

# A Study on Selecting Principle Component Variables Using Adaptive Correlation

Ko Myung-Sook<sup>†</sup>

## ABSTRACT

A feature extraction method capable of reflecting features well while maintaining the properties of data is required in order to process high-dimensional data. The principal component analysis method that converts high-level data into low-dimensional data and express high-dimensional data with fewer variables than the original data is a representative method for feature extraction of data. In this study, we propose a principal component analysis method based on adaptive correlation when selecting principal component variables in principal component analysis for data feature extraction when the data is high-dimensional. The proposed method analyzes the principal components of the data by adaptively reflecting the correlation based on the correlation between the input data. I want to exclude them from the candidate list. It is intended to analyze the principal component hierarchy by the eigen-vector coefficient value, to prevent the selection of the principal component with a low hierarchy, and to minimize the occurrence of data duplication inducing data bias through correlation analysis. Through this, we propose a method of selecting a well-presented principal component variable that represents the characteristics of actual data by reducing the influence of data bias when selecting the principal component variable.

Keywords : Principle Component Analysis, Correlation, Eigenvalue Graph, Eigenvector Coefficient

## 적응적 상관도를 이용한 주성분 변수 선정에 관한 연구

고 명 숙<sup>†</sup>

### 요 약

고차원의 데이터를 처리하기 위해서는 데이터의 성질을 유지하면서 특징을 잘 반영할 수 있는 특징 추출 방법이 필요하다. 주성분분석 방법은 고차원 데이터에 포함된 정보를 저차원의 데이터로 변환하여 원래 데이터의 변수 수보다 적은 수의 변수로 고차원 데이터를 표현 할 수 있는 방법으로 데이터의 특징 추출을 위한 대표적인 방법이다. 본 연구에서는 데이터가 고차원인 경우 데이터 특징 추출을 위한 주성분 분석에 있어서 주성분 변수 선정 시 적응적 상관도를 기반으로 한 주성분 분석 방법을 제안한다. 제안하는 방법은 입력 데이터간의 상관 관계를 기반으로 상관도를 적응적으로 반영하여 데이터의 주성분을 분석함으로써 다른 여러 변수에 중복적으로 상관도가 높은 변수와 주성분을 유도하는데 연관성이 적은 변수를 주성분 후보 대상에서 제외시키고자 한다. 고유벡터 계수 값에 의한 주성분 위계를 분석하고 위계가 낮은 주성분이 변수로 선정이 되는 것을 막고 또한 상관 분석을 통하여 데이터의 중복 발생이 데이터 편향을 유도하는 것을 최소화하 하고자 한다. 이를 통하여 주성분 변수 선정 시 데이터 편향성의 영향을 줄임으로써 실제 데이터의 특징을 잘 나타내는 주성분 변수를 선정하는 방법을 제안하고자 한다.

키워드 : 주성분 분석, 상관도, 고윳값 그래프, 고유벡터 계수

### 1. 서 론

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 중에서 주성분 분석 기

법이 많이 사용된다[1]. 주성분 분석(Principle Component Analysis; 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 상상을 통하여 주요성분들을 축으로 하여 선형 변환한다. 즉 다차원 데이터에 포함된 정보의 손실을 최소화하여 저차원의 데이터로 변환한다. 원래 데이터의 변수 개수보다 더 적은 개수의 변수를 이용하여 전체 고차원 데이터를 표현하는 방법이다. 주성분(Principle Components, PC)들은 통계적 의미로 서로 독립성을 가진다. 데이터가 내포하는 모든 주성분

\* 이 논문은 2020년 한국정보처리학회 춘계학술발표대회에서 “적응적 상관도를 이용한 주성분 분석에 관한 연구”의 제목으로 발표된 논문을 확장한 것이다.

<sup>†</sup> 정 회 원 : 부천대학교 경영과 교수

Manuscript Received : July 22 2020

Accepted : October 1, 2020

\* Corresponding Author : Ko Myung-Sook(kms@bc.ac.kr)

을 사용할 경우에는 정보의 손실은 zero 이다. 제1주성분이 데이터의 특성을 가장 잘 반영하는 주성분이며, 이후의 주성분은 제1주성분으로 설명되지 못한 데이터의 특성을 설명하는 주성분이며 제2주성분, 제3주성분, 제4주성분 등으로 위계 순서가 증가할수록 원 데이터에 대한 특성 반영도는 낮아진다[1-5].

