# 적응적 상관도를 이용한 주성분 분석에 관한 연구

고명숙*
*부천대학교 경영과
kms@bc.ac.kr

# A Study on PCA using Adaptive Correlation

Myung-Sook Ko*
*Dept. of Business Administration, Bucheon University

요 약

고차원의 데이터를 처리하기 위해서는 데이터의 성질을 유지하면서 특징을 잘 반영할 수 있는 특징 추출 방법이 필요하며 주성분분석 방법은 대표적인 특징 추출 방법이다. 본 연구에서는 데이터가 고차원인 경우 데이터 특징 추출을 위한 주성분 분석의 주성분 변수 선정시 적응적 상관도(Correlation)를 기반으로 한 주성분 분석 방법을 제안한다. 제안하는 방법은 입력 데이터간의 상관관계를 기반으로 상관도를 적응적으로 반영하여 데이터의 주성분을 분석함으로써 실제 데이터의 특징을 나타내는 세분화 변수 선정 시 데이터 편향성의 영향을 줄이기 위한 방법이다.

## 1. 서론

고차원의 데이터는 데이터 분포 형태를 알 수 없을 뿐만 아니라 많은 양의 메모리와 계산을 필요로 한다. 이러한 데이터를 처리하는 과정에서 차원을 낮추고 데이터의 특징을 추출하기 위해 사용하는 대표적인 기법 중 주성분 분석 기법이 많이 사용된다[1]. 주성분 분석(Principle Component Analsys; PCA)은 고차원의 데이터를 저차원의 데이터로 환원시켜 서로 연관 가능성이 있는 고차원 공간의 표본들의 공분산 행렬을 원 변수의 선형 결합을 이용하여 분석하는 방법으로서 선형 연관성이 없는 분산 기반 저차원 공간으로의 사상을 통하여 주요성분들을 축으로 하여 선형 변환한다[1,2]. 본 논문에서는 주성분 분석의 주성분 변수(또는 세분화변수) 선정시 상관도(Correlarion)를 적응적으로 적용하여 데이터 편향성의 반영도를 낮추고 연관성을 높이는 데이터 특성을 반영하여 고 차원 데이터의 특징을 더 잘 반영한 주성분 분석의 세분화 변수를 얻을 수 있는 방법을 제안하고자 한다.

## 2. 주성분 분석 및 세분화변수 선정

주성분 분석은 데이터의 특성을 찾아내는 가장 대표적인 방법 중의 하나로서 고차원 데이터의 정보 손실을 최소화하는 저차원의 세분화 변수를 통하여 전체 데이터를 표현하는 방법이다[1].

x와 y의 공분산(covariance) cov는 다음과 같이 정의될 수 있다.

$$cov(x,y) = E[(x-m_x)(y-m_y)] = E[xy] - m_x m_y$$

단, $m_x$는 x의 평균, $m_y$는 y의 평균, E[]는 기대값

x의 분산은 x값들이 평균을 중심으로 얼마나 흩어져 있는지를 나타내고, x와 y의 공분산은 x, y의 흩어진 정도가 얼마나 서로 상관관계를 가지고 흩어져 있는지를 나타낸다. 공분산 행렬(covariance matrix) C는 데이터의 좌표 성분들 사이의 공분산 값을 원소로 하는 행렬로서 데이터의 I번째 좌표와 j번째 좌표의 공분산 값을 행렬 i행 j열 원소값으로 하는 행렬을 말한다[1].

$$x = [x_1,...,x_n]^T : n\text{차원 } T\text{갯수의 열벡터}$$
$$y_1 = a_{11}x + a_{12}x + ... + a_{1p}x_p = a_1^T x$$
$$y_n = a_{p1}x1 + ax + ... + a_{pp}x_p = a_p^T x$$
$$C = E[(x_i - m_\xi)(x_j - m_{xj})] : n \times n\text{행렬}$$

아래 식에서 주성분 분석 세분화 변수 중 제1세분화 변수는 데이터의 특성을 가장 잘 반영하는 변

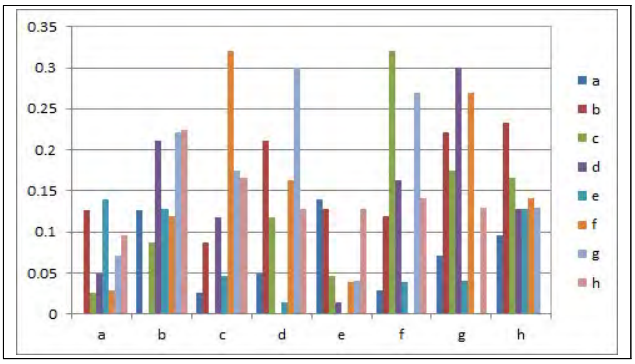
수이며(고유벡터 $e_1$), 제2세분화변수는 $e_2$로 다음과 같이 정의된다.

$$Ce_i = \lambda_i e_i$$
$$e_i : eigenvector of c$$
$$\lambda_i : eigenvalue, e_i 방향으로의 분산$$
$$\lambda_1 \geq ... \geq \lambda_n \geq 0$$
$$e_1 : 분산이 가장 큰 방향$$
$$e_2 : e_1 에 수직이면서 다음으로 가장 분산이 큰 방향$$
$$e_k : e_1, ... e_{k-1} 에 모두 수직이면서 가장 분산이 큰 방향$$

