NOTE **Communicated by Michael Lewicki**

A Note on Lewicki-Sejnowski Gradient for Learning Overcomplete Representations

Zhaoshui He
zhshhe@scut.edu.cn
School of Electronics and Information Engineering, South China University of Technology, Guangzhou, 510640, China, and Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Wako-shi, Saitama 351-0198, Japan

Shengli Xie
adshlxie@scut.edu.cn
School of Electronics and Information Engineering, South China University of Technology, Guangzhou, 510640, China

Liqing Zhang
zhang-lq@cs.sjtu.edu.cn
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China

Andrzej Cichocki
cia@brain.riken.jp
Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Wako-shi, Saitama 351-0198, Japan; System Research Institute, Polish Academy of Sciences (PAN), Warsaw, Poland; and Department of Electrical Engineering, Warsaw University of Technology, Warsaw, Poland

Overcomplete representations have greater robustness in noise environment and also have greater flexibility in matching structure in the data. Lewicki and Sejnowski (2000) proposed an efficient extended natural gradient for learning the overcomplete basis and developed an overcomplete representation approach. However, they derived their gradient by many approximations, and their proof is very complicated. To give a stronger theoretical basis, we provide a brief and more rigorous mathematical proof for this gradient in this note. In addition, we propose a more robust constrained Lewicki-Sejnowski gradient.

1 Introduction

Overcomplete coding, viewed as a generalization of independent component analysis (ICA), is a powerful tool in, for example, signal processing, data processing, and neural information processing (Lewicki &

Sejnowski, 2000; Girolami, 2001). Overcomplete coding not only provides more robust representations in the presence of noise, but also can be more flexible in matching the structure in the data. Lewicki and Sejnowski proposed a very efficient natural gradient (extension) for learning the overcomplete basis (or dictionary) and developed an overcomplete representation approach in 2000 (Lewicki & Sejnowski, 2000). Then Lee, Lewicki, Girolami, and Sejnowski (1999) successfully performed blind separation (BSS) of more sources than mixtures based on such overcomplete representations. However, they obtained their gradient by a series of approximations. Furthermore, their proof is very complicated and mathematically not very rigorous in some degree. In this note, we present a brief and rigorous mathematical proof for the Lewicki-Sejnowski gradient. Moreover, we propose the constrained Lewicki-Sejnowski gradient, which is more robust than the conventional Lewicki-Sejnowski gradient.

The basis learning model can be described as follows (Lewicki & Sejnowski, 2000; Girolami, 2001):

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}, \tag{1.1}$$

where the data vector $\boldsymbol{x} = (x_1, \ldots, x_m)^T$ is known, $\boldsymbol{s} = (s_1, \ldots, s_n)^T$ is unknown, and the $m \times n$ basis matrix $\boldsymbol{A}$ is also unknown. We assume that $s_1, \ldots, s_n$ are statistically independent. When $m < n$, the basis matrix $\boldsymbol{A}$ is overcomplete. The aim is to find a solution to equation 1.1 under some reasonable constraint(s) on $\boldsymbol{s}$ (Lewicki & Sejnowski, 2000).

2 The New Mathematical Proof of the Lewicki-Sejnowski Gradient

By maximizing the posterior distribution of $\boldsymbol{x}$, learning basis matrix $\boldsymbol{A}$ can be converted to solve the following optimization problem (Lewicki & Sejnowski, 2000):

$$\begin{cases} \max L(\boldsymbol{A}, \boldsymbol{s}) = \max\limits_{\boldsymbol{A},\boldsymbol{s}}\{\log[p(\boldsymbol{x})]\} = \max\limits_{\boldsymbol{A},\boldsymbol{s}}\{\log[p(x_1, \ldots, x_m)]\}, \\ \text{subject to: } \boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}, \end{cases} \tag{2.1}$$

where $p(\boldsymbol{x}) = p(x_1, \ldots, x_m)$ is the joint probability density function (PDF) of observation vector $\boldsymbol{x}$.