주성분을 추출했을 때 위계가 낮은 주성분이 고유벡터 계수 값이 높게 나타남으로써 주성분 변수 선정 시 최종 주성분 변수 결정에 side-effect 적인 영향을 끼치는 경우가 발생하게 된다. 이러한 영향을 최소화하기 위하여 첫째, 주성분 변수 선정 시 먼저 고유벡터 계수 값에 의한 주성분 위계를 분석한다. 둘째, 이를 통하여 위계가 낮은 주성분이 최종 변수로 선정이 되는 것을 막고 또한 상관 분석 결과를 반영한다. 이러한 단계를 통하여 데이터의 중복 발생이 데이터 편향을 유도하는 것을 최소화 시키고자 한다.

본 논문에서는 주성분 분석의 주성분 변수(또는 세분화변수) 선정 시 상관도(Correlation)를 적응적으로 적용하여 데이터 편향성의 반영도를 낮추고 고유벡터 계수 값에 의한 주성분 위계 분석을 통하여 연관성을 높이는 데이터 특성을 반영하여 고 차원 데이터의 특징을 더 잘 반영한 주성분 분석의 세분화 변수를 얻을 수 있는 방법을 제안하고자 한다. 제 2장에서는 주성분 분석 및 세분화변수 선정 방법에 대해 설명하고 제 3장에서는 상관도를 기반으로 한 세분화 변수를 선정하는 방법에 대해 기술한다. 제 4장에서는 제안한 주성분 분석 선정 방법에 대한 실험 결과 및 분석을 하고 제 5장에서 결론 및 향후 연구 과제에 대하여 기술한다.

**2. 주성분 분석 및 세분화변수 선정**

주성분 분석은 여러 개 ( $p \geq 2$ )의 양적변수(quantitative variable)들 사이의 분산-공분산 관계를 이용하여 이 변수들의 선형결합(linear combination)으로 표시되는 주성분(principal component)을 찾고, 이 중에서 중요한  $k(\leq p)$ 개(보통 2~3개)의 주성분으로 전체 변동(variance)의 대부분을 설명하고자 하는 다변량분석법이다. 주성분분석은 자료의 요약이나 선형관계식을 통하여 차수(dimension)를 감소시켜 해석을 용이하게 하는데 목적이 있으며(data reduction and interpretation), 주성분들은 서로 상관없이 없거나 또는 독립적인 새로운 변수들로서 정보의 손실이 최소화되도록 구한다. 변수들 사이의 관계를 분석하기 위해서는 주로 공분산행렬(covariance matrix) 또는 상관행렬(correlation matrix)을 사용하여 고차원의 자료를 2차원 또는 3차원 주성분 공간으로 사영(projection)시켜 저차원 공간에 그래프로 나타내어 자료가 갖는 특성을 찾는다. 고차원 공간의 다수개의 관측점을 가장 잘 근접하는 1차원 부공간을 찾는다. 즉, 가장 설명력 이 높은 새로운 축은 가장 큰 고유값에 해당되는 고유벡터에 의해 결정되며, ( $p \times 1$ ) 관측점  $x_1, x_2, x_3, \dots, x_n$  을 ( $p \times 1$ ) 단위벡터  $e(e^T e = 1)$ 에

사영시킨다. 구해진 주성분은 회귀분석에서 포함시킬 설명변수의 수를 결정하는데 사용되기도 하며 군집분석(cluster analysis) 등과 같은 분석법의 입력변수로서 이용되기도 한다. 특히 인자분석(factor analysis)에서는 인자를 구하는 방법 중의 하나로 이용할 수 있다. 중요한 주성분들의 산점도(scatter diagram) 또는 Q-Q plot은 정규성 검정에 이용되며, 중요하지 않은 주성분들의 산점도나 Q-Q plot은 이상점(특이점)의 탐색에 이용할 수 있다[6-10].

주성분 분석은 데이터의 특성을 찾아내는 가장 대표적인 방법 중의 하나로서 고차원 데이터의 정보 손실을 최소화하는 저차원의 주성분 변수를 통하여 전체 데이터를 표현하는 방법이며 변수들의 공분산 행렬을 원 변수의 선형 결합을 이용하여 설명하고자 하는 분석 방법이다. 공분산 행렬을 가장 많이 설명하는 선형 결합된 변수를 주성분(Principle Component)이라고 하며 변수들의 차원 축소 및 해석을 위해 사용한다[1].

