Available online at www.sciencedirect.com
ScienceDirect
Progress in Natural Science 19 (2009) 1027–1031
Progress in Natural Science
www.elsevier.com/locate/pnsc

Estimating the mixing matrix by using less sparsity

Guoxu Zhou [a,*], Zuyuan Yang [a], Xiaoxin Liao [b], Jinlong Zhang [a]

[a] School of Electrics & Information Engineering, South China University of Technology, Guangzhou 510641, China
[b] Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

Received 30 June 2008; received in revised form 20 August 2008; accepted 26 August 2008

Abstract

In this paper, the nonlinear projection and column masking (NPCM) algorithm is proposed to estimate the mixing matrix for blind source separation. It preserves the samples which are close to the interested direction while suppressing the rest. Compared with the existing approaches, NPCM works efficiently even if the sources are less sparse (i.e., they are not strictly sparse). Finally, we show that NPCM provides considerably accurate estimation of the mixing matrix by simulations.

Keywords: Sparse component analysis; Blind source separation; Particle swarm optimization

1. Introduction

Blind source separation (BSS), which arises from the cocktail party problem, has received much attention for more than two decades, and many results have been reported [1]. The linear instantaneous mixing model of BSS is as follows:

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \boldsymbol{N}(t), \quad t = 1, 2, \cdots, T \tag{1}$$

where $\boldsymbol{x}(t) = [x_1(t), x_2(t), \cdots, x_m(t)]^T$ and $s(t) = [s_1(t), s_2(t), \cdots, s_n(t)]^T$ are the observation vector and source vector at time instant t, respectively. $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is the mixing matrix. BSS tries to recover the sources $\boldsymbol{s}(t)$ only from their mixtures (or equivalently observations) $\boldsymbol{x}(t)$, whereas the mixing matrix $\boldsymbol{A}$ is unknown.

Independent component analysis (ICA) pioneers the early study of BSS [2]. Meanwhile, the case of $m \geq n$ is mainly studied. ICA-based approaches assume independence of the sources, and they reconstruct the sources by multiplying the observations from the left by the inverse of the mixing matrix $\boldsymbol{A}$ [1,2]. Based on the temporal predictability, Stone and Xie et al. developed a novel approach [3,4], which only uses the second-order statistics. However, the above-mentioned approaches fail in underdetermined BSS, i.e. $m < n$. For the case of underdetermined mixing, sparsity is a widely used assumption. Sparsity means that only one source is active or dominant at each time instant t. This problem is often referred to as the sparse component analysis (SCA), and some important results can be found in [5–9]. Consider an equivalent formulation of (1) (noise is neglected):

$$\boldsymbol{x}(t) = \sum_{i=1}^{n} \boldsymbol{a}_i s_i(t) \tag{2}$$

where $\boldsymbol{a}_i$ is the ith column of $\boldsymbol{A}$. If only one source, for example s_i, is active at time instant t_0, then $\boldsymbol{x}(t_0) = \boldsymbol{a}_i s_i(t_0)$. That is, the observation vector $\boldsymbol{x}(t_0)$ is collinear with $\boldsymbol{a}_i$. Consequently, the observations can be clustered in lines, and each cluster center is an estimation of one column of the mixing matrix $\boldsymbol{A}$. Due to the inherent scale indeterminacy of BSS, each column of $\boldsymbol{A}$ provides only the direction information, and thus the norm of each column of $\boldsymbol{A}$ can be normalized to be unit. In this sense, columns of the mixing matrix $\boldsymbol{A}$ and directions are simply the same notions.

* Corresponding author. Tel./fax: +86 20 87114709.
E-mail address: zhou.guoxu@mail.scut.edu.cn (G. Zhou).