For convenience, the Lewicki-Sejnowski gradient learning can be expressed as the following theorem:

Theorem 1 *(Lewicki & Sejnowski, 2000). For optimization problem 2.1, the Lewicki-Sejnowski gradient $\Delta\boldsymbol{A}$ can be used to learn the basis matrix $\boldsymbol{A}$. $\Delta\boldsymbol{A}$ is given by*

$$\Delta\boldsymbol{A} = -\boldsymbol{A} \cdot [\phi(\boldsymbol{s}) \cdot \boldsymbol{s}^T + \boldsymbol{I}], \tag{2.2}$$

where $\phi(\boldsymbol{s}) = (\phi(s_1), \ldots, \phi(s_n))^T$, $\phi(s_k) = \partial \log p(s_k)/\partial s_k$, $k = 1, \ldots, n$ *(Lewicki & Sejnowski, 2000). And the learning rule can be*

$$\boldsymbol{A} = \boldsymbol{A} + \mu \Delta \boldsymbol{A}, \tag{2.3}$$

where step size $\mu > 0$.

Proof. We prove this theorem in two cases: (1) $m = n$, matrix $\boldsymbol{A}$ is square and nonsingular, and (2) $m < n$, $\boldsymbol{A}$ is overcomplete.

Case 1: $m = n$, matrix $\boldsymbol{A}$ is square and nonsingular. In this case, $\boldsymbol{A}$ is invertible. Denote $\boldsymbol{W}$ is the inverse of $\boldsymbol{A}$: $\boldsymbol{W} = \boldsymbol{A}^{-1}$. So $\boldsymbol{s} = \boldsymbol{W} \cdot \boldsymbol{x}$. Then the joint PDF of $\boldsymbol{s}$ should be $p(\boldsymbol{x}) = |\det(\boldsymbol{W})| \cdot p(\mathbf{s})$. So optimization problem 2.1 is equivalent to the following problem:

$$\begin{cases} \max L(\boldsymbol{W}) = \max\limits_{\boldsymbol{W}} \{\log[p(\boldsymbol{x})]\} = \max\limits_{\boldsymbol{W}} \{\log[|\det(\boldsymbol{W})| \cdot p(\boldsymbol{s})]\}, \\ \text{subject to: } \boldsymbol{s} = \boldsymbol{W}x. \end{cases} \tag{2.4}$$

From Lee, Girolami, and Sejnowski (1999), we have

$$\frac{\partial L}{\partial \boldsymbol{W}} = (\boldsymbol{W}^T)^{-1} + \phi(\boldsymbol{s}) \cdot \boldsymbol{x}^T = (\phi(\boldsymbol{s}) \cdot \boldsymbol{s}^T + \boldsymbol{I}) \cdot (\boldsymbol{W}^T)^{-1}. \tag{2.5}$$

Since $\boldsymbol{W} = \boldsymbol{A}^{-1}$, from the matrix differential formula in lemma 1 of He, Xie, Ding, and Cichocki (2007), we have

$$\frac{\partial L}{\partial \boldsymbol{A}} = -(\boldsymbol{A}^T)^{-1} \cdot \frac{\partial L}{\partial \boldsymbol{W}} \cdot (\boldsymbol{A}^T)^{-1} = -(\boldsymbol{A}^T)^{-1} \cdot (\phi(\boldsymbol{s}) \cdot \boldsymbol{s}^T + \boldsymbol{I}). \tag{2.6}$$

So the natural gradient of $L(\cdot)$ with respect to $\boldsymbol{A}$ is

$$\Delta \boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^T \cdot \frac{\partial L}{\partial \boldsymbol{A}} = -\boldsymbol{A} \cdot (\phi(\boldsymbol{s}) \cdot \boldsymbol{s}^T + \boldsymbol{I}). \tag{2.7}$$

From lemma 2 of He et al. (2007), we know that the natural gradient 2.7, $\Delta \boldsymbol{A} = -\boldsymbol{A} \cdot [\phi(\boldsymbol{s}) \cdot \boldsymbol{s}^T + \boldsymbol{I}]$, can be used to learn basis matrix $\boldsymbol{A}$ in the case that $\boldsymbol{A}$ is square and nonsingular.

