Predicting Unknown Composition of a Mixture Using Independent Component Analysis

Hyeseon Lee[1)], Haesang Park[2)], Chi-Hyuck Jun[3)]
Department of Industrial and Management Engineering, POSTECH

Abstract

A suitable representation for the conceptual simplicity of the data in statistics and signal processing is essential for a subsequent analysis such as prediction, pattern recognition, and spatial analysis. Independent component analysis (ICA) is a statistical method for transforming an observed high-dimensional multivariate data into statistically independent components. ICA has been applied increasingly in wide fields of spectrum application since ICA is able to extract unknown components of a mixture from spectra. We focus on application of ICA for separating independent sources and predicting each composition using extracted components. The theory of ICA is introduced and an application to a metal surface spectra data will be described, where subsequent analysis using non-negative least square method is performed to predict composition ratio of each sample. Furthermore, some simulation experiments are performed to demonstrate the performance of the proposed approach.

KEY WORDS: Independent component analysis; Non-negative least square; non-Gaussian

1. 서론

독립성분분석(Independent component analysis; ICA)은 다변량 데이타로부터 서로 독립인 성분을 분리하는데 사용되는 통계적 기법이다. ICA 는 주성분분석(Principal Component Analysis), 요인분석(Factor Analysis), Projection pursuit 등과 같이 일종의 선형변환방법이라고 볼 수 있으며, 최근 영상의학, 통신, 이미지분석 등 다양한 분야에서 기저구조를 밝혀내기 위한 기법으로 적용되고 있다 (Ikeda and Toyama, 2000; Viga´rio et al., 1998). 유사한 목적으로 NMF (nonnegative matrix factorization) 등이 사용되기도 한다 (Lee and Seung, 1999 &, 2001).

---

1) 경북 포항시 남구 효자동 산 31 포항공과대학교 산업경영공학과 책임연구원
e-mail : hyelee@postech.ac.kr
2) 경북 포항시 남구 효자동 산 31 포항공과대학교 산업경영공학과 석박사통합과정
3) 경북 포항시 남구 효자동 산 31 포항공과대학교 산업경영공학과 교수

본 연구의 목적은 ICA 를 이용하여 여러가지 순수물질이 섞인 혼합물에 대하여 측정된 시그널 데이타로부터 구성성분들을 도출하고, 그 독립적 성분들의 구성비율을 비음최소자승법(non-negative least square)에 의해서 예측하는 것이다. 본 연구에서 제안하는 방법은 금속표면의 X-ray 회절강도 데이타에 적용하였으며, 제안하는 성분비율 예측 방법의 정확도를 평가하기 위해 시뮬레이션을 실시하였다.

2. 독립성분분석(Independent component analysis)

2.1 개요

ICA 방법은 데이타 X 로부터 서로 독립인 source 를 분리해내는 방법으로서 미지요인분리(Blind Source Separation) 기법이라고도 한다. 즉, 다변량데이타 X 를 p 개의 서로 독립인 source 가 합해진 대상 (이를 mixture 라 함)으로부터 관측된 것이라 할 때 X 로부터 p 개의 독립인 source 를 분리하고자 하는 것이다. 예를 들어서 칵테일파티장에 p 개의 음원이 있고 그 안에 설치된 m 개의 마이크로 녹음을 하여 파장데이타 (n 개의 주파수에 대한 진폭)를 얻었다고 할 때, 이 (m x n) 차원의 mixture 데이타를 바탕으로 p 개의 음원을 분리할 수 있다 (Hyvarinen, 1999; Hyvarinen and Oja, 2000).