$x$ 와  $y$ 의 공분산(covariance) cov는 Equation (1)과 같이 정의될 수 있다.

$$cov(x, y) = E[(x - m_x)(y - m_y)] = E[xy] - m_x m_y \quad (1)$$

단,  $m_x$ 는  $x$ 의 평균,  $m_y$ 는  $y$ 의 평균,  $E[\ ]$ 는 기댓값

$x$ 의 분산은  $x$ 값들이 평균을 중심으로 얼마나 흩어져 있는지를 나타내고,  $x$ 와  $y$ 의 공분산은  $x, y$ 의 흩어진 정도가 얼마나 서로 상관관계를 가지고 흩어져 있는지를 나타낸다. 공분산 행렬(covariance matrix)  $C$ 는 데이터의 좌표 성분들 사이의 공분산 값을 원소로 하는 행렬로서 데이터의  $i$ 번째 좌표와  $j$ 번째 좌표의 공분산 값을 행렬  $i$ 행  $j$ 열 원소값으로 하는 행렬을 말한다[1]. 이를 식으로 표현하면 Equation (2)와 같다.

$$\begin{aligned} x &= [x_1, \dots, x_n]^T : n\text{차원 } T\text{개수의 열벡터} & (2) \\ y_1 &= a_{11}x + a_{12}x + \dots + a_{1p}x_p = a_1^T x \\ &\dots \\ y_n &= a_{n1}x_1 + ax + \dots + a_{np}x_p = a_p^T x \\ C &= E[(x_i - m_{\epsilon})(x_j - m_{x_j})] : n \times n \text{행렬} \end{aligned}$$

단,  $a_i' = (a_{1i}, a_{2i}, \dots, a_{pi})$ ,  $i = 1, \dots, p$ 는 모두 크기가 1인 단위벡터이다. 즉,  $\sum_{j=1}^n a_{ij}^2 = 1, i = 1, \dots, p$ 로서 다음의 조건을 만족하도록 결정된다.

- ①  $y_1 : x$ 의 원소의 선형결합 가운데 최대분산을 갖는다.
- ②  $y_2 : y_1$ 과 무상관( $Cov(y_1, y_2) = 0$ )이면서  $x$ 의 원소의 선형결합 가운데 최대분산을 갖는다.
- ⋮
- ①  $y_i : y_1, \dots, y_{i-1}$ 과 무상관이면서  $x$ 의 원소의 선형결합 가운데 최대분산을 갖는다( $i = 1, 2, \dots, p$ )

여기서  $y_1 \sim y_p$ 는 다음과 같은  $p$ 개의 주성분을 의미하며  $p$

쪽으로 위계순서가 증가할수록 원 데이터에 대한 특성 반영도는 낮아지게 된다.

- $y_1$  : 제1주성분(1st principal component)
- $y_2$  : 제2주성분(2nd principal component)
- ...
- $y_p$  : 제p주성분(pth principal component)

주성분점수(principal component scores)  $a_{ij}$ 는 다음과 같이 정의된다.

- ①  $a_1$ 은 제일 큰 고유값  $\lambda_1$ 에 대응되는 고유벡터  $e_1$ 과 같다.
- ②  $a_2$ 는 두번째로 큰 고유값  $\lambda_2$ 에 대응되는 고유벡터  $e_2$ 와 같다.
- ⋮
- ①  $a_i$ 는  $i$ 번째로 큰 고유값  $\lambda_i$ 에 대응되는 고유벡터  $e_i$ 와 같다.

아래 식에서 주성분 분석의 결과로 얻을 수 있는 세분화 변수 중 제1세분화 변수는 데이터의 특성을 가장 잘 반영하는 변수이며(고유벡터  $e_1$ ), 제2세분화 변수는  $e_2$ 로 다음 Equation (3)과 같이 정의된다.

$$C\hat{e}_i = \lambda_i e_i \tag{3}$$

$e_i$  : eigenvector of  $c$   
 $\lambda_i$  : eigenvalue,  $e_i$  방향으로의 분산  
 $\lambda_1 \geq \dots \geq \lambda_n \geq 0$

- $e_1$  : 분산이 가장 큰 방향
- $e_2$  :  $e_1$ 에 수직이면서 다음으로 가장 분산이 큰 방향
- $e_k$  :  $e_1, \dots, e_{k-1}$ 에 모두 수직이면서 가장 분산이 큰 방향

제1세분화 변수 이후의 세분화 변수는 제1세분화 변수로 설명할 수 없는 자료의 변동을 설명하며 제k세분화 변수로 갈수록 원래 데이터에 대한 특징 반영도는 낮다고 볼 수 있다.

