Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2015, Article ID 386201, 8 pages
http://dx.doi.org/10.1155/2015/386201

Research Article

Independent Component Analysis Based on Information Bottleneck

Qiao Ke,[1] Jiangshe Zhang,[1] H. M. Srivastava,[2] Wei Wei,[3] and Guang-Sheng Chen[4]

[1] *School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China*
[2] *Department of Mathematics and Statistics, University of Victoria, Victoria, BC, Canada V8W 3R4*
[3] *School of Computer Science and Engineering, Xi'an University of Technology, Shaanxi Key Laboratory for Network Computing and Security Technology, Xi'an 710048, China*
[4] *Department of Construction and Information Engineering, Guangxi Modern Vocational Technology College, Hechi, Guangxi 547000, China*

Correspondence should be addressed to Qiao Ke; keqiao1989@stu.xjtu.edu.cn

Received 26 June 2014; Accepted 15 July 2014

Academic Editor: Hui Zhang

The paper is mainly used to provide the equivalence of two algorithms of independent component analysis (ICA) based on the information bottleneck (IB). In the viewpoint of information theory, we attempt to explain the two classical algorithms of ICA by information bottleneck. Furthermore, via the numerical experiments with the synthetic data, sonic data, and image, ICA is proved to be an edificatory way to solve BSS successfully relying on the information theory. Finally, two realistic numerical experiments are conducted via FastICA in order to illustrate the efficiency and practicality of the algorithm as well as the drawbacks in the process of the recovery images the mixing images.

1. Introduction

Information theory is found by Claude Elwood Shannon (1948) in one of his famous academic papers, "A Mathematical Theory of Communication," where he gave the definition of information and information entropy based on the probability theory which build a bridge between the information theory and the numerical mathematics. Some basic conceptions (entropy, negentropy, mutual information, and so on) in the information theory have been successfully used to elaborate the independent components (ICs) and to deal with the problems on the application of the blind source separation (BSS). In the past decades, the information theory has been applied successfully into many fields such as clustering [1], medical examination [2], independent component analysis [3], feature learning [4], and telecommunication [5–8]. The purpose of this paper is to use the information bottleneck to derive the maximum of the mutual information (MI) between the mixing data and the recovery data which is no more than the MI of the recovery data and the original sources.

The rest of the paper is organized as follows. In Section 2, we first explain the information theory and introduce some important formulas. In Section 3, based on the entropy, the mutual information (MI), and negentropy, information bottleneck is used to illustrate the equivalence of the two classical algorithms, informax [3] and FastICA [9]. At last, by a series of experiments of synthetic data, sonic data, and image in Section 4, it is easy to compare the accuracy and complexity of the two algorithms. However, the ambiguity of the direction and scale of the recovery matrix lead the results of the image to the opposite.

2. Information Theory

According to the explanation of communication theory by Warren Weaver, "information" is not related to what you

do say but to what you could say. That is, information is a measure of one's freedom of choice when one selects a message [6, 7, 10].

At first, people focused attention on the "meaningful" or "relevant" information, which is crucial in solving the problem of transmitting information. Then, some scholars argue that lossy source compression provides a natural quantitative approach to "relevant information" [11, 12].

So, information bottleneck, which is going to seek for a tradeoff between the compression and the representation and preserving meaningful information, could be decomposed into the following aspects:

(1) how to define the "meaningful" or "relevant" information;

(2) how to extract the efficient representation of relevant information in order to transmit it speedily;

(3) how to recover the information as exactly and comprehensively as possible only based on the efficient representation of relevant information.

