Estimation of underdetermined mixing matrix based on support vector machine*

Yang Zuyuan and Xie Shengli**

(School of Electrics & Information Engineering, South China University of Technology, Guangzhou 510641, China)

Accepted on May 31, 2007

Abstract In underdetermined blind source separation (BSS), a novel algorithm based on extended support vector machine (SVM) is proposed to estimate the mixing matrix in this paper, including the number of the active sources. Instead of traditional clustering algorithms, it mainly takes the modulus of observations and the number in each direction of arrival, without any prior knowledge about the sources except for sparsity, and it is not sensitive to the initial values. Simulations are given to illustrate availability and robustness of our algorithm.

Keywords: underdetermined BSS, SVM, clustering.

Nowadays, blind source separation (BSS) has attracted more and more attention because of its wide application in signal processing, including space signal, electroencephalography (EEG), electrocardiogram (ECG), speech signal, geophysical and chaos signals and so on. Conventional algorithms separate the signals under the assumption that the mixing matrix is invertible, such as classical independent component analysis (ICA)[1], and algorithms based on signal temporal predictability[2]. However, these algorithms fail when the matrix is ill-conditioned (e.g. the sources are more than sensors)[9,10]. It has been found that the separation quality seems to improve with the higher sparsity of the sources for the underdetermined case[6], and many source signals are sparse[3,4,7,9] in the time-domain. Furthermore, many natural signals have sparser representations in other transform domains, such as the Fourier transform, the wavelet transform and the modified discrete cosine transform (MDCT)[9]. Based on the sparsity of the sources, several methods are developed for BSS. For instance, the mixing matrix and sources are estimated using overcomplete representation[7], the maximum posterior approach, and maximum-likelihood approach[10]. However, they are quite complicated and time-consuming. And a two-stage method (TSM), i.e. a clustering-then-l^1-optimization approach, was proposed by Bofill, in which six sources are separated[9]. The necessary and sufficient condition of separability using TSM is introduced in Refs. [4, 6], together with the probability estimation for recoverability. TSM is an attractive method to recover the sources in the underdetermined case. However, the error from the estimation of the basis matrix will affect precision of separation inevitably. There exist several problems using the methods mentioned above to obtain basis matrix (mixing matrix): 1) the number of the sources is unknown, so it is hard to get the amount of the clustering centers, while the conventional clustering methods such as k-means clustering and K-EVD[3,5] depend on it quite a lot; 2) clustering algorithms are sensitive to the initial value which is random in reality.

In this study, SVM is used to estimate the mixing matrix and the number of active sources simultaneously, under the assumption that columns of mixing matrix are unit length.

The typical model for BSS with m sources and n sensors is

$$\boldsymbol{X}(t) = \boldsymbol{A}\boldsymbol{S}(t) + \boldsymbol{V}(t) \quad (1)$$

where $\boldsymbol{X}(t)$ are observations, $\boldsymbol{A} \in \mathbf{R}^{n\times m}$ is a mixing matrix, $\boldsymbol{S}(t)$ are sources and $\boldsymbol{V}(t)$ are additive noises. If $m > n$, it is the underdetermined model, and

* Supported by National Science Fund for Distinguished Young Scholars (Grant No. 60325310), Guangdong Province Science Foundation for Program of Research Team (Grant No. 04205783), Specialized Prophasic Basic Research Projects of Ministry of Science and Technology, China (Grant No. 2005CCA04100), and Key Program of National Natural Science Foundation of China (Grant No. U0635001)

** To whom correspondence should be addressed. E-mail: adshlxie@scut.edu.cn

Eq. (1) can be rewritten as follows neglecting noise:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix} s_1(t) + \cdots + \begin{bmatrix} a_{1m} \\ a_{2m} \\ \vdots \\ a_{nm} \end{bmatrix} s_m(t) \quad (2)$$

1 Learning theory of SVM

In the 1960s, the statistic learning theory was introduced by Vapnik and Cervonenkis, and as a new learning machine algorithm, SVM was proposed under the standard of structure risk minimum (SRM).