입력 데이터 집합에 대하여 주성분 분석을 수행한 후 입력 데이터에 대하여 고유벡터 계수에 의한 PCA 분석 결과를 세분화 변수 선정에 반영할 수 있다. 각 변수에 대하여 제1주성분, 제2주성분,.....에 대하여 높은 고유벡터 계수 값을 나타내는 경우에 대한 분석을 수행한다.

또한 스크리 산점도(scree plot)로 표현되는 고유값(eigenvalue) 그래프를 사용하여 세분화 변수 수를 결정한다 [1,2,11,12].

### 3. 상관도 기반 세분화 변수 선정

본 논문에서는 전체 데이터에 대한 주성분 분석 결과에 대하여 상관도를 기반으로 하여 주성분 변수를 결정하고자 한다. 또한, 변수간의 상관도를 적응적으로 고려함으로써 전체

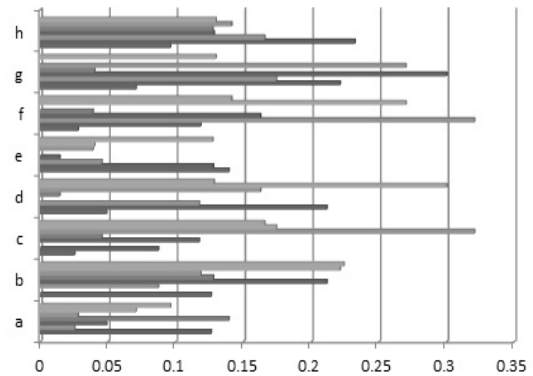


Fig. 1. Correlation Coefficient of Input Data

데이터에 대한 특성을 잘 반영할 수 있는 주성분 변수를 결정하고자 한다.

제안하는 방법은 주성분 분석의 주성분 변수(또는 세분화 변수) 선정 시 상관도(Correlation)를 데이터 편향성을 줄일 수 있도록 적용하고 고유벡터 계수 값에 의한 주성분 위계 분석을 통하여 연관성을 높이는 데이터 특성을 반영하고자 한다.

먼저, 상관 분석을 수행한 후 상관도(Correlation) 결과 값을 기반으로 상관도가 높은 데이터 변수를 제거해 나간다. 세분화 변수 수를 결정하기 위하여 고유 값(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.

Fig. 1은 8개의 변수에 대한 1000개 데이터(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 세분화 변수가 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 고유벡터 계수 값을 구한 후 두 데이터 변수간의 고유벡터 계수 값이 낮은 것을 차례로 제거해 나간다.

Fig. 2는 8개의 변수에 대한 PCA 결과를 보여주는 그래프이다. 여기서 eigen-vector 계수 값이 낮은 PC8, PC7 순으로 세분화 변수(변수a, 변수e)를 제거하는 방식으로 처리한다.

### 4. 결과 및 분석

본 논문에서는 입력 데이터의 주성분 분석의 주성분 변수 결정시 먼저, 고유벡터 계수 값에 의한 주성분 위계 분석을 통하여 연관성을 높이는 데이터 특성을 반영하였고 또한, 데