Case 2: $m < n$, $\boldsymbol{A}$ is overcomplete. In this case, $\boldsymbol{A}$ is a rectangular matrix because $m < n$. We can add extra $n - m$ virtual observed mixtures $\boldsymbol{x}^{\text{virtual}}_{(2n-m)\times 1}$ and have the following equations:

$$\boldsymbol{x}^{\text{new}} = \boldsymbol{A}^{\text{new}} \boldsymbol{s}, \tag{2.8}$$

where $\boldsymbol{x}^{\text{new}} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{x}^{\text{virtual}}_{(n-m)\times 1} \end{pmatrix}$, $\boldsymbol{A}^{\text{new}} = \begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{A}^{\text{virtual}}_{(n-m)\times n} \end{pmatrix}$, and $(n-m)\times n$ matrix $\boldsymbol{A}^{\text{virtual}}_{(n-m)\times n}$ associates with $\boldsymbol{x}^{\text{virtual}}_{(n-m)\times 1}$. Now $\boldsymbol{A}^{\text{new}}$ in equation 2.8 is an $n \times n$ nonsingular square matrix. From the discussion in case 1, the natural gradient of $L(\cdot)$ with respect to $\boldsymbol{A}^{\text{new}}$ is

$$\begin{aligned}
\Delta \boldsymbol{A}^{\text{new}} &= \begin{pmatrix} \Delta \boldsymbol{A} \\ \Delta \boldsymbol{A}^{\text{virtual}}_{(n-m)\times n} \end{pmatrix} = -\boldsymbol{A}^{\text{new}} \cdot [\phi(\boldsymbol{s}) \cdot \boldsymbol{s}^T + \boldsymbol{I}] \\
&= -\begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{A}^{\text{virtual}}_{(n-m)\times n} \end{pmatrix} \cdot [\phi(\boldsymbol{s}) \cdot \boldsymbol{s}^T + \boldsymbol{I}] \\
&= \begin{pmatrix} -\boldsymbol{A} \cdot [\phi(\boldsymbol{s}) \cdot \mathbf{s}^T + \boldsymbol{I}] \\ -\boldsymbol{A}^{\text{virtual}}_{(n-m)\times n} \cdot [\phi(\boldsymbol{s}) \cdot \boldsymbol{s}^T + \boldsymbol{I}] \end{pmatrix}. \qquad (2.9)
\end{aligned}$$

From equation 2.9, we still have $\Delta \boldsymbol{A} = -\boldsymbol{A} \cdot [\phi(\boldsymbol{s}) \cdot \boldsymbol{s}^T + \boldsymbol{I}]$ when $m < n$. The theorem is proved.

Lewicki-Sejnowski gradient $\Delta \boldsymbol{A}$ (see equation 2.2) is the extension of Amari's natural gradient (Amari, Cichocki, & Yang, 1996; Amari, 1998; Lee, Lewicki et al., 1999). When $\boldsymbol{A}$ is square, Lewicki-Sejnowski gradient $\Delta \boldsymbol{A}$ is exactly the natural gradient (Lewicki & Sejnowski, 2000). Note that $\Delta \boldsymbol{A} = -\boldsymbol{A} \cdot [\phi(\mathbf{s}) \cdot \boldsymbol{s}^T + \boldsymbol{I}] \neq \boldsymbol{A}\boldsymbol{A}^T \cdot \frac{\partial L}{\partial \boldsymbol{A}}$ when $m < n$. More exactly, expression 2.2 can be seen as the extended version of the natural gradient. Compared with the standard gradient, the (extended) natural gradient 2.2 is much more advantageous: (1) it can speed up convergence (Amari, 1998; Lee, Lewicki et al., 1999), and (2) the (extended) natural gradient 2.2 is simple and more efficient because of no matrix inverse calculations.

In addition, given the basis matrix $\boldsymbol{A}$ and the observation $\boldsymbol{x}$, we can estimate or update $\boldsymbol{s}$ by the maximum a posteriori (MAP) method. Then we can estimate $\boldsymbol{A}$ and $\boldsymbol{s}$ by alternatively updating $\boldsymbol{A}$ and $\boldsymbol{s}$ repeatedly until convergence (Lee, Lewicki et al., 1999; Lewicki & Sejnowski, 2000).

3 The Constrained Lewicki-Sejnowski Gradient

To fix the arbitrary scaling, we usually set the constraints $\|\boldsymbol{a}_i\|_2^2 = 1,\ i = 1, \ldots, n$ (Bofill & Zibulevsky, 2001; Li, Cichocki, & Amari, 2004; Parra & Spence, 2000), that is, $\sum_{j=1}^{m} a_{ji}^2 = 1,\ i = 1, \ldots, n$. To enforce the solutions to satisfy these constraints, here instead of Lewicki-Sejnowski gradient 2.2, we consider their projections onto the hyperplanes defined by $\|\boldsymbol{a}_i\|_2^2 = 1,\ i = 1, \ldots, n$ and derive the constrained natural gradient 3.1 or 3.3 in the same way as Parra and Spence's (2000) constrained gradient method.

