*The Korean Journal of Applied Statistics* (2020) DOI: http://dx.doi.org/10.5351/KJAS.2020.33.1.061
**33**(1), 61–74

# On principal component analysis for interval-valued data

Soojin Choi[a] · Kee-Hoon Kang[a,1]

[a]Department of Statistics, Hankuk University of Foreign Studies

(Received December 30, 2019; Revised January 6, 2020; Accepted January 7, 2020)

**Abstract**

Interval-valued data, one type of symbolic data, are observed in the form of intervals rather than single values. Each interval-valued observation has an internal variation. Principal component analysis reduces the dimension of data by maximizing the variance of data. Therefore, the principal component analysis of the interval-valued data should account for the variance between observations as well as the variation within the observed intervals. In this paper, three principal component analysis methods for interval-valued data are summarized. In addition, a new method using a truncated normal distribution has been proposed instead of a uniform distribution in the conventional quantile method, because we believe think there is more information near the center point of the interval. Each method is compared using simulations and the relevant data set from the OECD. In the case of the quantile method, we draw a scatter plot of the principal component, and then identify the position and distribution of the quantiles by the arrow line representation method.

Keywords: center method, quantile method, symbolic data, truncated normal distribution, vertices method

## 1. 서론

보통 통계적 분석 대상이 되는 자료의 형태는 하나의 값을 관측값으로 가지는 단일 값 자료(single-valued data)이다. 그러나 현대사회에서는 자료의 양이 점점 방대해지고 구조가 복잡해지면서 다양한 형태의 자료가 등장하고 있다. 예를 들어 사람의 혈압과 목소리는 자연적으로 변동을 가지기 때문에 단일 값으로 나타내기에 한계가 있다. 이처럼 각 관측값이 내부적으로 구조와 변동을 가질 때, 그 특성이 반영되도록 표현한 자료를 심볼릭 자료(symbolic data)라고 한다. 심볼릭 자료에 관한 자세한 소개는 Billard와 Diday (2006)을 참고하면 된다. 심볼릭 자료는 값의 본질적 특성에 의해 자연스럽게 생성될 수 있고 또는 연구의 목적에 따라 단일 값 자료를 심볼릭 자료로 변형시켜 얻을 수도 있다.

전자의 예로 수축·이완기 혈압과 목소리 주파수 영역이 있으며, 후자의 예로는 시간 단위로 측정된 단일 값 기온 자료를 최저기온과 최고기온이라는 구간의 형태로 변형시킨 일별 기온 자료가 있다. 심볼릭 자료에는 이와 같이 구간을 관측값으로 가지는 구간형 자료(interval-valued data)가 있으며, 단일 값 자료 역시 구간의 상한과 하한이 동일한 구간형 자료의 특별한 경우라고 할 수 있다. 이 외에도 단일 값들의

This research was supported by Hankuk University of Foreign Studies Research Fund of 2019 and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A1B03035026).

[1]Corresponding author: Department of Statistics, Hankuk University of Foreign Studies, 81 Oedae-ro, Yongin-si, Gyeonggi-do 17035, Korea. E-mail: khkang@hufs.ac.kr

집합을 관측값으로 가지는 다중 값 자료(multi-valued data), 구간과 각 구간에 대응되는 확률을 관측값으로 가지는 히스토그램 자료(histogram data) 등이 있다.

단일 값 형태가 아닌 심볼릭 자료는 관측값 내에 변동이 존재한다는 점에서 단일 값 자료와 구분되며, 이를 대상으로 한 통계적 분석기법이 개발되어 왔다. 주성분분석(principal component analysis; PCA)의 경우 자료에 내재되어 있는 분산 구조를 최대로 설명하는 것이 목적이기 때문에 심볼릭 관측값 내의 분산을 설명하기 위한 다양한 심볼릭 주성분분석(symbolic PCA; SPCA) 방법들이 연구되었다.

SPCA는 구간형 자료에 관한 연구로부터 시작되었다. Cazes 등 (1997)과 Chouakria (1998)가 가장 먼저 분석을 시도하였는데, 이들은 구간의 중심점을 추출하여 단일 값 자료로 변환한 뒤 주성분을 계산하는 중심점 방법(centers method; CPCA)과 구간들에 의해 형성된 초직사각형(hyperrectangle)에서 꼭짓점을 추출해 주성분을 계산하는 꼭짓점 방법(vertices method; VPCA)을 제안하였다. CPCA에서 구간의 변동성을 전혀 반영하지 않는다는 문제점을 개선하기 위해, Palumbo와 Lauro (2003)와 Lauro 등 (2008)은 구간의 중심점을 추출하는 것에서 나아가 구간의 범위의 절반값을 추출하여 이들을 두 개의 변수로 사용하는 중심점-반지름 방법(midpoint-radii method; MRPCA)을 제안하였다. VPCA의 경우는 모든 꼭짓점들이 서로 독립인 관측값이라고 가정하였는데, 동일한 초직사각형에서 추출된 꼭짓점들은 서로 독립이라고 보기 힘들다는 문제점이 있다. Chouakria 등 (2011)에서도 VPCA의 공분산 행렬은 구간의 전체 변동이 아닌 일부만 설명하고 있다는 것을 보였다.

Ichino (2011)는 각 구간이 균일분포를 따른다고 가정한 후 분위수를 추출하여 주성분을 계산하는 분위수 방법(quantile method; QM)을 제안하였다. 분위수의 단조 구조에 의해 높은 상관관계를 가지는 자료를 이용하여 분산을 더욱 잘 설명하는 주성분을 찾는 방법이다. 이 방법은 기존의 방법들과 다르게 구간형, 히스토그램, 다중 값 변수들이 동시에 존재하는 자료에 적용 가능하다는 장점이 있다.

Le-Rademacher와 Billard (2012)는 Billard (2008)에 의해 고안된 심볼릭 공분산 행렬을 이용하여 주성분을 계산하는 심볼릭 공분산 방법(symbolic covariance PCA)을 제안하였다. 심볼릭 공분산 행렬은 구간 내 분산과 구간 간 분산으로 분해 될 수 있는데, 이때 구간 간 분산은 CPCA에서 구한 공분산과 동일한 반면, 구간 내 분산은 VPCA와 MRPCA에서 구한 공분산과 다름을 보였다. 또한, Wang 등 (2016)은 정규분포를 관측값으로 가지는 분포형 자료에 관한 주성분 분석 방법을 제안하였다.

본 논문에서는 구간형 자료에 대해 앞에서 설명된 여러 주성분분석 방법을 고찰하고 분위수 방법의 변형으로 구간에서 분위수를 추출할 때 균일분포를 사용하는 것이 아니라 구간의 중심점 부근이 좀 더 많은 정보를 가지고 있는 것으로 보고 절단정규분포를 사용하는 방법을 제안하였다. 각 방법에 대한 개념 및 소개는 2장에 제시했으며, 3장에서는 모의실험을 통해 2장에서 소개하거나 제안된 방법들의 결과를 비교한다. 4장에서는 OECD의 실제 자료에 적용하여 그 결과를 살펴본다. 또한 단일 값 자료에서 인위적으로 변형된 구간형 자료에 대해, 주성분 분석의 결과로써 각 주성분이 설명하는 분산 크기뿐만 아니라 주성분 점수의 분포를 비교한다. 마지막으로, 5장에서는 간략하게 결론을 제시하고자 한다.

