주성분 분석 방법을 이용한 DNA mircoarray 분석

권 성우, 한 종훈

(sw74@postech.ac.kr)

포항 공과 대학교 화학 공학과

1. 서론

많은 양의 유전체 데이터를 분석하기 위해서는 새로운 방식의 분석 기술이 필요하다. 이러한 기술들 중에서 대표적인 것이 바로 Microarray 기술이다. Microarray 기술에는 DNA chip, protein chip, lab-on a chip 등이 있다. DNA chip은 고형체에 고정 시킨 DNA와 mRNA나 다른 DNA를 잡종형성 시켜 만들게 되는데 특정 상태의 유전자 발현 양상을 연구 할 때 사용하고 있다.

DNA microarray 데이터의 경우 각 유전자간의 상관 관계가 많고 변수로 사용되는 유전자의 개수가 5000개에서 2만 4000개 사이로 매우 많은 특징이 있다. 따라서 여러 패턴 인식 방법을 적용하기 전에 데이터 전 처리를 많이 한다. 즉, 상관 관계를 제거하고 변수의 개수를 감소 시키기 위해서 전 처리로 주성분 분석법을 많이 사용한다. 본 글에서는 주성분 분석법에 대해서 알아보고자 한다.

2. 주성분 분석

원래 Pearson에 의해 개발된 주성분 분석은 다변량 문제들에 있어서 변수들의 상호 관련 구조들을 분석하기 위하여 많이 사용하였다. 이 방법은 다변량 통계 기법에 관한 많은 책들에서 설명 되고 있다.

주성분 분석과 같은 모든 다변량 자료 해석에 있어서 출발점은 다음과 같은 X로 나타내는 자료 행렬에서 시작된다.

$$X = \begin{pmatrix} x^*_{11} & \cdots & x^*_{1k} \\ \vdots & \ddots & \vdots \\ x^*_{n1} & \cdots & x^*_{nk} \end{pmatrix}$$

이 행렬에서 N개의 행들은 DNA microarray의 각 표본이고 k개의 열은 유전자이다.

주성분 분석을 이해하는 방법에는 기하학적 방법과 수학적 방법이 있다. 먼저 기하학적으로 주성분 분석을 이해 한다는 것은 주성분과 그에 대한 정사영인 주성분 점수 그리고 상관 관계를 포함하여 이용되는 수학적 연산까지도 공간에서 일어나는 것으로 이해 하는 것이다. 주성분 분석은 데이터를 발생시킨 생물체나 혹은 질병을 분석하기 위하여 데이터 공간에 대해 분포가 넓은 축부터 시작하여 차례대로 서로 직교하도록 새롭게 축들을 하나씩 정의하고 이들에 대한 주성분 점수들을 구하는 것에서 출발한다. 이때 데이터의 상관 관계를 더 쉽게 분석하기 위해 새롭게 정의된 새로운 축들을 주성분이라 하며 데이터를 이들 축에 투영시켜 얻은 정사영 값들을 그 축에 대한 주성분 점수라고 한다. 그런데 주성분 분석 방법이 실질적으로 적용될 때는 정의될 수 있는 주성분과 그들에 대한 주성분 점수들을 이용하여 생물체나 혹은 질병을 분석하는 것이 아닌 주요한 a개의 축들과 그것들의 주성분 점수들만을 가지고 그들의 선형 합으로 생물체나 질병을 근사한 후 이 근사된 시스템을 분석하게 된다.

이런 a개의 주성분으로 근사하는 것을 NxK 자료 행렬에 대해 수식적으로 표현하면 아래와 같다 (그림 1).

