4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan

A ONE-BIT-MATCHING THEOREM BASED SIMPLIFIED LPM-ICA ALGORITHM

Zhi-Yong Liu, Kai-Chun Chiu and Lei Xu

Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, P.R. China

ABSTRACT

Although the idea that the model probability density function (pdf) is learned together with the de-mixing matrix $\mathbf{W}$ for independent component analysis (ICA) first proposed by Xu *etc.* (1996) was adopted in the learned parametric mixture based ICA (LPM-ICA) algorithm, theoretical guidance on how to design the learnable density model is not yet mentioned. Recently, the ICA one-bit-matching theorem stating that *all the sources can be separated as long as there is a one-to-one same sign correspondence between the kurtosis signs of all source pdf's and the kurtosis signs of all model pdf's* is theoretically proved under the assumption of zero skewness. In this paper, we propose a simplified LPM-ICA algorithm based on the theorem. Compared with the original algorithm which adopts mixture density as the model pdf, the simplified LPM-ICA has the advantage of improved computational efficiency by using only one free parameter for each model pdf.

1. INTRODUCTION

Independent component analysis (ICA) aims at blindly separating the independent sources $\mathbf{s}$ from their linear mixture $\mathbf{x} = \mathbf{A}\mathbf{s}$ via:

$$\mathbf{y} = \mathbf{W}\mathbf{x}, \mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n, \mathbf{W} \in \mathbb{R}^{m \times n} \qquad (1)$$

The recovered $\mathbf{y}$ is required to be as component-wise independent as possible where independence is defined as

$$q(\mathbf{y}) = \prod_{j=1}^{n} q(y_j). \qquad (2)$$

This effort is supported by the result of [2]. They showed that $\mathbf{y}$ recovers $\mathbf{s}$ up to constant scales and a permutation of components when the components of $\mathbf{y}$ become component-wise independent and at most one of them is gaussian. The problem is further formularized by Comon in 1994 [3] under the name ICA.

Although ICA has been studied from different perspectives [4, 5], in the case that $\mathbf{W}$ is invertible, all such approaches are equivalent to minimizing the following cost function:

$$\mathcal{L}(\mathbf{W}) = \ln \det \mathbf{W} + \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1} \log p_i(y_i(t)) \qquad (3)$$

where $p_i(y_i)$ is the pre-fixed model probability density function (pdf).

Conventionally, the model pdf $p_i(y_i)$ is either pre-fixed as sub-gaussian [5] or super-gaussian [4] according to the source property. However, this approach may not work in case the number of sub-gaussian and super-gaussian sources is not known a priori.

To solve the problem, Xu etc. [1] suggested to let $p_i(y_i)$ to be a flexibly adjustable density that is learned together with $\mathbf{W}$, in help of a parametric model (e.g., a mixture of parametric pdf's). As an example, the idea has been further implemented by the learned parametric mixture based ICA (LPM-ICA) algorithm [6], with successful results on the sources that can be either sub-gaussian or super-gaussian as well as any combination of the both types. However, the computational load the LPM-ICA is heavy because the mixture model usually introduces rather many free parameter for each model pdf. For instance, it has to estimate $3k - 1$ free parameters for each model pdf if the gaussian mixture is adopted by LPM-ICA algorithm, where k denotes the number of components for the mixture model.

Interestingly, it was also found that the model density $p_i(y_i)$, or cumulative distribution function (cdf), only needs to be learned loosely instead of precisely. For instance, a simple sigmoid function such as $tanh(x)$ seems to work well on the super-gaussian sources [4], and a mixture of only two or three gaussians may be enough already [6, 7] for the mixed sub- and super-gaussian sources. It led to the so-called one-bit-matching conjecture [7], which is usually stated as "*all the sources can be separated as long as there is a one-to-one same-sign-correspondence between the kurtosis signs of all source pdf's and the kurtosis signs of all model pdf's*". This conjecture has also been implied by some subsequent ICA studies [8, 9].

In literature, there were several theoretical studies closely related to the one-bit-matching conjecture [10, 11]. A proof [10] was given for the case of two sub-gaussian sources. However, it cannot be extended to a model either with more than two sources, or with mixed sub- and super-gaussian sources. The conditions for certain nonlinear function $\varphi_i(y_i) = -\frac{d}{dy_i}\log p_i(y_i)$ with stable and correct solutions were also theoretically studied [11]. However, it did not hint at the circumstances under which the algorithm would converge to a successful solution. Recently, Liu etc. [12] theoretically proved the conjecture under the assumption of zero skewness for both source and model pdf's.