1002-0071/$ - see front matter
doi:10.1016/j.pnsc.2008.08.014

The SCA-based BSS approaches are mainly divided into two categories: one consists of methods that estimate the mixing matrix and the sources jointly [10], and the other is the classic two-stage method that estimates the mixing matrix at first and then reconstructs the sources according to the estimated mixing matrix [5,8,11]. No matter which category is concerned, estimating the mixing matrix is significant: it is the first step of the methods in the second category and it can provide a good initial point of the mixing matrix, and consequently, it can accelerate the convergence of the methods in the first category [12]. Currently, *k*-means is a mainly used clustering algorithm for SCA [11–14]. *k*-Means is simple and easy to implement. However, it has disadvantages: first, it is sensitive to the initial values, but the initial values usually have to be generated randomly. Secondly, *k*-means performs well only if the sources are strictly sparse or nearly strictly sparse. Unfortunately, this restriction often cannot be satisfied in practice. Some authors proposed the potential function-based approaches to estimate the mixing matrix [5,15]. However, they work well if only two mixtures are involved. To overcome the above-mentioned shortcomings, a new approach, named nonlinear projection and column masking (NPCM), is proposed to estimate the mixing matrix. NPCM is free of the dimension of the observations and works effectively for less sparse sources.

The remainder of the paper is organized as follows: in Section 2, the motivation and the NPCM algorithm are presented. In Section 3, the particle swarm optimization (PSO) is introduced to optimize the objective function. Simulations are presented in Section 4. Finally, conclusions are made in Section 5.

2. Nonlinear projection and column masking

In this paper, the source signals are assumed to be sparse. Here, the sparsity means that there are sufficient time instants at which one and only one source is active or dominant. In other words, we need sufficient samples to express each direction (column) rather than require that the sources are strictly sparse. Generally, many un-sparse time instants are permitted in our algorithm, i.e., our approach works well for less sparse sources.

2.1. Motivation

Suppose that the interested direction is $\boldsymbol{w}$; define $y_t = \|\boldsymbol{x}_t\|\left|\cos(\widehat{\boldsymbol{w},\boldsymbol{x}_t})\right|$, where $\widehat{\boldsymbol{w},\boldsymbol{x}_t}$ is the angle between $\boldsymbol{w}$ and $\boldsymbol{x}_t$ (thus $0 \le \widehat{\boldsymbol{w},\boldsymbol{x}_t} \le \pi$). Thus, y_t is the length of the projection of the sample $\boldsymbol{x}_t$ on the vector $\boldsymbol{w}$. Classic principle component analysis (PCA) searches a vector $\boldsymbol{w}$ which maximizes the cost function $J(\boldsymbol{w}) = Var(y_t) = E[y_t^2]$ (assume that the sources are centered, and E is the math expectation operator). Fig. 1a shows the scatter plot of two sparse speech signals. From the figure, we see that the first principal component cannot indicate columns of the mixing matrix. But if the samples far away from the direction $\boldsymbol{a}_1$ are masked, $\boldsymbol{a}_1$ will become the principal component of the rest of the samples (see Fig. 1b). Motivated by this, we modify the objective function of PCA. The new objective function will sufficiently suppress the samples which are far away from the interested direction $\boldsymbol{w}$.

In order to suppress the samples which are far away from $\boldsymbol{w}$, an exponential function $f(v) = e^{-\rho v^2}$ is employed, where $\rho > 0$ is used to control the decaying speed. To ensure that f is maximized when $\boldsymbol{x}_t$ and $\boldsymbol{w}$ are collinear, the mapping $v = 1 - \cos^2(\widehat{\boldsymbol{w},\boldsymbol{x}_t})$ is applied. Note that $\cos(\widehat{\boldsymbol{w},\boldsymbol{x}_t}) = \frac{\boldsymbol{w}^T\boldsymbol{x}_t}{\|\boldsymbol{x}_t\|\cdot\|\boldsymbol{w}\|}$. Thus, the following cost function is defined:

$$\max J(\boldsymbol{w}) = \sum_t \|\boldsymbol{x}_t\| \exp\left(-\rho\left(1 - \frac{\boldsymbol{w}^T\boldsymbol{x}_t\boldsymbol{x}_t^T\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{w}\boldsymbol{x}_t^T\boldsymbol{x}_t}\right)^2\right) \tag{3}$$

Fig. 2a is the scatter plot of two linear mixtures of four speech signals. It is almost impossible to detect directions from the scatter plot directly. Let $\boldsymbol{w} = [\cos\theta, \sin\theta]^T$. Fig. 2b is the plot of $J(\theta)$ with $\rho = 10^5$. From Fig. 2b, four directions are indicated by the four maxima. Thus, the estimation of the mixing matrix is converted to search all local maxima of the objective function.