## 2. 구간형 자료의 주성분분석 방법

주성분 분석은 $p$개의 본래 변수들에 내재되어 있는 분산을 가장 많이 설명하도록 $p$보다 적은 $s$개의 새로운 변수로 자료의 차원을 축소하는 방법이다. 이 때, 주성분은 공분산행렬의 고윳값과 고유벡터를 계산하여 구할 수 있다. $p$개의 변수로 구성된 자료 행렬 $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$와 공분산행렬이 있을 때, $\lambda_1 > \lambda_2 > \cdots > \lambda_p$를 공분산행렬의 고윳값, $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$를 각 고윳값에 대응되는 고유벡터라고 하면,

$v$번째 주성분은 다음과 같이 구할 수 있다.

$$PC_v = e_{v1}X_1 + e_{v2}X_2 + \cdots + e_{vp}X_p,$$

여기서 $v = 1, \ldots, p$에 대해 $\boldsymbol{e}_v = (e_{v1}, e_{v2}, \ldots, e_{vp})$는 $v$번째 고유벡터이며 이에 대응되는 고윳값 $\lambda_v$는 $v$번째 주성분이 설명하는 분산 크기이다.

구간형 자료의 주성분분석은 구간 내에 존재하는 변동을 설명해야 하므로 일반적인 주성분 분석 방법을 변형해서 적용해야 한다. 이를 위해 두 가지 단계를 고려할 수 있는데, 첫 번째는 구간형 자료의 변동을 완전히 설명할 수 있는 올바른 공분산 구조를 찾아 주성분을 계산하는 것이다. 두 번째는 첫 번째에서 구한 단일 값 주성분을 구간형 주성분으로 표현하는 것이다. 구간형 자료의 공분산 구조를 찾는 방법 중 가장 간단한 것은 구간으로부터 단일 값을 추출하여 이로부터 공분산행렬을 계산하는 것이다. 따라서 이 경우에는 구간의 정보를 최대한 반영할 수 있는 단일 값을 추출하는 것이 중요하다고 할 수 있다.

### 2.1. 중심점 방법

$p$개의 구간형 변수와 $n$개의 관측값으로 구성된 구간형 자료 행렬 $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)^t$가 있다고 하자. 여기서 $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$는 $i$번째 관측값을 나타내며, $i$번째 관측값의 $j$번째 변수는 $X_{ij} = [a_{ij}, b_{ij}]$, $a_{ij} \le b_{ij}$, $j = 1, \ldots, p$이며, $n \times p$ 구간형 자료 행렬은 다음과 같다.

$$\boldsymbol{X} = \begin{bmatrix} [a_{11}, b_{11}] & \cdots & [a_{1p}, b_{1p}] \\ \vdots & \ddots & \vdots \\ [a_{n1}, b_{n1}] & \cdots & [a_{np}, b_{np}] \end{bmatrix}. \tag{2.1}$$

중심점 방법(CPCA)은 Cazes 등 (1997)과 Chouakria (1998)가 제안한 것으로 각 구간에서 중심점을 추출하여 이로부터 주성분을 계산하는 방법이다. 구간형 자료가 식 (2.1)과 같이 주어졌을 때, $i$번째 관측값의 $j$번째 변수의 구간의 중심점은 다음과 같이 구할 수 있다.

$$X_{ij}^c = \frac{a_{ij} + b_{ij}}{2}, \quad i = 1, \ldots, n, \ j = 1, \ldots, p. \tag{2.2}$$

식 (2.2)를 이용하여 모든 관측값에서 중심점을 추출하면 이를 관측값으로 가지는 $n \times p$ 중심점 행렬 $\boldsymbol{X}^c$는 다음과 같이 나타낼 수 있다.

$$\boldsymbol{X}^c = \begin{bmatrix} X_{11}^c & \cdots & X_{1p}^c \\ \vdots & \ddots & \vdots \\ X_{n1}^c & \cdots & X_{np}^c \end{bmatrix}.$$

즉, 구간형 자료 행렬 $\boldsymbol{X}$를 중심점 행렬 $\boldsymbol{X}^c$로 변환해주었다. 중심점 행렬은 단일 값으로 구성되어 있기 때문에 일반적인 주성분 분석법을 적용하여 주성분을 계산할 수 있다.

CPCA는 중심점 행렬을 구하여 일반적인 주성분 분석법을 이용한다는 점에서 비교적 계산이 간단하다는 장점이 있다. 그러나 예를 들어 두 개의 구간 $[0, 100]$, $[40, 60]$이 있을 때 CPCA는 이 두 구간 값을 구분하지 못하고 중심점이 50으로 동일한 자료라고 판단한다. 따라서 이 방법은 구간 내 변동성을 전혀 반영하지 못한다는 단점이 있다. 다음의 꼭짓점 방법은 이와 같은 문제점을 보완한 방법 중 하나이다.

### 2.2. 꼭짓점 방법

꼭짓점 방법(VPCA)은 CPCA와 함께 Cazes 등 (1997)과 Chouakria (1998)에 의해 제안된 방법이다. 구간형 변수로 구성된 자료 공간에서 각 관측값은 초직사각형을 형성하게 되는데, 이때 초직사각형의 모든 꼭짓점을 추출하여 주성분을 계산하는 방법이다.

식 (2.1)의 자료 행렬에서 $i$번째 관측값은 $\boldsymbol{X}_i = ([a_{i1}, b_{i1}], \ldots, [a_{ip}, b_{ip}])$, $i = 1, \ldots, n$이다. 이때, $a_{ij} \neq b_{ij}$인 변수의 수를 $m_i$라고 하자. 즉, $m_i$는 $i$번째 관측값에서 구간의 상한과 하한이 같지 않은 구간형 변수의 개수를 의미한다. $H_i$를 $i$번째 관측값으로부터 형성된 초직사각형이라고 할 때 $H_i$는 $n_i = 2^{m_i}$개의 꼭짓점을 가지게 된다. 예를 들어 $m_i = 0$인 경우 $2^0 = 1$개의 꼭짓점을 가지므로 자료 공간에서 점을 의미하고, $m_i = 1$이면 $2^1 = 2$개의 꼭짓점을 가지는 선, $m_i = 2$이면 $2^2 = 4$개의 꼭짓점을 가지는 직사각형을 의미한다. 즉, $H_i$는 $i$번째 관측값이 $p$차원 자료공간에서 차지하는 영역이라고 할 수 있다.

각 초직사각형 $H_i$는 각 행이 꼭짓점의 좌표값 $(x^i_{k1}, x^i_{k2}, \ldots, x^i_{kp})$로 구성된 $n_i \times p$ 꼭짓점 행렬 $\boldsymbol{X}^v_i$로 다음과 같이 나타낼 수 있다.

$$\boldsymbol{X}^v_i = \begin{bmatrix} x^i_{11} & \cdots & x^i_{1p} \\ \vdots & \ddots & \vdots \\ x_{n_i 1} & \cdots & x_{n_i p} \end{bmatrix},$$

여기서 $i = 1, \ldots, n$에 대해 꼭짓점 행렬 $\boldsymbol{X}^v_i$은 각 초직사각형의 꼭짓점 수에 따라 행의 수가 다를 수 있다.

