On Independent Component Analysis using Tsallis Mutual Entropy

Hiroki Suyari† and Masumi Dakemoto‡

†Department of Information and Image Sciences, Chiba University
‡Graduate School of Science and Technology, Chiba University
1-33 Yayoi-cho, Inage-ku, Chiba-shi, 263-8522 Japan
Phone:+81-43-290-3509, Fax:+81-43-290-3510
Email: †suyari@tj.chiba-u.ac.jp, ‡dakemoto@icsd3.tj.chiba-u.ac.jp

1. Introduction

Effects of using Tsallis mutual entropy in stead of Shannon mutual entropy for independent component analysis (ICA for short) are discussed.

ICA has been studied in applied statistics to separate mutually independent sources from given observed linearly mixed signals for many applications such as speech recognition, data communication and sensor signal processing [1, 2]. Some measures for statistical independency among source signals have been proposed in the applications of ICA. Such typical measures are nongaussianity [3] and Shannon mutual entropy [4]. In this paper, we consider the latter case in accordance with the algorithm [4, 5]. ICA using Shannon mutual entropy has some advantages over other measures for separation of sources with sub-Gaussian statistics such as images. However, the ICA algorithm using Shannon mutual entropy does not always succeed in ideal source separation in each application. It often depend on a choice of initial demixing matrix for ICA. Then we apply Tsallis mutual entropy to ICA in stead of Shannon mutual entropy.

Tsallis entropy was presented as one-parameter generalization of Shannon entropy in 1988 [6]. There have been too many generalizations of Shannon entropy since Shannon's famous papers [7], but only Tsallis entropy has succeed in a natural generalization of the usual Boltzmann-Gibbs statistical mechanics [8, 9]. Especially, the results derived from Tsallis entropy can present the nice interpretations for many physical phenomena such as long-range interactions and long-time memories, which cannot be explained systematically in the usual Boltzmann-Gibbs statistical mechanics [8, 9].

This paper presents advantages of applying Tsallis mutual entropy to ICA over using Shannon mutual entropy. Note that, in stead of the original Tsallis entropy, we use the normalized Tsallis entropy to satisfy the form invariance of the maximum entropy principle and that of additivity as similarly as Shannon entropy [10, 11]. We present some successful simulation results of the ICA for some initial demixing matrices which fail to separate source signals in case of using Shannon mutual entropy. Moreover, appropriate control of a new parameter $q \in [0,1]$ in the ICA can provide speed-up of the convergence to each desired demixing matrix.

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Encouragement of Young Scientists(B), 14780259, 2002.

2. Tsallis Entropy

Tsallis entropy is defined by

$$S_q^{\mathrm{org}}\left(p_1,\cdots,p_n\right) := \frac{1-\sum\limits_{i=1}^{n} p_i^q}{q-1}. \tag{1}$$

The maximization of the original Tsallis entropy (1) with respect to $\{p_i\}$ under the energy constraint $\sum_{i=1}^{n} p_i^q E_i \Big/ \sum_{j=1}^{n} p_j^q = E$ leads to

$$p_i = \frac{\exp_q\left[-\beta\left(E_i - E\right)\right]}{Z_q} \propto \left[1+(q-1)\beta\left(E_i-E\right)\right]^{\frac{1}{1-q}} \tag{2}$$

where β is a Lagrangian coefficient, Z_q is a normalized constant called partition function, and

$$\exp_q[x] := \left[1+(1-q)x\right]^{\frac{1}{1-q}}, \quad \lim_{q\to 1}\exp_q[x] = \exp[x]. \tag{3}$$

This generalized equilibrium state $\{p_i\}$ given by (2) represents power law and recovers the usual exponential expression when $q \to 1$. The expression (2) can lead to the successful generalization of the usual Boltzmann-Gibbs statistical mechanics and its applications [8, 9].

The above formula (1) is the original Tsallis entropy, but we apply the following formula to ICA in order to satisfy the form invariance of additivity as similarly as

Shannon entropy [11].

$$S_q(p) := \frac{S_q^{\mathrm{org}}(p)}{\sum_{j=1}^{n} p_j^q} = \frac{1-\sum_{i=1}^{n} p_i^q}{(q-1)\sum_{j=1}^{n} p_j^q}. \quad (4)$$

Note that the original Tsallis entropy (1) is concave for any $q \in R^+$ but the normalized Tsallis entropy (4) is concave only for any $q \in [0,1]$ [10, 11]. These Tsallis entropies (1) and (4) is one-parameter generalizations of Shannon entropy in the sense that

$$\lim_{q\to 1} S_q^{\mathrm{org}} = \lim_{q\to 1} S_q = S_1 : = -\sum_{i=1}^{n} p_i \ln p_i. \quad (5)$$

Moreover, when two random variables X and Y are mutually independent, the normalized Tsallis entropy (4) satisfies the following characteristic property.

$$S_q(X,Y) = S_q(X) + S_q(Y) + (q-1)\, S_q(X)\, S_q(Y) \quad (6)$$

The property (6) called pseudoadditivity represents nonextensive additivity of Tsallis entropy, i.e. Tsallis entropy is a nonextensive entropy. When $q = 1$ the pseudoadditivity (6) reverts to the standard (extensive) additivity, so that Shannon entropy is considered to be extensive entropy in contrast to Tsallis entropy.

