**2장- 이론 설명 절 (하나씩 추가 예정)**

***주성분 분석 방법론 소개***

주성분 분석 (PCA, Principal component analysis)는 가장 보편적인 차원 축소 방법론으로 다차원 데이터를 압축 및 시각화하여 현상의 핵심적 원인을 규명하는 방법론이다. 또한, 변수 차원 (descriptor space)의 데이터를 정보 차원 (Information space)로 투영하여, 가장 주요한 변수를 파악할 수 있고, 데이터 분산 정도를 파악해 이상점 (outlier) 제거에 효과적이다. 이처럼 정보의 특이성 식별에 효과적인 까닭에 오늘날에는 신경 분포 조사, 재료 물성 분석, 유전자 유사도 등 다양한 분야에서 응용된다.

주성분 분석의 원리는 저차원의 초평면 (Hyperplane)에 데이터 분산이 최대가 되는 벡터를 찾는 것이다. 아래 데이터의 공분산 (Covariance) 식을 유클리드 공간에 가정하여 데이터 분산을 계산할 수 있다.

*mx*와 *my*는 임의의 *x*와 *y*에대한 평균이다. 따라서 공분산 행렬 *C*는 다음과 같이 정의된다.

Eigen-decomposition 방법론을 사용하여 직교 행렬 *P*와 대각 행렬 *P*T를 사용하여 Eigenvector *e*와 Eigenvalue *λ*를 정의할 수 있다.

여기서 Eigenvector와 Eigenvalue는 각각 주성분 분석을 통해 얻은 Score와Loading 값이다. Eigenvector는 곧 정보 차원의 주 성분이며 데이터의 위치를 설명하는 좌표축이다. Score는 정보 차원에서의 데이터 좌표이며 Loading은 임의의 변수의 정보 차원 형성 기여도이다.

정보 차원은 주성분으로 구성되며 PC (Principal component)로 표현된다. 이를테면 첫번째 차원은 PC1, n번째 차원은 PCn으로 표현할 수 있다. 데이터의 분산을 가장 잘 설명하는 주성분은 PC1으로 첫번째 성분이 된다. PC1은 가장 중요한 차원으로 주성분 중 가장 높은 데이터 설명력 (Dimension coverage)을 갖는다. 따라서 차원 축소를 위해서는 데이터 설명력과 변수의 차원 형성 기여도를 함께 확인하는 것이 일반적이다.