모든 관측값에 대해 총 꼭짓점의 수를 $N = \sum_{i=1}^n n_i = \sum_{i=1}^n 2^{m_i}$라고 하면 모든 초직사각형의 꼭짓점 좌표를 가지는 $N \times p$ 꼭짓점 행렬은 $\boldsymbol{X}^v = (\boldsymbol{X}^v_1, \ldots, \boldsymbol{X}^v_n)^t$으로 표현할 수 있다. 이제 구간형 자료 행렬 $\boldsymbol{X}$를 꼭짓점 행렬 $\boldsymbol{X}^v$로 변환해주었다. 중심점 행렬과 마찬가지로 꼭짓점 행렬은 단일 값으로 구성되어 있기 때문에 일반적인 주성분 분석법을 적용하여 주성분을 계산할 수 있다.

꼭짓점 행렬 $\boldsymbol{X}^v$는 각 꼭짓점의 좌표값을 나타내는 행렬로써 구간의 상한과 하한의 정보를 담고 있다. 따라서 VPCA는 CPCA와 달리 구간의 크기를 설명하고 있는 방법이다. 그러나 구간의 양 끝 값만을 이용한다는 점에서 여전히 구간형 자료의 전체 변동성을 설명하지 못한다는 한계가 있다. 또한 이 방법은 모든 꼭짓점이 서로 독립이라는 가정을 하였으나 동일한 초직사각형을 구성하는 꼭짓점들은 서로 독립이라고 보기 힘들다는 문제점이 있다.

### 2.3. 균일분포를 이용한 분위수 방법

분위수 방법(QM)은 Ichino (2011)에 의해 제안된 방법으로 각 구간에서 분위수를 추출하여 주성분을 계산하는 방법이다. 분위수를 추출하기 위해 모든 구간에 대해 균일분포(uniform distribution)를 가정하여 구간 내 모든 값이 동일한 확률로 발생한다고 보았다(QM-Uniform).

식 (2.1)의 자료 행렬에서 $i$번째 관측값의 $j$번째 변수는 $X_{ij} = [a_{ij}, b_{ij}]$, $a_{ij} \leq b_{ij}$로 관측된다. 각 구간이 균일분포를 따른다고 가정할 때, 구간을 $m$등분하는 위치를 나타내는 $(m-1)$개의 분위수 값을 다음과 같이 구할 수 있다.

$$Q_{kij} = a_{ij} + (b_{ij} - a_{ij}) \times \frac{k}{m}, \quad k = 1, \ldots, m-1. \tag{2.3}$$

구간의 상한과 하한은 각각 최댓값과 최솟값이 되며, 식 (2.3)을 이용하여 $(m-1)$개의 분위수 값을 구하였다면 각 구간의 $m$분위수를 나타내는 $(m-1)$-집합(tuple)을 다음과 같이 정의할 수 있다.

$$(a_{ij}, Q_{1ij}, Q_{2ij}, \ldots, Q_{(m-1)ij}, b_{ij}), \quad i = 1, \ldots, n, \ j = 1, \ldots, p.$$

즉, 상한과 하한 두 개의 값으로 표현된 구간을 $(m+1)$개의 분위수 값으로 표현한다. 예를 들어 사분위수의 경우 $m=4$이므로 5-집합을 만들 수 있으며 $(a, Q_1, Q_2, Q_3, b)$ 형태로 구성될 것이다.

모든 구간을 $(m-1)$-집합으로 표현하였다면 다음으로 $i$번째 관측값에 대해 $k$번째 분위수 벡터(quantile sub-object)를 다음과 같이 정의할 수 있다.

$$(Q_{ki1}, Q_{ki2}, \ldots, Q_{kip}), \quad i = 1, \ldots, n, \ k = 1, \ldots, m-1. \tag{2.4}$$

이는 $i$번째 관측값의 각 변수에서 동일한 분위수에 해당하는 값을 추출한 것이다. 즉, $(m+1)$-집합은 하나의 구간으로부터 $(m+1)$개의 분위수를 추출하여 구간을 재표현한 것이며, $k$번째 분위수 벡터는 각 변수의 $(m+1)$-집합로부터 $k$번째 분위수에 해당하는 값을 추출한 것이다.

대표적으로 최솟값과 최댓값을 추출한 것을 각각 최솟값 벡터(minimum sub-object), 최댓값 벡터(maximum sub-object)라고 하며 $i$번째 관측값에서 다음과 같이 구성된다.

$$(a_{i1}, a_{i2}, \ldots, a_{ip}), (b_{i1}, b_{i2}, \ldots, b_{ip}), \quad i = 1, \ldots, n. \tag{2.5}$$

식 (2.4)와 (2.5)를 통해 $(m+1)$개의 분위수 벡터를 구하였다면 이들을 행으로 가지는 분위수 행렬 $\boldsymbol{X}_i^q$를 정의할 수 있다. 각 관측값으로부터 $(m+1) \times p$ 분위수 행렬 $\boldsymbol{X}_i^q$를 구하면 모든 관측값에 대한 $(n(m+1)) \times p$ 분위수 행렬 $\boldsymbol{X}^q$는 다음과 같다.

$$\boldsymbol{X}^q = (\boldsymbol{X}_1^q, \ldots, \boldsymbol{X}_n^q)^t.$$

이제 구간형 자료로부터 구한 분위수 행렬 $\boldsymbol{X}^q$에 대해 일반적인 주성분분석 방법을 적용하여 주성분을 계산할 수 있다.

분위수 행렬로부터 구한 상관행렬은 분위수의 단조 구조에 의해 비대각 원소의 절댓값이 1에 가깝다. 이러한 상관행렬을 이용하여 고윳값과 고유벡터를 구했을 때 큰 고윳값을 얻을 수 있으며, 이는 주성분이 설명하는 분산 크기가 증가한다는 의미이므로 자료의 분산을 더 많이 설명하는 주성분을 찾는 것이 가능하도록 한 것이다. QM-Uniform은 자료의 분포함수(distribution function)를 이용하기 때문에 히스토그램 변수, 다중 값 변수 등에서도 분포함수를 적용한다면 분위수를 추출할 수 있으며 여러 형태의 변수가 동시에 포함된 자료에도 적용이 가능하다는 장점이 있다.

### 2.4. 절단정규분포를 이용한 분위수 방법

이 방법은 구간형 자료로부터 분위수를 추출한다는 점에서 2.3절과 동일하지만 분위수를 추출하기 위한 분포를 균일분포가 아닌 절단정규분포(truncated normal distribution)를 가정한다(QM-TNormal). 절단정규분포는 확률변수 $X$가 평균이 $\mu$, 분산이 $\sigma^2$인 정규분포를 따를 때, $-\infty \le a < b \le \infty$인 구간 $(a, b)$에서 $X$의 조건부 분포를 의미한다. 절단정규분포의 확률밀도함수는 다음과 같다.

$$f(x; \mu, \sigma, a, b) = \frac{\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}, \quad a \le x \le b.$$

이때, 본 논문에서는 2.3절의 QM-Uniform과 비교를 하기 위해 정규분포의 평균과 분산으로써 균일분포의 평균과 분산을 사용하였다. 균일분포가 아닌 정규분포를 사용한 것은 변동성이 있는 구간형 자료에서 일반적으로 중심점 부근의 정보가 더 대표성을 띌 수 있으며, 구간의 상한과 하한은 이상값에 의해 실제 구간의 범위보다 넓어질 수 있기 때문에 구간의 끝으로 갈수록 확률이 감소하는 것이 바람직하다는 예상에 근거한 것이다. 또한, 정규분포가 아닌 절단정규분포를 사용한 이유는 각 관측값은 주어진 구간에 속해있어야 하기 때문이다.

