl(n)$, in the present case. It is a Lie group, which is not a Euclidean but Riemannian space.

Let dW be a small change of W at point W. We want to have an invariant metric to define the squared length of dW. We map W to the unit matrix I by multiplying W^{-1} from the right. Then $W + dW$ is mapped to

$$(W + dW)W^{-1} = I + dWW^{-1} = I + dX, \tag{27}$$

so that dW at W corresponds to

$$dX = dWW^{-1} \tag{28}$$

at I. We postulate that squared lengths of dX and dW are the same,

$$\|dX\|_I^2 = \|dW\|_W^2 . \tag{29}$$

Moreover, because of the isotropy of I, it is natural to define

$$\|dX\|_I^2 = tr\left(dX^T dX\right). \tag{30}$$

Then, we have

$$\|dW\|^2 = tr\left(W^{-T} dW^T dW W^{-1}\right). \tag{31}$$

This is an invariant Riemannian metric in the space of W.

The steepest direction of function $l(\boldsymbol{y})$ is given by the natural gradient, which, in this case, is given by

$$\tilde{\nabla} l(\boldsymbol{y}) = \nabla l(W) W^T W. \tag{32}$$

This is the Riemannian gradient, called the natural gradient.[10] The learning algorithm based on the natural gradient in general takes the form

$$\Delta W_t = -\eta_t \left(I - \boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{y}^T\right) \boldsymbol{W}, \tag{33}$$

where η_t is learning constant, $\boldsymbol{\varphi}(\boldsymbol{y}) = \left(\varphi_1\left(y_1\right), \cdots, \varphi_n\left(y_n\right)\right)^T$,

$$\varphi_i\left(y_i\right) = -\frac{d}{dy_i} \log q_i\left(y_i\right). \tag{34}$$

The performance of the natural gradient is equivariant,[11] that is, its dynamic behaviour is the same whichever the true W is. Hence, even when W is close to a singular matrix, it works well.

4. MATHEMATICAL FOUNDATION BY ESTIMATING FUNCTIONS

4.1. Estimating functions

An unbiased estimating function[12] gives a universal technique to obtain an estimator in a semiparametric statistical model. Consider a matrix function $F(\boldsymbol{x}, W)$ of $\boldsymbol{x}$ and W, which does not depend on unknown $r_1, \cdots, r_n$. When it satisfies

$$E_A\left[F(\boldsymbol{x}, W)\right] = 0, \tag{35}$$

when $W = A^{-1}$, and is not equal to 0 when W is different from A^{-1} in a neighbourhood of the true A^{-1}, whichever $r_1, \cdots, r_n$ one chooses, the matrix function $F(\boldsymbol{x}, W)$ is called an unbiased estimating function.[13] Here, E_A denotes expectation is taken with respect to $p\left(\boldsymbol{x}; A^{-1}, r\right)$.

When an estimating function exists, given observations $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t$, we have the estimating equation

$$\sum_{i=1}^{t} F\left(\boldsymbol{x}_i, W\right) = 0, \tag{36}$$

because the right-hand side approximates t times the expectation with respect to the true distribution. We can also have the related learning algorithm

$$\Delta W_t = -\eta_t F\left(\boldsymbol{x}_t, W_t\right). \tag{37}$$

It is easily proved that

$$F(\boldsymbol{x}, W) = I - \boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{y}, \tag{38}$$

where $\boldsymbol{y} = W\boldsymbol{x}$, is an estimating function.

There are many estimating functions. Let $R(W)$ be a reversible linear operator which maps a matrix to a matrix depending only on W. Then,

$$\tilde{F}(\boldsymbol{x}, W) = R(W) \circ F(\boldsymbol{x}, W) \tag{39}$$

is again an estimating function.[14] The estimating equations are the same for both F and $\tilde{F}$, but the dynamic properties are quite different for the learning algorithms using F and $\tilde{F}$. In particular, for some R, the true solution is a stable equilibrium while it is unstable (saddle) for other R.

4.2. Error and stability analysis

We first show the error of estimation. Let $\hat{W}$ be the solution of the estimating equation, and W be the true A^{-1}, and denote the error by

$$\Delta W = \hat{W} - W. \tag{40}$$

We further put

$$\Delta X = \Delta W W^{-1}, \tag{41}$$

which is convenient, because the error in the recovered signal is given by

$$\Delta \boldsymbol{s} = \Delta W \boldsymbol{x} = \Delta X \boldsymbol{s}. \tag{42}$$

From the Taylor expansion, we have

$$0 = \sum F\left(\boldsymbol{x}_i, W + \Delta X W\right) = \sum F\left(\boldsymbol{x}_i, W\right) + \frac{\partial F}{\partial W}\Delta X W \tag{43}$$

