The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.3.385
28(3), 385–392

Principal Component Analysis with Coefficient of Variation Matrix

Ji-Hyun Kim[a,1]

[a]Department of Statistics and Actuarial Science, Soongsil University

(Received January 5, 2015; Revised March 2, 2015; Accepted March 21, 2015)

Abstract

Principal component analysis (PCA), a dimension-reduction technique, is usually implemented after the variables are standardized when the measurement unit of variables are different. To standardize a variable we divide it by its standard deviation. But there is another way to transform a variable to be independent of its measurement unit. It is to divide it by its mean rather than standard deviation. Implementing PCA on standardized variables is equivalent to implementing PCA with a correlation matrix of original variables. Similarly, implementing PCA on the transformed variables divided by their means is equivalent to implementing PCA with a matrix related to the coefficients of variation of the original variables. We explain why we need to implement PCA on the variables transformed by their means.

Keywords: principal components, correlation matrix, coefficient of variation, ratio scale

1. 서론

주성분분석(principal component analysis)은 상관관계가 있는 변수들 사이의 변동을 상관관계가 없는 주성분이라 부르는 새로운 변수들로 설명하고자 하는 차원축소 기법이다. 주성분은 원래 변수들의 선형결합이 되며 공분산행렬의 고유벡터가 선형결합을 위한 계수 벡터가 된다.

주성분분석에 관한 책이나 소프트웨어를 보면 공분산행렬과 상관행렬을 이용하는 두 가지 선택 방법에 대해서만 다룬다. 상관행렬로 주성분분석을 하는 것은 각 변수를 그 변수의 표준편차로 나누는 변환을 한 다음 변환된 변수의 공분산행렬로 주성분분석을 하는 것으로 간주할 수 있다. 본 연구에서는 각 변수를 표준편차가 아닌 평균으로 나누는 변환을 한 다음 변환된 변수의 공분산행렬로 주성분분석하는 방법을 (본 논문에서 이 방법을 간략하게 '변동계수행렬'로 주성분분석하는 방법이라 부르기로 함) 제안하고자 한다. 그리고 이 방법이 공분산행렬이나 상관행렬을 이용하는 방법과 함께 보완적으로 실시되어야 할 필요성을 밝히고자 한다.

평균으로 나누어 변수를 변환하는 것은 변수를 변환하는 여러 가지 방법 중의 하나이고, 변수를 변환한 다음 주성분분석을 하는 것은 새로운 방법이 아니므로 일부 사례 연구에서 이 방법이 이미 실시되었을 수도 있다. 하지만 그 필요성과 타당성에 대해 본격적인 연구를 한 참고문헌은 찾을 수 없었다.

다음 절에서는 변동계수행렬로 주성분분석을 하는 것이 공분산행렬이나 상관행렬로 분석하는 것과 어떤 차이가 있으며 왜 필요한가를 인위적인 자료와 실제 자료를 통해 밝히고자 한다. 3절에서는 연구 내용을 요약하고 이 방법을 사용함에 있어 주의해야 할 점 등을 서술한다.

[1]Department of Statistics and Actuarial Science, Soongsil University, Sangdo-Ro 369, Dongjak-Gu, Seoul 156-743, Korea. E-mail: jxk61@ssu.ac.kr

2. 변동계수행렬을 이용한 주성분분석의 필요성

변수들이 거리와 시간, 또는 인구수와 소득 같이 측정단위가 다르거나 분산의 불균형이 심할 경우 상관행렬로 주성분분석을 실시할 것이 권장된다. SPSS나 SAS 같은 상용 소프트웨어에는 주성분분석에서 상관행렬이 자동설정(default)으로 되어있다 (반면에 R (2010)의 princomp 함수에는 공분산행렬이 자동설정으로 되어있다.). 상관행렬로 주성분분석하는 것은 변수를 표준화한 다음 변환된 변수의 공분산행렬로 주성분분석하는 것과 동일하다. 표준화는 변수를 표준편차로 나누어주는 변환 x/s_x인데, (주성분분석에서 두 변환 $(x-\bar{x})/s_x$와 x/s_x는 동일한 공분산행렬을 가지므로 동등한 변환으로 간주할 수 있다.) 변수를 표준화하면 모든 변수의 분산이 1이 되어 동일하게 된다. 하지만 서로 다른 변수들 사이의 관련성과 함께 각 변수의 변동의 크기도 중요한 정보인데 이를 인위적으로 같게 만드는 것이 과연 타당한가에 대한 의문이 든다.