The process of training the SVM equals solving a quadratic programming (QP) problem[12, 13, 15]:

For a group of training samples

$$\{(\boldsymbol{x}_i, y_i) \mid i = 1, \cdots, T; \boldsymbol{x}_i \in \mathbf{R}^n, y_i \in \{-1, +1\}\} \quad (3)$$

the aim is to get the decision-making function below:

$$f(\boldsymbol{x}) = sign\left\{ \sum_{i=1}^{T} \alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) - b \right\} \quad (4)$$

where $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ *is the kernel function*, α *is the Lagrange multiplier*, b *is the threshold and* $\{\alpha_i\}_{i=1}^{T}$ *are results of the following QP problem*[13]

$$\begin{cases} \max\limits_{\alpha} \quad W(\alpha) = \sum_{i=1}^{T} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{T} \alpha_i Q(i,j) \alpha_j \\ \text{S.T.} \quad \sum_{j=1}^{T} \alpha_i y_j = 0, \ 0 \leqslant \alpha_i \leqslant C, \ i = 1, \cdots, T \end{cases} \quad (5)$$

where $Q(i,j) = y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathbf{R}$, and C is a regularizing parameter.

The classical method sequential minimal optimization (SMO) in Ref. [11] will be adopted to solve the problem above. However, the traditional SVM can only be used for binary classification and it will be extended into multi-class classification as follows, that is K-SVM:

Algorithm 1: Training K-SVM

```
for i=1:N
  for j=i+1:N
    selecting the ith class and the jth class of
    data set, and training binary SVM using
    SMO
  end
end
```

where N is the number of the classes in the training sets.

There will be $N(N-1)/2$ SVMs when the algorithm ends, and K-SVM is composed of these SVMs. The corresponding principle of classification is based on directed acyclic graph (DAG) in Ref. [14] (Fig. 1).

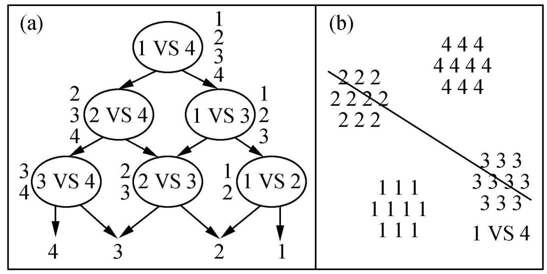

Fig. 1. Classification using K-SVM. (a) K-SVM with four classes; (b) illustration of the SVM generated for the 1 vs 4 subproblem.

2 Estimation of mixing matrix

Based on the sparsity of the sources, we can conclude that observations are distributed mainly in several directions with the largest number of the samples and each principal direction is decided by samples with larger modulus (Fig. 2). Before introducing the algorithm, some definitions are given as follows:

Definition 1. For given vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbf{R}^n$, the angle of $\boldsymbol{a}, \boldsymbol{b}$ is

$$\alpha = \arccos \frac{\boldsymbol{a}^{\mathrm{T}} \circ \boldsymbol{b}}{\|\boldsymbol{a}\|_2 \circ \|\boldsymbol{b}\|_2} \quad (6)$$

And the corresponding normalized angle is

$$\theta(\boldsymbol{a}, \boldsymbol{b}) = \begin{cases} \alpha & \text{if } \alpha \in \left[0, \frac{\pi}{2}\right] \\ \pi - \alpha & \text{else} \end{cases} \quad (7)$$

Definition 2. For given vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbf{R}^n$, $\boldsymbol{b} \in \boldsymbol{O}(\boldsymbol{a}, \varepsilon)$ if and only if $\theta(\boldsymbol{a}, \boldsymbol{b}) \leqslant \varepsilon$, i.e.

$$\boldsymbol{b} \in \boldsymbol{O}(\boldsymbol{a}, \varepsilon) \Leftrightarrow \theta(\boldsymbol{a}, \boldsymbol{b}) \leqslant \varepsilon \quad (8)$$

Definition 3. Classify the jth class of data set containing M_j samples into $K (0 < K \leqslant M_j)$ subclasses using K-SVM. If the number of the samples which belong to the ith subclass is M_{ji}, then the index q_{ji} of the data set is

$$q_{ji} = \frac{M_{ji}}{M_j} \quad (9)$$

The algorithm to estimate the mixing matrix and the active source number is given as follows.