As a by-product, the one-bit-matching theorem motivates us on the simplification of model pdf design [1]. As there is only one bit of information, i.e., the sign of kurtosis corresponding to either sub- or super-gaussian density, needed to be learned, we expect that only one free parameter is enough for each model pdf.

In this paper, under the guidance of the one-bit-matching theorem we propose a simplified LPM-ICA algorithm with only one free parameter that can smoothly switch between sub- and super-gaussian. This can be contrasted with the heuristic soft switching algorithm [13] which is given without any theoretical basis.

Remainder of this paper is organized in the following manner. The one-bit-matching theorem and LPM-ICA algorithm are briefly reviewed in section 2 and 3 respectively. A simplified LPM-ICA algorithm is introduced in section 4, followed by the experimental demonstration in section 5. Section 6 is devoted to a discussion on issue related to learning rates. Section 7 concludes this paper.

2. THE ONE-BIT-MATCHING THEOREM

The one-bit-matching theorem proved in [12] is shown below.

Theorem 2.1 *Provided that the skewness of source and model pdf's are both zero. All the sources can be separated as long as there is a one-to-one same sign correspondence between the kurtosis signs of all source pdf's and the kurtosis signs of all model pdf's.*

3. BRIEF REVIEW ON THE LPM-ICA ALGORITHM

The LPM-ICA approach models each marginal pdf $p_i(y_i)$ via a mixture density, such as the gaussian mixture:

$$p_i(y_i|\xi_\mathbf{i}) = \sum_{j=1}^{k} G(y_i|m_{ij},\sigma_{ij}^2)\alpha_{ij} \qquad (4)$$

where $\alpha_{ij} > 0$, $\sum_{j=1}^{k}\alpha_{ij} = 1$, $G(y_i|m_{ij},\sigma_{ij}^2)$ denotes a gaussian pdf with mean m_{ij} and variance σ_{ij}^2. The set of free parameters to be learned is $\Xi = \{\xi_\mathbf{1},...,\xi_\mathbf{m}\}$ with $\xi_\mathbf{i} = \{m_{ij},\sigma_{ij},\alpha_{ij}\}_{j=1}^{k}$. The algorithm framework given in [6] is as follows:

step 1: fix Ξ, update $\mathbf{W}$ along the natural gradient [5]:

$$\mathbf{W}^{new} = \mathbf{W}^{old} + \eta(\mathbf{I} + \phi(\mathbf{y})\mathbf{y}^T)\mathbf{W} \qquad (5)$$

where $\phi(\mathbf{y}) = [\phi_1(y_1),...,\phi_k(y_k)]^T$, $\phi_i(y_i) = \frac{d\ln p_i(y_i|\theta)}{dy_i}$, η is learning rate.

step 2: fix $\mathbf{W}$, update Ξ:

$$\xi_\mathbf{i}{}^{new} = \xi_\mathbf{i}{}^{old} + \zeta\Delta\xi_\mathbf{i} \qquad (6)$$

where $\Delta\xi_\mathbf{i}$ denotes the derivative of

$$\ln\sum_{j=1}^{k} G(y_i|m_{ij},\sigma_{ij}^2)\alpha_{ij}$$

with respective to $\xi_\mathbf{i}$ and ζ is learning rate.

4. THE SIMPLIFIED LPM-ICA ALGORITHM

In this section, by adopting the theorem as a necessary condition, we design a simplified LPM-ICA algorithm in help of gaussian mixture with only one parameter that can smoothly switch between sub- and super-gaussian. Thus, we first obtain the higher order statistics of gaussian mixture.

4.1. Higher Order Statistics of Gaussian Mixture

The success of LPM-ICA for separating mixed sub- and super-gaussian sources implies that gaussian mixture can behave as a sub-gaussian or a super-gaussian via adapting its kurtosis κ. Normally $\kappa > 0$ implies super-gaussian while $\kappa < 0$ implies sub-gaussian. To derive the higher order statistics of gaussian mixture, we first obtain its moment generating function (mgf) [14] via

$$\begin{aligned}\varphi(\tau) &= \int_{-\infty}^{+\infty} e^{i\tau y}p(y)dy \\ &= \int_{-\infty}^{+\infty} e^{i\tau y}\sum_{j=1}^{k}\alpha_j G(y|m_j,\sigma_j^2)dy \\ &= \sum_{j=1}^{k}\alpha_j \exp\left\{\tau m_j i - \frac{\tau^2\sigma_j^2}{2}\right\} \qquad (7)\end{aligned}$$

where $i \triangleq \sqrt{-1}$.