만약 변수의 값이 양수라면, 측정단위에 무관하게 변수를 변환하는 방법으로 표준편차 대신에 평균을 나누어주는 변환도 생각해 볼 수 있다. 원래 변수에 평균을 나누어주는 변환 $x/\bar{x}$을 한 후 변환된 변수의 분산을 구하면 원래 변수의 변동계수(coefficient of variation)의 제곱인 $(s_x/\bar{x})^2$이 된다. 각 변수를 해당 변수의 평균으로 나누어주는 변환을 하면 표준화 변환과는 달리 변수의 변동성이라는 유용한 정보를 잃지 않게 된다는 장점이 있다 (한편 각 변수의 값을 그 변수의 평균으로 나누어도 표준편차를 나누었을 때와 마찬가지로 변수들 사이의 상관계수는 변하지 않는다.).

p차원 확률벡터(random vector) $\boldsymbol{x}$의 (표본)공분산행렬을 $S=\{s_{ij}\}$ (s_{ij}는 i번째 변수와 j번째 변수의 공분산)라고 하고 D를 대각행렬(diagonal matrix)이라고 할 때, 변수변환을 $\tilde{\boldsymbol{x}}=D\boldsymbol{x}$로 표현할 수 있다. 표준화 변환은 $D=\{1/\sqrt{s_{ii}}\}$이고, 평균으로 나누어주는 변환은 $D=\{1/\bar{x}_i\}$이다. 변환된 확률벡터 $\tilde{\boldsymbol{x}}$의 공분산행렬은 DSD인데, 표준화된 변수의 공분산행렬은 원래 변수의 상관행렬인 $R=\{s_{ij}/\sqrt{s_{ii}s_{jj}}\}$이 되며 대각원소는 모두 1이 된다. 그리고 평균으로 나누어주는 변환을 한 변수의 공분산행렬은 $V=\{s_{ij}/(\bar{x}_i\bar{x}_j)\}$로서 대각원소는 변동계수의 제곱이 된다 ($V$의 원소는 변동계수가 아니라 변동계수의 제곱이거나 공분산을 평균의 곱으로 나눈 값이지만 편의상 V를 '변동계수행렬'이라 부르기로 한다. 그리고 평균으로 나누어주는 변환을 한 후 변환된 변수의 공분산행렬로 주성분분석을 하는 것을 '변동계수행렬을 이용한 주성분분석'이라고 줄여서 부르기로 한다.).

2.1. 정성적 비교

변동계수행렬을 이용한 주성분분석이 왜 필요한가를 주성분의 해석이라는 정성적 기준에서 설명해보고자 한다. 먼저 상관행렬을 이용한 주성분분석이 필요한 상황이라면 변동계수행렬을 이용한 주성분분석을 보완적으로 실시할 필요가 있다. 그 이유는 측정단위에 무관하게 하는 변환 중에서 표준화 변환이 아닌 다른 변환의 경우 어떤 결과가 나오는지 살펴볼 필요가 있기 때문이다. 2005년 헬싱키 올림픽의 남자 육상기록 자료 (Johnson과 Wichern, 2007, Table 8.6)의 경우, 100미터, 200미터, 400미터 기록은 초 단위로 측정하였고 나머지 다섯 종목은 분 단위로 측정하였다. 이 자료의 경우 주성분분석을 공분산행렬로 실시하는 것은 적절하지 않다. 측정 단위도 다를뿐더러 최대 분산을 갖는 마라톤의 분산과 최소분산을 갖는 800미터의 분산의 비가 약 170^2배로서 지나치게 큰데, 이렇게 되면 800미터 종목은 주성분을 결정하는 데 거의 아무런 역할을 할 수 없는 반면에 마라톤은 지나치게 큰 역할을 하게 된다. 8개 종목에 대한 변동계수의 제곱의 최댓값과 최솟값의 비는 9.6^2인데, 각 종목의 기록에 평균을 나누어주면 측정 단위의 문제도 없고 변동성이 큰 종목이 그에 상응하는 역할을 할 수 있게 된다. 반면에 표준편차를 나누어주면 모든 변수의 분산이 1이 되므로 변수의 변동성에 대한 정보를 잃어버리게 되고 변수들 사이의 상관성만 남아있게 된다.