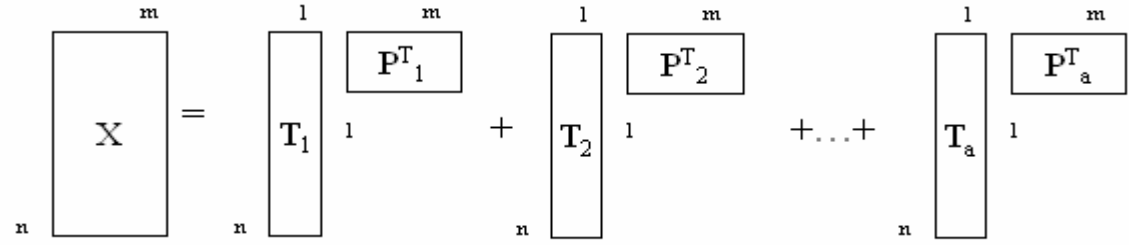

그림 1. 자료 행렬의 근사

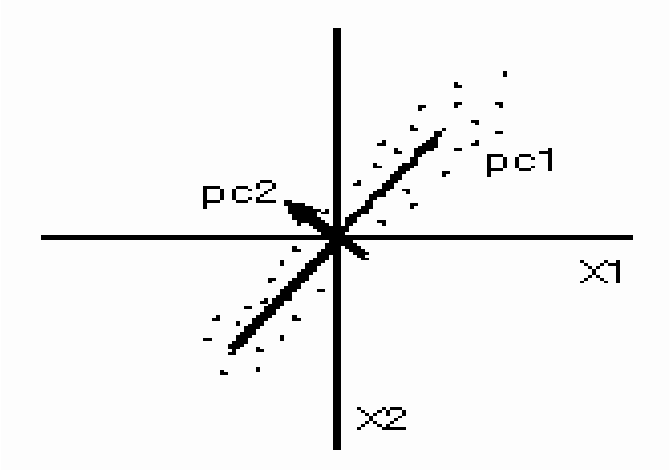

그림 2. 주성분 분석의 이해

한편, 자료 행렬에 존재하는 변수들의 모든 관계를 하나의 대표적인 관계로 근사한 것이 첫번째 주성분이라는 것을 알 수가 있다 (그림 2). 만약 a개의 주성분들이 있다면 첫번째 주성분은 자료 행렬에 존재하는 변수들의 모든 관계를 가장 잘 설명 할 수 있는 벡터이고 두번째 주성분은 그 다음으로 잘 설명하는 관계를 의미하며 a번째 주성분은 a번째로 변수들의 관계를 잘 설명할 수 있는 벡터를 의미한다. 또한 주성분 점수 (T)는 a개의 주성분들에 투영시킨 값들로 해석하는 것이 일반적이다. 따라서 a가 충분히 크다면 이들 변수들 간의 가장 대표적인 a개의 관계들과 주성분 값들을 가지고 이것들의 선형 합으로서 자료 행렬을 근사할 수가 있다는 것이다 (그림 1).

두번째로 주성분 분석을 수학적으로 이해하려면 선형 대수적으로 설명해야 한다. 자료 행렬을 normalization하고 난 후 $\frac{1}{N}X^T X$ 와 같은 행렬 곱을 하여 상관 행렬을 얻을 수 있는데 이 행렬의 열 공간이 바로 데이터 공간이 된다. 따라서 이 상관 행렬의 열 공간을 span하는 고유 벡터들이 바로 데이터 공간을 span하는 주성분들이 되며 각각의 고유 벡터에 대한 고유 값이 그 고유 벡터 방향으로 움직임의 대소를 결정한다. 다시 말해 고유 값이 크면 데이터 공간에서 고유 값 방향으로 데이터 분포의 분산이 크다는 것을 의미 한다. 따라서 주성분들을 정의할 때 고유 값이 큰 것부터 차례대로, 첫번째 주성분, 두번째 주성분등과 같이 정하게 된다. 이렇게 해서 고유 값이 큰 것부터 우리가 원하는 a개의 주성분을 구해 자료 행렬을 근사하게 된다. 이것을 수식적으로 표현하면 다음과 같다.

$$X = T_a P_a^T + E, \text{ (E: residual)}$$
$$E = X - T_a P_a^T$$

실제로 이와 같이 주성분 분석을 적용하여 자료 행렬을 근사할 때 중요하게 대두되는 문제 중에 하나가 몇 개 (a개)까지의 주성분을 사용할 것인가 하는 것이다.