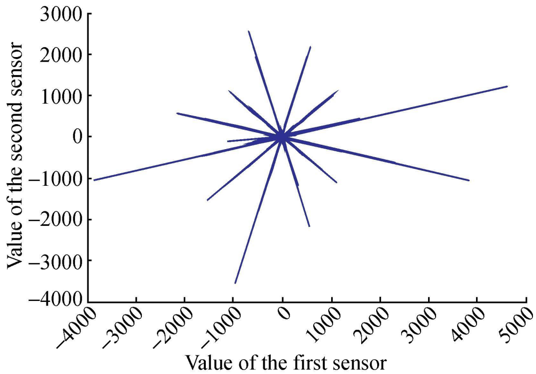

Fig. 2. Mixtures of the sparse signals measured by two sensors.

Algorithm 2: Estimate the source number and the mixing matrix.

1) **Initialization**

Transform the observations into the corresponding sparse domain, and delete the samples with short modulus to reduce the effect of overlapping sources. Split the rest samples, which format the testing set, into N atomic classes G_i ($i=1, \cdots, N$) according to the property of modulus and angle, such that $\forall S \in G_i \Rightarrow S \in O(Z_i, \varepsilon)$, $Z_i \notin O(Z_j, \varepsilon)$, $i \neq j$, where Z_i is the sample with the largest modulus in G_i, and ε is a small threshold.

2) **Begin**

Step 1: Select two classes from the testing set, where one class is with the most number of samples, the other is what makes the normalized angle θ of the principal direction of these two classes be the largest, and push them into the training set.

Step 2: Train K-SVM with the samples in the training set.

Step 3: Test the rest classes G_j ($j=1, \cdots, N-K$) in the testing set using K-SVM above, and calculate the property index q_{ji} ($i=1, \cdots, k$) of the class G_j; let

$$\begin{cases} P_{\min} = \min_j \max_i q_{ji} \\ P_{\max} = \max_j \max_i q_{ji} \end{cases} \tag{10}$$

If $p_{\min} < \eta_1$, then push the data in the jth class into the training set as a new class, where $j = \arg_j P_{\min}$, and let $K = K+1$. If $K \geqslant N$, goto Step 4; else, goto Step 2.

If $p_{\max} > \eta_2$, then push the data in the jth class into the ith class in the training set, where $[j, i] = \arg_{j,i} P_{\max}$, and let $N = N-1$. If $K \geqslant N$, goto Step 4; else, goto Step 2.

Step 4: If the sample numbers in the main classes ($\widehat{K}$ classes with the most samples) in the training set are changeless, stop the algorithm; else, push all of the classes in training set into the testing set, and adjust η_1, η_2 as

$$\begin{cases} \eta_1 = \eta_1 + \sigma_1 \\ \eta_2 = \eta_2 - \sigma_2 \end{cases} \tag{11}$$

3) **goto Begin**

In the algorithm above, the initial values of η_1, η_2 are $\eta_1 = 0.4$, $\eta_2 = 0.9$ and σ_1, σ_2 are small positive parameters.

When the algorithm ends, let K be the total number of the classes, and the number $\widehat{K}$ of the main classes be the estimated number of active sources. Columns of the mixing matrix correspond to the principal directions of the main classes.

3 Simulations

In the simulation, six sources (Fig. 3) in Ref. [9] are used and the mixing matrix is as follows

$$A = \begin{bmatrix} 0.5000 & 0.8892 & 0.2382 & 0.8892 & 0.4640 & 0.2167 \\ 0.5000 & 0.2382 & -0.8892 & -0.2382 & -0.4640 & 0.8086 \\ 0.5000 & 0.3244 & -0.2177 & 0.2177 & -0.5952 & -0.5462 \\ 0.5000 & 0.2177 & 0.3244 & 0.3244 & 0.4640 & 0.0306 \end{bmatrix} \tag{12}$$

From the observations (Fig. 4), sources are not sparse in time domain, and they are transformed into frequency domain via FFT.