Table 2.1. The first two principal components for track records (The values less than 0.1 in absolute value are suppressed.).

		주성분을 나타내는 고유벡터							
		100미터	200미터	400미터	800미터	1500미터	5천미터	1만미터	마라톤
	(표준편차)	(0.221)	(0.549)	(1.439)	(0.052)	(0.152)	(0.761)	(1.679)	(8.952)
	(변동계수)	(2.166)	(2.670)	(3.140)	(2.966)	(4.154)	(5.587)	(5.885)	(6.707)
공분산행렬	주성분1			0.114				0.175	0.974
	주성분2		−0.253	−0.916			−0.117	−0.209	0.167
상관행렬	주성분1	0.332	0.346	0.339	0.353	0.366	0.370	0.366	0.354
	주성분2	−0.529	−0.470	−0.345		0.154	0.295	0.334	0.387
변동계수행렬	주성분1	0.140	0.181	0.213	0.220	0.327	0.462	0.486	0.545
	주성분2	−0.384	−0.478	−0.530	−0.240	−0.204		0.185	0.450

공분산행렬과 상관행렬, 변동계수행렬로 각각 분석했을 때 얻게 되는 처음 두 개의 주성분을 나타내는 고유벡터를 Table 2.1에 정리하였다. 상관행렬을 이용한 분석의 경우 첫 번째 주성분을 보면 각 변수에 대응하는 고유벡터 원소의 값이 비슷하지만(0.332∼0.370) 변동계수행렬을 이용한 분석의 경우 변동성을 나타내는 변동계수의 크기에 따라 달라진다(0.140∼0.545). 두 번째 주성분을 보면 상관행렬과 변동계수행렬의 두 경우 모두 장거리와 단거리 기록의 차이를 나타낸다는 점은 같으나 각 변수에 대응하는 고유벡터 원소의 값은 다소 다르다 (한편, 주성분분석을 실시해서 처음 두 개의 주성분의 분산이 '전체 분산'에서 차지하는 비율을 보면, 변동계수를 이용한 분석의 경우 89.9%, 94.6%이고, 상관계수를 이용한 경우 83.8%, 91.8%이다. 하지만 이 값의 크기가 두 방법에 대한 비교 기준이 될 수 없다. 왜냐하면 '전체 분산'이라고 했지만 두 방법에서 얘기하는 전체 분산이 각각 상관행렬과 변동계수행렬이라는 서로 다른 행렬의 대각선의 합이기 때문이다. 공분산행렬로 주성분분석을 하는 경우 98.3%, 99.6%가 나오는데, 이 비율이 높다고 해서 공분산행렬로 분석하는 것이 상관행렬로 분석하는 것보다 더 낫다고 얘기할 수 없는 것과 같은 이유이다.).

2차원 확률벡터의 상관 행렬 $\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$의 큰 고유값은 $1+r$이고, 큰 고유값에 대응하는 고유벡터는 상관계수 r의 값에 관계없이 $(1/\sqrt{2}, 1/\sqrt{2})$이다. 이로부터 표준화를 통해 분산을 인위적으로 같게 하여 첫 번째 주성분을 얻으면 두 변수의 기여도가 같아짐을 알 수 있다. 이 성질은 두 변수가 어떠한 분포를 갖든 관계없이 상관행렬을 이용한 주성분이 항상 갖게 되는 성질인데 바람직한 성질은 아니다. 3차원 확률벡터의 경우 상관행렬을 $\begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{pmatrix}$로 표현할 때, $r_{12} = r_{13} = r_{23}$이면 첫 번째 주성분을 나타내는 고유벡터는 상관계수의 값에 관계없이 $(1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$가 된다 (상관행렬을 이용했을 때 첫 번째 주성분에 대응하는 고유벡터의 모든 원소가 같아진다는 사실은 모든 차원에서 성립한다.). 세 상관계수의 값이 달라지면 첫 번째 고유벡터의 원소가 모두 같게 되지는 않지만 비슷할 것으로 예상된다. 두 변수 사이의 상관계수 r_{ij}가 달라짐에 따라 첫 번째 고유벡터의 원소들의 상대적 크기가 어떻게 달라지는가를 보기 위해 간단한 모의실험을 실시하였다.