S_i 를 i 번째 source (혹은 pure component)를 나타내는 확률변수라 할 때 j-번째 (mixture 에 대한) 변수인 X_j 는 다음과 같이 p 개의 source 의 결합에 의하여 이루어졌다고 표현할 수 있다.

$$X_j = a_{j1}S_1 + a_{j2}S_2 + ... + a_{jp}S_p = \sum_{i=1}^{p} a_{ji}S_i \quad (1)$$

source vector 를 $\mathbf{s} = (S_1, ..., S_p)^T$ 와 같이 정의하면 식(1)은 다음과 같이 기술할 수 있다.

$$\mathbf{x} = A\mathbf{s} \quad (2)$$

여기서 $A = (a_{ji})$ 는 (m×p) 행렬로서 mixing matrix 라 한다. 식(2)를 data matrix 에 대하여 다시 쓰면 다음과 같다.

$$\mathbf{X} = A\mathbf{S} \quad (3)$$

ICA 는 관측된 X 로부터 S 를 구하는 것으로 S_i 들이 서로 통계적으로 독립이 되어야 한다는 제약조건을 부가한 것이다. 즉,

$$f(s_1, ..., s_p) = f_1(s_1)...f_p(s_p) \quad (4)$$

2.2 비정규성척도

ICA 에서의 중요조건은 pure component 들이 비정규분포 (non-Gaussian)를 따른다는 것이다. ICA 에서는 비정규성 척도로서 negentropy 를 사용하고 있다. 식(1)에서 하나의 pure component 는 다음과 같은 형태의 식으로부터 얻을 수 있다고 볼 수 있다.

$$S = w_1X_1 + ... + w_mX_m = w^T x \quad (5)$$

결국 ICA 에서는 독립적인 S 를 도출하기 위하여 $w^T x$ 의 비정규성을 최대화시키는 weight vector w 를 찾고자 한다. 확률변수 벡터 y 의 pdf 가 f(y)일 때 이에 대한 entropy 는 다음과 같이 정의된다.

$$H(y) = -\int f(y) \log f(y) dy \tag{6}$$

y 가 정규분포를 따를 때 동일한 분산을 갖는 확률변수 중 위의 entropy 가 최대가 됨을 이용하여 negentropy J 를 다음과 같이 정의한다.

$$J(y) = H(y_{gauss}) - H(y) \tag{7}$$

위에서 y_{gauss} 는 y 와 동일한 분산-공분산 행렬을 갖는 정규 확률변수벡터이다. 즉, 위의 negentropy 는 y 가 정규분포를 따를 때 0 가 되며 다른 분포를 따를 때는 0 보다 큼을 알 수 있다. 따라서 negentropy 가 클수록 비정규성을 띠게 된다고 하겠다.

ICA 에서는 negentropy 를 근사시키기 위하여 다음과 같은 식을 사용한다.

$$J(y) \propto \{E[G(y)] - E[G(z)]\}^2 \tag{8}$$

여기서 z 는 평균 0, 분산 1 인 정규확률변수를 의미하며, 함수 G 로는 임의의 non-quadratic 함수를 사용할 수 있으나 통상 아래와 같은 contrast 함수를 사용한다.

$$G_1(u) = \frac{1}{a_1} \log \cosh a_1 u \tag{9a}$$

$$G_2(u) = -\exp(-u^2 / 2) \tag{9b}$$

식(9a)에서 a_1 은 주로 $1 \le a_1 \le 2$ 의 값이 사용된다.

2.3 FastICA 알고리듬

널리 사용되는 FastICA 는 식(5)에서의 weight vector w 를 구하기 위하여 식 (8)의 negentropy 를 최대화시킨다. 즉,

$$\max_w J(w^T x) \tag{10a}$$

$$\text{subject to } \|w\| = 1 \tag{10b}$$

위의 최적화문제 해결을 위해 fixed point algorithm 을 사용하며 주로 Newton's method 가 이용된다. 결국 하나의 w 를 구하기 위하여는 다음과 같은 과정이 진행된다. 알고리즘의 유도는 Hyvarinen and Oja (2000) 참조할 수 있다.