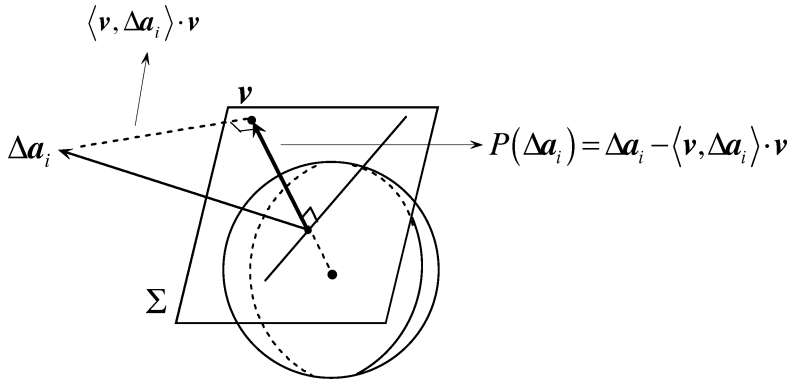

Figure 1: Gradient projection on the hyperplane Σ defined by the constraint $\|\boldsymbol{a}_i\|_2^2 = 1$.

Theorem 2. *The projection operator P for the ith column $\Delta \boldsymbol{a}_i$ of $\Delta \boldsymbol{A}$ onto the hyperplane Σ defined by the constraint $\|\boldsymbol{a}_i\|_2^2 = 1$ is*

$$P(\Delta \boldsymbol{a}_i) = \left(\boldsymbol{I} - \boldsymbol{a}_i \cdot \boldsymbol{a}_i^T\right) \cdot \Delta \boldsymbol{a}_i, \text{ or } P: \Delta \boldsymbol{a}_i \to \left(\boldsymbol{I} - \boldsymbol{a}_i \cdot \boldsymbol{a}_i^T\right) \cdot \Delta \boldsymbol{a}_i. \quad (3.1)$$

Proof. The normal vector of the hyperplane Σ defined by the constraint $f(\boldsymbol{a}_i) = \|\boldsymbol{a}_i\|_2^2 - 1 = 0$ is $\boldsymbol{v} = \frac{\partial f(\boldsymbol{a}_i)}{\partial \boldsymbol{a}_i} / \| \frac{\partial f(\boldsymbol{a}_i)}{\partial \boldsymbol{a}_i} \| = \boldsymbol{a}_i$. Incidentally, the hyperplane equation at the point $\boldsymbol{a}_i = \boldsymbol{a}_i^{(0)}$ is $\langle \frac{\partial f(\boldsymbol{a}_i)}{\partial \boldsymbol{a}_i}|_{\boldsymbol{a}_i = \boldsymbol{a}_i^{(0)}}, \boldsymbol{a}_i - \boldsymbol{a}_i^{(0)} \rangle = 0$. So the projection (of the vector $\Delta \boldsymbol{a}_i$) on $\boldsymbol{v}$ is $\langle \boldsymbol{v}, \Delta \boldsymbol{a}_i \rangle \cdot \boldsymbol{v}$ (see Figure 1). Thus, the projection operator P for $\Delta \boldsymbol{a}_i$ is as follows:

$$\begin{aligned} P(\Delta \boldsymbol{a}_i) &= \Delta \boldsymbol{a}_i - \langle \boldsymbol{v}, \Delta \boldsymbol{a}_i \rangle \cdot \boldsymbol{v} = \Delta \boldsymbol{a}_i - (\boldsymbol{v}^T \cdot \Delta \mathbf{a}_i) \cdot \boldsymbol{v} \\ &= \Delta \boldsymbol{a}_i - \left(\boldsymbol{a}_i^T \cdot \Delta \boldsymbol{a}_i\right) \cdot \boldsymbol{a}_i = \Delta \boldsymbol{a}_i - \boldsymbol{a}_i \cdot \left(\boldsymbol{a}_i^T \cdot \Delta \boldsymbol{a}_i\right) \\ &= \Delta \boldsymbol{a}_i - \left(\boldsymbol{a}_i \cdot \boldsymbol{a}_i^T\right) \cdot \Delta \boldsymbol{a}_i = \left(\boldsymbol{I} - \boldsymbol{a}_i \cdot \boldsymbol{a}_i^T\right) \cdot \Delta \boldsymbol{a}_i, \end{aligned} \quad (3.2)$$

that is, $P(\Delta \boldsymbol{a}_i) = (\boldsymbol{I} - \boldsymbol{a}_i \cdot \boldsymbol{a}_i^T) \cdot \Delta \boldsymbol{a}_i$.