이때는 매번 주성분이 하나씩 추가될 때마다 통계에서 모델 상호 교차 검사 (cross-validation)시 많이 사용하는 F-test를 하여 잔차(E)가 random error라고 판단될 때까지의 주성분 수를 구하면 된다.

주성분 분석을 통해 자료 행렬을 a개만의 주성분을 사용하여 생물체나 질병을 근사하므로 데이터 전체를 다룰 때에 비해 단순화와 차원 축소가 된다 (그림 3). 또한 축소된 차원을 이용하여 군집화 분석 이나 분류 작업등을 할 수가 있고 (그림 4) 새로운 입력 데이터에 대해서 출력 데이터를 예측할 수도 있으며 주성분 점수 그림 (T_1 vs T_2)을 해보면 간단한 분류를 할 수가 있다 (그림 5). 더욱이 어느 군집에도 속하지 않는 이상치를 찾아낼 수 있으며 고유 벡터 그림 (P_1 vs P_2)과 비교해 보면 어느 유전자에 의해서 이러한 이상치가 발생했는지 알 수가 있다. 이를 통해 이상 진단과 특정 유전자 선택이 가능한 것이다.

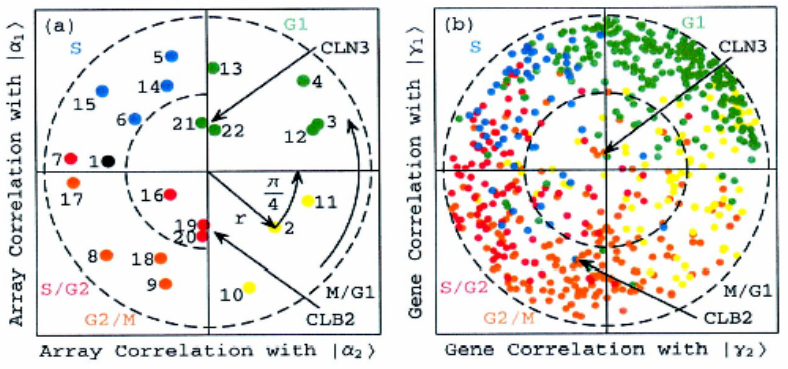

그림 3. 효모 분열의 DNA microarray 데이터를 주성분 분석법으로 분석 한 예이다. 첫 번째 주성분과 두 번째 주성분을 그림으로 표시한 것으로 효모 분열이 진행됨에 따라서 반 시계 방향으로 과다 발현 하는 유전자들이 바뀌는 것을 알 수가 있다.