Step 1. 초기 w 벡터를 랜덤하게 선택

Step 2. $w \leftarrow E[xg(w^T x)] - wE[g'(w^T x)]$

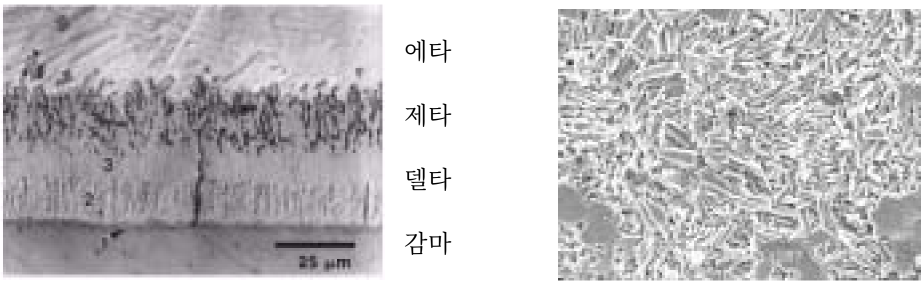
Step 3. $w \leftarrow w / \|w\|$

Step 4. If not converged, go back to Step 2.

위에서 기대치를 산출할 때 관측에 대한 산술평균을 사용한다.

3. 금속표면 상(phase)별 분율 예측

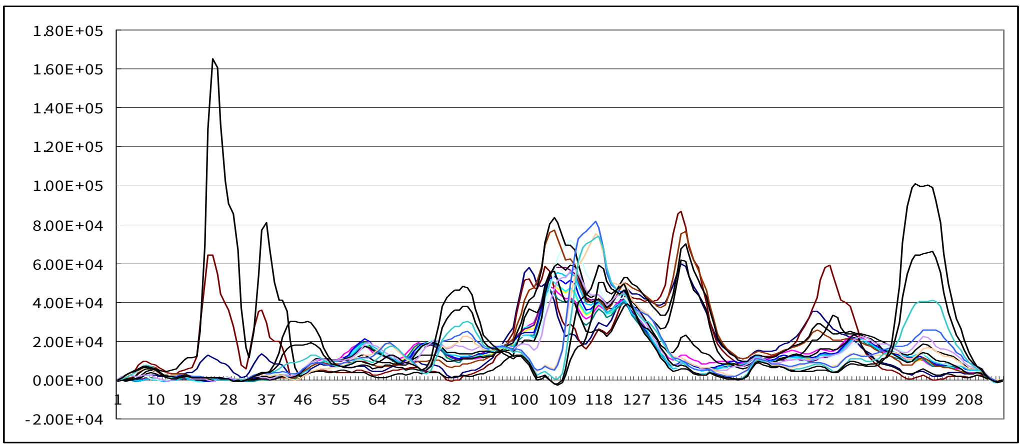
본 연구에서는 금속표면의 X-선 회절강도값을 이용하여 그 물질의 여러가지 상(phase)에 대한 분율을 예측하고자 한다. X-ray 회절법은 물질의 구조를 밝혀내는데 적용되는 분광기술이다. 금속의 회절 패턴으로부터 미지시료내의 물질을 확인할 수 있고 혹은 이미 확인된 물질의 원자 scale 의 구조를 분석할 수도 있다. 본 연구에서 사용된 시료는 네가지 상(phase)- 에타, 제타, 델타, 감마상-으로 구성되어 있고, 각 상의 피이크 (peak)는 어느 각도에서 나오는지 알려진 것이다. 여기서 문제가 되는 것은 회절강도값의 피이크의 규모에 따라 적분하여 원하는 상의 분율을 구하는 방법은 측정조건에 따라 피이크의 스케일이 달라지므로 일관성있는 결과를 얻을수 없으며, [그림 1]의 (b)와 같이 표면의 이미지 사진을 찍어서 이미지 분석을 하는 것도 각 상의 모양의 구분과 영역이 뚜렷하지 않아 어렵다는 것이다. 표면의 단면상 혹은 표면상으로는 품질특성치인 델타상 비율을 적절히 산정하는 것이 어렵다는 점에서 ICA 를 이용하여 source 인 네가지 상의 벡터를 얻고 원래의 X-선 회절강도값에 회귀시키는 방식으로 각 분율을 예측하게 되었다.



