Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Neurocomputing 61 (2004) 503–512

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Letters

Expectation–Maximization approaches to independent component analysis

Mingjun Zhong[a,b], Huanwen Tang[a], Yiyuan Tang[b,c,d,*]

[a] Institute of Computational Biology and Bioinformatics, Dalian University of Technology, Dalian 116023, People's Republic of China

[b] Institute of Neuroinformatics, Dalian University of Technology, Dalian 116023, People's Republic of China

[c] Laboratory of Visual Information Processing, The Chinese Academy of Sciences, Beijing 100101, People's Republic of China

[d] Key Lab for Mental Health, The Chinese Academy of Sciences, Beijing 100101, People's Republic of China

Available online 20 August 2004

Abstract

Expectation–Maximization (EM) algorithms for independent component analysis are presented in this paper. For super-Gaussian sources, a variational method is employed to develop an EM algorithm in closed form for learning the mixing matrix and inferring the independent components. For sub-Gaussian sources, a symmetrical form of the Pearson mixture model (Neural Comput. 11 (2) (1999) 417–441) is used as the prior, which also enables the development of an EM algorithm in fclosed form for parameter estimation.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Independent component analysis; Overcomplete representations; EM algorithm; Variational method

1. Introduction

Blind source separation (BSS) by independent component analysis (ICA) has received great attention due to its potential signal processing applications

*Corresponding author. Institute of Neuroinformatics, Dalian University of Technology, Dalian 116023, China. Tel.: +86-411-470-6039; fax: +86-411-470-9304.

E-mail addresses: sunxl_zhong@yahoo.com (M. Zhong), yy2100@163.net (Y. Tang).

0925-2312/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2004.06.007

[1,2,4–6,8,10]. In particular, the following ICA model is considered in this paper

$$x = As + \varepsilon, \tag{1}$$

where $A \in R^{N\times M}$ is the mixing matrix, x is an N-dimensional data vector, the elements of s which is an M-dimensional random vector define the independent components, and ε is the noise which is modelled as Gaussian with zero mean and covariance matrix Σ. In ICA, the elements of s are assumed mutually statistically independent denoting that the joint probability distribution of s is factorable, i.e., $p(s) = \prod_{m=1}^{M} p(s_m)$, where p represents the probability density function (p.d.f.). The aim of ICA is as follows: Given T observed data samples $\{x_t\}_{t=1}^{T}$, recover the mixing matrix A, the original source sequences $\{s_t\}_{t=1}^{T}$, and the noise covariance matrix Σ.

Several researchers have proposed various methods for estimating the mixing matrix and the noise covariance matrix, in which the posterior moments are estimated by various approximation techniques [2,5,10,12,13]. In contrast to these methods, this paper presents a combined estimation method for the source signals, the mixing matrix and the noise covariance matrix based on the Expectation–Maximization (EM) algorithm [3]. For super-Gaussian sources, a variational method enables the posterior analytically tractable [4,7], which formulates an EM algorithm in closed form for parameter estimation. For sub-Gaussian sources, the Pearson mixture model [8] is employed to be the source density, which naturally gives an EM algorithm in closed form for parameter estimation.

2. Parameter estimation methods

In ICA, it has been shown that there are only two density models for the source priors, i.e., the super-Gaussian density which has a positive kurtosis and the sub-Gaussian density which has a negative kurtosis.[1] In this section, we will derive EM algorithms in accordance with these two source densities, respectively.

2.1. Super-Gaussian density model

In ICA, a super-Gaussian density which has a positive kurtosis is placed on the independent components. In particular, since the elements of s are assumed mutually statistically independent we employ the following factorable super-Gaussian density as the source model [8]

$$p(s) = \prod_{m=1}^{M} \frac{1}{Z(\beta)} G_{s_m}(0,1)\cosh^{-2/\beta}(\beta s_m), \tag{2}$$

where the notation $G_{s_m}(0,1)$ denotes a normal distribution computed at s_m with zero mean and unit variance, β is a constant and $Z(\beta)$ is the normalizing coefficient irrelevant to s. This prior renders the posterior, i.e., $p(s|x, A, \Sigma)$, analytically

[1] For a scalar random variable y, kurtosis is defined in the zero-mean case by the equation $\mathrm{kurt}(y) = E\{y^4\} - 3(E\{y^2\})^2$ (For more details, see [6].)

intractable. In the following, we will derive a strict lower bound on this prior, which gives a closed form for parameter estimation in an EM framework.