세 변수의 분산을 1, 4, 100으로 두고, 변동계수는 1, 1.5, 2가 되도록 평균을 정한 다음 이 값들을 고정시켰다. 그리고 상관계수 r_{12}, r_{13}, r_{23}의 값은 독립적으로 생성한 0.1과 0.9 사이의 균일 난수로 정하였다. 이 값들로부터 공분산행렬과 상관행렬, 변동계수행렬을 각각 구할 수 있다. 그리고 세 가지 방법에 의한 주성분 중에서 첫 번째 주성분을 나타내는 고유벡터를 각각 구한 다음 고유벡터의 원소 중에서 제일 큰 값과 작은 값의 비(ratio)를 구한다 (첫 번째 주성분보다 두 번째 주성분이 해석상 더 관심을 끄는 경우도 있다. 하지만 모의실험에서는 분산이 가장 큰 첫 번째 주성분에 대해서만 비교하였다.). 이 작업을 10000번 반복해서 구한 값들의 평균을 구해보면, 공분산행렬과 상관행렬, 그리고 변동계수행렬을 이

Table 2.2. The first principal component for a simulated case with three variables where $(\mu_1, \mu_2, \mu_3) = (1, 4/3, 5)$, $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 4, 100)$, $r_{12} = 0.517, r_{13} = 0.294, r_{23} = 0.770$

	주성분을 나타내는 고유벡터		
	x_1	x_2	x_3
공분산행렬	0.030	0.155	0.987
상관행렬	0.474	0.651	0.594
변동계수행렬	0.191	0.567	0.801

Table 2.3. The first principal component for KOSPI200 data

	주성분을 나타내는 고유벡터	
	영업이익률	부채비율
공분산행렬	−0.003	1.000
상관행렬	−0.707	0.707
변동계수행렬	−0.143	0.990

용하는 세 방법에 대해 각각 28.31 (17.09), 1.40 (0.33), 3.55 (1.51)이 된다 (괄호 안의 값은 표준편차). 1.40이라는 값이 의미하듯, 상관행렬에서 구한 첫 번째 고유벡터의 세 원소는 크기가 비슷해 각 변수가 첫 번째 주성분에 기여하는 정도가 비슷함을 알 수 있다. 세 변수의 변동의 상대적 크기에 관계없이 성립하는 이러한 성질은 바람직한 성질은 아닌데, 변동계수행렬을 이용한 주성분은 3.55라는 값이 의미하듯 그렇지 않았다. 한편 첫 번째 고유벡터의 원소 중에서 분산이 100으로서 제일 큰 세 번째 변수의 계수를 나타내는 원소와, 분산이 4로서 그 다음으로 큰 두 번째 변수의 계수를 나타내는 원소의 비의 경우 각각 13.10 (8.75), 1.05 (0.32), 1.95 (0.69)인데, 13.10과 1.95이라는 값이 의미하듯 공분산행렬을 이용하면 분산이 상대적으로 큰 하나의 변수가 주성분을 거의 결정하지만 변동계수행렬을 이용하면 그렇지 않다는 것을 알 수 있다. Table 2.2에 10000번의 모의실험 중 하나의 경우에 대한 결과를 보고하였다.

