독립성분 분석을 이용한 순수 스펙트럼 분리

Pure-spectrum separation From simulated IR spectral data using Independent-component analysis

한 상 준, 조 혜 민, 윤 길 원
Medical Application Team, Samsung Advanced Institute of Technology,
P.O. Box 111, Suwon 440-600, Korea
sjhahn@sait.samsung.co.kr

중적외선 분광학(mid-IR spectrocopy) 에서는 2.5 ~ 15 μm 의 파장을 갖는 적외선을 시료에 조사하여 absorption spectrum 을 측정하는데 이 대역은 주로 분자들의 fundamental band 가 나타나는 영역으로 각 분자의 characteristic absorption band 가 비교적 깨끗하게 분리되어 있어 특정 compound 의 존재 유무를 확인하는데 쓰여 왔다.(1) 그러나 보통의 경우에 시료에는 한 가지 화합물만이 존재하는 것이 아니고 여러 가지 화합물이 섞여 있는 경우가 많다. 특히 관심있는 분자의 spectrum 이 혼합물을 구성하는 다른 성분의 spectrum 과 겹치게 되는 경우에는 화합물의 유무를 확인하기가 매우 어렵게 된다. 이 때 측정된 혼합물 시료의 스펙트럼으로 부터 구성성분의 순수 스펙트럼을 분리해 낼 수 있다면 특정 유기화합물의 존재 유무를 확인하는데 큰 도움이 될 것이다. IR spectrum 으로 부터 특정 성분의 농도를 추정하기 위하여 흔히 쓰이고 있는 Multivariate statistical method 에는 MLR, PCR, PLS 등이 있다. 그러나 이 방법들은 혼합물에서 구성성분의 순수 스펙트럼을 분리해내기 보다는 혼합물에 섞여 있는 특정성분의 농도를 추정하기 위하여 개발되었고 regression vector를 구하기 위한 calibration 을 하기 위해서는 사전적으로 calibration set 의 농도에 관한 정보를 필요로 한다. 또한 calibration 으로 부터 얻어지는 regression vector 도 관심 있는 성분 이외의 성분의 pure spectrum 과 contravariant 하지만 관심 있는 성분의 pure spectrum 과 일치 하지는 않는다. 그래서 이들 방법은 혼합물의 구성성분의 순수스펙트럼을 구하지 않고 농도를 추정한다는 의미에서 implict method 라고 불리운다. 반면 1990 초반에 개발되어 현재 biomedical signal processing, speech recognition, image separation 등의 분야에서 활발히 응용되고 있는 Independent Component Analysis(ICA) 는 PCR 이나 PLS 와는 달리 사전적으로 주어지는 농도에 관한 정보 대신 spectral data의 higher order statistics 를 이용하여 아무런 사전 정보 없이 pure spectrum 을 분리할 수 있다.(2) 본 논문에서는 Independent Component Analysis(ICA) 분석법을 이용하여 혼합물의 스펙트럼으로 부터 각 성분의 순수 스펙트럼을 분리해내는 방법에 대하여 연구하였다. 컴퓨터 시뮬레이션을 통하여 두 종류 화합물 A, B로 구성된 혼합물 100 개의 simulated IR spectra를 만들었다. 이때 혼합물을 구성하는 두 화합물의 농도는 통계적으로 서로 independent 하도록 구성하였다. 이렇게 만들어진 simulated IR spectra에 대해서 우선 주성분 분석(PCA)과 F-test를 통하여 혼합물에 섞여있는 두 성분과 관계 있는 두 개의 score 와 factor를 구한다. 그러나 일반적으로 PCA를 통하여 얻은 score 와 factor 는 혼합물을 구성하고 있는 성분의 농도와 순수 스펙트럼과 일치하지는 않는다. 즉, A를 혼합물의 스펙트럼을 나타내는 matrix라고 할 때 Beer's law 에 의하면

A 는 혼합물의 구성성분의 순수스펙트럼을 나타내는 matrix M 과 혼합물의 농도를 표현하는 matrix C 의 곱으로 표현된다.

$$A = M\ C$$

한편 PCA 에서는 A를 SVD(Singular Value Decomposition) 하여 score 와 factor 의 곱으로 표현한다.

$$A = \text{Factor Score}$$

그런데 일반적으로 M ≠ Factor, C ≠ Score 이다. 그림 1은 PCA로 부터 얻은 factor 와 구성성분의 스펙트럼을 보이고 있다. PCA 로부터 얻은 factor 가 순수스펙트럼과 일치하기 위해서는 다음의 관계를 만족하는 적당한 matrix W를 구해야 한다.

$$A = \text{Factor Score} = (\text{Factor W})\ (W^{-1}\ \text{Score}) = M\ C$$

W를 구하기 위해서 C가 통계적으로 independent 하다는 점을 이용한다. PCA 로부터 얻어진 score 들은 통계적으로 uncorrelated 되어 있지만 independent 하지는 않다. ICA 에서는 PCA 로부터 얻어진 score가 서로 independent 하도록 score 의 higher-order statistics를 분석하여 W를 구하게 된다. 즉, ICA 에서는 score를 다음과 같이 분해한다.

$$\text{Score} = W\ C'$$

ICA에서 구한 C′ 은 실제 혼합물의 농도 C 와 부호와 scale factor 가 다를 수 있다. PCA에서 구한 Factor와 ICA에서 구한 W를 이용하여 구성성분의 순수 스펙트럼과 농도를 추정할 수 있었다. (그림 2)

$$M = \text{Factor W}$$

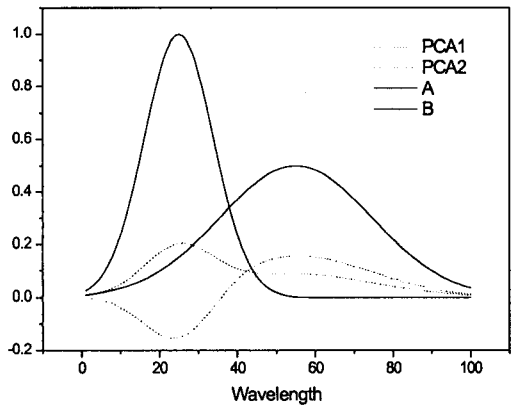

그림 1. 구성성분의 순수 스펙트럼과 PCA에서 얻어진 factor

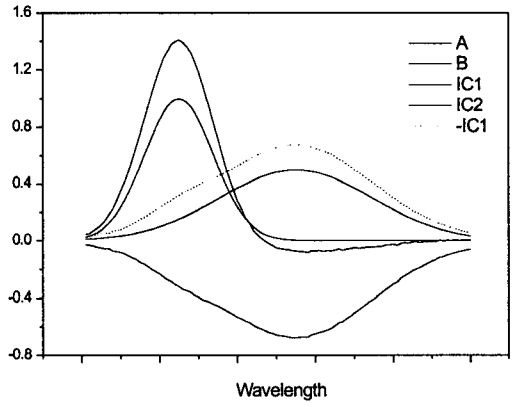

그림 2. 구성성분의 순수 스펙트럼과 ICA 로부터 추정된 순수스펙트럼

T
B

본 연구는 과학기술부 국가지정연구실(2000년 생체분광학연구실)에 의해 지원되었음.

1. Skoog, Holler, and Nieman "Principles of Instrumental Analysis", 5th ed.
2. Aapo Hyvarinen, Juha Karhunen, and Erkki Oja "Independent Component Analysis"