The prior density in Eq. (2) has a strict lower bound over the additional variational parameter $\xi = (\xi_1, \xi_2, \ldots, \xi_M)^T$ such that (see Appendix)

$$p(s) \geqslant p(s|\xi) = \prod_{m=1}^{M} \varphi(\xi_m) G_s(0, \Lambda), \tag{3}$$

where

$$\Lambda = \mathrm{diag}\left(\frac{\xi_1}{\xi_1 + 2\tanh(\beta\xi_1)}, \ldots, \frac{\xi_M}{\xi_M + 2\tanh(\beta\xi_M)}\right).$$

Note that as $\beta \to \infty$, the density in Eq. (2) becomes the following factorable Laplacian:

$$p(s) \propto \prod_{m=1}^{M} G_{s_m}(0,1)\exp(-2|s_m|) \geqslant \prod_{m=1}^{M} \varphi(\xi_m) G_s(0, \Lambda), \tag{4}$$

where

$$\Lambda = \mathrm{diag}\left(\frac{|\xi_1|}{|\xi_1| + 2}, \ldots, \frac{|\xi_M|}{|\xi_M| + 2}\right),$$

which is used as the interesting prior in learning sparse and overcomplete representations [10,4]. It should be noted that the variational lower bound on the prior derived in this paper is similar to the one presented by Girolami [4], who introduced the variational method into learning sparse and overcomplete representations. (For a specific application of the variational method to machine learning, see [7].) Fig. 1 shows ranges of the variances of the lower bounds in the Gaussian form presented in this paper and those proposed by Girolami [4], i.e., the variances of the lower bounds in this paper belong to $[0, 1)$ and the ones presented in [4] belong to $[0, \infty)$, reflecting that the prior employed in this paper has a smaller variance than the one used in [4]. Furthermore, given the previous strict lower bound on the prior, the corresponding posterior can be represented as a strict lower bound in Gaussian form

$$\begin{aligned} p(s|x, A, \Sigma) \geqslant p(s|x, A, \Sigma, \xi) &= \frac{p(x|s, A, \Sigma)p(s|\xi)}{\int p(x|s, A, \Sigma)p(s|\xi)\,\mathrm{d}s} \\ &= \frac{G_x(As, \Sigma)G_s(0, \Lambda)}{\int G_x(As, \Sigma)G_s(0, \Lambda)\,\mathrm{d}s}. \end{aligned} \tag{5}$$

Thus, a normal form of the following variational moments can be easily obtained based on the posterior $p(s|x, A, \Sigma, \xi)$:

$$E\{s|x_t, \xi_t\} = (A^T\Sigma^{-1}A + \Lambda_t^{-1})^{-1}A^T\Sigma^{-1}x_t, \tag{6}$$

$$E\{ss^T|x_t, \xi_t\} = (A^T\Sigma^{-1}A + \Lambda_t^{-1})^{-1} + E\{s|x_t, \xi_t\}E\{s|x_t, \xi_t\}^T. \tag{7}$$

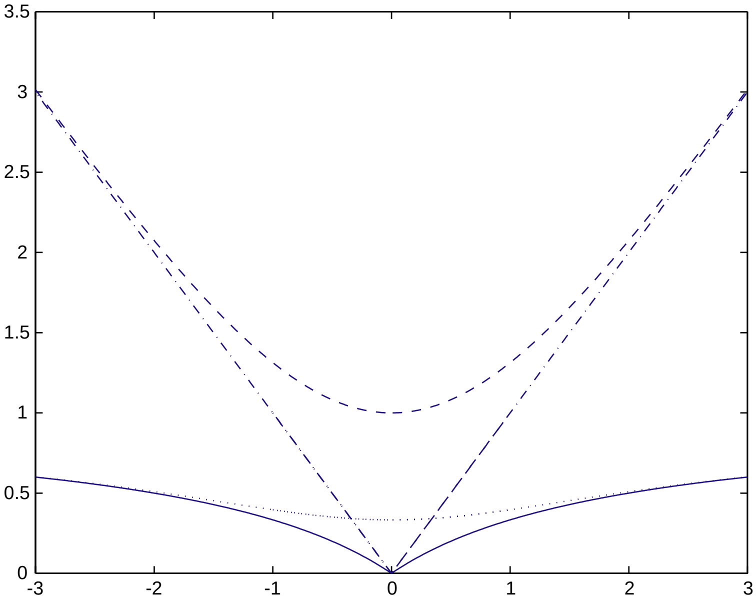

Fig. 1. Variance of the variational lower bound in Gaussian form. The variational parameter ξ runs along the horizontal axis, with the vertical axis giving the value of variance. The plots of the presented variances

$$\sigma^2 = \frac{\xi}{\xi + 2\tanh(\xi)} \quad \text{and} \quad \sigma^2 = \frac{|\xi|}{|\xi| + 2}$$

are, respectively, denoted by the dot line and the solid line, comparing the plots of the variances

$$\sigma^2 = \frac{\xi}{\tanh(\xi)} \quad \text{and} \quad \sigma^2 = |\xi|$$

[4] being, respectively, denoted by the dash line and the dot-dash line.