We use the following notation,

$$\frac{\partial F}{\partial X} = \frac{\partial F}{\partial W} W^T, \tag{44}$$

and

$$K = E\left[\frac{\partial F}{\partial X}\right] \tag{45}$$

which is a quantity having four indices, because it is the derivative of a matrix with respect to a matrix. We then have

$$\Delta X = -\frac{1}{\sqrt{t}} K^{-1} \left\{ \frac{1}{\sqrt{t}} \sum F\left(\boldsymbol{x}_i, W\right) \right\}. \tag{46}$$

Since $E\left[F\left(\boldsymbol{x}_i, W\right)\right] = 0$,

$$\frac{1}{\sqrt{t}} \sum F\left(\boldsymbol{x}_i W\right) \tag{47}$$

converges to the Gaussian distribution with mean 0 and variance-convariance matrix (having four indices)

$$G = E\left[F(\boldsymbol{x}, W) F(\boldsymbol{x}, W)\right], \tag{48}$$

and hence, the error is calculated as

$$E\left[\Delta X \Delta X\right] = \frac{1}{t} K^{-1} G K^{-1}. \tag{49}$$

Now we use the Newton method to solve the estimating equation adaptively. We use K, and put

$$F^*(\boldsymbol{x}, W) = K^{-1} F(\boldsymbol{x}, W), \tag{50}$$

which is called the standard estimating function.[14] We have

$$K^* = \frac{\partial F^*}{\partial X} = I. \tag{51}$$

Then, the error is given by

$$D\left[\Delta X \Delta X\right] = \frac{1}{t} E\left[F^* F^*\right], \tag{52}$$

and the true W is stable.

The operator K is calculated explicitly for (21). Let us define the following quantities

$$n_i = E\left[s_i^2 \varphi_i'\left(s_i\right)\right], \tag{53}$$
$$k_i = E\left[\varphi_i'\left(s_i\right)\right], \tag{54}$$
$$\sigma_i^2 = E\left[s_i^2\right], \tag{55}$$

under the scale condition

$$E\left[s_i \varphi\left(s_i\right)\right] = 1. \tag{56}$$

Then, the components of the operator K is given by

$$K_{ij,kl} = E\left[\varphi_i'\left(s_i\right) s_j^2\right] \delta_{jl}\delta_{ik} + \delta_{il}\delta_{jk}. \tag{57}$$

We can invert K, and the standard estimating function is given by

$$F_{ij}^*(\boldsymbol{x}, W) = c_{ij}\left\{k_j \sigma_i^2 \varphi\left(y_i\right) y_j - \varphi\left(y_j\right) y_i\right\}, \tag{58}$$
$$c_{ij} = \frac{1}{k_i k_j \sigma_i^2 \sigma_j^2 - 1}. \tag{59}$$

The algorithm is always stable[14, 15] whatever φ_i we choose, but we need to determine k_i and σ_i^2 adaptively from data.

Finally, we remark that the method of estimating function is useful even in the case where $\boldsymbol{s}_t$ has temporal correlation.[16] The joint diagonalization method such as JADE[17] is a special example of the estimating function.

4.3. Nonholonomic algorithm

For an estimating function $F(\boldsymbol{x}, W)$, its diagonal term F_{ii} is used to determine the magnitute of the recovered signals. For example, in the case of (33), the magnitude of s_i is determined from

$$E\left[s_i \varphi\left(s_i\right)\right] = 1. \tag{60}$$

However, the magnitudes of the source signals are not identifiable, and can be determined arbitrarily. Therefore we may choose the diagonal terms of F or F^* arbitrarily. What will, then, happen if we put $F_{ii} = 0$. Then, the magnitudes are not fixed, and fluctuate arbitrarily depending on the observed $\boldsymbol{x}$ and the current state. This is convenient for time varying source signals. Some source $s_i(t)$ becomes very small or even 0 at some time interval. If we force even such a small or 0 signal to have a fixed magnitude, this instabilizes the algorithm. Hence, it looks better to let the magnitudes be free.

The constraints $F_{ii} = 0$ correspond to

$$\left(\Delta X\right)_{ii} = \left(\Delta W W^{-1}\right)_{ii} = 0 \tag{61}$$

in the algorithm. However, these constraints are "nonholonomic", and allow free motions of the amplitudes.[18]

It is shown by computer simulations that the non-holonomic algorithm works well when the number of sources is unknown and is smaller than that of sensors.

5. SPARSE COMPONENT ANALYSIS

5.1. Basis of sparse representation

Given signals $\boldsymbol{x}$, we consider the following problem of minimizing

$$\sum_t \left[\left| \boldsymbol{x}_t - \sum_{i=1}^{n} s_i(t)\boldsymbol{a}_i \right|^2 - \lambda \sum_{i=1}^{n} \log\left(1 + (s_i(t))^2\right) \right]. \tag{62}$$

The first term is the squared error of representations of $\boldsymbol{x}$ with basis $(\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n)$, while the second term requires $|s_i|$ to be small, hopefully to be 0. Under this condition, Olshausen and Field[19] search for the basis. Here, λ is the Lagrangean multiplier, and when $\lambda = 0$ the problem reduces to PCA. It is possible to give the Bayesian interpretation of this criterion.