Based on the cumulant generating function (cgf) $\phi(\tau) = \ln(\varphi(\tau))$, we then compute the cumulants c_n by

$$c_n = (-i)^n \frac{d^n \phi(\tau)}{d\tau^n} \tag{8}$$

Specifically:

$$\begin{aligned} c_1 &= m_1 = \mu_{10} \\ c_2 &= m_2 = \mu_{20} + \mu_{02} - \mu_{10}^2 \\ c_3 &= m_3 = \mu_{30} + 3\mu_{12} + 2\mu_{10}^3 - 3\mu_{10}\mu_{02} - 3\mu_{10}\mu_{20} \\ c_4 &= m_4 - 3m_2^2 \\ &= \mu_{40} + 6\mu_{22} + 3\mu_{04} + 12\mu_{10}^2\mu_{02} + 12\mu_{10}^2\mu_{20} \\ &\quad -12\mu_{10}\mu_{12} - 4\mu_{10}\mu_{30} - 3\mu_{02}^2 - 3\mu_{20}^2 \\ &\quad -6\mu_{02}\mu_{20} - 6\mu_{10}^4 \end{aligned} \tag{9}$$

where m_n refer to the n-order moment and $\mu_{pq} \triangleq \sum_{j=1}^{k} \alpha_j m_j^p \sigma_j^q$. Actually, c_1, c_2, c_3, c_4 respectively represent the mean, variance, skewness, and kurtosis of the gaussian mixture.

4.2. The Algorithm

The gaussian mixture based LPM-ICA involves estimating three free parameters for each model pdf. However, according to the one-bit-matching theorem, there is only one bit of information needed to be specified for the pdf. Below we discuss how we can take advantage of the one-bit-matching theorem to simplify the original LPM-ICA algorithm.

Based on the assumption of the one-bit-matching theorem, the absolute skewness of the designed density function should be as small as possible. Based on the higher-order statistics of gaussian mixture we design the following pdf with only one parameter:

$$p(y|\theta) = \frac{1}{3}G(y|-\theta,1) + \frac{1}{3}G(y|0,4) + \frac{1}{3}G(y|\theta,1); 0 \le \theta \le 2 \tag{10}$$

It follows that

$$c_3 = 0 \tag{11}$$

$$c_4 = -\frac{2}{3}\theta^4 - 4\theta^2 + 6 \tag{12}$$

The changes of c_3 and c_4 as θ varies are shown in Fig. 1. Note with a zero skewness how the density changes smoothly from super-gaussian to sub-gaussian as θ varies from 0 to 2. Fig. 2 gives a pictorial view of the density $p(y)$ versus θ.

According to the two steps for LPM-ICA, the algorithm for the simplified LPM-ICA is as follows:

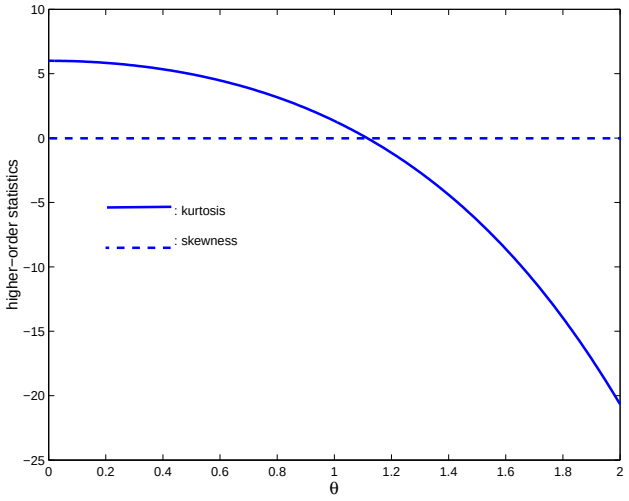

Figure 1: skewness and kurtosis vs θ

1. Fix $\mathbf{\Theta} = [\theta_1, \theta_2, ..., \theta_m]^T$, update $\mathbf{W}$ according to (5), with

$$\phi_i(y_i) = -r_{i1}(y_i + \theta_i) - \frac{r_{i2}y_i}{4} - r_{i3}(y_i - \theta_i) \tag{13}$$

where $r_{i1} = \frac{G(y_i|-\theta_i,1)}{3p(y_i|\theta_i)}$, $r_{i2} = \frac{G(y_i|0,4)}{3p(y_i|\theta_i)}$, $r_{i3} = \frac{G(y_i|\theta_i,1)}{3p(y_i|\theta_i)}$.