Now given these conditional moments for the lower bound on the posterior, it is an easy task to derive an EM algorithm in closed form for parameter estimation. For each data sample, there is a corresponding M-dimensional variational parameter associated with it. Given T data samples $\{x_t\}_{t=1}^{T}$, we will estimate the parameters A, Σ and $\{\xi_t\}_{t=1}^{T}$ by maximizing the data log-likelihood

$$\log p(x|A, \Sigma, \xi) = \sum_{t=1}^{T} \log\{p(x_t|A, \Sigma, \xi_t)\}.$$

Conveniently, the standard forms of M-step for the parameters A, Σ and ξ can be represented as (for derivations, see [4])

$$\xi_t^{\text{new}} = \pm\sqrt{\text{diag}[E\{ss^T|x_t, \xi_t\}]}, \tag{8}$$

$$A^{\text{new}} = \left\{ \sum_{t=1}^{T} x_t E\{s|x_t, \xi_t\}^T \right\} \left\{ \sum_{t=1}^{T} E\{ss^T|x_t, \xi_t\} \right\}^{-1}, \tag{9}$$

$$\Sigma^{\text{new}} = \frac{1}{T} \sum_{t=1}^{T} \{x_t x_t^T - A^{\text{new}} E\{s|x_t, \xi_t\} x_t^T\}. \tag{10}$$

Note that

$$M_t = (A^T \Sigma^{-1} A + \Lambda_t^{-1})^{-1} = \Lambda_t - \Lambda_t A^T (A \Lambda_t A^T + \Sigma)^{-1} A \Lambda_t. \tag{11}$$

This transformation of the matrix inversion allows a really hard inverse to be converted into an easy inverse especially when considering BSS of more sources than mixtures, i.e., A has many columns but few rows. Inserting the terms for the variational posterior moments into Eq. (8)–(10) gives the following updates for the parameters A, Σ, and ξ:

$$\xi_t^{\text{new}} = \pm \sqrt{\text{diag}[M_t \{I + A^T \Sigma^{-1} x_t x_t^T \Sigma^{-1} A M_t\}]}, \tag{12}$$

$$A^{\text{new}} = \left\{ \sum_{t=1}^{T} x_t x_t^T \Sigma^{-1} A M_t^T \right\} \left\{ \sum_{t=1}^{T} M_t \{I + A^T \Sigma^{-1} x_t x_t^T \Sigma^{-1} A M_t\} \right\}^{-1}, \tag{13}$$

$$\Sigma^{\text{new}} = \frac{1}{T} \sum_{t=1}^{T} x_t x_t^T - A^{\text{new}} \frac{1}{T} \sum_{t=1}^{T} M_t A^T \Sigma^{-1} x_t x_t^T. \tag{14}$$

Eq. (12) serves to improve the variational data likelihood $p(x|A, \Sigma, \xi)$ to the true data likelihood, and the convergence properties of the EM algorithm [3] ensure that $p(x|A, \Sigma) \geqslant p(x|A, \Sigma, \xi^{\text{new}}) \geqslant p(x|A, \Sigma, \xi^{\text{old}})$.

2.2. Sub-Gaussian density model

For sources with sub-Gaussian densities, the following Pearson mixture model in the univariate case is employed in this paper [8]:

$$p(s) = \tfrac{1}{2}(G_s(\mu, \sigma^2) + G_s(-\mu, \sigma^2)). \tag{15}$$

This mixture model is a symmetric strictly sub-Gaussian density and may serve as a suitable density for computing the score function of symmetric sub-Gaussian sources. Lee et al. [8] had derived a learning rule for complete ICA without additive noise. Actually, this Pearson mixture model makes the posterior analytically tractable and thus it is not required to employ various approximations for it such that