People regard the possible results of the uncertainty or fuzzy as the surprise or information [13], and the smaller probability of the results occurring, the bigger surprise or the more information people obtained. So, entropy $H(X)$, a measure of the chaotic degree, is defined to measure the uncertainty of information. Assume that X is a discrete random variable, and probability density function (pdf) $p(x) = \Pr(X = x)$, $x \in \chi$; then, entropy is defined as

$$H(X) = -\sum_{x\in\chi} p(x)\log\left(p(x)\right). \tag{1}$$

Moreover, it is easy to generalize it to more than two random variables, the joint entropy. On the other hand, mutual information (MI), a measure of dependency of two different random variable sets, is regarded as the reduction of uncertainty of the random variable, given the other random variable. Consider two random variables X and Y, with the joint pdf $p(x, y)$ and marginal pdfs $p(x)$ and $p(y)$, respectively. MI can be written as follows:

$$\begin{aligned} I(X;Y) &= \sum_{x\in\chi}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) \\ &= E_{p(x,y)}\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right). \end{aligned} \tag{2}$$

It is easy to prove the following equations about MI based on information entropy:

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(Y) - H(Y \mid X) \\ &= H(X) - H(Y \mid X). \end{aligned} \tag{3}$$

According to the last two terms, we can find the relationship between MI and entropy. if and only if X and Y are irrelevant.

3. The Equivalence of the Two ICA Algorithms Based on the IBN

Information bottleneck (IBN) [14] is used to make sure to recover the compressed information X, which is presumed to be good representation or compression of the original information S, to the recipient in terms of Y in the following type:

$$S \Longrightarrow X \Longrightarrow Y. \tag{4}$$

Now, in the terminology of information theory and optimizing theory, there are two inconsistent optimal problems that, on the one hand, we would make sure to minimize MI between the original information S and the compressed information X and, on the other hand, we want to capture the maximum of mutual information between Y and X. Obviously, the amount of information about Y in $\widetilde{X}$ is given by

$$I(X;Y) = \Sigma_y\Sigma_{\tilde{x}} p(y,x)\log\left(\frac{p(y,x)}{p(y)\,p(x)}\right) \leqq I(S;Y), \tag{5}$$

while the mutual information between the independent sources and the mixing signals is determinate but unknown with the precondition of ICA.

ICA is studied to find the independent sources as y_i, $Y = (y_1, y_2, \ldots, y_n)$, which is equal to the original independent sources, $S = (s_{i_1}, s_{i_2}, \ldots, s_{i_n})$ ignoring the ambiguity of the direction and scale. Furthermore, the independent sources are the most concise, while any linear transformation of the independent sources obtains the redundance information

$$\begin{aligned} I(S;Y) &= H(S) - H(S \mid Y) \\ &\leqq H(S). \end{aligned} \tag{6}$$

If and only if Y are the independent sources the equation is true. That is to say, we need to find the recovery matrix W, $Y = WX$, in order to obtain the independent sources. Because of the precondition of the unknown independent sources and mixing matrix, the optimal problem of ICA is written by IBN [14] as follows:

$$\max_{Y} I(X;Y), \tag{7}$$

where $H(S)$ is an theoretic maximum and $H(Y)$ is an approximate maximum.

3.1. Infomax Method. Infomax method [3] is used to tackle the problem of separating the mixture signals X, attempting to look for the weight matrix W without both the mixture matrix A and the original signal S. We attempt to illustrate BBS in the following:

$$\mathbf{S} \xrightarrow[\text{mixture}]{X=AS} \mathbf{X} \xrightarrow[\text{recovery}]{Y=WX} \mathbf{Y}. \tag{8}$$

According to the optimization problem of ICA (16), we could rewrite it as follows:

$$\max_{W} I(X;Y) = H(Y) - H(YX). \tag{9}$$

That is also regarded as

$$\max_{W} H(Y). \tag{10}$$

The equation can be differentiated with respect to a parameter W, involved in the mapping from X to Y:

$$\frac{\partial}{\partial W} I(X;Y) = \frac{\partial}{\partial W} H(Y). \tag{11}$$

Therefore, MI between the mixtures X and recoveries Y could be maximized by maximizing the entropy of the recoveries alone. And then the gradient method is used to obtain the learning method [3].

Considering the slow convergence and nonprecise and low accuracy, *Hyvärinen* gave a FastICA based on the maximum of the negentropy, which is regarded as the measure for the independency of the signal.