Alternatively, formula 3.1 can be rewritten as the following constrained Lewicki-Sejnowski gradient:

$$\Delta \boldsymbol{A}|_{\|\boldsymbol{a}_i\|_2^2 = 1,\ i = 1, \ldots, n} = \Delta \boldsymbol{A} - \boldsymbol{A} \cdot \operatorname{diag}(\boldsymbol{A}^T \cdot \Delta \boldsymbol{A}). \quad (3.3)$$

It should be noted that since the constraints $\|\boldsymbol{a}_i\|_2^2 = 1,\ i = 1, \ldots, n$ can absorb the scaling ambiguity, these constraints play an important role in learning the basis matrix $\boldsymbol{A}$ for overcomplete sparse representation (see Bofill & Zibulevsky, 2001; Li et al., 2004). The constrained Lewicki-Sejnowski

gradient 3.3 is more robust than the conventional Lewicki-Sejnowski gradient 2.2 because it can search the solutions under the constraints $\|\boldsymbol{a}_i\|_2^2 = 1,\ i = 1, \ldots, n$, especially when the number of the sources is much larger than the number of the mixtures (see example 2 in section 4).

4 Experiments

Similar to Lee, Lewicki et al.'s (1999) experiment, here we assume the source s_k follows Laplacian density $p(s_k) \propto \exp(-\alpha|s_k|)$ (a sparse distribution) in the considered domain (time domain, frequency domain, time frequency domain). So $\phi(s_k) = \partial \log p(s_k)/\partial s_k \propto -\text{sign}(s_k)$. To get robust solutions, we set the initial basis matrix $\boldsymbol{A}^{(0)}$ such that $\|\boldsymbol{a}_i^{(0)}\|_2^2 = 1,\ i = 1, \ldots, n$ in the following examples. To evaluate the quality of the separations, we use the SIR (signal-to-interference ratio), which is the same as the performance index S/N in Bofill and Zibulevsky (2001).

Example 1. First we tested, Lee, Lewicki et al.'s (1999) "two mixtures and three sources" BSS example. The same three sources and mixing matrix were used. The Lewicki-Sejnowski gradient 2.2 and the constrained Lewicki-Sejnowski gradient 3.3 were respectively employed. By many tests and trials, we found that both gradients 2.2 and 3.3 could well estimate $\boldsymbol{A}$ and $\boldsymbol{s}$, and they almost always produced the same results in each trial for this example. To compare these two gradients further, we tested a more challenging example.

Example 2. We used Bofill and Zibulevesky's (2001) "two mixtures and six sources" BSS example. The sources $\boldsymbol{s}$ are six flute signals (32,768 samples) and are from the experiment SixFlutes I in Bofill and Zibulevsky (2001). The 2×6 mixing matrix $\boldsymbol{A}$ is

$$\boldsymbol{A} = \begin{pmatrix} 0.9659 & 0.7071 & 0.2588 & -0.2588 & -0.7071 & -0.9659 \\ 0.2588 & 0.7071 & 0.9659 & 0.9659 & 0.7071 & 0.2588 \end{pmatrix}.$$

Then two mixtures were produced by model $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$. The sources $\boldsymbol{s}$ are very sparse in frequency domain (Bofill & Zibulevsky, 2001), so we performed blind separation in frequency domain.

We conducted 100 Monte Carlo runs to evaluate their robustness of (extended) natural gradients 2.2 and 3.3. In each Monte Carlo run, the step size was set as $\mu = 0.01$, and both approaches started from the same initial matrix randomly generated. The Lewicki-Sejnowski gradient 2.2 nearly failed to estimate $\boldsymbol{A}$ and $\boldsymbol{s}$ in each Monte Carlo run, while the constrained Lewicki-Sejnowski gradient 3.3 succeeded in all Monte Carlo runs. For example, $\boldsymbol{A}$ was randomly initialized and followed by normalization as

Table 1: SIRs of the Estimated Sources.