Because of the second term, the selected basis gives such $\boldsymbol{s}$ of which most components are zero or small, and only a small number of components are significant. Hence, this called the sparse representation. It is used nonlinear denoising and others, opening a new field of signal representation.

5.2. Overcomplete basis and various solutions[3]

The sparse representation poses an interesting problem. Let us fix a basis $\{\boldsymbol{a}_i\}$, which is overcomplete, that is it includes dependent basis vectors. For the overcomplete basis $\{\boldsymbol{a}_i\}$, the decomposition of $\boldsymbol{x}$,

$$\boldsymbol{x} = \sum s_i \boldsymbol{a}_i = A\boldsymbol{s} \tag{63}$$

is not unique. Indeed, let $\boldsymbol{r}$ be a vector belonging to the null space N of A,

$$N = \{\boldsymbol{r} | A\boldsymbol{r} = 0\}. \tag{64}$$

Then, for a solution $\boldsymbol{s}$,

$$\boldsymbol{s} = \boldsymbol{s}_0 + \boldsymbol{r}, \quad \boldsymbol{r} \in N \tag{65}$$

is also a solution.

Among all the solutions, we search for the one that minimizes the p-norm,

$$\boldsymbol{s}_p = \arg\min_{A\boldsymbol{s}=\boldsymbol{x}} \|\boldsymbol{s}\|_p. \tag{66}$$

The solution $\boldsymbol{s}_2$ is given by

$$\boldsymbol{s}_2 = A^\dagger \boldsymbol{x}, \tag{67}$$

where $A^\dagger$ is the generalized inverse of A. The $\boldsymbol{s}_\infty$ is the one that minimizes the maximum absolute value of the components of $\boldsymbol{s}$. The $\boldsymbol{s}_0$ is the one that minimizes

$$\sum |s_i|^0 = \text{the number of non-zero components of } \boldsymbol{s} \tag{68}$$

Hence, $\boldsymbol{s}_0$ gives the sparsest solution.

It is interesting to know how $\boldsymbol{s}_p$ changes depending on p. Let us consider the case where $s_{pi} \geq 0$, that is, $\boldsymbol{s}_p$ lies in the first quadrangle (the other case can be treated in the same way). In this case, the set M of solution,

$$M = \{\boldsymbol{s} | A\boldsymbol{s} = \boldsymbol{x}\}, \tag{69}$$

is a subspace passing through the first quadrangle, and whose normal vectors $\boldsymbol{n}$ satisfy $n_i \geq 0$.

Observations. The solution $\boldsymbol{s}_\infty$ is the intersection of the 45° line $s_1 = s_2 = \cdots = s_n$ and M. As p decreases $\boldsymbol{s}_p$ moves continuously, and $\boldsymbol{s}_2$ is the orthogonal projection of the origin to M. As p further decreases, $\boldsymbol{s}_p$ moves toward the corner of M at which many $s_i = 0$. When $p = 1$, it reaches the corner. As p further

decreases, it stays at that corner, but there may exist a number of local minima at various corners, and the $\boldsymbol{s}_1$ can be one of the local minima of $\boldsymbol{s}_p$ $(p < 1)$.

This is understood from the following. Let us consider the p-indicatrix

$$s_1^p + \cdots + s_n^p = c. \tag{70}$$

This is convex for $p \geq 1$, but is concave for $p < 1$. When c is small, it lies below M (in the side of origin). As c increase, it eventually touch M. This first contact point is $\boldsymbol{s}_p$.

5.3. Sparse solution

The $\boldsymbol{s}_p$ $(p > 1)$ is not sparse, because all the components of $\boldsymbol{s}_p$ are not zero. However, $\boldsymbol{s}_p$ $(p \leq 1)$ is sparse in the sense that many components are zero. When A is $n \times m$ matrix $(m > n)$, at most n components are non-zero. The $\boldsymbol{s}_1$ is easy to obtain by solving the LP problem or by the gradient descent method.