2. Fix $\mathbf{W}$, update $\mathbf{\Theta}$:

$$\mathbf{\Theta}^{new} = \mathbf{\Theta}^{old} + \zeta\Delta\mathbf{\Theta} \tag{14}$$

where $\Delta\mathbf{\Theta} = [\Delta\theta_1, \Delta\theta_2, ..., \Delta\theta_m]^T$, with

$$\Delta\theta_i = -r_{i1}(y_i + \theta_i) + r_{i3}(y_i - \theta_i) \tag{15}$$

$\theta_i = \frac{2}{1+\exp(\varsigma_i)}$ is introduced to constrain $0 \le \theta \le 2$.

5. EXPERIMENTAL ILLUSTRATION

In the section, we use two experiments to illustrate the simplified LPM-ICA algorithm on synthetic and real data respectively.

5.1. On Synthetical Data

The 200 data samples used in this experiment are mixed from four sources: two sub-gaussian sources from uniform distribution and two super-gaussian sources from $\sinh(x)$ with x from standard gaussian distribution. The algorithm accuracy performance is measured by the following error metric [5]:

$$\begin{aligned} E &= \sum_{i=1}^{d}\left(\sum_{j=1}^{d}\frac{\mid p_{ij} \mid}{\max_k \mid p_{ik} \mid} - 1\right) \\ &\quad + \sum_{j=1}^{d}\left(\sum_{i=1}^{d}\frac{\mid p_{ij} \mid}{\max_k \mid p_{kj} \mid} - 1\right) \end{aligned} \tag{16}$$

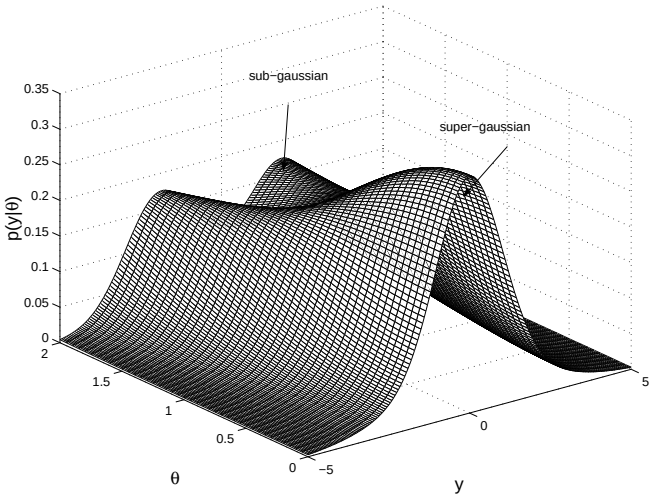

Figure 2: Density variation with θ

where $\mathbf{P} \triangleq \mathbf{WA}$.

The learning processes for the four pdf's are shown in Fig. 3 and the corresponding error metric is shown in Fig. 6. When the algorithms converged, the matrix $\mathbf{P} = \mathbf{WA}$ are

$$\mathbf{P} = \begin{pmatrix} -0.0054 & -0.0280 & 0.0220 & \mathbf{1.5889} \\ 0.0026 & \mathbf{2.0626} & -0.0078 & 0.0110 \\ \mathbf{2.0752} & 0.0202 & 0.0521 & -0.0675 \\ -0.0168 & 0.0197 & \mathbf{1.5422} & -0.0830 \end{pmatrix}$$