Based on the normalized Tsallis entropy (4), the nonextensive Kullback-Leibler entropy D_q between the random variable Y of m-dimensions and the product of its components $\prod_{i=1}^{m} Y_i$ is derived as [11]

$$D_q\left(Y \,\middle|\, \prod_{i=1}^{m} Y_i\right) = -S_q(Y) + E_q\left[\ln_q \prod_{i=1}^{m} f_{Y_i}(y_i)\right] \quad (7)$$

where m is the number of the components of independent sources, and $\ln_q$ and E_q are defined by

$$\ln_q x := \frac{x^{1-q}-1}{1-q}, \quad E_q[X] := \frac{\int x f(x)^q\, dx}{\int f(x')^q\, dx'} \quad (8)$$

In this paper the minimization of D_q using the truncated Gram-Charlier expansion at the sixth order cumulant is applied to a ICA algorithm in accordance with [4, 5].

3. Independent Component Analysis using Tsallis Mutual Entropy

Let us denote by $\vec{s}(t) = (s_1(t), \cdots, s_m(t))^T$ $(t = 0, 1, \cdots)$ unknown source signals which are mutually independent. The observation vector $\vec{x}(t) = (x_1(t), \cdots, x_m(t))^T$ $(t = 0, 1, \cdots)$ is given by

$$\vec{x}(t) = A \cdot \vec{s}(t) \quad (9)$$

where $A \in R^{m\times m}$ is an unknown non-singular mixing matrix. The only informations available to us are the observation vector $\vec{x}(t)$ and the knowledge which the source signals $\vec{s}(t)$ are mutually independent. Under these conditions, we want to recover the source signals $\vec{s}(t)$ from the observation vector $\vec{x}(t)$ without knowledge of the source signals $\vec{s}(t)$ and the mixing matrix A.

For a given observation vector $\vec{x}(t)$, the problem is to find a demixing matrix W such that the source signals $\vec{s}(t)$ can be recovered from $\vec{x}(t)$.

$$\vec{y}(t) = W \cdot \vec{x}(t) \quad (10)$$

Ideally $\vec{y}(t) = \vec{s}(t)$ is desired, but there remains two uncertainties such as permutation of indices and scale in each components $s_i(t)$.

The ICA algorithm used in this paper is given in accordance with [4]. For simplicity, we consider the case of $m = 2$. It is easy to extend to the case of any $m \in N$, but the calculation becomes very complicated.

Step 1: Initialization of the demixing matrix W

Step 2: Compute $\vec{y} = W \cdot \vec{x}$

Step 3: Compute dW by

$$(dW)_{jk} = \frac{\partial}{\partial w_{jk}} D_q = -\frac{\partial}{\partial w_{jk}} S_q(Y) + \left(1 + (q-1) S_q\left(\tilde{Y}_{3-j}\right)\right) \frac{\partial}{\partial w_{jk}} S_q\left(\tilde{Y}_j\right) \quad (j = 1, 2) \quad (11)$$

Step 4: $W(n+1) = W(n) + \eta dW$ where η is a learning rate.

Step 5: Repeat the steps $2 \sim 4$ until all components in dW are sufficiently small.

Each term of dW in Step 3 is given in the appendix.

4. Simulation

In our simulation, we take some images with sub-Gaussian statistics. The elements of the mixing matrix A and the demixing matrix W are randomly chosen in $[0,1]$. The learning rate η is set to be 0.01.

There exist some images such that ICA using the Shannon mutual entropy cannot obtain successful extraction of information sources for any initialization in Step 1. However, ICA using the Tsallis mutual entropy can obtain successful extraction among them.

Note that when the elements of the mixing matrix A are set to be $[0,1]$ some elements of A^{-1} are not in $[0,1]$. Thus any initial demixing matrix $W \in [0,1]^{2\times 2}$ does not include A^{-1}. Therefore successful separation using the Tsallis mutual entropy under the same condition as above means the extending of the initial range of W for the success.