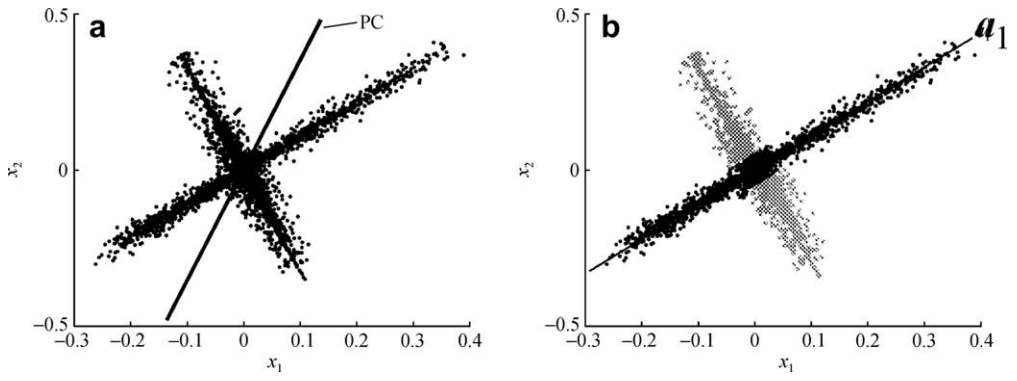

Fig. 1. (a) PCA cannot indicate columns of the mixing matrix. (b) The principal component indicates a column of the mixing matrix when some samples are masked.

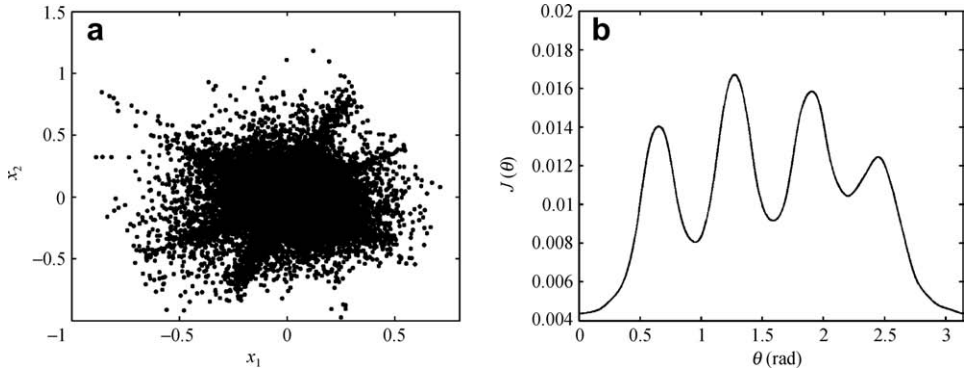

Fig. 2. (a) Scatter plot of two linear mixtures of four sources in the time domain. (b) Plot of $J(\theta)$ where $\theta \in [0, \pi]$.