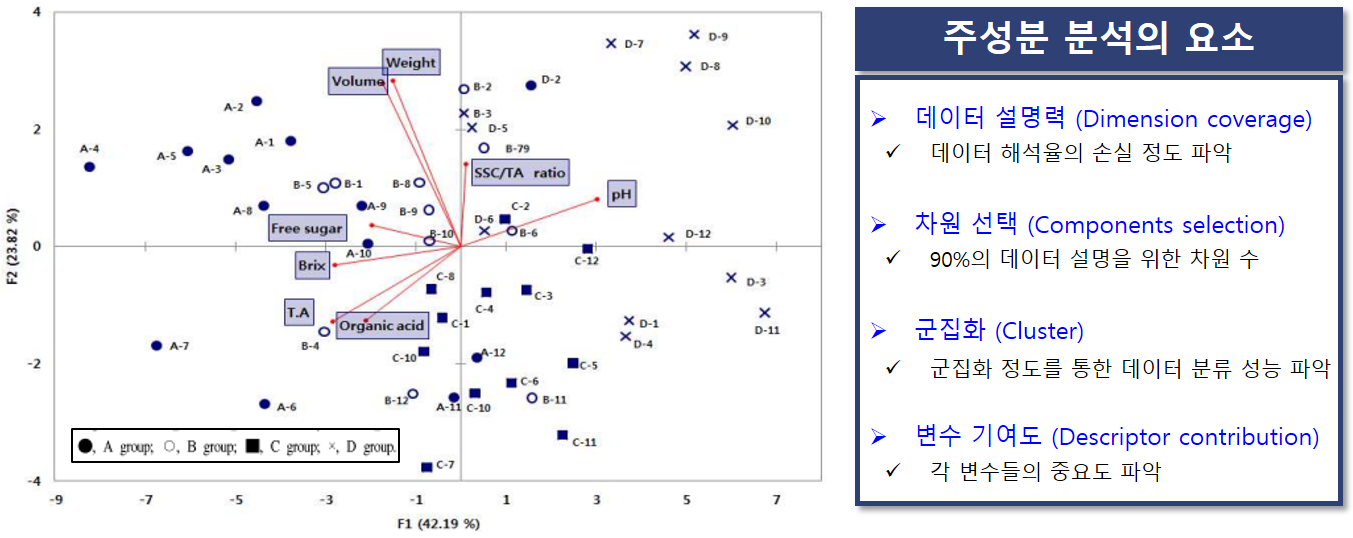


그림 1.6. 주성분 분석의 예시와 요소

주성분 분석을 활용한 데이터 분석에는 크게 4가지 요소 (데이터 설명력, 차원 선택, 군집화, 변수 기여도)가 있는데 이를 사과 분류 데이터를 활용해 설명한다(그림 1.6). 사과는 8가지 차원 (weight, volume, free sugar, brix, T.A, Organic acid, SSC/TA ratio, pH)이 존재하고 사과의 분류를 해내는 것이 예시 분석의 목적이다.

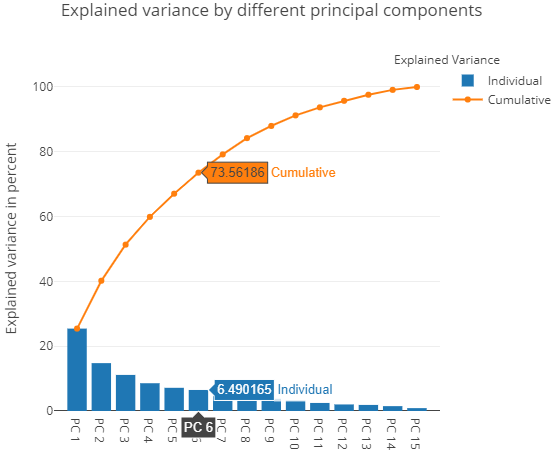


그림 1.7. 주성분 수 결정을 위한 Scree plot과 Cumulative plot

데이터 설명력 (Dimension coverage)이란, 정보 차원으로 변환 과정에서 데이터 손실 정도와 각 주성분의 데이터 분산 설명 정도를 표현한 척도이다. 일반적으로 차원 변환 시에 낮은 차원으로의 시각화 과정에서 데이터의 왜곡이 발생하기 때문에 데이터의 손실이 발생한다. 이때 데이터 설명력은 시각화된 정보와 원본 데이터와의 괴리감을 수치적으로 표현한다. 그림 1.6에서 x축과 y축 옆에 표시된 (42.19%)와 (23.82%)이 데이터 설명력으로 PC1 (F1)이 전체 데이터 분산의 42.19%, PC2 (F2)가 전체 데이터 분산의 23.82%를 설명한다는 의미이다. 데이터 시각화를 위해서 보통 2~3개의 주성분이 채택되지만, 최소한의 데이터 설명력을 확보하기위한 차원 선택 (Dimension selection) 과정에 Kaiser-Harris 기준이 있다. Kaiser-Harris 기준에 의하면 데이터 설명력의 합은 80~90%가 넘어야하며, 이를 위해 Scree plot과 Cumulative plot으로 필요한 주성분의 수를 파악할 수 있다. 사과의 분류 데이터에서는 80% 수준의 데이터 설명력을 갖기 위해 최소 PC1부터 PC8까지 사과의 분류를 위해 채택되야한다 (그림 1.7).