## 3. 모의실험

2절에서 소개한 4가지 방법(CPCA, VPCA, QM-Uniform, QM-TNormal)을 모의실험을 통하여 결과를 비교해 보고자 한다. 모든 방법에서 피어슨 상관행렬을 이용하여 주성분 분석을 진행하며 특히 QM에 기반한 방법들은 스피어만 상관행렬을 이용한 경우를 추가하여 총 6가지 방법을 비교한다. 이는 QM의 경우 분위수들이 크기 순서대로 행을 구성하므로 단조 구조를 가지고 있으며 따라서 순위(rank order)를 이용한 경우와 비교해보기 위함이다. 또한 분위수는 모든 자료에서 공통적으로 사분위수를 사용하기 때문에 5-집합(min, Q1, Q2, Q3, max)을 형성하여 분위수 방법에 적용하였다.

### 3.1. 모의실험 자료 생성

모의실험을 위한 자료는 다음과 같은 절차로 생성한다.

Step1: $n \times p$ 단일 값 자료 행렬 $\boldsymbol{X}$를 평균벡터가 $\mu$이고 분산공분산 행렬이 $\Sigma$인 정규분포로부터 생성한다.

Step2: 구간형 자료의 행의 수 $m$을 선택한다.

Step3: Step1에서 생성한 단일 값 자료를 첫 행부터 차례대로 $n/m$개의 행씩 취합하여 각 변수별로 최솟값과 최댓값을 추출한다.

Step4: Step3의 최솟값과 최댓값을 이용하여 $m \times p$ 구간형 자료 행렬 $\boldsymbol{X}^I$를 생성한다.

CPCA와 QM은 변수의 개수와 상관없이 항상 동일한 수의 행을 가지는 중심점 행렬과 분위수 행렬을 생성한다. 하지만 VPCA의 경우 변수의 수가 증가할수록 초직사각형을 구성하는 꼭짓점의 수가 증가하므로 꼭짓점 행렬의 행의 수가 달라진다. 따라서 변수의 개수가 다를 때 결과에 영향을 미치는지 확인해보기 위하여 변수의 개수 $p$에 따라 다음과 같이 실험을 설계한다.

실험1: $n = 1000,\ m = 10,\ p = 3,\ \mu = (0,0,0)^t$,

$$\Sigma = \begin{pmatrix} 1.61 & -0.40 & 0.70 \\ -0.40 & 0.95 & 0.27 \\ 0.70 & 0.27 & 1.45 \end{pmatrix}$$

실험2: $n = 1000,\ m = 10,\ p = 5,\ \mu = (0,0,0,0,0)^t$,

$$\Sigma = \begin{pmatrix} 0.37 & -0.04 & 0.22 & 0.25 & -0.50 \\ -0.04 & 0.61 & -0.06 & 0.10 & -0.03 \\ 0.22 & -0.06 & 1.00 & 0.28 & -0.61 \\ 0.25 & 0.10 & 0.28 & 0.49 & -0.23 \\ -0.50 & -0.03 & -0.61 & -0.23 & 1.01 \end{pmatrix}$$

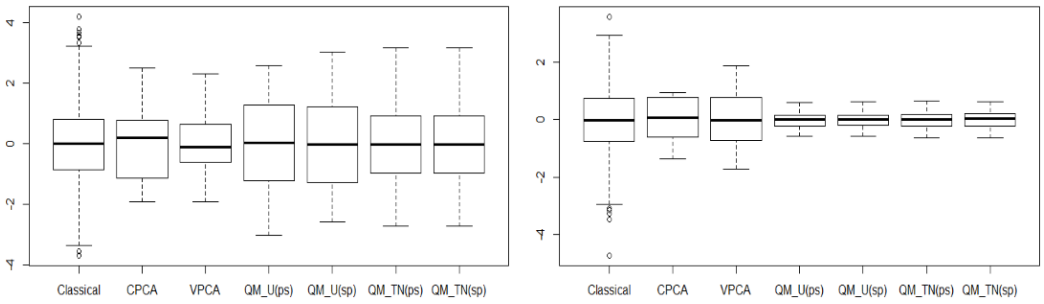
실험3: $n = 1000,\ m = 10,\ p = 7,\ \mu = (0,0,0,0,0,0,0)^t$,

$$\Sigma = \begin{pmatrix} 1.83 & 0.98 & 0.74 & 0.74 & 0.85 & -0.13 & -0.25 \\ 0.98 & 2.74 & 0.80 & -0.43 & -0.69 & -0.31 & 0.08 \\ 0.74 & 0.80 & 1.54 & -0.12 & 0.50 & 0.33 & 0.79 \\ 0.74 & -0.43 & -0.12 & 2.12 & -0.11 & -1.13 & 0.27 \\ 0.85 & -0.69 & 0.50 & -0.11 & 2.59 & 0.96 & -0.10 \\ -0.13 & -0.31 & 0.33 & -1.15 & 0.96 & 1.83 & -0.03 \\ -0.25 & 0.08 & 0.79 & 0.27 & -0.10 & -0.03 & 1.64 \end{pmatrix}$$

**Table 3.1.** Proportions of variance explained by the first three principal components of each method in experiment 1

| | PC1 | PC2 | PC3 |
|---|---|---|---|
| CPCA | 48.43 | 32.09 | 19.48 |
| VPCA | 33.65 | 33.30 | 33.05 |
| QM-Uniform(ps) | 97.24 | 1.72 | 1.04 |
| QM-Uniform(sp) | 97.50 | 1.48 | 1.02 |
| QM-TNormal(ps) | 97.00 | 1.87 | 1.12 |
| QM-TNormal(sp) | 96.94 | 1.77 | 1.28 |
| Classical | 48.22 | 40.90 | 10.88 |

ps = Pearson, sp = Spearman.




(a) PC1 (b) PC2

**Figure 3.1.** Box plots of PC scores in experiment 1.

모의실험을 위해 오픈 소스인 R을 사용하였으며 MASS 패키지의 mvrnorm 함수를 이용하여 다변량정규분포를 따르는 자료를 생성하였다. 각 실험에서 가정한 평균과 분산을 따르는 단일 값 자료를 먼저 생성한 뒤 이 자료를 바탕으로 구간형 자료를 생성하였다. 일반적으로 단일 값 자료는 각 관측단위의 정보를 담고 있으며 구간형 자료는 집합 단위의 정보를 담고 있기 때문에 그 결과가 완전히 일치한다고 할 수는 없다. 그러나 구간형 자료를 생성하는데 기반이 된 자료가 단일 값 자료이므로 단일 값 자료로 일반적인 주성분 분석법을 적용한 것과 구간형 자료로 2장에서 소개한 주성분분석 방법들을 비교해보기로 한다.

비교를 위한 측도로는 각 주성분의 분산설명 비율(proportion of variance)과 주성분 점수(pc score)를 이용한다. PCA를 실시한 후 주성분을 새로운 축으로 하여 계산된 주성분 점수는 차원이 축소된 새로운 자료의 값이 되기 때문에 주성분 점수의 분포를 확인할 필요가 있다.