2.2. Nonlinear projection and column masking

Now suppose that a maximum of model (3) has been obtained, i.e., the first column $\boldsymbol{w}_1$ has been estimated and now the next column is to be estimated. Let $\theta_0 = \max_{i \neq j} |\cos(\widehat{\boldsymbol{a}_i, \boldsymbol{a}_j})|$ and assume that θ_0 can be determined roughly. If $|\cos(\widehat{\boldsymbol{w}_l, \boldsymbol{x}_t})| \geq \delta_0 = \cos(\theta_0)$, $\boldsymbol{x}_t$ is useless for estimating the next column. So, we can mask the samples near the vector $\boldsymbol{w}_1$. More generally, assuming that the columns $\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots \boldsymbol{w}_p$ have been estimated, we define the masking vector M:

$$M_t = \begin{cases} 0 & |\cos(\widehat{\boldsymbol{w}_i, \boldsymbol{x}_t})| > \delta_0, \ i = 1, 2, \ldots, p \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

And the new objective function is

$$\max \quad H(\boldsymbol{w}) = \sum_t M_t \|\boldsymbol{x}_t\| \exp\left(-\rho\left(1 - \frac{\boldsymbol{w}^T \boldsymbol{x}_t \boldsymbol{x}_t^T \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{w} \boldsymbol{x}_t^T \boldsymbol{x}_t}\right)^2\right) \tag{5}$$

Note that $\delta_0 = \cos(\theta_0)$ and $\theta_0 = \max_{i \neq j} |\cos(\widehat{\boldsymbol{a}, \boldsymbol{a}_j})|$. One may question how to determine the threshold θ_0 or, equivalently, δ_0. Generally, it can be prior knowledge of the environment. If this prior knowledge is unavailable, benefiting from the nonlinear projection, it can also be approximate in our algorithm. Experientially, if the sources are less sparse or the environment is noisy, a larger value for δ_0 is recommended. The algorithm is described as follows:

1. Nonlinear projection and column masking algorithm (NPCM)
2. Step 1: Initialize $\boldsymbol{w}$.
3. Step 2: Update M by (4), solve model (5).
4. Repeat step 2, until all columns are estimated.

The optimization of model (5) will be discussed in Section 3.

2.3. Parameter selection

Setting the parameter ρ properly is important to the success of the algorithm. To get the value of ρ, we solve the following equation roughly:

$$\exp(-\rho(1 - \cos^2 \theta_0)^2) \leq \varepsilon \tag{6}$$

Therefore,

$$\rho > -\frac{\log(\varepsilon)}{\sin^4 \theta_0}. \tag{7}$$

where $\varepsilon < \frac{1}{T}$ is commonly recommended and T is the number of samples. $\varepsilon < \frac{1}{T}$ guarantees that even if there are many samples $\boldsymbol{x}_t'$ which satisfy that $|\cos(\widehat{\boldsymbol{x}, \boldsymbol{w}})| > \theta_0$, the sum $\sum_t \|\boldsymbol{x}_t'\| \exp(-\rho(1 - \cos^2(\widehat{\boldsymbol{x}_t', \boldsymbol{w}}))^2)$ will not affect the value of $H(\boldsymbol{w})$ or $J(\boldsymbol{w})$ evidently. So $\varepsilon < \frac{1}{T}$ is a rough criterion. Generally, the parameters in NPCM can be set roughly.

It is usually recommended that ρ possibly takes a larger value. However, the objective function is a sum relevant to discrete samples. If the number of samples is relatively small but ρ is too large, the objective function will have a saw-tooth outline (so a large number of local maxima exists), which is a nightmare for optimization algorithms. Generally, if the sources are sufficiently sparse and noise is mild, a large value should be set for ρ to achieve a higher direction resolution. And in contrast, a smaller value is recommended for ρ to make the function smoother, which leads the algorithm to be more robust to noise. For example, setting $\theta_0 = 5^\circ \approx 0.09$ radian, $T = 65449$, according to (7), we have $\rho = 3.48 \times 10^5$.

3. Optimization of the objective function

In the NPCM approach, the global maxima of the objective function are generally desired, and they correspond to the columns of the mixing matrix. As mentioned above, if ρ is too large, many local maxima exist. Consequently, the algorithms depending on the local properties of the objective function, for example gradient-based algorithms, generally cannot provide desirable results. In fact, it is still a challenge to develop algorithms in global convergence. Since the global maxima are crucial to NPCM, particle swarm optimization (PSO) is introduced here.