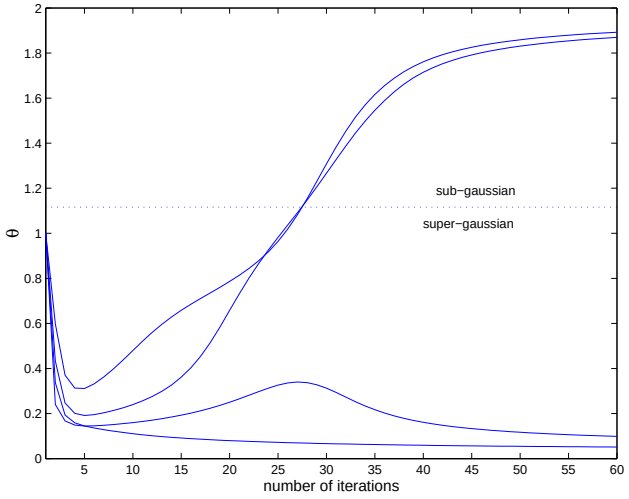

Figure 3: Learning curve of the θ for the simplified LPM-ICA algorithm

From Fig. 3 we can notice that the algorithm automatically detect the source properties via the learned θ,

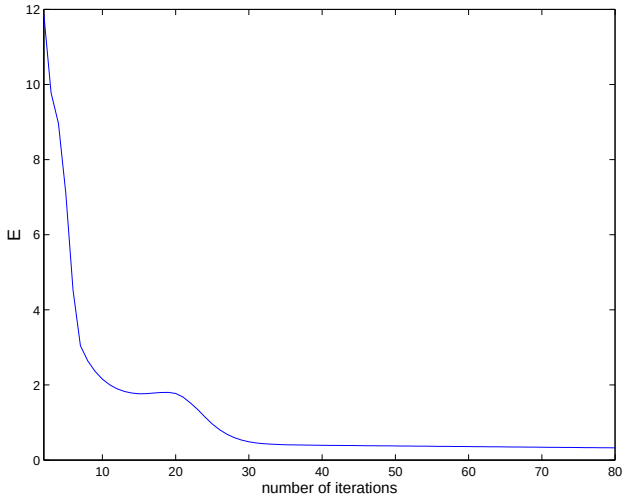

Figure 4: Learning curve of the error metric

and thus results in the successful recovery as witnessed by Fig. 6 and matrix $\mathbf{P}$.

5.2. On Real Data

The purpose of this experiment is to test the simplified LPM-ICA algorithm on real data. The 50,000 10-dimensional samples are linearly mixed from 10 source signals, of which 7 are super-gaussian speech tracks and 3 are synthetic sub-gaussian signals[1].

The variances of θ and error metric E along with the number iterations are shown in Fig. 5 and Fig. 6 respectively, and the matrix $\mathbf{P} = \mathbf{WA}$ after convergence is shown in Fig. 7. Fig. 5 illustrates how the algorithm automatically learns the source properties, and it is evident by direct inspection of $\mathbf{P}$ and Fig. 6 that the algorithm succeeds in recovering the original sources.

6. ISSUE RELATED TO THE CHOICE OF LEARNING RATES

The theoretical basis of the simplified LPM-ICA algorithm comes from the one-bit-matching theorem. However, the precondition of one-to-one same sign correspondence for the one-bit-matching theorem cannot be guaranteed for the algorithm. Actually, such a precondition is what the simplified LPM-ICA needs to learn to adapt to.

The simplified LPM-ICA algorithm introduced above involves two critical learning rates, i.e., η for $\mathbf{W}$ and ζ for θ. In fact, improperly set learning rates for the