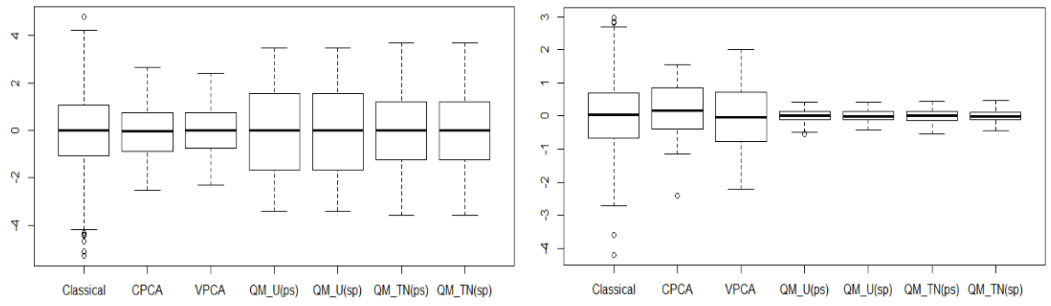
### 3.2. 모의실험 결과

실험1에 대한 모의실험 결과는 Table 3.1에 제시하였다. CPCA는 단일 값 자료의 경우(classical)와 유사한 주성분의 분산비율을 가지며, VPCA는 PC1, PC2, PC3의 분산비율이 거의 일정하였다. QM에 기반한 4가지 경우 모두 PC1에서 매우 높은 분산비율을 가지며 PC2에서 급격히 감소하는 팔꿈치 지점(elbow point)을 보였다. 주성분분석의 목적인 자료에 내재되어 있는 소수의 주성분으로 분산을 가장 많이 설명하는 측면에서 QM이 가장 좋은 성능을 보인다고 할 수 있다. 그러나 SPCA의 경우 각 방법별로 구간에서 추출한 단일 값 행렬이 다르기 때문에 분산비율만으로 성능을 평가하기 힘들 수 있다. 따라서 이에 더하여 주성분 점수의 분포를 상자그림을 그려 비교해보았다. 실험1의 주성분 점수는 Figure 3.1과 같다.

**Table 3.2.** Proportions of variance explained by the first five principal components of each method in experiment 2

| | PC1 | PC2 | PC3 | PC4 | PC5 | Cum.Var (PC2) |
|---|---|---|---|---|---|---|
| CPCA | 40.74 | 30.83 | 16.16 | 7.46 | 4.80 | 71.57 |
| VPCA | 20.25 | 20.11 | 19.97 | 19.87 | 19.80 | 40.86 |
| QM-Uniform(ps) | 96.91 | 1.28 | 1.04 | 0.44 | 0.33 | 98.19 |
| QM-Uniform(sp) | 96.60 | 1.48 | 1.11 | 0.54 | 0.28 | 98.08 |
| QM-TNormal(ps) | 96.70 | 1.37 | 1.11 | 0.48 | 0.34 | 98.07 |
| QM-TNormal(sp) | 96.45 | 1.53 | 1.11 | 0.61 | 0.30 | 97.98 |
| Classical | 51.38 | 21.68 | 14.13 | 12.18 | 0.63 | 73.06 |

ps = Pearson, sp = Spearman.




(a) PC1 (b) PC2

**Figure 3.2.** Box plots of PC scores in experiment 2.

각 상자그림은 왼쪽부터 단일 값 자료, CPCA, VPCA, QM-Uniform(ps), QM-Uniform(sp), QM-TNormal(ps), QM-TNormal(sp)에 의한 결과를 나타낸다. Figure 3.1의 PC1 점수를 보면 모두 0 근방에서 중앙값을 가지며 구간형 자료에 기반한 6가지 방법 중 QM의 4가지 경우에서 주성분 점수들이 가장 넓게 분포하였다. 또한 절단정규분포를 이용한 경우가 균일분포를 이용한 경우보다 사분위범위가 좁은 것을 확인할 수 있다. 이는 절단정규분포가 중심점 부근의 정보를 더 많이 활용하였기 때문임을 알 수 있다. Figure 3.1의 PC2 점수를 보면 특히 QM에서 매우 밀집된 분포 형태를 확인하였다. 이는 QM이 PC1에서 많은 분산을 설명하고 있기 때문에 PC2에서 작은 분산 값을 가지게 되었음을 나타낸다. 변수의 수가 5개로 증가한 실험2의 결과는 Table 3.2와 같으며 PC1과 PC2 점수의 상자그림은 Figure 3.2에 나타내었다.

변수의 개수가 증가하여도 VPCA는 모든 주성분에서 분산비율이 동일한 수준으로 나타났다. QM은 Table 3.2에 의하면 앞의 경우와 마찬가지로 하나의 주성분 PC1만으로도 대부분의 분산을 설명하고 있다. Figure 3.2에 의하면 PC1에서는 주성분 점수가 넓게 분포하며, PC2에서는 매우 밀집되었음을 알 수 있다. 절단정규분포를 이용한 QM의 경우가 균일분포를 이용한 경우보다 좁은 사분위범위를 가지면서 넓은 수염(whisker) 범위를 가지고 있다.

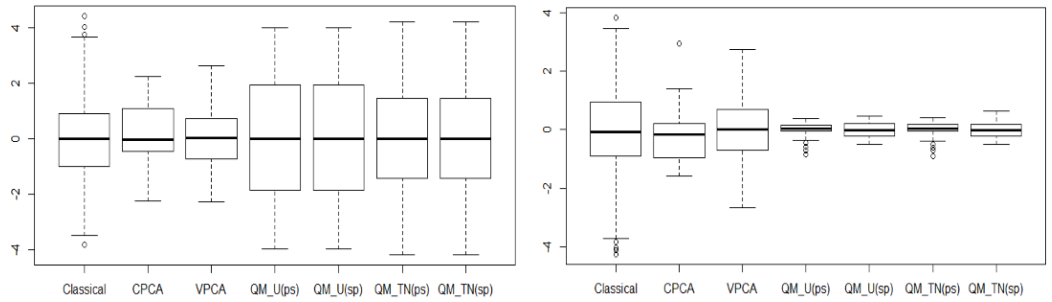
변수의 개수가 가장 많은 실험3에서도 같은 결과가 나타났다. QM의 경우가 PC1 점수에서 단일 값 자료의 분포와 가장 유사하면서 넓은 분포를 가졌다. PC2에서는 QM의 주성분 점수가 가장 밀집되게 나타났다.

심사위원께서 분위수방법의 경우 분위수의 개수에 따라 결과가 차이가 있는지, 어떻게 변화하는지에 대한 확인을 요청하였고, 이를 위하여 $m$을 4, 5, 7, 9로 하여 5, 6, 8, 10-집합의 경우를 살펴보았다. 앞에

**Table 3.3.** Proportions of variance explained by the first seven principal components of each method in experiment 3

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | Cum.Var (PC2) |
|---|---|---|---|---|---|---|---|---|
| CPCA | 38.44 | 24.25 | 12.86 | 10.27 | 8.00 | 3.85 | 2.33 | 62.69 |
| VPCA | 14.58 | 14.41 | 14.28 | 14.25 | 14.23 | 14.15 | 14.11 | 28.99 |
| QM-Uniform(ps) | 97.15 | 1.03 | 0.65 | 0.51 | 0.37 | 0.19 | 0.10 | 98.18 |
| QM-Uniform(sp) | 96.33 | 1.19 | 0.80 | 0.66 | 0.57 | 0.26 | 0.18 | 97.52 |
| QM-TNormal(ps) | 96.96 | 1.12 | 0.67 | 0.54 | 0.39 | 0.20 | 0.11 | 98.08 |
| QM-TNormal(sp) | 96.32 | 1.28 | 0.79 | 0.68 | 0.46 | 0.28 | 0.19 | 97.60 |
| Classical | 29.36 | 25.08 | 18.96 | 18.16 | 4.19 | 3.32 | 0.93 | 54.44 |

ps = Pearson, sp = Spearman.