The PSO, first introduced by Kennedy and Eberhart [16], is a stochastic optimization technique that can be likened to the behavior of a flock of birds or the sociological behavior of a group of people. These population-based living beings utilize two kinds of important knowledge when

they are moving, hunting for food and so on: one is from their own personal experience; the other is from the population. In optimization, denote a possible solution to the optimization model at hand by a particle, and let $\boldsymbol{w}^{(i)}$ be the ith particle; then $\boldsymbol{w}^{(i)}$ is updated by the following rule:

$$\begin{aligned} \boldsymbol{w}_{k+1}^{(i)} &= \boldsymbol{w}_k^{(i)} + \boldsymbol{v}_{k+1} \\ \boldsymbol{v}_{k+1} &= \boldsymbol{\omega}_k \boldsymbol{v}_k + c_1 \boldsymbol{r}_{k,1}(\boldsymbol{w}_l^{(i)} - \boldsymbol{w}_k) + c_2 \boldsymbol{r}_{k,2}(\boldsymbol{w}_g - \boldsymbol{w}_k) \end{aligned} \tag{8}$$

where $\boldsymbol{\omega}_k$ is the inertia weight, this value is typically setup to vary linearly from 1 to 0 during the course of a training run. $\boldsymbol{w}_l^{(i)}$ is the best solution found by the particle $\boldsymbol{w}^{(i)}$ so far, denoting the individual experience. $\boldsymbol{w}_g$ is the global best solution discovered so far by any of the particles in the swarm, i.e. it denotes the population experience. c_1 and c_2 denote the acceleration coefficients and are generally 2, $r_{k,i} \sim U(0,1)$. PSO cannot guarantee the global convergence, but it can find a relatively better solution in high probability, which is enough for our task. Further development of PSO can be found in [17–20], and for convenience, the standard PSO is employed and all the parameters are set to default values in this paper.

4. Simulations

In simulations, PSO uses the following settings: the size of the population is 20, and the iteration number is 100. The accuracy of direction estimation is measured by $D_{simi} = |\cos(\widehat{\boldsymbol{a}_i, \widetilde{\boldsymbol{a}}_i})|$, where $\boldsymbol{a}_i, \widetilde{\boldsymbol{a}}_i$ denote a column of the mixing matrix and its estimator, respectively. If $D_{simi} = 1$, the column is estimated accurately up to a scale.

In the first simulation, the sources named SixFlutes are used. SixFlutes consist of six sources which are not sparse in the time domain but sparse in the frequency domain. Fig. 3a is the scatter plot of two linear mixtures of these six sources in the transform domain. The 'sparse' transform is FFT, and the real coefficients of FFTs are used to estimate the mixing matrix. The mixing matrix is:

$$\boldsymbol{A} = \begin{bmatrix} 0.6105 & 0.9549 & -0.8969 & -0.6253 & 0.9975 & 0.2985 \\ 0.7920 & -0.2968 & -0.4422 & 0.78040 & 0.0708 & 0.9544 \end{bmatrix}$$

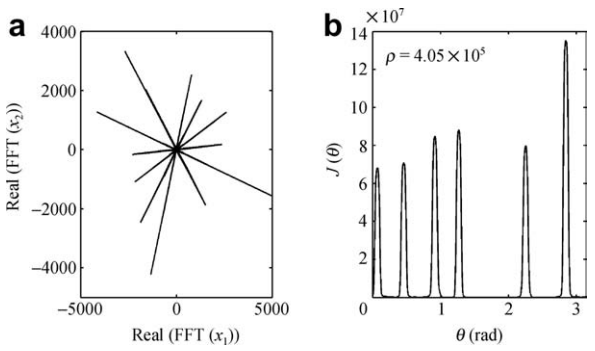

Fig. 3. (a) Scatter plot of two linear mixtures of six sources in the transform domain. (b) Plot of J where $J(\boldsymbol{w}) = J(\theta)$ is defined by (3), $\boldsymbol{w} = [\cos\theta, \sin\theta]^T$, $\theta \in [0, \pi]$.