(a) 단면전자상　　　　　　　　(b) 표면전자상

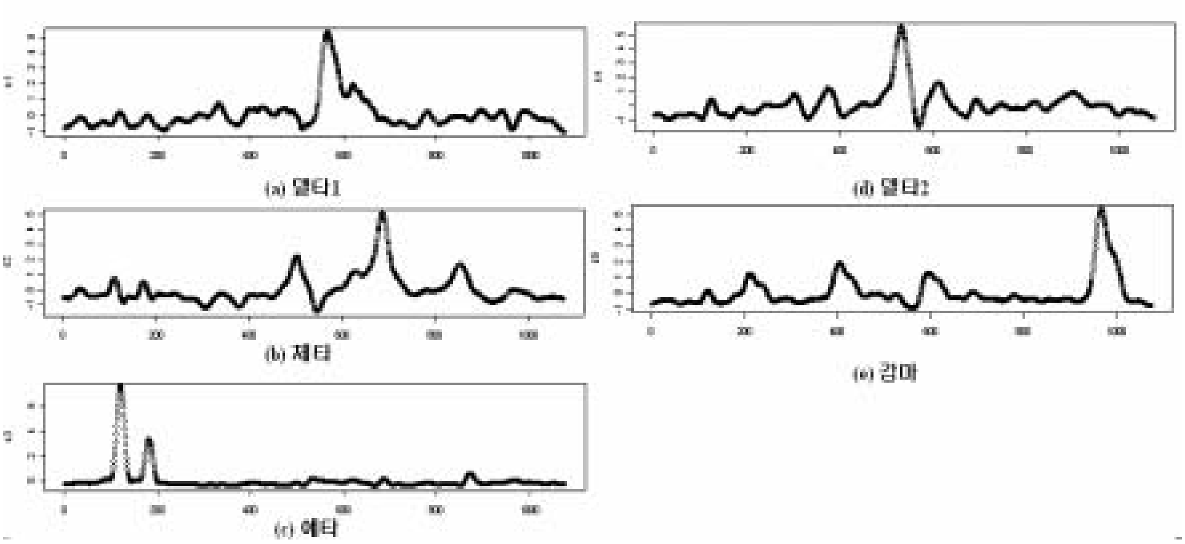
[그림 1] 금속의 단면과 표면상

본 연구에서 사용한 20 개 시료에 대해 일정간격으로 얻은 X-선 회절강도값은 [그림 2]와 같다.



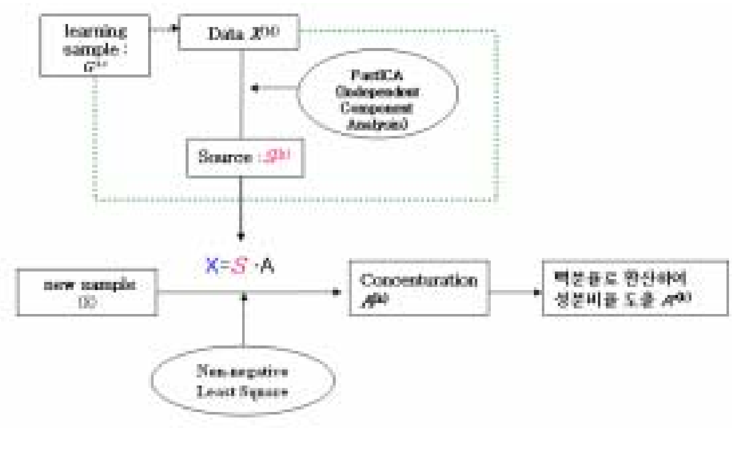
[그림 2] X-선 회절강도

이로부터 ICA 를 이용하여 얻은 소오스 스펙트럼은 [그림 3]에서 보는 바와 같이 5 가지로 분리되었는데 델타상이 두 개의 소오스로 구성되어 있음이 확인되었다. 다른 여러 개의 실험데이타로부터 일관성있게 [그림 3]의 소오스가 얻어짐으로써 ICA 적용의 적합성이 검증되었다. 분석프로그램은 R(version1.9)과 Matlab 을 사용하였다.