측정단위가 같은 경우에도 공분산행렬을 이용하는 분석과 함께 변동계수행렬을 이용한 분석을 보완적으로 실시할 필요가 있다. 예를 들어 기업의 영업이익률과 부채비율이라는 두 변수의 경우 측정단위는 동일하지만 평균과 분산은 크게 다르다. 코스피200 시가총액기준 상위 50개 기업의 2013년 연간 영업이익률과 부채비율의 평균은 각각 8.4, 254.5(단위: %), 분산은 $6.5^2, 370.1^2$이며 상관계수는 −0.20이다. Table 2.3에서 볼 수 있는 것과 같이 공분산행렬로 분석하면 첫 번째 고유벡터가 (−0.003, 1.000)이 되어 첫 번째 주성분은 곧 부채비율이 된다. 반면에 평균으로 나누어주는 변환을 하면 변환된 두 변수의 분산, 즉 변동계수의 제곱은 각각 $0.78^2, 1.45^2$이 되며, 첫 번째 고유벡터가 (−0.143, 0.990)가 되어 공분산행렬로 분석했을 때와 달리 영업이익률도 첫 번째 주성분에 일정 부분 기여를 하게 된다.

2.2. 정량적 비교

지금까지 변동계수행렬을 이용한 주성분분석을 왜 보완적으로 실시할 필요가 있는지를 주성분의 해석과 관련해서 세 방법을 비교하여 설명하였다. 주관적인 판단을 요구하는 이러한 정성적인 비교에 더해서 객관적인 판단을 할 수 있는 정량적인 비교에 대해 지금부터 논의하고자 한다.

주성분의 효용성을 크게 해석과 차원축소로 나눠볼 수 있다. 주성분의 해석을 통해 변수들의 관련성에 대한 새로운 통찰을 얻기도 하지만, 주성분분석에 이은 후속연구에서 관련성이 있는 여러 변수들을 관련성이 없는 몇 개의 주성분으로 대체함으로써 얻게 되는 효과도 있다. 판별분석과 회귀분석의 사례를 통해 주성분의 차원축소 효과가 주성분을 구하는 세 가지 방법에 따라 어떻게 달라지는지를 정량적으로 비교하고자 한다.

Table 2.4. Misclassification rates by n-fold cross-validation for iris data

주성분 추출 행렬	분류규칙에 쓰인 주성분의 수	오분류율
공분산행렬	1	0.067
	2	0.047
상관행렬	1	0.073
	2	0.073
변동계수행렬	1	0.04
	2	0.027

판별분석의 대표적 자료인 붓꽃자료는 (Fisher (1936), R에서 iris라는 데이터셋(dataset)에 내장되어 있음) 꽃받침과 꽃잎의 길이와 너비라는 네 개의 변수와 붓꽃의 품종을 나타내는 변수로 구성되어 있다. 손쉽게 측정할 수 있는 네 개의 변수로 붓꽃의 품종을 판별하기 위한 유용한 규칙을 찾을 수 있다는 것을 보여주는 자료이다. 우리는 이 자료에서 네 개의 변수를 한 개의 주성분이나 두 개의 주성분으로 대체하되, 주성분을 구하는 세 개의 방법에 따라 성능이 어떻게 달라지는가를 보고자 한다. 성능을 판단하는 기준으로 교차타당성(cross-validation)에 의한 오분류율(misclassification rate)을 이용한다 (이 때 교차타당성은, 한 개의 자료를 제외하고 규칙을 만들고 제외했던 자료로 규칙의 정확성을 평가하되, 이 작업을 n개의 모든 자료에 대해 실시하는 n절 교차타당성(n-fold cross-validation)을 의미한다.).

먼저 네 개의 변수를 그대로 이용하여 판별분석을 실시하면 오분류율은 $3/150 = 0.02$이다. 이 때 판별분석은 선형판별분석(linear discriminant analysis)으로서 R의 MASS 팩키지 (Venables와 Ripley, 2002)에 있는 함수 lda( )를 이용하였다. 이제 네 개의 변수 대신에 주성분 한 개만 써서 오분류율을 추정해보면, 공분산행렬, 상관행렬, 변동계수행렬의 경우 각각 $10/150 \doteq 0.067, 11/150 \doteq 0.073, 6/150 = 0.04$이며, 두 개의 주성분을 썼을 때는 각각 $7/150 \doteq 0.047, 11/150 \doteq 0.073, 4/150 \doteq 0.027$이다 (Table 2.4 참조). 이 결과는 붓꽃자료에서 네 개의 변수를 한 개나 두 개의 주성분으로 대체하는 경우 변동계수행렬에 의한 주성분을 쓰면 다른 방법에 의한 주성분을 썼을 때보다 오분류율을 더 낮출 수 있다는 것을 보여준다. 참고로, 네 변수의 분산은 각각 $0.83^2, 0.44^2, 1.77^2, 0.76^2$이고, 변동계수의 제곱은 $0.14^2, 0.14^2, 0.47^2, 0.64^2$이다.