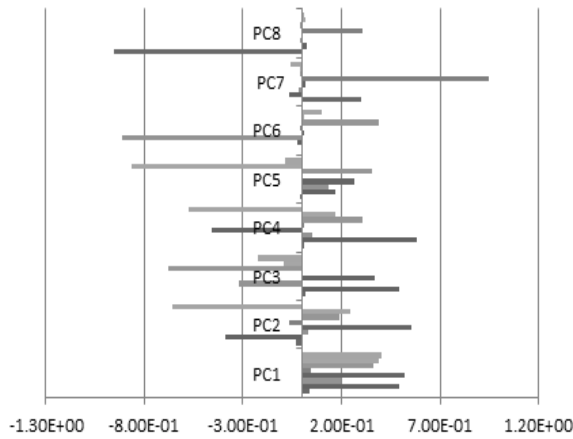


Fig. 2. PCA of Input Data

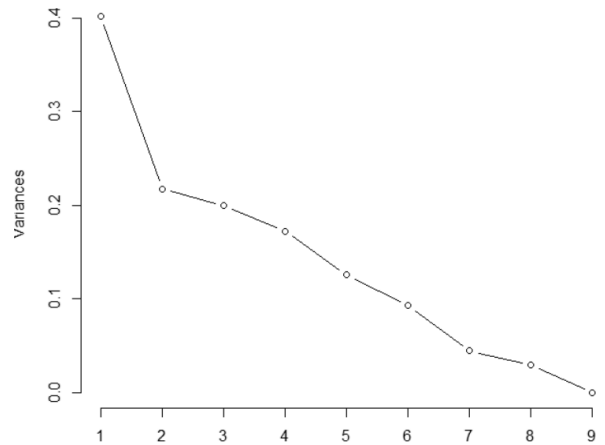


Fig. 3. Eigenvalue Graph of Input Data

Table 1. Correlation of Input-Data

Cor	a	b	c	d	e	f	g	h
a	1.00000000	0.12621495	0.02540530	0.04911064	0.13928146	0.02793495	0.07069212	0.09607586
b	0.12621495	1.00000000	0.08721985	0.21149036	0.12777402	0.11849164	0.22130311	0.23242330
c	0.02540530	0.08721985	1.00000000	0.11738195	0.04559758	0.32062510	0.17420436	0.16557232
d	0.04911064	0.21149036	0.11738195	1.00000000	0.01448669	0.16251003	0.29996837	0.12821601
e	0.13928146	0.12777402	0.04559758	0.01448669	1.00000000	0.03900155	0.04014108	0.12746063
f	0.02793495	0.11849164	0.32062510	0.16251003	0.03900155	1.00000000	0.26983426	0.14127926
g	0.07069212	0.22130311	0.17420436	0.29996837	0.04014108	0.26983426	1.00000000	0.12966124
h	0.09607586	0.2242330	0.16557232	0.12821601	0.12746063	0.14127926	0.12966124	1.00000000

이터 편향성의 영향을 줄이기 위해 여러 변수와 중복적으로 상관도가 높은 것으로 분석된 변수를 적응적으로 제거하는 방법을 제안하였다.

8개의 입력 변수에 대하여 입력 데이터에 대한 상관 분석 결과는 Table 1과 같다(Cor은 Correlation을 의미함).

먼저, 8개의 변수에 대한 PCA 결과를 보여주는 Fig. 2 그래프에 의해서 고유벡터 계수 값이 낮은 PC9, PC8 순으로 세분화 변수a와 변수e를 제거해 나간다.

주성분 분석 결과인 PC1, PC2 등의 주성분(데이터 특징)을 추출 시 상관관계를 기반으로 상관도가 높은 변수들을 차례로 제거한 후 8개 변수에 대하여 필요한 주성분의 개수를 결정하기 위하여 스크리 산점도(scree plot)를 이용한다. m 값을 결정하기 위하여 산점도를 이용하면 j번째부터 감소하는 추세가 완만해 질 경우에  $m=j-1$ 을 선택하는데, 고유값(eigenvalue) 그래프를 도식하면 Fig. 3과 같다.

그래프에서 고유치는 각 주성분의 분산에 해당하므로 각 주성분이 설명하는 정도의 크기를 알 수 있다. 주성분 분석의 결과인 주성분의 개수를 정하기 위해 고유값 그래프를 이용하여 결정할 수 있다.