3.2. FastICA Method. According to the informax method and IBN, the BBS is equal to the optimal problem as follows:

$$\max_{Y} H(Y). \tag{12}$$

Then, the optimal problem can be adapted as

$$\min_{Y} I(Y) = \sum_{i=1}^{n} H(y_i) - H(Y), \tag{13}$$

where $Y = (y_1, y_2, \ldots, y_n)$ and y_i $(i = 1, 2, \ldots, n)$ are the ICs. How can we identify and measure the independence of the recovery data? The equivalence of the non-Gaussian random variables and negentropy is illustrated based on the Central-Limit Theorem [9].

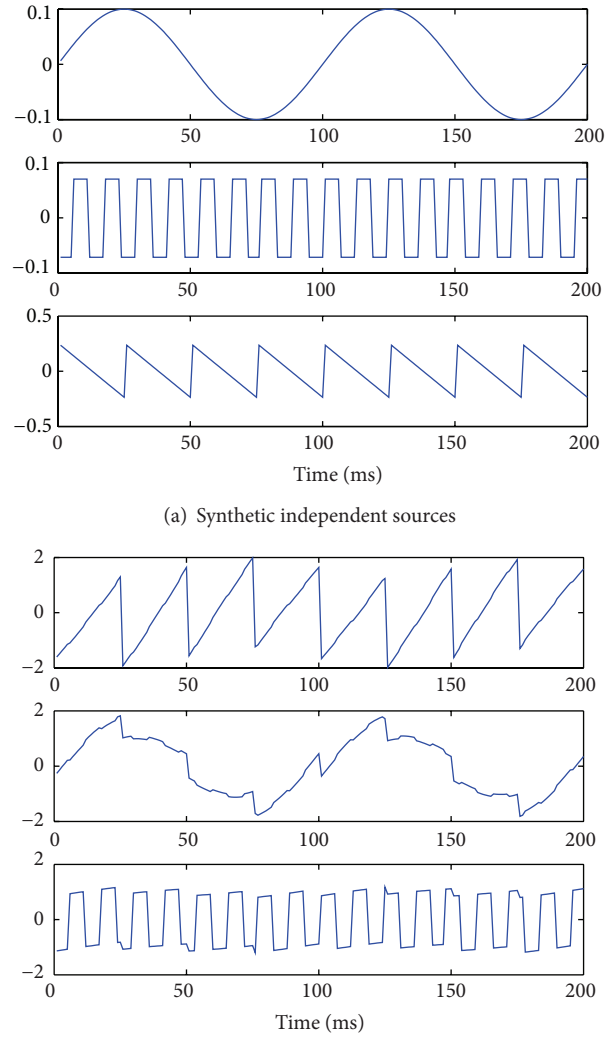
Theorem 1 (Central-Limit Theorem [15]). *Given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed.*

According to the Central-Limit Theorem, if and only if the recovery data Y is a permutation of the original independent sources X, the non-Gaussian random variables reaches the maximum. Consider

$$\begin{aligned} I(Y) &= \sum_{i=1}^{n} H(y_i) - H(X) - \log\det(W) \\ &= \left(-H(X) - \log\det(W)\right) \\ &\quad + \sum_{i=1}^{n} \left(H\left(y_{i\text{gauss}}\right) - J(y_i)\right) \end{aligned} \tag{14}$$

via the normalization of mixing data:

$$E\left\{yy^T\right\} = WE\left\{xx^T\right\}W^T = I, \tag{15}$$



(a) Synthetic independent sources

(b) Recovery data

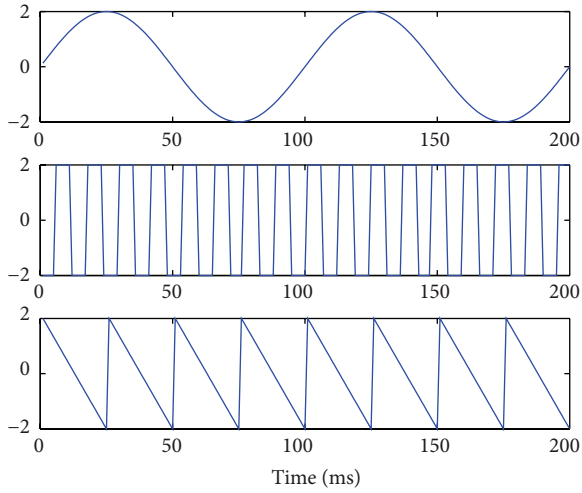
Figure 1: ICA. The synthetic independent data are plotted in (a), and the recovery data are shown in (b) corresponding to the matrix wa in (17). In terms of every column of the matrix wa, the substantial entry, wa_{ij}, is almost a reflection of the transformation between the original data (i) and the recovery data (j) by multiplying the substantial entry wa_{ij}, accompanied by the nonzero entries, $w_{it}, t \neq j$. For example, in the first result of the ICA experiments, $wa_{21} = -0.6804$ is just a proof that the original data (1) is recovered into the recovery data (2) with a multiplicator wa_{21} and some noises based on the minor numbers of wa_{22} and wa_{23}.

where I is the *n_order* identical matrix. So, (10) is rewritten as

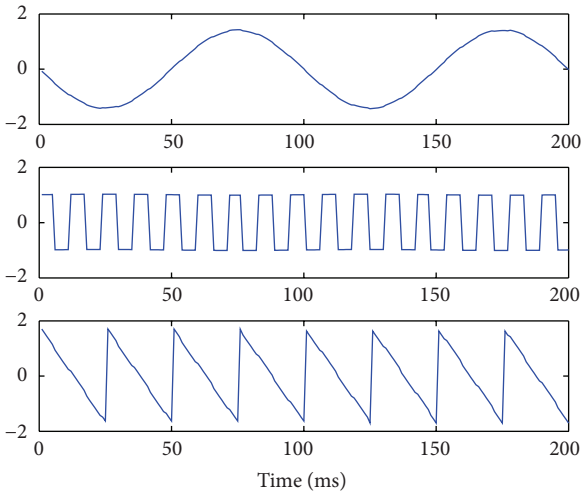
$$\begin{aligned} \max_{Y} \quad & J(Y) \\ \text{s.t.} \quad & \left\|W^2\right\| = 1 \end{aligned} \tag{16}$$

which is equivalent to (7) and (10). So, we can obtain the equivalence of the two classical algorithms in the point of IBN.

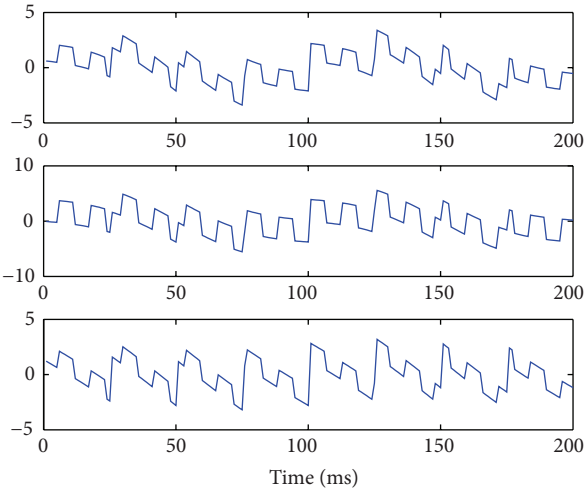
The approximation of negentropy and fixed-point algorithm are applied to derive the learning rule [9].