(a) PC1 (b) PC2

**Figure 3.3.** Box plots of PC scores in experiment 3.

서 서술한 바와 결과가 비슷하고 지면 관계상 구체적인 결과표를 본 논문에 수록하진 않았다. 다만, 여기서 확인한 것은 $t$-tuple에서 $t$를 증가시킬수록, 스피어만 상관행렬을 이용한 경우 PC1의 분산비율은 매우 소폭이지만 증가하고 PC2는 소폭 감소하였으며, 반대로 피어슨 상관행렬의 경우는 PC1에서 소폭 감소하고 PC2에서 소폭 증가하였다. 이는 $t$가 증가할수록 분위수 행렬의 단조성이 증가함으로 인해 스피어만 상관행렬을 이용할 때 PC1에서 더 많은 분산을 설명하고 팔꿈치 지점이 명확해지는 것으로 이해할 수 있다. 하지만 그 변화량은 유의미한 정도라고 보기에 충분치는 않은 정도이며 실제 문제에서는 대략 5-집합을 선택하면 된다고 하겠다.

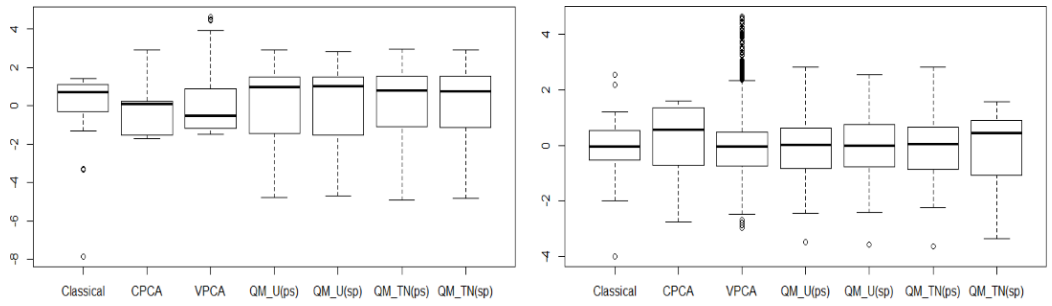
## 4. 실제 자료 적용

이 절에서는 OECD 국가의 실제 자료를 이용하여 주성분 분석 결과를 비교해본다. 자료는 통계청(http://kostat.go.kr)의 국제기구통계에서 확인할 수 있으며 각 연도별로 OECD 국가의 주요통계지표를 보여주고 있다. OECD 통계(https://stats.oecd.org)에서 추가적인 지표 자료를 제공하고 있다. 본 논문에서 사용한 자료는 2016년 자료로 36개국의 국내총생산, 1인당 GDP, 경제성장률, 수출, 수입, 실업률, 소비자 물가지수, 조강생산량의 8가지 정보를 포함하고 있다. OECD는 세계경제의 공동발전을 위해 설립된 경제협력개발기구로써 그 자료를 국가 단위로 분석하는 것 보다 대륙 단위로 취합하여 구간형 자료로 분석하는 것이 의미가 있을 것으로 판단된다. 따라서 36개국을 아시아, 북아메리카, 남아메리카, 유럽, 오세아니아 5개 대륙의 구간형 자료로 변형하여 분석을 진행하였다. 남아메리카의 경우 칠레 1개국의 자료만 포함되어 있어 칠레는 2015년 자료를 이용하여 전체 변수에서 단일 값이 되지 않도록 해주었다. 따라서 최종적으로 사용한 OECD 주요통계지표 데이터는 행렬의 단일 값 자료를

**Table 4.1.** Proportions of variance explained by the first eight principal components of each method for OECD data

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | Cum.Var (PC2) |
|---|---|---|---|---|---|---|---|---|---|
| CPCA | 42.65 | 40.39 | 11.63 | 5.33 | 0.00 | 0.00 | 0.00 | 0.00 | 94.67 |
| VPCA | 24.77 | 18.52 | 12.28 | 11.17 | 9.68 | 8.91 | 7.67 | 7.00 | 55.57 |
| QM-Uniform(ps) | 57.96 | 23.26 | 14.52 | 2.59 | 1.26 | 0.37 | 0.04 | 0.00 | 95.74 |
| QM-Uniform(sp) | 65.24 | 15.07 | 13.72 | 3.80 | 1.15 | 0.66 | 0.35 | 0.02 | 94.03 |
| QM-TNormal(ps) | 56.89 | 23.92 | 14.77 | 2.72 | 1.30 | 0.36 | 0.04 | 0.00 | 95.58 |
| QM-TNormal(sp) | 64.16 | 15.56 | 14.15 | 3.99 | 1.18 | 0.63 | 0.30 | 0.03 | 93.87 |
| Classical | 44.65 | 16.95 | 16.12 | 8.28 | 7.20 | 4.47 | 2.27 | 0.05 | 77.72 |

ps = Pearson, sp = Spearman.




(a) PC1 (b) PC2

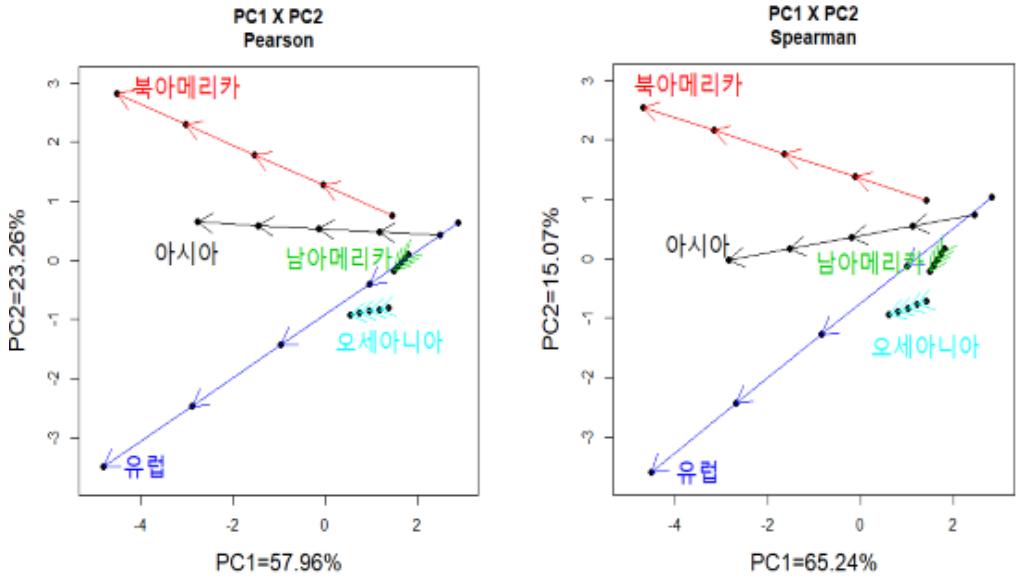
**Figure 4.1.** Box plots of PC scores for OECD data.