Here are such examples. The observation is given by the linear superposition of two images called “milk crown”. The results of ICA in $q = 1$ (ICA using the Shannon mutual entropy) and $q = 0.5$ (ICA using the Tsallis mutual entropy) can be obtained as follows.

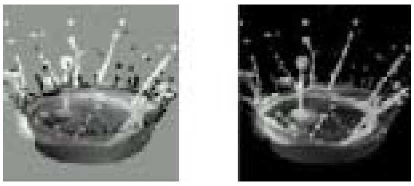


Figure 1: results of ICA when $q = 1$

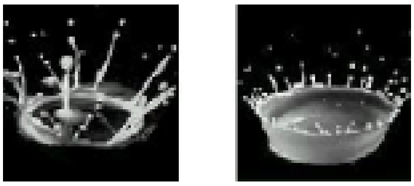


Figure 2: results of ICA when $q = 0.5$

In both cases, the demixing matrix W converges. However, clearly the results of ICA using the Shannon mutual entropy fails in separation, but those of ICA using the Tsallis mutual entropy succeeds.

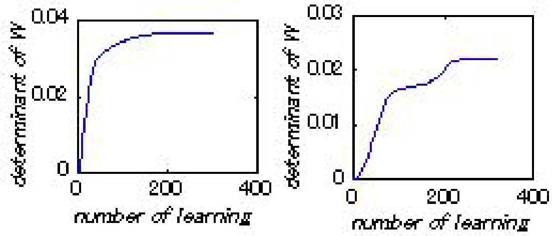
In any successful examples when Tsallis mutual entropy is applied, we can observe one more interesting property on convergence. On the right hand of (11), we have to compute $\frac{\partial}{\partial w_{jk}} S_q(Y)$ which is given by

$$\frac{\partial}{\partial w_{jk}} S_q(Y) = \left\{ 1 - \frac{(1-q)\cdot\left(S_q(X) + \ln_q |\det(W)|\right)}{|\det(W)|^{1-q}} \right\} \left(W^{-T}\right)_{jk}. \tag{12}$$

In applying Tsallis mutual entropy to ICA the computation of $\det(W)$ is needed, which differs from the case of Shannon mutual entropy. (Note that the computation of the determinant $\det(W)$ becomes a problem when the number of sources increases. This is clearly a fault when Tsallis mutual entropy is applied to ICA.)

The change of $\det(W)$ in learning can be observed in the following graphs (Fig.3 and Fig.4).

In the four graphs below, x-axis represents the number of learning (the number of iterations of Step2 $\sim$ Step4)



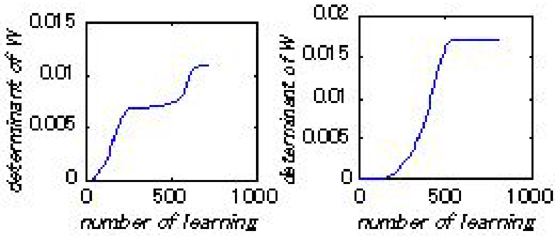


Figure 4: dW when q=0.9, 1.0

and y-axis the determinant $\det(W)$. Clearly when q decreases the convergence becomes very fast. However, when $q < 0.5$ the algorithm does not converge for any initial setting. On the other hand, when $q > 1$ the algorithm does not converge for lack of concavity of the normalized Tsallis entropy (4) as stated in section 2. In any examples such that ICA using Tsallis mutual entropy can succeed the similar change of $\det(W)$ as above can be observed. This reveals that the applying of the new parameter q makes the convergence of the algorithm very fast. The similar effect of the new parameter q as above can be found in the simulated annealing [12].

5. Conclusion

In the present paper the Tsallis mutual entropy, one-parameter generalization of Shannon mutual entropy, is applied to ICA, and its effect of the new parameter q is discussed. We have presented maily two appealing points.

(1) ICA using the Tsallis mutual entropy can extend the initial range of the demixing matrix W for successful separation. Thus for a given initial demixing matrix W there are some cases such that ICA using the Shannon mutual entropy cannot obtain successful separation but ICA using the Tsallis mutual entropy can succeed.

(2) The new parameter q makes the convergence of the ICA algorithm very fast.

However, there remains obvious faults in the generalized ICA. For example, the computation of dW in Step 3 becomes very complicated especially in case of $m \geq 3$.

Moreover, the new computation task such as the determinant of W in each iteration is added in the generalized ICA, which differs from the usual ICA using the Shannon mutual entropy. These faults should be overcome, but actually it is very difficult.

As our future work on the generalized ICA, we have to answer the question: under what condition does the generalized ICA have advantages over the usual ICA ? The answer to this natural question is still unsolved in our work. Moreover, the effect of dynamical control of the new parameter $q(t)$ in learning will be an interesting issue in this ICA, while in the present simulation $q(t)$ is set to be a constant q. The stability of the generalized ICA algorithm should be also discussed in the same line as [13].