[그림 3] 추출된 5 가지 상별 독립 성분

다음으로 상별 구성비율 (분율)을 예측하기 위하여 다음과 같은 방법을 사용하였다. ICA 에 의해 S 를 얻은 후 선형회귀모형에서 적합시키듯이 원래의 X 에 대해 회귀시키면 각 시료에 대한 계수를 얻을 수 있는데, 여기서는 상의 분율이므로 계수가 0 보다 커야한다는 제한조건을 충족해야 한다. 따라서 ICA 로부터 얻은 독립적 구성성분, S 와 원래의 X 에 대해 비음최소자승법 (nonnegative least squares)을 적용하여 계수를 얻고 합이 100%가 되도록 하여 분율을 예측하고자 한다. 이와 같은 방법을 본 데이타에 적용한 결과는 [표 1]과 같다. 여기서, 델타 1 과 델타 2 는 분율산정시 합산된다. 위와 같이 제안된 방법을 정리하여 도식화하면 [그림 4]와 같다.



[그림 4] ICA 와 NNLS 에 의한 성분비율 예측과정

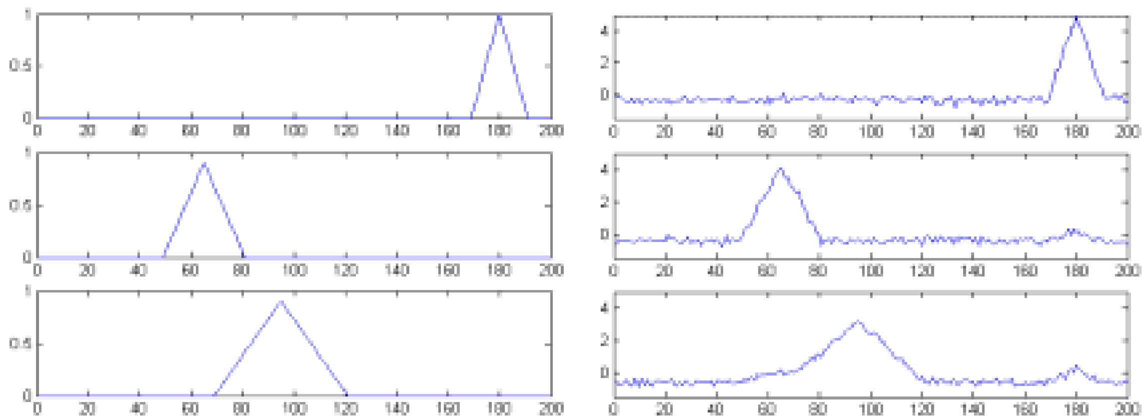
[표 1] 예측된 성분비율 (단위 %)

시료번호	감마	제타	에타	델타
1	0	46	10	44
2	0	11	0	89
3	0	0	0	100
4	0	0	0	100
5	0	25	64	11
6	0	41	0	59
7	2	8	0	90
8	4	0	0	96
9	6	0	0	94
10	0	16	0	84
11	0	34	0	66
12	0	18	56	26
13	0	27	0	73
14	1	0	0	99
15	6	0	0	94
16	10	0	0	90
17	18	0	0	82
18	38	0	0	62
19	61	0	0	39
20	82	0	0	18

4. 시뮬레이션

본 제안방법을 검증하기 위하여 시뮬레이션을 수행하였다. 시뮬레이션을 위해 [그림 5(a)]와 같은 세가지 종류의 소오스 S를 생성하였는데 여기서 S_2, S_3 은 일부 피크가 서로 겹치는 부분이 존재하도록 하였다. mixture 데이타 X 는 식(11)에 의거 생성하되, [표 2]에서 보는 바와 같은 비율로 각각의 소오스, S_1, S_2, S_3 를 곱하여 생성하였다. 소오스 구성비율의 합은 1(100 %)이다. 오차항 ε은 평균이 0 이고 편차가 1 인 정규분포로부터 랜덤하게 추출한 값에 소오스 데이타의 범위를 고려하여 0.01 을 곱해주었다. 데이타수는 19 개, 차원 (변수수)은 200 으로 하였다.

$$X = A \times S + \varepsilon \tag{11}$$



(a) original source (b) source by ICA

[그림 5] original source 와 ICA 에 의한 extracted source

시뮬레이션 데이타에 Fast ICA 알고리즘를 적용하여 얻은 소오스는 [그림 5(b)]와 같이 원래의 소오스와 유사한 형태이다. 세개의 소오스를 비음최소자승법에 적합시켜 얻은 구성비율은 [표 2]와 같다. 세가지 상, S_1, S_2, S_3 에 대한 실제 혼합비율과 예측비율의 차이는 [표 2 (b)]에서 보는바와 같이 각 표본의 세가지 비율의 오차에 절대값을 취하여 합한 값이 0.5 를 넘지 않는다. 전체 표본에 대한 평균오차는 0.21 이다.