$$p(s|x, A, \Sigma) = \frac{p(x|s, A, \Sigma)p(s)}{\int p(x|s, A, \Sigma)p(s)\,\mathrm{ds}} = \frac{G_x(As, \Sigma)G_s(\mu, \mathrm{V}) + G_x(As, \Sigma)G_s(-\mu, \mathrm{V})}{\int G_x(As, \Sigma)G_s(\mu, \mathrm{V})\,\mathrm{ds} + \int G_x(As, \Sigma)G_s(-\mu, \mathrm{V})\,\mathrm{ds}}, \tag{16}$$

where $\mathrm{V} = \sigma^2\mathrm{I}$. Note that $W = (A^T\Sigma^{-1}A + \mathrm{V}^{-1})^{-1} = \mathrm{V} - \mathrm{V}A^T(A\mathrm{V}A^T + \Sigma)^{-1}A\mathrm{V}$. This posterior conveniently gives the following moments:

$$E\{s|x_t\} = WA^T\Sigma^{-1}x_t, \tag{17}$$

$$E\{ss^T|x_t\} = W + W(A^T\Sigma^{-1}x_tx_t^T\Sigma^{-1}A + \mathrm{V}^{-1}\mu\mu^T\mathrm{V}^{-T})W^T. \tag{18}$$

Inserting these moments into Eqs. (9) and (10) gives the following updates:

$$A^{\mathrm{new}} = \left\{\sum_{t=1}^{T} x_tx_t^T\Sigma^{-1}AW^T\right\} \times \left\{\sum_{t=1}^{T} W\{I + (A^T\Sigma^{-1}x_tx_t^T\Sigma^{-1}A + \mathrm{V}^{-1}\mu\mu^T\mathrm{V}^{-T})W^T\}\right\}^{-1}, \tag{19}$$

$$\Sigma^{\mathrm{new}} = \frac{1}{T}\sum_{t=1}^{T} x_tx_t^T - A^{\mathrm{new}}\frac{1}{T}\sum_{t=1}^{T} WA^T\Sigma^{-1}x_tx_t^T. \tag{20}$$

3. Simulations

To compare this proposed method with Girolami's EM algorithm [4], where the prior was set to be $p(s) = \cosh^{-1}(s)$, the error measure used in [8] was computed in the first experiment for standard BSS with three speech sources (see Fig. 2). The result shows that the performance of the proposed method is similar to Girolami's EM algorithm, though the proposed method slightly increased the convergence speed in this experiment.

The second experiment shows the ability of the proposed EM algorithm to perform blind separation of binary sources with more sources than observations. Fig. 3 shows the result. The complexity of Pajunen's method [11] grows exponentially with the number of the sources, whereas the complexity of this EM algorithm scales as $\mathcal{O}(N^3)$.

The third experiment shows the ability of the method to learn sparse representation for speech data which were obtained from the TIMIT database, using the speech of a single speaker, speaking 10 different example sentences. Speech segments with 64 samples were randomly selected from the 10 speech signals, where each of the speech signals has 16 bits per sample at the sampling frequency of 16 000 HZ. Both a complete (64 basis vectors) and an overcomplete basis (128 basis

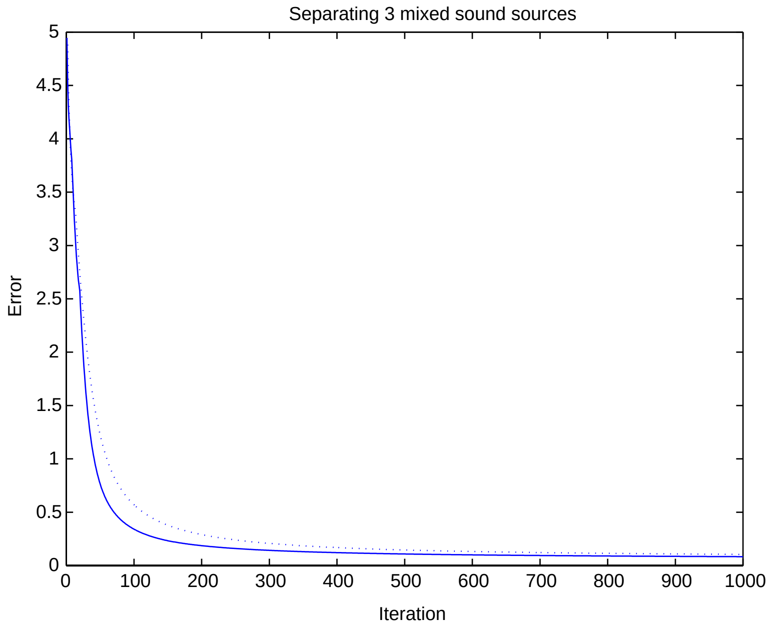

Fig. 2. Error measure [8] for the separation of three sound sources. The solid line plots the performance for the EM algorithm outlined in Eqs. (12), (13) and (14), and the dot line plots the performance for the method proposed by Girolami [4].

vectors) were learned. Fig. 4 shows the learned complete basis vectors which are referred to as the columns of the mixing matrix in the linear ICA model in Eq. (1). This result is similar to those reported in [9,10].