According to (7), we set $\rho = 4 \times 10^5$ (ρ is estimated by setting $\varepsilon = \frac{1}{T}$, where T is the number of samples). The plot of J is shown in Fig. 3b. From Fig. 3b, we see that, because the sources are very sparse in the transform domain, the six columns are indicated clearly by the local maxima of J. Actually, the accuracy of the estimated directions is greater than 0.9999.

In the following simulation, an example with less sparsity is exploited. The sources are four speech signals, and they are mixed by a 3×4 matrix. To investigate the robustness to sparsity of NPCA, no sparse transforms are applied. Fig. 4a shows the scatter plot of the three observations. It is almost impossible to detect the directions from the scatter plot. Fig. 4b shows the landscape plot of the objective function ($\rho = 10^4$): each peak corresponds to one source and the peak location corresponds to a column of the mixing matrix.

After one direction $\boldsymbol{w}_k$ is obtained, set $M_t = 0$ if $|\cos(\widehat{\boldsymbol{w}_k, \boldsymbol{x}_t})| < 0.9$. Table 1 shows the accuracies obtained by k-means and NPCM in the estimation of the mixing matrix. Compared with k-means, NPCM achieves more accurate estimation of the mixing matrix. What is more, k-means often estimates only three directions correctly. In fact, k-means is well known to be extremely sensitive to initial cluster centers. A bad choice of initial cluster centers will result in wrong clusters. However, benefiting from the global search ability of PSO, NPCM almost always succeeds in estimating all directions and is barely affected by initial values. In other words, compared with the k-means, NPCM is more robust to the initial settings.

5. Conclusion

SCA is a powerful tool to solve the underdetermined BSS problem. Under the assumption of sparsity, the mixing matrix can be estimated via the linear clustering. However, the existing approaches often assume that the sources are strictly sparse. This paper proposed a new algorithm named NPCM to estimate the mixing matrix for less sparse sources. NPCM estimates the directions by suppressing the samples far away from the interested direction. This method is proved to be robust and efficient.

Global maxima play key roles in the success of our algorithm, and PSO is employed to optimize the objective function. Although PSO cannot guarantee global convergence, it is able to arrive at a relatively better position, which is enough for our task. In fact, a desired solution is always achievable when the numbers of the population and the iteration are large. Simulations show that NPCM is feasible and is very robust to sparsity. An algorithm with global convergence will improve the reliability and robustness of the algorithm, but currently, global optimization is still full of challenges. We look forward to more efficient tools to deal with this kind of optimization model.

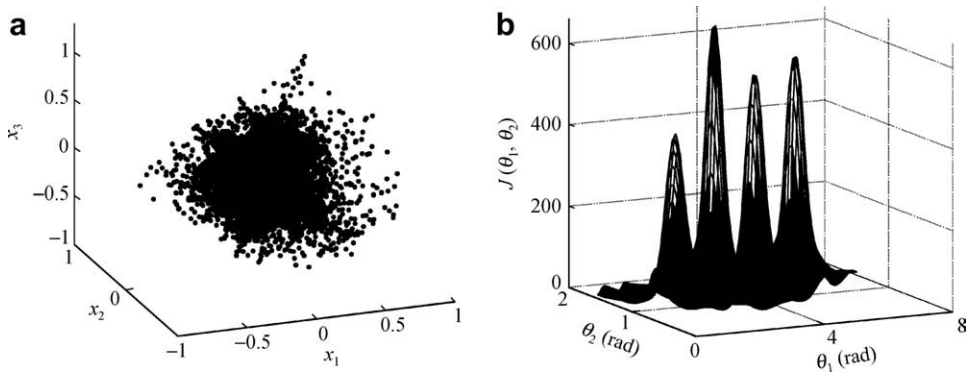

Fig. 4. (a) Scatter plot of three linear mixtures of four sources in the time domain. (b) Landscape plot of the objective function. Each peak corresponds to one source, and the peak locations correspond to the associated source's mixing parameters.