Angle error (AE) and beeline degree (BD) below are performing indices.

Given vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbf{R}^n$, the AE of $\boldsymbol{a}, \boldsymbol{b}$ is

$$AE(\boldsymbol{a}, \boldsymbol{b}) = \frac{180}{\pi} \arccos\left(\frac{\boldsymbol{a}^{\mathrm{T}} \boldsymbol{b}}{\| \boldsymbol{a} \|_2 \circ \| \boldsymbol{b} \|_2} \right) \tag{13}$$

The BD of the direction of arrival of the source signal is[8]

$$BD(s(t)) = 1 - \frac{\sqrt{\lambda_2}}{\sqrt{\lambda_1}} \tag{14}$$

where λ_i ($i=1, \cdots, m$) correspond to the first m nonzero eigenvalues (ranking from large to small) of the covariance matrix of the sources.

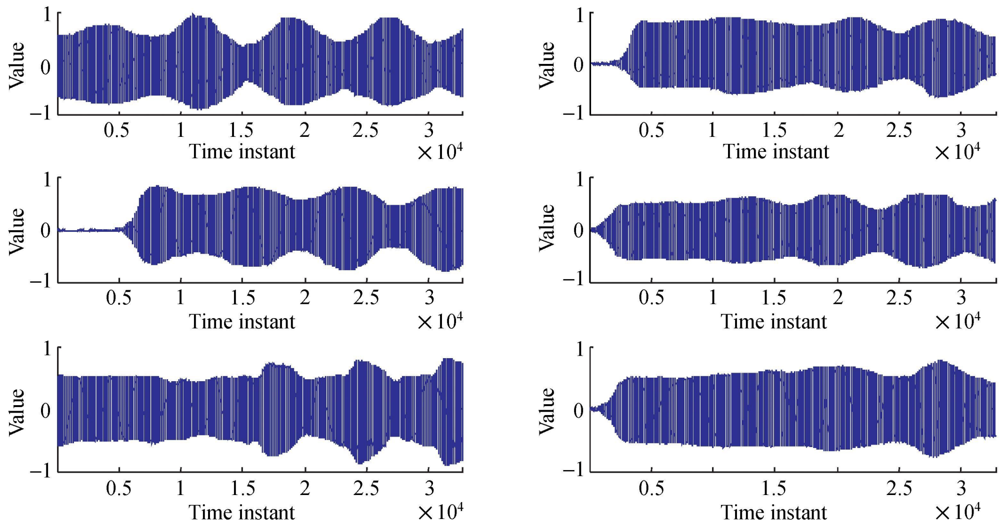

Fig. 3. Six sources in time domain.

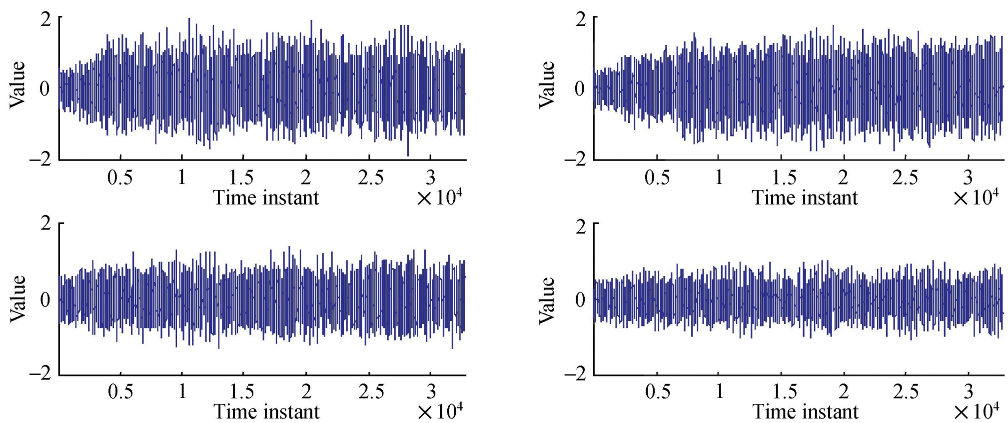

Fig. 4. Four mixing signals of the six sources.