4. Conclusions

This paper has presented a variational method for blind separation of super-Gaussian sources. For sub-Gaussian sources, the Pearson mixture model is employed as the prior, which naturally gives an EM algorithm in closed form for parameter estimation. Simulation results show that the proposed method can perform blind separation of both super-Gaussian and sub-Gaussian sources.

Appendix

For the univariate heavy-tailed distribution in Eq. (2), it is desirous to consider the following log-density:

$$h(s) = \log\{p(s)\} = C - \tfrac{1}{2}s^2 - 2\beta \log\{e^{\beta s} + e^{-\beta s}\}, \tag{21}$$

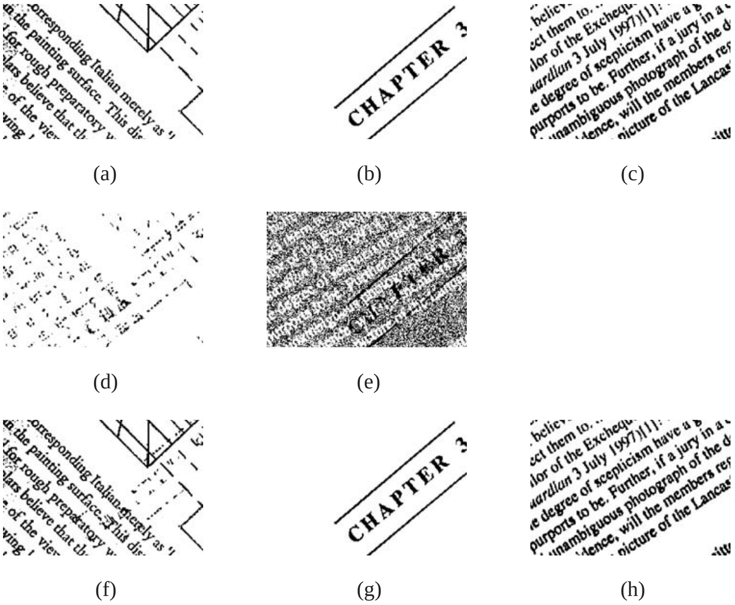

Fig. 3. (a–c) The three original binary images. (d,e) The noisy observations. (f–h) The inferred images using the EM method outlined in Eqs. (19) and (20).

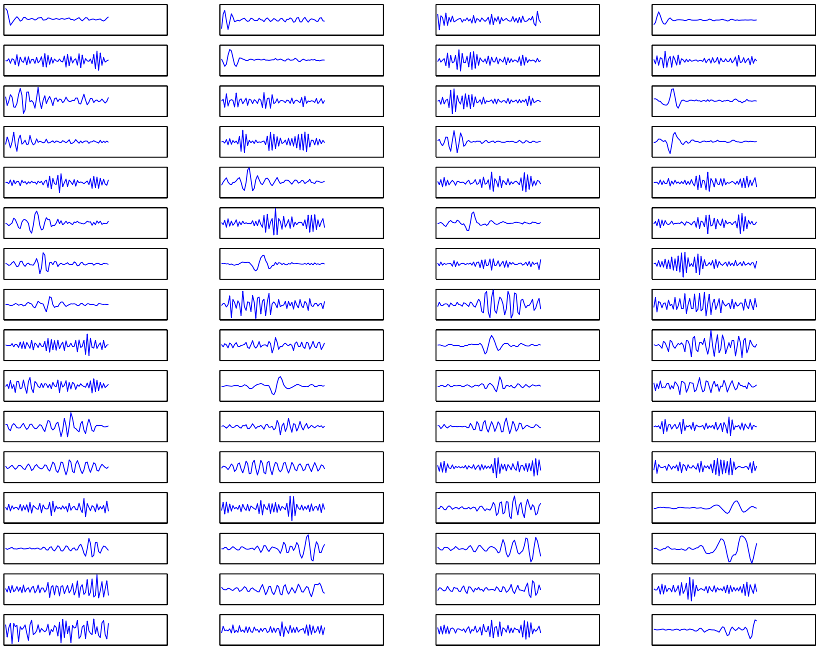

Fig. 4. Sixty-four basis vectors learned from the segments of natural speech which consisted of 64 samples in duration.