Table 1
Comparison of estimation accuracy between k-means and NPCM.

	D_{simi1}	D_{simi2}	D_{simi3}	D_{simi4}
k-Means	0.9222	0.9955	0.9952	0.8579
NPCM	0.9990	0.9999	0.9980	0.9736

Acknowledgements

This work was supported by the Key Program of the National Natural Science Foundation of China (Grant No. U0635001) and (Grant No. 60674033 and 60774094).

References

[1] Cichocki A, Amari S. Adaptive blind signal and image processing: learning algorithms and applications. John Wiley & Sons, Ltd; 2002.
[2] Hyvarinen A, Karhunen J, Oja E. Independent component analysis. New York: Wiley; 2001.
[3] Stone JV. Blind deconvolution using temporal predictability. Neurocomputing 2002;49(1):79–86.
[4] Xie SL, He ZS, Fu YL. A note on Stone's conjecture of blind signal separation. Neural Comput 2005;17(2):321–30.
[5] Bofill P, Zibulevsky M. Underdetermined blind source separation using sparse representations. Signal Process 2001;81(11):2353–62.
[6] He ZS, Xie SL, Ding SX, et al. Convolutive blind source separation in the frequency domain based on sparse representation. IEEE Trans Audio Speech Language Process 2007;15(5):1551–63.
[7] He ZS, Xie SL, Zhang LQ, et al. A note on Lewicki–Sejnowski gradient for learning overcomplete representations. Neural Comput 2008;20(3):636–43.
[8] Li YQ, Amari SI, Cichocki A, et al. Underdetermined blind source separation based on sparse representation. IEEE Trans Signal Process 2006;54(2):423–37.
[9] Li YQ, Cichocki A, Amari S. Analysis of sparse representation and blind source separation. Neural Comput 2004;16(6):1193–234.
[10] Fevotte C, Godsill SJ, Wolfe PJ. Bayesian approach for blind separation of underdetermined mixtures of sparse sources. Independent Compon Anal Blind Signal Separ 2004;3195:398–405.
[11] Saab R, Yilmaz O, McKeown MJ, et al. Underdetermined anechoic blind source separation via l(q)-basis-pursuit with $q < 1$. IEEE Trans Signal Process 2007;55(8):4004–17.
[12] Araki S, Sawada H, Mukai R, et al. Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. Signal Process 2007;87(8):1833–47.
[13] Abdeldjalil ABA, Nguyen LT, Karim AM, et al. Underdetermined blind separation of nondisjoint sources in the time-frequency domain. IEEE Trans Signal Process 2007;55(3):897–907.
[14] O'Grady PD, Pearlmutter BA. Soft-LOST: EM on a mixture of oriented lines. Fifth int. conf. on independent component analysis, vol. 3195. Granada: Springer-Verlag; 2004. p. 430–6.
[15] Zhang W, Liu J, Sun J, et al. A new two-state approach to underdetermined blind source separation using sparse representation. IEEE Int Conf Acoust Speech Signal Process 2007;3:953–6.
[16] Kennedy J, Eberhart RC. Particle swarm optimization. In: Proc. IEEE int. conf. on neural networks; 1995. p. 1942–8.
[17] Jiang M, Luo YP, Yang SY. Stochastic convergence analysis and parameter selection of the standard particle swarm optimization algorithm. Inform Process Lett 2007;102(1):8–16.
[18] Kadirkamanathan V, Selvarajah K, Fleming PJ. Stability analysis of the particle dynamics in particle swarm optimizer. IEEE Trans Evolution Comput 2006;10(3):245–55.
[19] Liang JJ, Qin AK, Suganthan PN, et al. Comprehensive learning particle swarm optimizer for global optimization of multimodal functions. IEEE Trans Evolution Comput 2006;10(3):281–95.
[20] Trelea IC. The particle swarm optimization algorithm: convergence analysis and parameter selection. Inform Process Lett 2003;85(6):317–25.