회귀분석에서 관련성이 높은 설명변수들이 있을 때 다중공선성(multicollinearity)의 문제가 발생할 수 있다. 이 때 대응방안 중의 하나가 관련성이 높은 변수들을 관련성이 없는 주성분으로 대체하는 것이다. 이 때 주성분을 구하는 방법에 따라 성능이 달라질 수 있는데, 결정계수라는 성능기준으로 비교하여 보았다. 다중공선성과 관련해서 많이 쓰이는 실제 자료가 Hald (1952) 자료이다. 이 자료에서 반응변수는 시멘트 응고과정에서 나오는 열량이고 네 개의 설명변수는 시멘트 제조과정에 쓰이는 원료이다. 네 개의 설명변수의 분산은 각각 $5.9^2, 15.6^2, 6.4^2, 16.7^2$이고, 변동계수의 제곱은 각각 $0.79^2, 0.32^2, 0.52^2, 0.56^2$이다. 상관계수 중에서 (절댓값이) 큰 것은 $r_{13} = -0.82, r_{24} = -0.97$이며, 나머지 네 개의 상관계수는 절댓값이 0.25보다 작다.

공분산행렬, 상관행렬, 변동계수행렬로 구한 한 개의 주성분을 네 개의 설명변수 대신으로 썼을 때, 결정계수는 각각 0.701, 0.965, 0.644이었으나 (Table 2.5 참조), 두 개의 주성분을 썼을 때는 각각 0.954, 0.965, 0.982로서 변동계수행렬을 이용했을 때 가장 높았다 (설명변수의 개수까지 고려하는 수정결정계수의 값을 기준으로 보면 네 개의 설명변수를 다 썼을 때 0.974이었으나, 변동계수행렬을 이용한 주성분 두 개를 썼을 때 0.978로서 더 높았다.). 0.954, 0.965, 0.982에서 보이는 차이가 비록 큰 차이는 아니지만 아무런 추가적인 노력 없이 단지 변수변환만으로 얻을 수 있는 성능의 향상이므로, 주성분을 구할 때 변동계수행렬을 이용한 방법을 보완적으로 실시할 필요가 있다는 것을 잘 보여준다.

Table 2.5. Coefficients of determination R^2 of regression model for Hald data

주성분 추출 행렬	회귀모형에 쓰인 주성분의 수	결정계수	수정결정계수
공분산행렬	1	0.701	0.674
	2	0.954	0.944
상관행렬	1	0.965	0.962
	2	0.965	0.958
변동계수행렬	1	0.644	0.611
	2	0.982	0.978

3. 요약과 첨언

주성분분석을 할 때 기존의 공분산행렬과 상관행렬을 이용한 분석에 변동계수행렬을 이용한 분석을 추가적으로 실시해서 결과가 어떻게 달라지며 주어진 자료에 더 적절한 방법이 무엇인가를 고민해보는 과정이 필요함을 역설하였다. 본 연구자는 주성분분석을 위한 소프트웨어에서 변동계수행렬을 이용한 분석을 공분산행렬이나 상관행렬과 함께 선택사항으로 제공할 것을 제안한다. 변동계수행렬로 주성분분석하는 것은 각 변수의 값을 그 변수의 평균으로 나눈 다음 변환된 변수의 공분산행렬로 주성분분석하는 것이 되므로 굳이 새로운 선택사항으로 둘 필요가 없다고 할 수도 있다. 하지만 사용자가 상관계수행렬을 이용한 주성분분석을 굳이 표준화 변환을 하지 않고도 메뉴선택이나 명령문의 옵션 지정을 통해 쉽게 할 수 있게 하듯이 변동계수행렬을 이용한 주성분분석을 변수변환 없이 선택적으로 실시할 수 있게 한다면 사용자의 편의성이 높아질 것이다.