(a) Independent sources

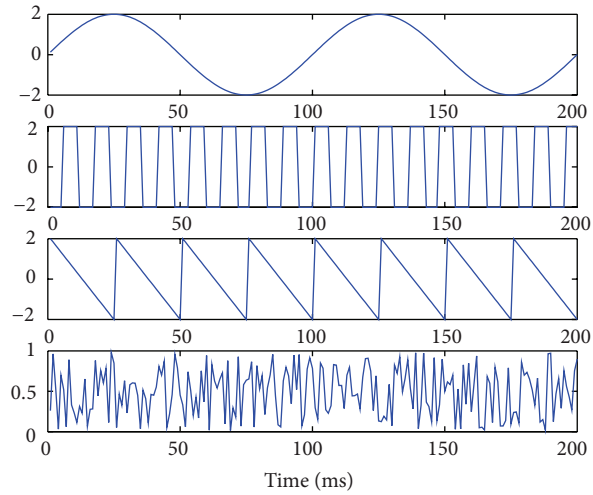


(b) Recovery data

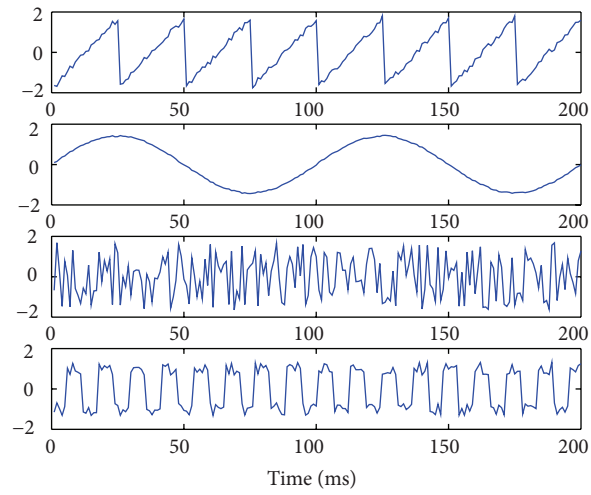


(c) Mixing data

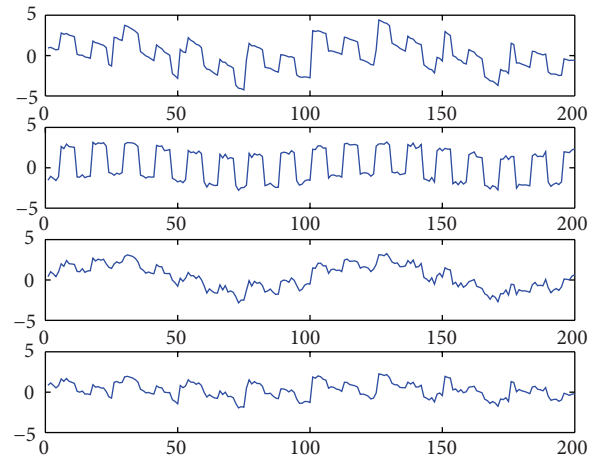
Figure 2: Based on the FastICA algorithm, we separate the randomly mixing data of the sinusoid, the rectangular curve, and the sawtooth curve successfully. At the same time, the product matrix wa of the separation matrix W and the random mixing matrix A is presented in (18).