There still remain some ambiguities in the foundation of Tsallis' statistical mechanics such as the meaning of the new parameter q in each physical situation and the method to obtain q concretely. These ambiguities are deeply ralted to our future work given above.

References

[1] A. Hyvärinen and J. Karhunen, E. Oja, "*Independent Component Analysis*," John Wiley & Sons, 2001.

[2] A. Cichocki and S. Amari, "*Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*," John Wiley & Sons, 2002.

[3] A. Hyvärinen and E. Oja. "A fast fixed-point algorithm for independent component analysis," Neural Computation, vol.9, pp. 1483-1492, 1997.

[4] S. Amari, A. Cichocki and H. H. Yang, "A new learning algorithm for blind signal separation," Advances in Neural Information Processing Systems, vol. 8, MIT Press, pp. 757-763, 1996.

[5] S. Haykin, *Neural Networks* (second edition), (Prentice-Hall, 1999).

[6] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," J. Stat. Phys., vol. 52, pp. 479-487, 1988.

[7] C.E. Shannon, "A mathematical theory of communication," Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, 1948.

[8] C. Tsallis et al., *Nonextensive Statistical Mechanics and Its Applications*, edited by S. Abe and Y. Okamoto (Springer-Verlag, Heidelberg, 2001); see also the comprehensive list of references in http://tsallis.cat.cbpf.br/biblio.htm

[9] S. Abe, "Tsallis' nonextensive statistical mechanics", Mathematical Sciences, vol. 439-442, 2000 (in Japanese).

[10] A.K. Rajagopal and S. Abe, "Implications of form invariance to the structure of nonextensive entropies", Phys. Rev. Lett., vol. 83, pp. 1711-1714, 1999.

[11] H. Suyari, "Nonextensive entropies derived from form invariance of pseudoadditivity," Phys. Rev. E, vol. 65, 06618, 2002.

[12] C. Tsallis and D.A. Stariolo, "Generalized simulated annealing," Physica A, vol. 233, pp. 395-406 1996.

[13] S. Amari, T. Chen, A. Cichocki "Stability analysis of learning algorithms for blind source separation," Neural Networks, vol. 10, no. 8, pp. 1345-1351, 1997.

Appendix

The following formulas are required in Step 3.

$$
\begin{aligned}
&1 + (q-1) S_q(\tilde{Y}_{3-j}) \\
&\quad = \frac{1}{\frac{1}{\sqrt{q}} \cdot (2\pi)^{(1-q)/2} \cdot (A_{3-j})^{q-2} \cdot B_{3-j}}
\end{aligned}
\tag{13}
$$

$$
\frac{\partial}{\partial w_{jk}} S_q\left(\tilde{Y}_j\right) = \frac{1}{q-1} \frac{-\frac{\partial}{\partial w_{jk}} \int_{-\infty}^{\infty} \left(f_{\tilde{Y}_j}\left(y_j\right)\right)^q dy_j}{\left(\int_{-\infty}^{\infty} \left(f_{\tilde{Y}_j}\left(y'_j\right)\right)^q dy'_j\right)^2}
\tag{14}
$$

$$
\begin{aligned}
&\frac{\partial}{\partial w_{jk}} \int_{-\infty}^{\infty} \left(f_{\tilde{Y}_j}\left(y_j\right)\right)^q dy_j \\
&= \frac{1}{\sqrt{q}} \cdot (2\pi)^{(1-q)/2} \cdot (A_j)^{q-3} \left\{ 2(q\ 2)\left(2y_j^7 - y_j^5\right) \cdot B_j \right. \\
&\left. + A_j \cdot \left[-\frac{(q-1)}{4} \cdot 3y_j^2 - 4y_j^{15}\ \ + 2y_j^5 + 7y_j^{13} - \frac{8}{3} y_j^{11} \right] \right\} x_k
\end{aligned}
\tag{15}
$$

$$
\left(\int_{-\infty}^{\infty} \left(f_{\tilde{Y}_j}(y_j)\right)^q dy_j\right)^2 = \frac{1}{q} \cdot (2\pi)^{(1-q)} \cdot (A_j)^{2(q-2)} \cdot (B_j)^2
\tag{16}
$$

$$
\begin{aligned}
A_j &= 1 + \frac{y_j^8}{2} - \frac{y_j^6}{3} \\
B_j &= 1 - (q-1)\frac{y_j^3}{4} - \frac{y_j^{16}}{4} + y_j^6 \left(\frac{1}{3} + \frac{y_j^8}{2} - \frac{2y_j^6}{9} \right)]
\end{aligned}
$$