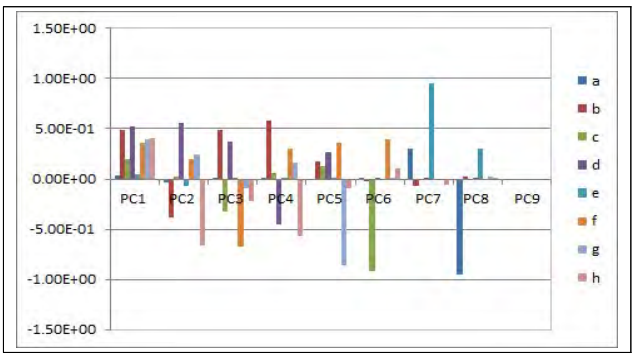
제1세분화변수 이후의 세분화변수는 제1세분화변수로 설명할 수 없는 자료의 변동을 설명하며 제k세분화변수로 갈수록 원래 데이터에 대한 특징 반영도는 낮다고 볼 수 있으며 고유값(eigenvalue) 그래프를 사용하여 세분화 변수 수를 결정한다[1-4].

## 3. 상관도 기반 세분화 변수 선정

본 논문에서는 전체 데이터에 대한 주성분분석 결과에 대하여 상관도를 기반으로 하여 세분화변수를 결정하고자 한다. 또한, 변수간의 상관도를 적응적으로 고려함으로써 전체 데이터에 대한 세분화 변수를 결정하고자 한다. 먼저, 상관 분석을 수행한 후 상관도(Correlation) 결과 값을 기반으로 상관도가 높은 데이터 변수를 제거해 나간다. 세분화 변수 수를 결정하기 위하여 고유값(eigenvalue) 그래프를 사용하여 세분화 변수 결정시에 서로 밀접하게 영향을 끼칠 수 있는 변수들의 영향을 최소화하기 위하여 상관도가 높은 변수들을 차례로 제거한 후 고유값 그래프를 완성함으로써 적응적으로 세분화 변수를 개수 결정에 반영하고자 한다. 다음 그림 1은 8개 변수 1000개(변수a~변수h, generatedata.com)에 대하여 변수 간의 상관도를 계산한 결과이며, 여기에서 (b,d), (c,f), (b,g), (d,g), (f,g), (b,h) 의 상관도가 높게 나타남을 알 수 있다. 상관도가 높음으로 인하여 데이터 주성분 분석의 세분화변수가 편향되게 결정되는 것을 막기 위하여 앞에서 나타난 여러 변수와 중복적으로 상관도가 높은 변수들 중 (b,d), (b,g), (b,h)쌍에서 변수 b와 (b,g), (d,g), (f,g) 쌍에서 변수 g가 세분화 변수 선정 시 편향화를 초래할 수 있으므로 이 두 변수 b와 g를 제거 대상으로 선정한다. 다음 단계로 전체 데이터에 대하여 PCA eigenvector 계수 값을 구한 후 두 데이터 변수간의 eigenvector 계수 값이 낮은 것을 차례로 제거해 나간다. 다음 그림 2는 8개의 변수에 대한 PCA 결과를 보여주는 그래프이다. 여기서 eigenvector 계수 값이 낮은 PC9, PC8 순으로 세분화 변수(변수a, 변수e)를 제거하는 방식으로 처리한다.




(그림 1) Correlation coefficient of input data




(그림 2) PCA of input data

## 4. 분석 및 결론

본 논문에서는 입력 데이터의 주성분 분석의 세분화변수 결정시 데이터 편향성의 영향을 줄이기 위해 여러 변수와 중복적으로 상관도가 높은 변수로 분석된 변수를 적응적으로 제거하는 방법을 제안하였다. 주성분 분석 결과인 PC1, PC2 등의 주성분(데이터 특징)을 추출 시 상관 관계를 기반으로 상관도가 높은 변수들을 차례로 제거한 후 8개 변수에 대하여 고유값(eigenvalue) 그래프를 도식하면 주성분분석의 세분화 변수는 적응적 상관도 적용 전 b,f,h,d에서 a,c,d,e로 선정되었음을 알 수 있으며 편향성 유도 변수는 세분화 변수에 포함되지 않았음을 알 수 있다.

## 참고문헌

[1] I.T. Jollife, "Principle Component Analysis" Springer-Verlag, New York, 1986.
[2] B .J. Kim, et al, "On-line Nonlinear Principal Component Analsys for Nonlinear Feature Extraction", The Journal of KISS, 31(3) pp.361-368, 2004.
[3] Y. J. Kim,"Evaluation of Urban Lakes Water Quality Using Principle Component Analysis",The Journal of KSEA, 9(2) pp.197-203, 2003.
[4] H.J. Joo, N.H. Kim et al,"A Study on Data Types ans Visualization for Traffic Congestion and Accidents", Proceeding of IEIE, 2019, pp1011-1013.