행렬(공분산행렬, 상관행렬, 변동계수행렬)의 관점이 아니라 측정단위에 무관하게 하는 또 다른 변수변환이라는 관점에서 보면 변동계수행렬을 이용한 방법을 보완적으로 실시할 필요가 있음을 쉽게 납득할 수 있다. 분산이 유난히 큰 변수가 있는 경우 공분산행렬로 분석을 하면 주성분이 그 하나의 변수에 의해서만 결정된다. 이 때 상관행렬로 분석하게 되면 변수의 변동성이라는 유용한 정보를 잃게 되는 단점이 있다. 절충적이면서 보완적인 방법으로 변동계수행렬을 이용한 방법이 있는데, 세 가지 방법을 주성분의 해석이라는 기준에서 실제 자료와 인위적 자료를 이용해 비교하고자 하였다. 하지만 이 정성적 비교는 주어진 자료에 어떤 방법이 더 나은 방법인지를 객관적으로 알려주지는 않는다.

객관적인 정량적 비교를 위해 판별분석과 회귀분석의 예를 들었다. 어떤 방법으로 주성분분석을 실시하는 것이 좋은지를 판단하는 기준으로 회귀분석의 경우 결정계수를, 판별분석의 경우 오분류율을 썼다. 변동계수행렬로 분석했을 때 성능이 더 좋아지는 실제 자료를 각각 제시하였다. 본 연구에서 변동계수행렬을 이용한 방법이 항상 좋다는 것을 주장하는 것이 아니라 보완적으로 실시되어야 할 필요가 있음을 보이고자 하는 것이므로 이런 자료가 존재하는 것을 보여주는 것으로 충분하다. 또한 정량적 비교에서 쓰인 두 실제 자료에서 변수들의 분산의 불균형이 그다지 심하지 않았다. 그럼에도 불구하고 주성분분석 방법에 따라 성능 차이가 발생한다는 점에 주목할 필요가 있다.

이 방법을 쓸 때 주의해야 할 점이 있는데, 변수가 음수와 양수 값을 동시에 갖는 경우 평균이 0에 가까워질 수 있어 평균으로 나누어주는 것이 좋지 않은 변환이 될 수 있다는 점이다. 변동계수가 양수 값을 갖는 비척도(ratio scale) 자료에 쓰이는데, 이 방법도 비척도 자료에 안심하고 적용할 수 있다.

References

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-188.

Hald, A. (1952). Statistical Theory with Engineering Applications. Wiley, New York.

Johnson, R. A. and Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. Sixth Edition. Pearson Prentice Hall.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Venables, W. N. and Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition. Springer, New York.

변동계수행렬을 이용한 주성분분석

김지현[a,1]

[a]숭실대학교 정보통계보험수리학과

(2015년 1월 5일 접수, 2015년 3월 2일 수정, 2015년 3월 21일 채택)

요 약

주성분분석은 차원축소를 위한 대표적 기법이다. 주성분분석에서 변수들이 측정단위가 다르거나 분산의 불균형이 심할 경우 흔히 변수를 표준화한 다음 분석할 것이 권장된다. 표준화 변환은 표준편차를 나누어주는 변환인데, 측정단위에 무관하게 만들기 위해서라면 평균을 나누어주는 변환도 고려해볼 수 있다. 표준화 변환을 한 다음 주성분분석하는 것은 상관행렬로 주성분분석하는 것과 같은데, 평균을 나누어주는 변환을 한 후 주성분분석하는 것은 변동계수와 관련된 행렬로 주성분분석하는 것과 같음을 보이고, 그렇게 변환을 한 다음 주성분분석을 실시하는 것이 왜 필요한가를 설명하였다.

주요용어: 주성분, 상관행렬, 변동계수, 비적도

[1](156-743) 서울시 동작구 상도로 369, 숭실대학교 정보통계보험수리학과. E-mail: jxk61@ssu.ac.kr