(a) Synthetic independent source

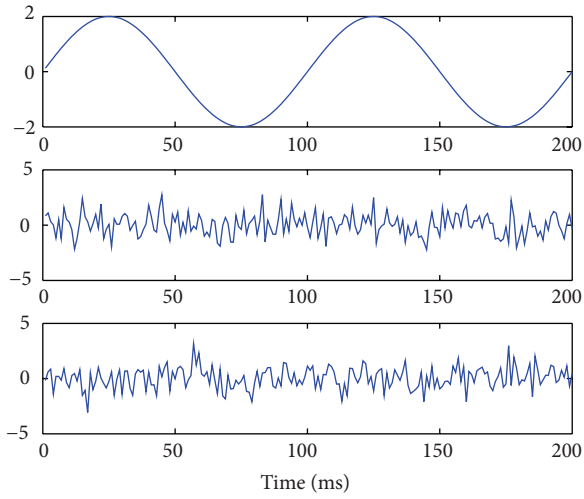


(b) Recovery data

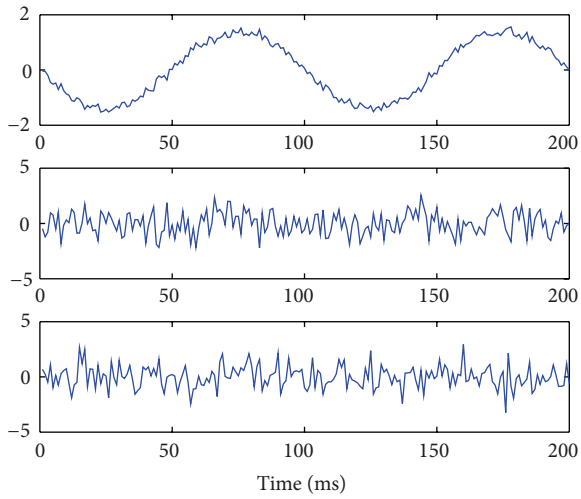


(c) Mixing data

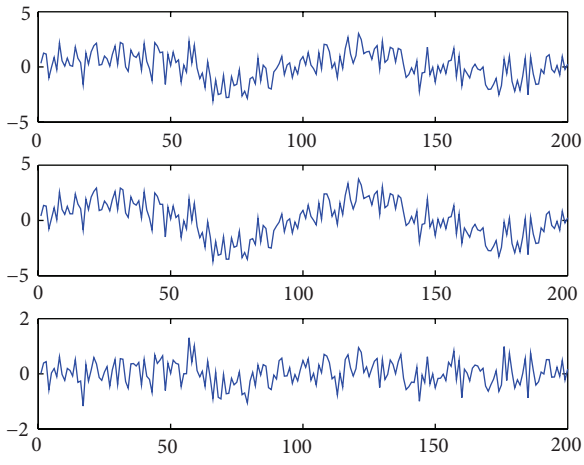
Figure 3: In this numerical experiment, we add the Gaussian into the above experiment and succeed in blindly separating the mixing data. The product matrix wa is given in (19).