평균오차는 식 (12)과 같이 산정하였다.

$$\text{평균오차} = \frac{\sum_{i=1}^{n}\left(\sum_{j=1}^{k}|S_j - \hat{S}_j|\right)_i}{n} \quad ,k=1,\ldots,3 \quad ,n=1,\ldots 19 \qquad (12)$$

[표 2] 실제 혼합비율과 예측 혼합비율 (단위 %)

(a) 실제 혼합비율

표본	S_1 비율	S_2 비율	S_3 비율
1	0	100	0
2	0	95	5
3	0	90	10
4	0	85	15
5	0	83	17
6	0	80	20
7	0	70	30
8	0	65	35
9	0	60	40
10	40	10	50
11	20	30	50
12	20	20	60
13	0	40	60
14	30	0	70
15	10	20	70
16	10	10	80
17	0	20	80
18	0	0	100
19	0	10	90

(b) 예측 혼합비율

S_1 예측	S_2 예측	S_3 예측	∑|오차|
0.21	99.67	0.13	0.24
0	94.92	5.09	0.07
0	89.74	10.26	0.21
0.12	84.5	15.38	0.37
0	82.95	17.05	0.04
0.3	79.65	20.05	0.26
0	69.54	30.46	0.37
0.16	64.31	35.52	0.51
0	59.87	40.13	0.11
39.93	10.06	50.02	0.05
19.95	30.14	49.92	0.1
20.05	19.9	60.05	0.07
0	39.85	60.15	0.13
30.19	0.11	69.7	0.21
9.89	19.89	70.22	0.16
9.95	9.93	80.12	0.08
0.18	20.21	79.61	0.28
0.31	0.5	99.19	0.58
0.12	10.03	89.85	0.11

5. 결론

ICA 는 오디오 프로세스, 의학영상 시그널분석, 이미지 분석, 통신분야, 계량경제분야 등 다양한 분야에서 그 적용범위가 넓어지고 있는 통계적기법으로서, 소오스가 미지인 경우 혹은 소오스를 계량화하여 이차적 분석이 필요한 경우 유용한 다변량기법이다.

본 연구에서는 미지의 구성성분을 분리하고, 분리된 성분을 이용하여 성분간 비율을 예측하는 방안을 제시하고, 금속표면의 X-회절강도 데이타에 적용하여 분율을 예측한 사례를 보였다. ICA 에 의한 성분도출과 NNLS 를 이용한 비율이 적합하게 추정되는지 평가하기 위하여 시뮬레이션을 수행한 결과 비교적 만족스러운 결과를 얻었다.

참고문헌

Hyvarinen, A. (1999), “Survey on independent component analysis”, *Neural Computing Surveys*, 2, 94-128.

Hyvarinen, A.,and Oja, E. (2000), “Independent component analysis: algorithms and applications”, *Neural Network*s, 13(4-5), 411-430.

Ikeda, S., and Toyama, K. (2000), “Independent component analysis for noisy data-MEG data analysis”, *Neural Networks* 13, 1063-1074.

Lee, D. D., and Seung, H. S.(1999), “Learning the parts of objects by non-negative matrix factorization”, *Nature*, vol. 401, 788-791.

Lee, D. D., and Seung, H. S. (2001), "Algorithms for nonnegative matrix factorization," in *Advances in Neural Information Processing Systems 13*. Cambridge, MA: MIT Press, 556-562

Viga´rio, R., Jousma¨ki, V., Ha¨ma¨la¨inen, M., Hari, R., & Oja, E. (1998), Independent component analysis for identification of artifacts in magnetoencephalographic recordings, *Advances in Neural Information Processing Systems*, Vol. 10. Cambridge, MA:MIT Press (pp. 229–235).