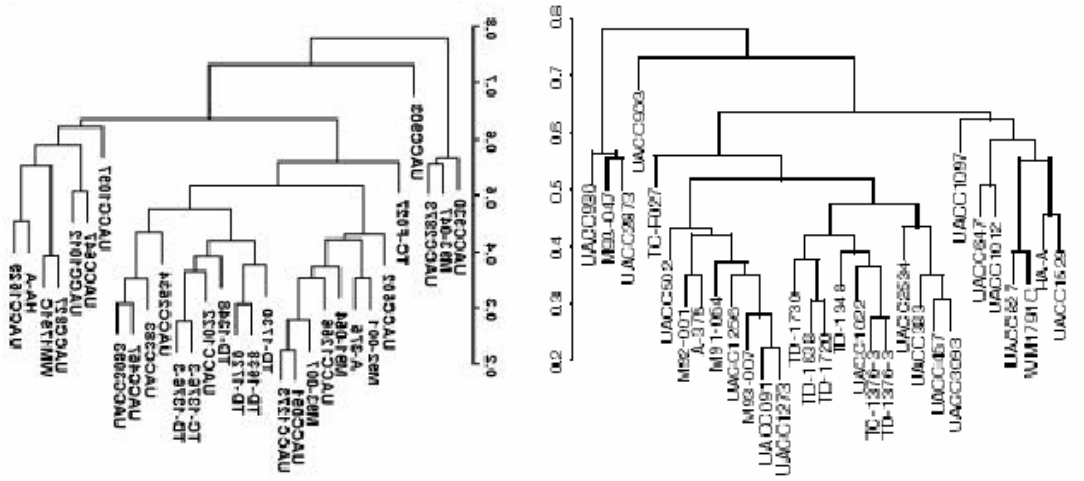

그림 4. 주성분 분석법에 기초한 암 환자 군집화와 원래 유전자를 이용한 암 환자

군집화 분석의 비교이다. 왼쪽이 주성분 분석을 수행한 후 주성분 점수를 이용하여 군집화 분석을 한 것이고 오른쪽은 원래 변수를 이용하여 군집화 분석을 이용한 것이다. 결과가 둘 다 매우 유사한 것을 알 수가 있다.

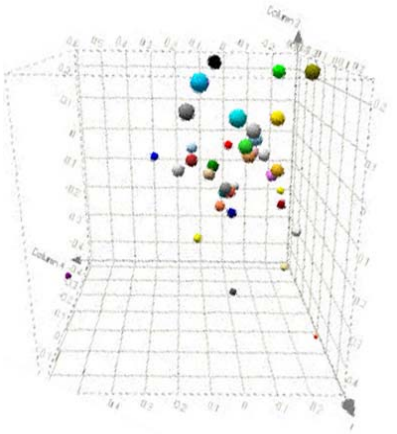

그림 5. 주요한 세 개의 주성분으로 암 환자 표본을 표시한 예이다. 각 색깔은 암의 종류를 나타내고 있다.

4. 주성분 분석시 주의 사항

주성분 분석시 자료 행렬의 정보를 가장 많이 설명하는 주성분을 찾게 된다. 암 환자 DNA microarray의 경우 암 분류가 목적이기 때문에 자료 행렬의 정보에서도 각 환자 분류와 관련된 정보만이 필요 하다. 하지만 주성분 분석만을 수행하면 각 환자의 분류와 관련된 정보뿐만 아니라 분류와 상관 없는 자료 행렬의 정보를 가장 많이 설명하는 것을 주성분으로 결정한다 (그림 6). 이러한 경우 주성분 분석 외에 ICA (independent component analysis)와 같은 다른 분석 방법을 사용하거나 주성분 분석과 SDA (stepwise discriminant analysis)와 같은 변수 선택 방법을 동시에 사용한다.

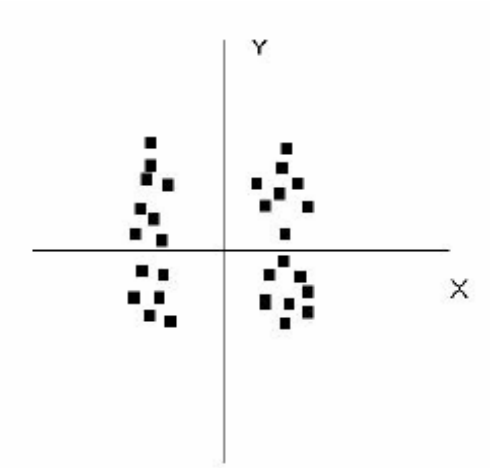

그림 6. 주성분 분석과 분류. 왼쪽 자료와 오른쪽 자료가 각기 다른 종류 암에 걸린 환자 표본이다. 만일 주성분 분석을 하게 된다면 각 암을 분류하는데 주요한 X축을 주성분으로 선택하지 않고 자료의 분산를 최대화하는 Y축으로 주성분을 선택하게 된다. 따라서 선택된 주성분으로는 암 환자의 분류가 더욱 어려워 지게 된다.

참고 문헌

M. Bittner et al., (2000). Molecular classification of Cutaneous malignant melanoma by gene expression profiling. *Nature*. 406. 536-540.

Orly. A et al., (2000). Singular value decomposition for genome-wide expression data processing and modeling. *PNAS*. 97. 10101-10106.