(a) Independent sources

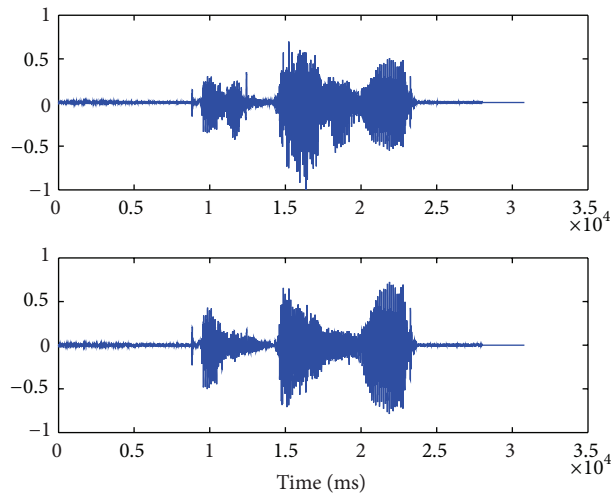


(b) Recovery data

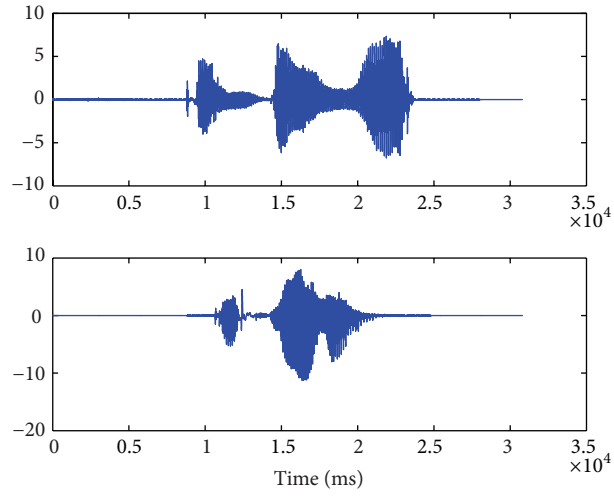


(c) Mixing data

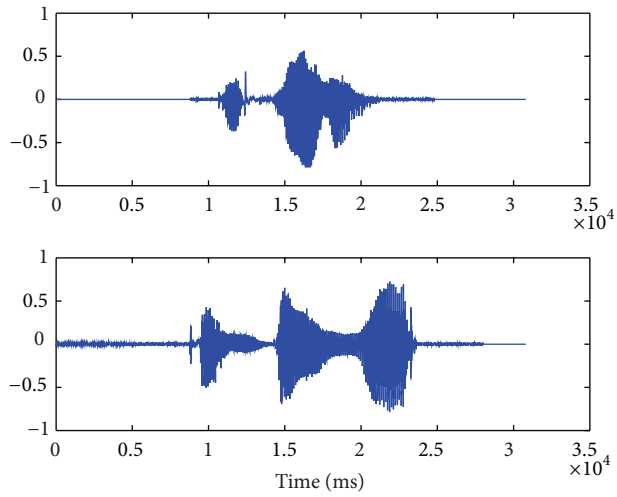
Figure 4: The algorithm is not very efficient in separating the sinusoid from the two Gaussian signals in the mixing data. The more Gaussian variables there are, the more difficult it is to recover the original data.



(a) Mixing data



(b) Recovery data



(c) Independent sources

Figure 5: Real sonic data (FastICA). Using the real sonic data from the website, we also can get the recovery data and the product matrix (20).

(a) Independent images

(b) Mixing images

(c) Recovery images

FIGURE 6: Using the picture from the website and Gaussian noise mixing with the matrix $A = [2, 3; 2, 1]$, the Gaussian noise and the original picture are shown in the first line, the two mixing pictures in the second line, and the recovery pictures in the third line.

Figure 7: The recovery matrix W is not the exact inverse of the mixing matrix A, while the recovery data y has the different orders with s and is very accurately estimated, up to multiplicative signs (FastICA).

4. Experiments

Based on the infomax learning rule, the experiments presented here were obtained using the synthetic data as the original data plotted in Figure 1(a). The result by infomax is listed in Figure 1(b) corresponding to the recovery matrices (17). Obviously, $wa = W * A$ is the product of the recovery matrix and the mixture matrix so that it would be the permutation of the approximate diagonal matrix. Then, we can easily find that only one substantial entry (boxed) exists in each row and column

$$wa = \begin{pmatrix} -0.191 & -0.015 & \boxed{-0.800} \\ \boxed{0.680} & -0.020 & -0.223 \\ 0.023 & \boxed{0.500} & -0.063 \end{pmatrix}. \tag{17}$$

In order to illustrate the efficiency of FastICA algorithm and the limitation on no more than one Gaussian variable, we list some numerical results on the blind mixing signals shown in Figures 2, 3, and 4, using the nonquadratic function G_1 to approximate the negentropy. Consider

$$wa = \begin{pmatrix} \boxed{-0.707} & -0.008 & 0.001 \\ 0.009 & \boxed{-0.500} & 0.000 \\ 0.001 & -0.018 & \boxed{0.833} \end{pmatrix}. \tag{18}$$

Figure 2 is an obvious proof to declare the efficiency of the algorithm separating the randomly mixing data of the sinusoid, the rectangular curve, and the sawtooth curve successfully. And (18) revealed that the matrix W is an elementary transformation of the approximative inverse of the mixing matrix A

$$wa = \begin{pmatrix} -0.00 & 0.03 & \boxed{-0.83} & -0.14 \\ \boxed{0.71} & 0.01 & 0.00 & 0.20 \\ -0.06 & 0.01 & -0.05 & \boxed{1.50} \\ -0.01 & \boxed{0.50} & 0.01 & -0.08 \end{pmatrix}. \tag{19}$$

Table 1: The comparison of the different data on the iterative steps.

Original data set	The iterative steps
Figure 2	Less than 15 steps
Figure 3	Less than 20 steps
Figure 4	Not stationary

Then, it is necessary and meaningful to add the Gaussian variable into the original data to prove the efficiency of the algorithm so that the result is shown in Figure 3 and the product matrix wa is in (19). At last, based on the two Gaussian signals in the mixing data, the algorithm is not efficient to separate the two Gaussian signals apart shown in Figure 4.