[1]Data can be downloaded from http://sweat.cs.unm.edu/~bap/domos.html

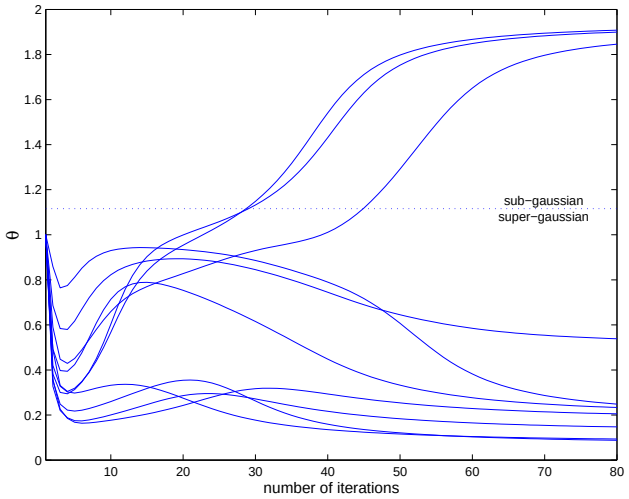

Figure 5: Learning curve of the θ for the simplified LPM-ICA algorithm

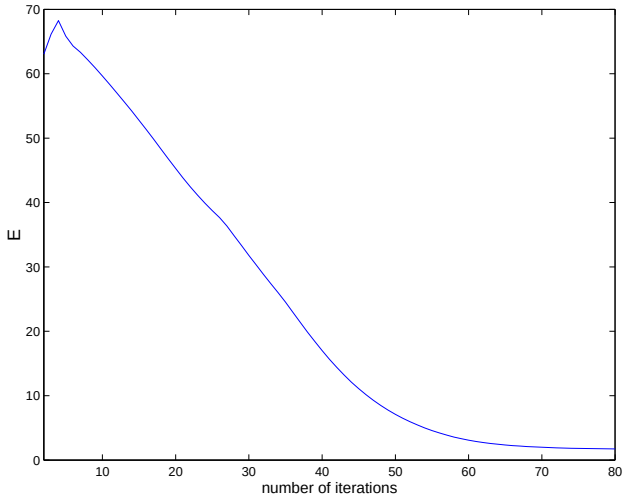

Figure 6: Learning curve of the error metric

algorithm may lead to violation of the precondition of the one-bit-matching theorem and consequently ruin the task of source separation.

Consider a typical task of separating two sub-gaussian sources. Assume the learning rate for $\mathbf{W}$ is much slower than θ such that $\eta \ll \zeta$. Also assume the initial $\mathbf{W}$ causes the two outputs y_1 and y_2 to be super-gaussian. As the simplified LPM-ICA algorithm seeks to maximize the likelihood between y_i and model density $p_i(y_i)$ subject to certain structure constraint, this would cause each model density $(p_i(y_i))$ to be super-gaussian as well and thus violates the precondition of the one-bit-matching theorem.

In fact, it is simple to show that successful separation is not guaranteed if the precondition no longer holds. Continue with the above example. Denote the two-dimensional orthonormal matrix $\mathbf{R}=\begin{pmatrix} \cos\theta, & -\sin\theta \\ \sin\theta, & \cos\theta \end{pmatrix}$ with only one parameter, and denote $k_{s_1} = \alpha k_{s_2} < 0$, $k'_{m_1} = \beta k'_{m_2} > 0$, where, according to the assumption, $\alpha, \beta > 0$, k_{s_i} denotes the kurtosis of source i and k'_{m_j} denotes a variable with the same sign as the kurtosis k_{m_j} of model pdf j. Then, according to the proof of the one-bit-matching theorem [12], minimizing

$$D = -H(\mathbf{y}) - \sum_{i=1}^{n} \int p_{y_i}(y_i) \log p_i(y_i) dy_i \qquad (17)$$

where $H(\mathbf{y}) = -\int p(\mathbf{y}) \log p(\mathbf{y}) d\mathbf{y}$ is the entropy of $\mathbf{y}$, and $p_i(y_i)$ is the pre-fixed model pdf, can be written as minimizing the following cost function

$$\hat{J}(\theta) = \cos^4\theta(1+\alpha\beta) + \sin^4\theta(\alpha+\beta) \qquad (18)$$

due to opposite signs between k'_m and k_s. The first-order and second-order derivatives of (18) with respect to θ can be obtained as follows:

$$\frac{d\hat{J}}{d\theta} = 4\sin^3\theta\cos\theta(\alpha+\beta) - 4\cos^3\theta\sin\theta(1+\alpha\beta) \qquad (19)$$

$$\frac{d^2\hat{J}}{d\theta^2} = 4(\alpha+\beta)(3\sin^2\theta\cos^2\theta - \sin^4\theta) - 4(1+\alpha\beta)(\cos^4\theta - 3\cos^2\theta\sin^2\theta) \qquad (20)$$

From (19) we can get the three local extrema:

$$\sin\theta = 0 \Rightarrow \frac{d^2\hat{J}}{d\theta^2} = -4(1+\alpha\beta) \qquad (21)$$

$$\cos\theta = 0 \Rightarrow \frac{d^2\hat{J}}{d\theta^2} = -4(\alpha+\beta) \qquad (22)$$

$$\tan^2\theta = \frac{1+\alpha\beta}{\alpha+\beta} \Rightarrow \frac{d^2\hat{J}}{d\theta^2} = \frac{8(1+\alpha\beta)(\alpha+\beta)}{1+\alpha+\beta+\alpha\beta} \qquad (23)$$

Since $\alpha > 0, \beta > 0$, (21) and (22) imply two local maxima, and (23) is a local minimum. Thus, minimizing (18) results in θ in the form of (23), which is not correct for arbitrary $\alpha > 0, \beta > 0$.