The $\boldsymbol{s}_0$ is the sparsest solution. A question naturally arises: What is the condition that guarantees $\boldsymbol{s}_1 = \boldsymbol{s}_0$. There are lot of interesting theories concerning this problem.[20] We are searching for the following problem (Li, Amari and Cichocki[21]):

Let A be a random $n \times m$ matrix $(m > n)$, and let $\boldsymbol{s}^*$ be a randomly chosen sparse vector which has k non-zero components. Let $\boldsymbol{s}_1$ be the L_1-solution of the problem

$$A\boldsymbol{s} = \boldsymbol{x}, \tag{71}$$

where $\boldsymbol{x} = A\boldsymbol{s}^*$. What is the probability

$$P(n, m, k) = \mathrm{Prob}\left\{\boldsymbol{s}_1 = \boldsymbol{s}^*\right\}. \tag{72}$$

When $m = \alpha n$, $k = \beta n$, $n \to \infty$, how is the asymptotic result,

$$P(\alpha, \beta) = \lim_{n\to\infty} P(n, \alpha n, \beta n). \tag{73}$$

6. NON-NEGATIVE MATRIX FACTORIZATION

When the source signals are limited in the first quadrangle, $s_1 \geq 0, \cdots, s_n \geq 0$, and the probability density is positive, we can use another technique (NMF) to determine A from observed $\boldsymbol{x}$'s $(\boldsymbol{x} = A\boldsymbol{s})$.

The first quadrangle of the space of $\boldsymbol{s}$ is mapped by A to the inside of the cone spanned by $\boldsymbol{a}_i$ in the space of $\boldsymbol{x}$. This is clear from

$$\boldsymbol{x} = \sum s_i \boldsymbol{a}_i, \quad s_i \geq 0. \tag{74}$$

Hence, observing a number of $\boldsymbol{x}$, we can estimate $\boldsymbol{a}_i$'s from the distribution of $\boldsymbol{x}$'s without assuming independency of s_i. There have been proposed a number of algorithms.

7. CONCLUSIONS

We have overviewed mathematical structures of ICA, and also new techniques of sparse component analysis and non-negative matrix factorization. Information geometry plays a fundamental role for elucidating the mathematical structures. We have given a unified standpoint of analyzing the techniques of ICA searching for a general framework for its error analysis and stability analysis. New techniques of sparse component analysis and non-negative matrix factrization are touched upon.

REFERENCES

1. A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*, John Wiley, West Sussex, 2002.
2. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley, New York, 2001.
3. S. Chen, D. Donoho, and M. Sandners, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.* **20**, pp. 33–61, 1999.
4. D. Lee and H. Seung, "Learning of the parts of objects by non-negative matrix factorization," *Nature* **401**, pp. 788–791, 1999.
5. S. Amari and H. Nagaoka, *Methods of Information Geometry*, AMS and Oxford University Press, 1999.
6. D.-T. Pham, "Blind separation of instantaneous mixture of sources via and independent component analysis," *IEEE Trans. Signal Processing* **44**(11), pp. 2768–2779, 1996.
7. A. Bell and T. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation* **7**(6), pp. 1129–1159, 1995.
8. J.-F. Cardoso, "High-order contrasts for independent component analysis," *Neural computation* **11**(1), pp. 157–192, 1999.
9. S. Amari, "Natural gradient works efficiently in learning," *Neural Computation* **10**, pp. 271–276, 1998.
10. S. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind signal separation," in *Advances in Neural Information Processing Systems 1995*, M. C. M. D. S. Touretzky and M. E. Hasselmo, eds., **8**, pp. 757–763, MIT Press, (Cambridge, MA), 1996.
11. J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing* **44**(12), pp. 3017–3030, 1996.
12. S. Amari and M. Kawanabe, "Information geometry of estimating functions in semiparametric statistical models," *Bernoulli* **3**.
13. S. Amari and J.-F. Cardoso, "Blind source separation —semi-parametric statistical approach," *IEEE Trans. on Signal Processing* **45**(11), pp. 2692–2700, 1997.
14. S. Amari, "Super-efficiency in blind source separation," *IEEE Trans. on Signal Processing* , 1997.
15. S. Amari, T.-P. Chen, and A. Cichocki, "Stability analysis of adaptive blind source separation," *Neural Networks* **10**(8), pp. 1345–1351, 1997.
16. S. Amari, "Estimating function of independent component analysis for temporally correlated signals," *Neural Computation* **12**(9), pp. 2083–2107, 2000.
17. J.-F. Cardoso and A. Souloumiac, "Jacobi angles for simultaneous diagonalization," *SIAM Journal Mat. Anal. Appl.* **17**(1), pp. 161–164, 1996.
18. S. Amari, T.-P. Chen, and A. Cichocki, "Non-holonomic constraints in learning algorithms for blind source separation," *Neural Computation* **12**, pp. 1463–1484, 2000.
19. B. Olshausen and D. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature* **381**, pp. 607–609, 1996.
20. D. L. Donoho and M. Elad, "Maximal sparsity representation via l_1,"
21. Y. Li, A. Cichocki, and S. Amari, "Analysis of sparse representation and blind source separation," *Neural Computation* **16**, pp. 1193–1204, 2004.