고유값 그래프에서 기울기가 완만해지는(감소하는 추세가 완만해지는) 부분이  $n \sim m$  일 때 여기서는 5~7가된다. 따라

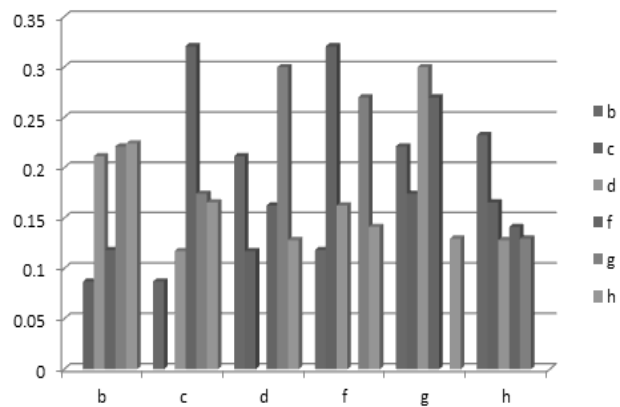


Fig. 4. Correlation of Reduced Input-Data

서 고유값 그래프에 기반한 주성분의 개수는  $(n-1)$ 개에서  $(m-1)$ 개이고 Fig. 3 그래프에서는 (5-1)개~(7-1)개가 되므로 4개~6개로 주성분의 개수를 예측할 수 있다.

Table 1로부터 고유값 계수 값이 낮은 PC8, PC7 순으로 세분화 변수a와 변수e를 제거된 축소된 입력 데이터에 대한 그래프는 Fig. 4와 같다.

고유벡터 계수 값이 낮은 순으로 세분화 변수a, 변수e가 제거된 입력 데이터에 대한 상관 분석 결과는 Table 2와 같

Table 2. Pairs of High Correlation Value

Pair of variable	Correlation Value
(b,d)	0.21149036
(c,f)	0.32062510
(d,g)	0.29996837
(b,g)	0.22130311
(f,g)	0.26983426
(b,h)	0.23242330

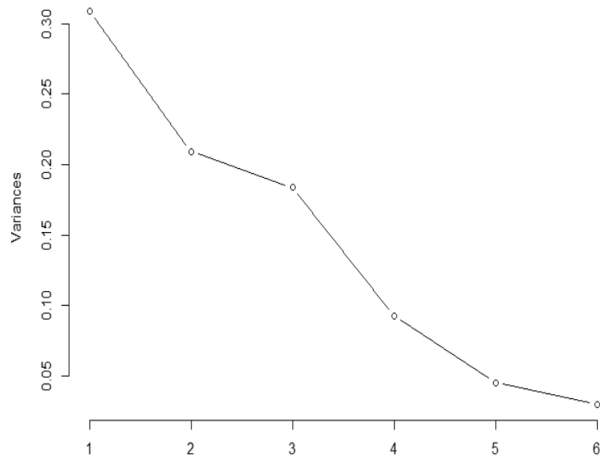


Fig. 5. Eigenvalue Graph of Reduced Input Data

다. Table 1의 상관 분석 결과로부터 상관도가 높은 변수 쌍을 구할 수 있으며 그 결과는 Table 2와 같다.

Table 2와 Fig. 4로부터 변수 b는 변수 d, f, g와 높은 상관도를 보임을 알 수 있으며, 변수 g는 변수 b, d, f와 높은 상관도를 보임을 알 수 있다. 이로부터 변수 b와 변수 g가 제거 대상 변수로 선정됨을 알 수 있다.

변수 b, g가 제거되었을 때의 고유값 그래프는 Fig. 5와 같으며 그래프에 기반한 주성분의 개수는 (n-1)개에서 (m-1)개이고 Fig. 4에서는 (4-1)개~(5-1)개가 되므로 3개~4개로 주성분의 개수를 예측할 수 있다. 따라서 Fig. 3과 Fig. 5의 고유값 그래프로부터 주성분 개수는 4개로 결정되어짐을 알 수 있다.

### 5. 결 론

주성분 분석은 고차원의 데이터를 저차원의 데이터로 환원시켜 서로 연관 가능성이 있는 고차원 공간의 표본들의 공분산 행렬을 원 변수의 선형 결합을 이용하여 분석하는 방법이다. 본 논문에서는 주성분 분석의 주성분 변수(또는 세분화변수) 선정 시 상관도(Correlation)를 적응적으로 적용하여 데이터 편향성의 반영도를 낮추고 고유벡터 계수 값에 의한 주성분 위계를 분석을 통하여 연관성을 높이는 데이터 특성을 반영하여 고 차원 데이터의 특징을 더 잘 반영한 주성분 분석

의 세분화 변수를 얻을 수 있는 방법을 제안하였다.