Two experiments with the same signals will be shown as follows:

Experiment 1: When the number of the initial atomic class is 29, the numbers of samples in every class are 276, 224, 194, 120, 116, 94, 48, 42, 25, 20, 20, 16, 16, 14, 14, 10, 6, 6, 4, 4, 4, 2, 2, 2, 2, 2, 2, 2, 2.

When the algorithm in this paper stops, the results are:

The number of total classes: $K=9$, the numbers of samples in every class: 124, 124, 258, 248, 143, 296, 26, 50, 20; the number of main classes: $\bar{K}=6$, the numbers of samples in every main class: 124, 124, 258, 248, 143, 296; the corresponding mixing matrix which is composed of the normalized principal direction vector is

$$\hat{A}=\begin{bmatrix} 0.4989 & 0.8888 & 0.2384 & 0.8878 & 0.4619 & 0.2174 \\ 0.5009 & 0.2395 & -0.8895 & -0.2426 & -0.4658 & 0.8080 \\ 0.5013 & 0.3251 & -0.2159 & 0.2192 & -0.5962 & -0.5466 \\ 0.4989 & 0.2168 & 0.3245 & 0.3238 & 0.4628 & 0.0317 \end{bmatrix} \tag{15}$$

Experiment 2: When the number of the initial atomic class is 20, the numbers of samples in every class are 310, 234, 190, 162, 142, 102, 43, 22, 20, 14, 14, 10, 8, 4, 4, 2, 2, 2, 2, 2.

When the algorithm stops, the results are:

The number of total classes: $K=11$, the numbers of samples in every class: 104, 142, 248, 224, 162, 310, 43, 22, 20, 10, 4; the number of main classes: $K=6$, the numbers of samples in every class: 104, 142, 248, 224, 162, 310; the corresponding mixing matrix which is composed of the normalized principal direction vector is

$$\tilde{A}=\begin{bmatrix} 0.4999 & 0.8878 & 0.2393 & 0.8881 & 0.4606 & 0.2175 \\ 0.4999 & 0.2425 & -0.8894 & -0.2417 & -0.4668 & 0.8077 \\ 0.5004 & 0.3265 & -0.2152 & 0.2193 & -0.5969 & -0.5471 \\ 0.4997 & 0.2153 & 0.3247 & 0.3237 & 0.4622 & 0.0319 \end{bmatrix} \quad (16)$$

end.

Comparing experiment 1 and 2, we can see that the result of the mixing matrix is not sensitive to the number of the total classes at the beginning and the

The estimated mixing matrices $\boldsymbol{B}$, $\tilde{\boldsymbol{B}}$ via k-means and K-EVD respectively are

$$\boldsymbol{B}=\begin{bmatrix} 0.4437 & 0.8634 & 0.2395 & 0.8865 & 0.5556 & 0.2224 \\ 0.6077 & 0.2231 & -0.8831 & -0.2657 & -0.4219 & 0.8220 \\ 0.5661 & 0.2841 & -0.2614 & 0.1618 & -0.5246 & -0.5235 \\ 0.3368 & 0.3523 & 0.3073 & 0.3425 & 0.4880 & 0.0278 \end{bmatrix} \quad (17)$$

$$\tilde{\boldsymbol{B}}=\begin{bmatrix} 0.4806 & 0.9092 & 0.2399 & 0.8883 & 0.4930 & 0.2129 \\ 0.5270 & 0.2102 & -0.8943 & -0.2535 & -0.4644 & 0.8153 \\ 0.5590 & 0.2947 & -0.2139 & 0.2120 & -0.5789 & -0.5374 \\ 0.4228 & 0.2352 & 0.3114 & 0.3189 & 0.4540 & 0.0334 \end{bmatrix} \quad (18)$$