	SIR (dB)					
Sources	s_1	s_2	s_3	s_4	s_5	s_6
Lewicki-Sejnowski gradient 2.2	12.8135	9.0512	0	14.3304	10.4519	0
Constrained Lewicki-Sejnowski gradient 3.3	50.5133	52.1840	49.1864	43.4974	49.1196	51.9210

follows:

$$\boldsymbol{A}^{(0)} = \begin{pmatrix} -0.2514 & 0.3994 & -0.6935 & 0.9995 & 0.8823 & -0.2491 \\ -0.9679 & 0.9168 & 0.7204 & -0.0316 & 0.4708 & 0.9685 \end{pmatrix}.$$

After 200 iterations, both solutions, using the above extended natural gradients 2.2 and 3.3, respectively, were convergent. We obtained the estimations of $\boldsymbol{A}$ and $\boldsymbol{s}$. The results are shown in Table 1, and two estimations $\hat{\boldsymbol{A}}_{\text{LSG}}$ and $\hat{\boldsymbol{A}}_{\text{CLSG}}$ of $\boldsymbol{A}$ from two gradients are, respectively, as follows:

$$\hat{\boldsymbol{A}}_{\text{LSG}} = \begin{pmatrix} 6.1008 & 3.2397 & 0.0000 & -1.3476 & -4.5548 & -0.0620 \\ 1.4805 & 4.2478 & 0.0000 & 5.4940 & 3.8266 & 0.0165 \end{pmatrix},$$

$$\hat{\boldsymbol{A}}_{\text{CLSG}} = \begin{pmatrix} 0.9667 & 0.7096 & 0.2579 & -0.2597 & -0.7055 & -0.9657 \\ 0.2560 & 0.7046 & 0.9662 & 0.9657 & 0.7087 & 0.2595 \end{pmatrix}.$$

From Table 1, we can see that Lewicki-Sejnowski gradient 2.2 failed, but the constrained Lewicki-Sejnowski gradient 3.3 succeeded in obtaining the correct solutions.

5 Conclusion

In this note, we rigorously and briefly proved the Lewicki-Sejnowski gradient and presented the more robust constrained Lewicki-Sejnowski gradient.

Acknowledgments

This work is in part supported by the National Natural Science Foundation of China (grants 60774094, U0635001, and 60505005), the Natural Science Fund of Guangdong Province, China (grants 04205783 and 05103553), SPSF (grant 20070410237), and the National Basic Research Program of China (grant 2005CB724301).

References

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, *10*, 251–276.

Amari, S., Cichocki, A., & Yang, H. H. (1996). A new learning algorithm for blind signal separation. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing systems* (Vol. 8, pp. 757–763). Cambridge, MA: MIT Press.

Bofill, P. & Zibulevsky, M. (2001). Underdetermined blind source separation using sparse representations. *Signal Processing*, *81*, 2353–2362.

Girolami, M. (2001). A variational method for learning sparse and overcomplete representations. *Neural Computation*, *13*(11), 2517–2532.

He, Z. S., Xie, S. L., Ding, S. X., & Cichocki, A. (2007). Convolutive blind source separation in the frequency domain based on sparse representation. *IEEE Trans. Audio, Speech, and Language Processing*, *15*(5), 1551–1563.

Lee, T. W., Girolami, M., & Sejnowski, T. (1999). Independent component analysis using an extended informax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Computation*, *11*(2), 609–633.

Lee, T. W., Lewicki, M. S., Girolami, M., & Sejnowski, T. J. (1999). Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letter*, *6*(4), 87–90.

Lewicki, M.S., & Sejnowski, T. J. (2000). Learning overcomplete representations. *Neural Computation*, *12*(2), 337–365.

Li, Y. Q., Cichocki, A., & Amari, S. (2004). Analysis of sparse representation and blind source separation. *Neural Computation*, *16*, 1193–1234.

Parra, L., & Spence, C. (2000). Convolutive blind separation of non-stationary sources. *IEEE Trans. Speech and Audio Processing*, *8*(3), 320–327.

Received July 31, 2006; accepted April 4, 2007.