주성분 분석의 세분화 변수는 먼저 주성분 분석 결과에 의해 PC1, PC2, PC3, PC4...등의 주성분을 유도하는데 연관성이 적은 변수 (a, e)가 제거 대상이 되었고 두 번째로 상관 분석 결과의 상관도가 높은 변수들 중에서 다른 여러 변수들과 중복적으로 상관도가 높은 변수 (b, g)가 제거 대상이 되었으며 결과적으로 적응적 상관도를 적용 하였을 때 (c, d, f, h)로 선정되었음을 알 수 있으며 편향성 유도 변수는 세분화 변수와 주성분 유도 연관성이 적은 변수는 주성분 변수에 포함되지 않았음을 알 수 있다. 또한 산점도로 고유값(eigenvalue) 그래프를 통하여 주성분의 개수는 4개임을 알 수 있었다. 구해진 주성분은 회귀분석에서 포함시킬 설명변수의 수를 결정하는데 사용되기도 하며 군집분석등과 같은 분석법의 입력변수로서 이용될 수 있다. 앞으로의 연구 방향은 결정된 주성분 개수를 활용하여 K-means 군집 분석을 통한 원 데이터를 분석 및 해석하는 방법에 대한 연구이다.

### References

- [1] I. T. Jolliffe, "Principle Component Analysis," Springer-Verlag, New York, 1986.
- [2] B. J. Kim, J. Y. Sim, C. H. Hwang, and I. G. Kim, "On-line Nonlinear Principal Component Analysis for Nonlinear Feature Extraction," *The Journal of Korean Institute of Information Scientists and Engineers*, Vol.31, No.3, pp.361-368, 2004.
- [3] J. H. Ha, S. W. Park, and G. M. Kim, "Development of Principal Component Analysis algorithms for detecting leak in water pipe networks," *Proceeding of Korean Society of Civil Engineering*, pp.52-53, 2018.
- [4] B. J. Kim, "Feature Extraction on High Dimensional Data Using Incremental PCA," *The Journal of The Korean Institute of Information and Communication Engineering*, Vol.8, No.7, pp.1475-1479, 2004.
- [5] B. J. Kim and Y. J. Shim, "On-line Nonlinear Principal Component Analysis for Nonlinear Feature Extraction," *The Journal of The Korean Institute of Information Scientists and Engineers*, Vol.31, No.3, pp.361-368, 2004.
- [6] Principle Component Analysis [Internet], <https://stat.snu.ac.kr/time>
- [7] S. Y. Kim, C. K. Kim, Y. H. Kang, C. Y. Yun, G. S. Jang, and H. G. Kim, "Verification of Typical Meteorological Year Data Structure Applying Principal Component Analysis," *Proceeding of The Korean Solar Energy Society*, p.159, 2018.
- [8] J. Y. Lim and P. S. Ji, "Developing of Fault Diagnosis Algorithm using Correlation Analysis and ELM," *The Journal of The Korean Institute of Electrical Engineers*, Vol.65, No.3, pp.204-209, 2016.

[9] H. J. Joo, N. H. Kim, R. Navin, H. P. Zhao, B. Sovit, and H. Kim, "A Study on Data Types and Visualization for Traffic Congestion and Accidents," *Proceeding of The Institute of Electronics and Information Engineers*, pp.1011-1013, 2019.

[10] W. S. Hwang, H. K. Kim, Y. S. Kim, and H. C. Kim, "Event-based Correlation Coefficient for Analyzing the Categorical Features in Industrial Control Systems," *Proceeding of The Korean Institute of Information Scientists and Engineers*, pp.812-814, 2019.

[11] Y. J. Kim, "Evaluation of Urban Lakes Water Quality Using Principle Component Analysis," *The Journal of Korean Society of Environmental Administration*, Vol.9, No.2, pp.197-203, 2003.

[12] H. J. Joo, N. H. Kim, R. Navin, H. P. Zhao, B. Sovit, and H. Kim, "A Study on Data Types and Visualization for Traffic Congestion and Accidents," *Proceeding of The Institute of Electronics and Information Engineers*, pp.1011-1013, 2019.



### 고 명 숙

<https://orcid.org/0000-0002-7445-6882>

e-mail : kms@bc.ac.kr

1989년 이화여자대학교 물리학과(이학사)

1993년 고려대학교 컴퓨터학과(이학석사)

1998년 고려대학교 컴퓨터학과(이학박사)

2001년~현재 부천대학교 경영과 교수

관심분야 : Algorithms, Big-Data Management