In short, improperly set learning rates may lead to violation of the precondition of the one-bit-matching theorem and thus lead to unsuccessful source separation. We find experimentally $\zeta \approx 50\eta$ a suitable combination. Of course, the more free parameters, the more difficult finding suitable learning rate combination. The effect of learning rates on separation performance should not be overlooked.

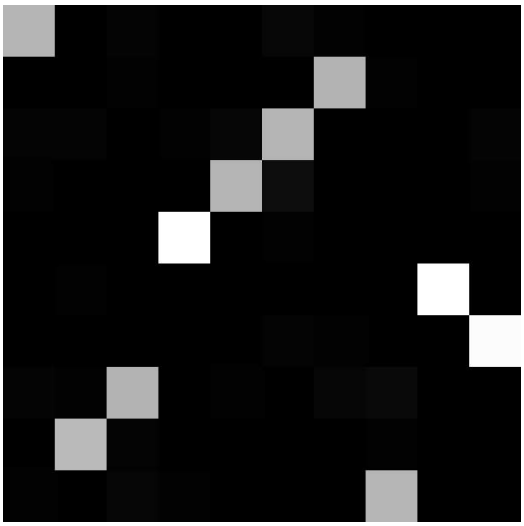

Figure 7: the resulted $P = WA$ matrix for the real-world data

7. CONCLUSION

In this paper, under the guidance of the recently proved one-bit-matching theorem for ICA, we proposed a simplified LPM-ICA algorithm that can smoothly switch between sub- and super-gaussian via only one free parameter. Empirical performance of the algorithm is studied via two experiments with success. The issue with regard to the choice of learning rates is also discussed.

8. REFERENCES

[1] L. Xu, H. H. Yang, and S. I. Amari, "Signal source separation by mixtures accumulative distribution functions or mixutre of bell-shape density distribution functions," *Research Proposal, Presented at FRONTIER FORUM, organized by Amari, S. I. etc.*, 1996.

[2] L. Tong, Y. Inouye, and R. Liu, "Waveform-preserving blind estimation of multiple independent sources," *IEEE Trans. on Signal Processing*, vol. 41, pp. 2461–2470, 1993.

[3] P. Comon, "Independent component analysis: a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 19943.

[4] A.J. Bell and T.J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[5] S.I. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind separation of sources," *Advances in Neural Information Processing*, vol. 8, pp. 757–763, 1996.

[6] L. Xu, C. C. Cheung, and S. I. Amari, "Learned parametric mixture based ica algorithm," *Neurocomputing*, vol. 22, pp. 69–80, 1998.

[7] L. Xu, C. C. Cheung, and S. I. Amari, "Furthere results on nonlinearity and separtion capability of a liner mixture ica method and learned lpm," *Proceedings of the I&ANN'98, Editor: C. Fyfe*, pp. 39–45, 1998.

[8] R. Everson and S. Roberts, "Independent component analysis: A flexible nonlinearity and decorrelating manifold approach," *Neural Computation*, vol. 11, pp. 1957–1983, 1999.

[9] T. W. Lee, M. Girolami, and T. J. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources," *Neural Computation*, vol. 11, pp. 417–441, 1999.

[10] L. Xu, C. C. Cheung, J. Ruan, and S. I. Amari, "Nonlinearity and separation capability: Futher justification for the ica algorithm with a learned mixture of parametric densities," *Proc. ESANN97*, pp. 291–296, 1997.

[11] S.I. Amari, T. P. Chen, and A. Cichocki, "Stability analysis of adaptive blind source separation," *Neural Networks Letter*, vol. 10, pp. 1345–1351, 1997.

[12] Z. Y. Liu, K. C. Chiu, and L. Xu, "Independent component analysis: the one-bit-matching conjecture and a simplified lpm-ica algorithm," *to appear in Advances in Data Mining and Modeling*, World Scientific, 2003.

[13] M. Welling and M. Weber, "A constrained em algorithm for independent component analysis," *Neural Computation*, vol. 13, pp. 677–689, 2001.

[14] A. Stuart and J.K. Ord, *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*, Edward Arnold, 1994.