변형한 행렬의 구간형 자료이다. 이 자료를 이용하여 3절에서와 같이 구간형 자료를 이용한 6가지 주성분분석 방법의 경우 각 주성분이 설명하는 분산 크기와 상위 세 개의 주성분의 누적 분산비율을 Table 4.1에 나타내었다.

PC1의 분산비율은 CPCA가 단일 값 자료의 분산비율과 가장 유사하였으나 PC2의 분산비율과 PC1에 비해 PC2가 설명하는 분산비율의 감소량을 종합하여 고려하면 QM이 자료의 분산을 설명하는 측면에서 가장 좋은 성능을 보인다. QM에서도 QM-Uniform(sp)과 QM-TNormal(sp)이 QM-Uniform(ps)과 QM-TNormal(ps)보다 더 많은 분산을 설명하고 있다. 이는 분위수 행렬은 값이 작은 순서대로 추출되어 행을 구성하므로 단조구조를 가지고 있으며 따라서 스피어만 상관행렬을 이용하는 것이 바람직함을 의미한다.

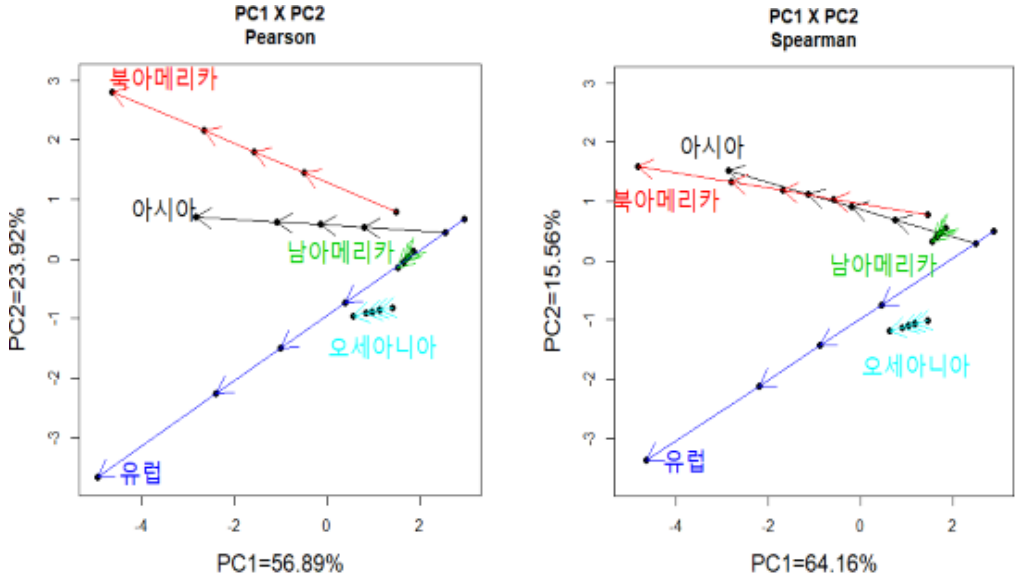
다음으로 주성분 점수를 살펴보자. Figure 4.1은 각 방법을 이용한 PC1과 PC2의 주성분 점수의 상자그림이다. 좌측의 PC1 상자그림에서 QM의 결과가 단일 값 자료의 결과와 유사하게 1 근방에서 중심점을 가지며 가장 넓은 분포를 가지고 있다. 또한 QM-TNormal의 경우가 좁은 사분위범위와 넓은 수염을 가지고 있기 때문에 PC1에서 자료의 분산을 많이 설명하고 있음을 알 수 있다. PC2의 상자그림에서 QM은 분포가 넓게 나타났으나 단일 값 자료와 함께 0 근방의 중심점을 가지며 아래 수염 바깥쪽에서 동일하게 이상점 한 개가 존재하였다. 또한 상자그림의 분포가 다른 방법의 것과 크게 다르지 않기 때문에 PC1과 PC2의 주성분 점수의 분포를 종합적으로 보았을 때 QM에서 자료의 분산을 많이 설명하는 주성분을 찾아 주성분 점수를 구할 수 있는 것으로 보인다.

QM에 기반한 4가지 경우는 주성분 공간에 산점도를 그리고 화살표 표현법(arrow line representa-




(a) Pearson (b) Spearman

**Figure 4.2.** Arrow line representation for OECD data by using QM-Uniform method.




(a) Pearson (b) Spearman

**Figure 4.3.** Arrow line representation for OECD data by using QM-TNormal method.

tion)을 통해 분위수들을 표현할 수 있다. PC1 × PC2 공간에서 동일한 구간에서 추출된 분위수들을 (min, Q1, Q2, Q3, max) 순서대로 화살표를 이용하여 이어줌으로써 분위수들을 해당 구간에 따라 구분되도록 한 것이다. 또한, 만약 구간의 상·하한의 정보만을 이용하여 주성분 점수를 계산한다면 PC1 ×

**Table 4.2.** Eigenvectors of PC1 for OECD data

| | QM_Uniform (ps) | QM_Uniform (sp) | QM_TNormal (ps) | QM_TNormal (sp) |
|---|---|---|---|---|
| $X_1$ (국내총생산) | −0.342 | −0.399 | −0.345 | −0.406 |
| $X_2$ (1인당GDP) | −0.358 | −0.323 | −0.356 | −0.320 |
| $X_3$ (경제성장률) | −0.345 | −0.323 | −0.341 | −0.318 |
| $X_4$ (수출) | −0.440 | −0.417 | −0.444 | −0.418 |
| $X_5$ (수입) | −0.392 | −0.399 | −0.396 | −0.404 |
| $X_6$ (실업률) | −0.318 | −0.288 | −0.314 | −0.288 |
| $X_7$ (소비자물가지수) | −0.221 | −0.225 | −0.211 | −0.206 |
| $X_8$ (조강생산량) | −0.372 | −0.408 | −0.374 | −0.410 |

ps = Pearson, sp = Spearman.

**Table 4.3.** Eigenvectors of PC2 for OECD data

| | QM_Uniform (ps) | QM_Uniform (sp) | QM_TNormal (ps) | QM_TNormal (sp) |
|---|---|---|---|---|
| $X_1$ (국내총생산) | 0.450 | 0.347 | 0.441 | 0.241 |
| $X_2$ (1인당GDP) | −0.392 | −0.430 | −0.400 | −0.582 |
| $X_3$ (경제성장률) | −0.461 | −0.531 | −0.462 | −0.551 |
| $X_4$ (수출) | 0.093 | 0.198 | 0.091 | 0.113 |
| $X_5$ (수입) | 0.312 | 0.319 | 0.304 | 0.166 |
| $X_6$ (실업률) | −0.479 | −0.457 | −0.478 | −0.141 |
| $X_7$ (소비자물가지수) | 0.139 | −0.031 | 0.154 | 0.417 |
| $X_8$ (조강생산량) | 0.278 | 0.247 | 0.280 | 0.255 |

ps = Pearson, sp = Spearman.