where $C = \log\frac{4}{\sqrt{2\pi}}$. It can be seen that this log-prior is convex in s^2 due to $-\log\{e^{\beta s} + e^{-\beta s}\}$ being convex in s^2 [7]. Thus, the variational representation for $h(s)$ has the following form [7]:

$$h(s) = \max_{\xi}\{(s^2 - \xi^2)\nabla_{\xi^2} h(\xi) + h(\xi)\}, \tag{22}$$

where ∇_{ξ^2} denotes the gradient with respect to ξ^2. It should be indicated that the maximum in the above representation is naturally attained for $s^2 = \xi^2$. This conveniently gives the following expression:

$$p(s) = \max_{\xi}[\varphi(\xi)G_s(0, \sigma_{\xi}^2)], \tag{23}$$

where

$$\varphi(\xi) = p(\xi)\exp\left\{\frac{\xi^2 + 2\xi\tanh(\beta\xi)}{2}\right\}\sqrt{2\pi\sigma_{\xi}^2} \quad \text{and} \quad \sigma_{\xi}^2 = \frac{\xi}{\xi + 2\tanh(\beta\xi)}.$$

Acknowledgements

We are indebted to Hongjun Chen for his helpful comments on the manuscript. The work was supported by NSFC (30170321, 90103033), MOST (2001CCA00700) and MOE (KP0302).

References

[1] A.J. Bell, T.J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (6) (1995) 1129–1159.

[2] A. Belouchrani, J.-F. Cardoso, Maximum likelihood source separation by the expectation—maximization technique: deterministic and stochastic implementation, in: Proceedings of NOLTA, 1995, pp. 49–53.

[3] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. B 39 (1) (1977) 1–38.

[4] M. Girolami, A variational method for learning sparse and overcomplete representations, Neural Comput. 13 (11) (2001) 2517–2532.

[5] P.A.d.E.R. Hojen-Sorensen, O. Winther, L.K. Hansen, Mean-field approaches to independent component analysis, Neural Comput. 14 (4) (2002) 889–918.

[6] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, Wiley, New York, 2001.

[7] T.S. Jaakkola, Variational methods for inference and estimation in graphical models, Unpublished Doctoral Dissertation, MIT, 1997.

[8] T.-W. Lee, M. Girolami, T.J. Sejnowski, Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources, Neural Comput. 11 (2) (1999) 417–441.

[9] M.S. Lewicki, Efficient coding of natural sounds, Nat. Neurosci. 5 (4) (2002) 356–363.

[10] M.S. Lewicki, T.J. Sejnowski, Learning overcomplete representations, Neural Comput. 12 (2) (2000) 337–365.

[11] P. Pajunen, Blind separation of binary sources with less sensors than sources, in: Proceedings of 1997 International Conference on Neural Networks, vol. 3, 1997, pp. 1994–1997.

[12] Z.-W. Shi, H.-W. Tang, W.-Y. Liu, Y.-Y. Tang, Blind source separation of more sources than mixtures using generalized exponential mixture models, Neurocomputing (2004), in press.

[13] M.-J. Zhong, H.-W. Tang, H.-J. Chen, Y.-Y. Tang, An EM algorithm for learning sparse and overcomplete representations, Neurocomputing 57 (2004) 469–476.

Mingjun Zhong studied mathematics and obtained his Ph.D. degree at the Dalian University of Technology (China) in 2004. He was working at the Institute of Neuroinformatics and the Institute of Computational Biology and Bioinformatics of the Dalian University of Technology. His research interests include independent component analysis, sparse coding, bioinformatics, and neuroinformatics.

Tang Huanwen graduated from the Dalian University of Technology in 1963, and currently a professor of mathematics and management engineering at the Dalian University of Technology. His research interest includes computational models and algorithm of human cognition and neural information coding, bioinformatics and neuroinformatics. (Home page: http://brain.dlut.edu.cn)

Tang Yiyuan graduated from the Jinlin University in 1987, and currently a professor of neuroinformatics and neuroscience at the Dalian University of Technology. His research interest includes neuroimaging, cognition and emotion interaction, neuroinformatics. (Home page: http://brain.dlut.edu.cn)