Furthermore, the average iterative steps on the first three numerical experiments based on FastICA algorithm are shown in Table 1.

After the experiments on the synthetic data, the algorithm is also efficient on the real sonic data in Figure 5 and image data in Figure 6. In the process of separating the image data, the picture in Figure 7 can be obtained, because the matrix W, which alters the picture in the opposite color, is not the exact inverse of A

$$wa = \begin{pmatrix} -0.376 & \boxed{9.318} \\ \boxed{14.193} & 0.175 \end{pmatrix}, \tag{20}$$

$$wa = \begin{pmatrix} 0.069 & \boxed{0.798} \\ \boxed{0.898} & 0.015 \end{pmatrix}. \tag{21}$$

5. Conclusion

The algorithm of independent component analysis is enlightened from BSS, which is a very successful application of the information theory in speech recognition, image separation without knowing the linear transformation. But, there are also some disadvantages. For example, there exist the strong preconditions that the original data should be independent and the transformation should be linear.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This investigation was supported by National Basic Research Program of China (973 Program) under Grant no. 2013CB329404, the Major Research Project of the National Natural Science Foundation of China under Grant no. 912300101, the National Natural Science Foundation of China under Grant no. 61075006, the Key Project of the National Natural Science Foundation of China under Grant no. 111311006, the Scientific Research Program Funded by Shaanxi Provincial Education Department (Program no. 2013JK1139), the China Postdoctoral Science Foundation (no. 2013M542370), and the Specialized Research Fund for

the Doctoral Program of Higher Education of the People's Republic of China (Grant no. 20136118120010).

References

[1] E. Gokcay and J. C. Principe, "Information theoretic clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 158–171, 2002.

[2] A. Kraskov, H. Stögbauer, R. G. Andrzejak et al., "Hierarchical clustering using mutual information," *Europhysics Letters*, vol. 70, no. 2, pp. 278–284, 2005.

[3] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[4] A. Hyvärinen, J. Hurri, and P. O. Hoyer, *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*, Springer, 2009.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 2012.

[6] H. Zhang, X. Liu, J. Wang et al., "Robust H_∞ sliding mode control with pole placement for a fluid power electrohydraulic actuator (EHA) system," *The International Journal of Advanced Manufacturing Technology*, vol. 73, no. 5-8, pp. 1095–1104, 2014.

[7] H. Zhang, Y. Shi, and J. Wang, "On energy-to-peak filtering for nonuniformly sampled nonlinear systems: a Markovian jump system approach," *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 1, pp. 212–222, 2014.

[8] H. Zhang and J. Wang, "Combined feedback-feedforward tracking control for networked control systems with probabilistic delays," *Journal of the Franklin Institute*, vol. 351, no. 6, pp. 3477–3489, 2014.

[9] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[10] H. Zhang, X. Zhang, and J. Wang, "Robust gain-scheduling energy-to-peak control of vehicle lateral dynamics stabilisation," *Vehicle System Dynamics*, vol. 52, no. 3, pp. 309–340, 2014.

[11] W. Wei and Y. Qi, "Information potential fields navigation in wireless Ad-Hoc sensor networks," *Sensors*, vol. 11, no. 5, pp. 4794–4807, 2011.

[12] W. Wei, P. Shen, Y. Zhang, and L. Zhang, "Information fields navigation with piece-wise polynomial approximation for high-performance OFDM in WSNs," *Mathematical Problems in Engineering*, vol. 2013, Article ID 901509, 9 pages, 2013.

[13] Z. Shuai, H. Zhang, J. Wang et al., "Lateral motion control for four-wheel-independent-drive electric vehicles using optimal torque allocation and dynamic message priority scheduling," *Control Engineering Practice*, vol. 24, pp. 55–66, 2014.

[14] N. Tishby, F. C. Pereira, and W. Bialek, "The information Bottleneck method," in *Proceedings of the 37th annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, September 1999.

[15] J. Rice, *Mathematical Statistics and Data Analysis*, Cengage Learning, 2006.