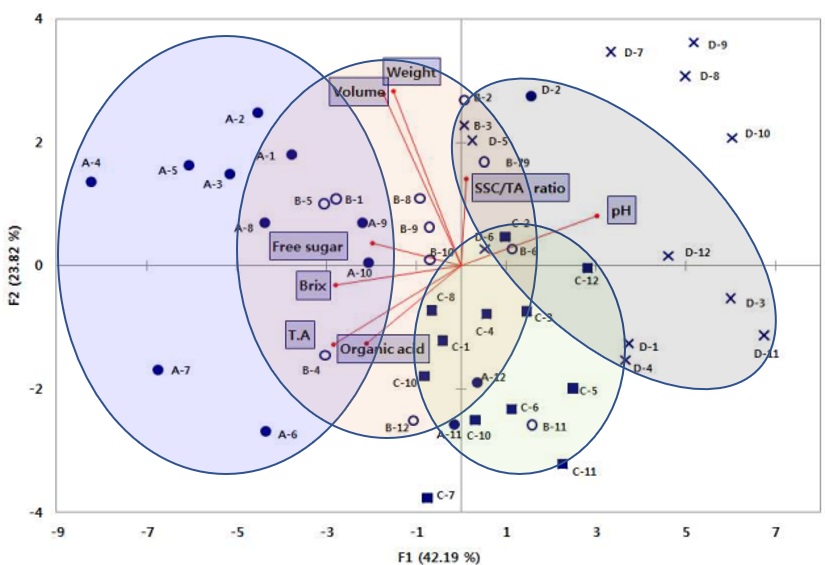
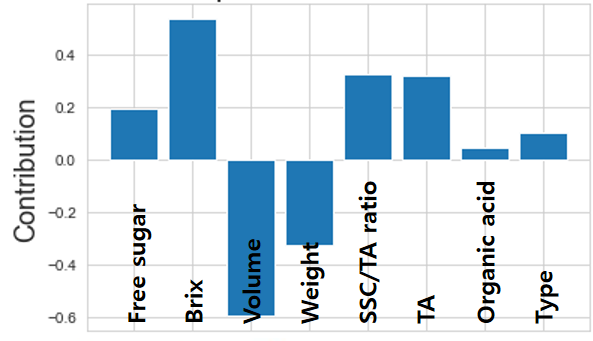
 

그림 1.8. (좌) 데이터 분류 성능 파악을 위한 군집화, (우) 변수 기여도 파악

군집화 (Clustering)정도의 파악은 정보 차원이 현실 분류와 괴리를 파악하여 시각화가 적절히 이루어졌는지 파악하기 위한 과정이다. 각 독립 변수의 영향력을 색이나 형태로 시각화 하므로서 종속변수 변화에 가장 민감한 독립변수를 하악할 수 있다. 이를테면 사과의 분류 데이터에 질량이나 부피와 같은 독립변수를 색으로 나타내 해당 독립변수의 종속변수에 대한 영향력과 경향을 파악할 수 있다 (그림 1.8). 변수 기여도 (Loading)은 변수들이 각 차원을 형성하는데 기여한 정도를 의미한다. 사과의 분류 데이터에서 가장 중요한 주 성분은 PC1 (F1)과 PC2 (F2) 이므로 해당 주 성분이 형성되는데 가장 기여를 많이 한 변수가 중요한 변수로 판단한다. 변수 기여도는 각 변수가 주성분에 절댓값만큼 기여한다. 이를테면 Free sugar가 Type에 두배 더 큰 수 이므로 두배 더 중요하는 것을 알 수 있다. 이처럼 주성분 분석을 통해 고차원의 복잡계 (Complex system) 데이터를 해석할 수 있고, 변수 기여도 파악 및 주요 변수 식별 등 다양한 데이터 분석 방법이 가능하다. 그러므로 본 실습에서는 주성분 분석 방법론을 이용하여 유기용매 나노여과 분리막을 분석해보도록 한다.