For simplicity, only the results of Experiment 1 are shown in Table 1, including BDs, AEs^1 (columns of $\boldsymbol{A}$ and $\boldsymbol{B}$), AEs^2 (columns of $\boldsymbol{A}$ and $\tilde{\boldsymbol{B}}$), AEs (columns of $\boldsymbol{A}$ and $\hat{\boldsymbol{A}}$), and AEs^* (columns of $\tilde{\boldsymbol{A}}$ and $\hat{\boldsymbol{A}}$).

Table 1. Values of BDs, AEs^1, AEs^2, AEs, and AEs^*

S	BDS	AES^1	AES^2	AES	AES^*
s_1	0.9991	12.2813	5.8880	0.1271	0.1064
s_2	0.9994	8.2357	2.6466	0.1017	0.2160
s_3	0.9992	2.7121	0.8368	0.1053	0.0665
s_4	0.9996	3.7213	0.9854	0.2803	0.0549
s_5	0.9998	7.1913	1.9881	0.1819	0.1078
s_6	0.9998	1.5522	0.6907	0.0852	0.0356

4 Conclusions

This paper studies a novel algorithm based on SVM which is extended into multi-class classification to estimate the mixing matrix for underdetermined BSS model, including the number of active sources. The information of observations is full utilized, and the algorithm is superior to k-means and K-EVD clustering which are sensitive to the initial values. Additionally, the result is neither sensitive to the number N of the initial partitioned atomic classes nor to the number K of the total classes.

References

1 Comon P. Independent component analysis, A new concept? Signal Processing, 1994, 36: 287—314

2 Xie SL, He ZS and Fu YL. A note on Stone's conjecture of blind signal separation. Neural Computation, 2005, 17: 321—330

3 Li YQ, Cichocki A and Amari S. Analysis of sparse representation and blind source separation. Neural Computation, 2004, 16: 1193—1234

4 Li YQ, Amari S, Cichocki A, et al. Underdetermined blind source separation based on sparse representations. IEEE Trans Signal Processing, 2006, 54(2): 423—437

5 He ZS and Cichocki A. K-EVD clustering and its applications to sparse component analysis. In: Lecture Notes in Computer Science (LNCS), 2006, 3889: 90—97

6 Li YQ, Amari S, Cichocki A, et al. Probability estimation for recoverability analysis of blind source separation based on sparse representation. IEEE Trans Information Theory, 2006, 52(7): 3139—3152

7 Michael SL and Terrence JS. Learning overcomplete representations. Neural Computation, 2000, 12: 337—365

8 Xie SL, He ZS and Gao Y. Adaptive Theory of Signal Processing (in Chinese), 1st ed. Beijing: Chinese Science Press, 2006, 183—203

9 Bofill P and Zibulevsky M. Underdetermined blind source separation using sparse representations. Signal Processing, 2001, 81: 2353—2362

10 Anthony JB and Terrence JS. An information-maximization approach to blind separation and blind deconvolution. Neural Computation, 1995, 7: 1129—1159

11 Platt JC. Sequential Minimal Optimization—a Fast Algorithm for Training Support Vector Machines. Cambridge: MIT Press, 1998, 185—208

12 Vapnik V. The Nature of Statistical Learning Theory. New York: Spring Verlag, 1995

13 Burges C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998, 2: 121—167

14 Ana CL and Carvalho CPLF. Comparing techniques for multiclass classification using binary SVM predictors. LNAI, 2004, 2972: 272—281

15 Asa BH, David H, Hava TS, et al. Support vector clustering. Journal of Machine Learning Research, 2001, 125—137

16 He ZS, Xie SL and Fu YL. Sparse representation and blind source separation of ill-posed mixtures. Science in China (Series F-Information Sciences), 2006, 49: 639—652