PC2 공간에서 구간형 자료는 직사각형으로 표현하게 된다. 그러나QM에서는 분위수들의 주성분 점수를 계산하기 때문에 구간 내 분위수들의 위치 뿐만 아니라 최솟값부터 최댓값까지의 방향의 변화를 구체적으로 파악할 수 있다.

Figures 4.2와 4.3은 각각 균일분포와 절단정규분포를 이용한 QM의 주성분 점수의 산점도를 나타낸 것이다. 왼쪽은 피어슨 상관행렬, 오른쪽은 스피어만 상관행렬을 이용한 경우이다. 각 축에는 주성분의 분산설명비율을 나타내었다.

Figure 4.2에서는 모든 분위수들이 구간의 상·하한의 주성분 점수 사이에서 동일한 간격으로 분포되어 있다. 전체적으로 PC1 방향에서 분위수들이 이어지는 것으로 보아 PC1이 구간 내 변동성을 가장 많이 설명하는 변수로 보이며 '유럽'과 '북아메리카'는 PC2에서도 주성분 점수의 분포가 넓게 나타났다. 두 산점도는 매우 유사하였으나 '아시아'의 경우 피어슨에서는 좌상향이었고 스피어만에서는 좌하향으로 나타나는 차이점이 보였다. 이에 대한 해석을 위해 Tables 4.2와 4.3의 PC1과 PC2의 고유벡터의 값(loading)을 보면 PC2의 $X_7$(소비자물가지수) 고유벡터 값이 반대 부호의 값을 가진 것을 확인할 수 있고 이것이 이유임을 짐작할 수 있다.

Figure 4.3에서는 Figure 4.2와 달리 제 1, 3사분위수들이 중심점에 가깝게 분포되었다. 또한 왼쪽과 오른쪽의 두 산점도는 분위수의 위치와 화살표의 방향에서 차이를 보였다. 피어슨의 경우 '아시아'와 '북아메리카'의 분위수들은 겹치는 부분이 없었으나 스피어만의 경우 분위수의 위치와 방향이 매우 유사해지면서 겹치게 표현되었다. 즉, 나머지 3개의 산점도와 다르게 Figure 4.3의 스피어만 상관행렬을 이용한 QM-TNormal 방법에서 차이점을 보였다. 이는 Table 4.3의 고유벡터 값을 부호를 제외하고 절댓값으로 비교할 때 다른 경우에 비해 QM-TNormal(sp)에서 $X_5$(수입)와 $X_6$(실업률)는 감소하였고 $X_7$

(소비자물가지수)는 증가하였기 때문으로 짐작할 수 있다. 고유벡터의 값은 본래 변수들이 각 주성분에 기여하는 정도로 볼 수 있으므로 세 변수의 중요도에서 큰 차이가 나타났음을 확인할 수 있다.

## 5. 결론

본 논문에서는 구간형 자료의 주성분 분석법 세 가지를 소개하고 추가적으로 절단정규분포를 이용하는 분위수 방법을 제시하였다. 세 가지 방법을 간략하게 정리하면 CPCA는 구간의 중심점을 추출하여 일반적인 주성분 분석을 적용하는 방법으로 구간의 변동성을 전혀 반영하지 못한다는 문제점이 있다. 이를 보완하는 VPCA는 자료 공간에서 구간형 관측값이 형성하는 초직사각형의 꼭짓점을 추출하여 주성분 분석을 적용하는 방법으로써 구간의 상·하한의 정보를 이용하기 때문에 구간 내 변동성을 일부 반영하고 있다. QM-Uniform과 QM-TNormal은 각각 구간에 균일분포와 절단정규분포를 가정하여 분위수를 추출하여 주성분 분석을 실시한다.

각 방법들의 성능을 비교해보기 위해 모의실험과 OECD 관련 실제 데이터에 적용하여 결과를 비교해보았다. 주성분의 분산비율을 살펴보았을 때 PC1의 높은 분산비율과 PC2에서 매우 명확한 팔꿈치 지점이 보이는 것을 고려했을 때 QM이 적은 주성분 개수로 자료의 분산을 가장 많이 설명하는 것으로 판단된다. 주성분 점수의 분포를 살펴보았을 때에도 PC1에서 가장 넓게 분포되어 있으며 PC2에서 매우 밀집되어 분포된 것으로 보아 자료의 분산을 잘 설명하는 것을 확인하였다. 또한 균일분포에 비해 중심점 부근의 정보를 더 포함하고 있는 절단정규분포를 이용한 경우에서 사분위수 범위가 더 좁으며 PC1 점수에서 더 넓은 분포를 가지고 있었음을 확인할 수 있었다. 결론적으로 분위수를 이용한 심볼릭 주성분분석법은 우수한 성질을 보이며, 구간형 자료뿐만 아니라 히스토그램 변수, 다중값 자료 및 이들 변수가 동시에 존재하는 자료에도 적용할 수 있는 장점이 있다.

## References


Billard, L. (2008). Sample covariance functions for complex quantitative data. In Mizuta M. and Nakano J. (Eds), *Proceedings of the International Association of Statistical Computing*, 157–163, Yokohama.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Wiley, Chichester.

Cazes, P., Chouakria, A., Diday, E., and Schektman, Y. (1997). Extension de l'analyse en composantes principales à des données de type intervalle, *Revue de statistique appliquée*, **45**, 5–24.

Chouakria, A. (1998). Extension des méthodes d'analyse factorielles à des données de type intervalle, Ph.D. Dissertation, Universite Paris-Dauphine.

Chouakria, A., Billard, L., and Diday, E. (2011). Principal component analysis for interval-valued observations, *Statistical Analysis and Data Mining*, **4**, 229–246.

Ichino, M. (2011). The quantile method for symbolic principal component analysis, *Statistical Analysis and Data Mining*, **4**, 184–198.

Lauro, N. C., Verde, R., and Irpino, A. (2008). Principal component analysis of symbolic data described by intervals. In Diday, E. and Noirhomme-Fraiture, M. (Eds), *Symbolic Data Analysis and the SODAS Software*, Wiley, Chichester, 279–311.

Le-Rademacher, J. and Billard, L. (2012). Symbolic Covariance Principal Component Analysis and Visualization for Interval-Valued Data, *Journal of Computational and Graphical Statistics*, **21**, 413–432.

Palumbo, F. and Lauro, N. C. (2003). A PCA for interval-valued data based on midpoints and radii. In Yanai, H., Okada, A., Shigemasu, K., Kano, Y. and Meulman, J. (Eds), *New Developments in Psychometrics*, 641–648.

Wang, H., Chen, M., Shi, X., and Li, N. (2016). Principal component analysis for normal-distribution-valued symbolic data, *IEEE Transactions on Cybernetics*, **46**, 356–365.

# 구간형 자료의 주성분 분석에 관한 연구

최수진[a] · 강기훈[a,1]

[a]한국외국어대학교 통계학과

(2019년 12월 30일 접수, 2020년 1월 6일 수정, 2020년 1월 7일 채택)

**요 약**

심볼릭 자료 중 하나인 구간형 자료는 모든 관측값에서 단일 값이 아닌 구간을 값으로 취하며, 관측값 내에 변동이 존재한다는 특징을 갖는다. 주성분 분석은 자료의 분산을 최대로 설명하여 자료의 차원을 축소하는 방법이므로 구간형 자료의 주성분 분석은 관측값 간의 분산 뿐만 아니라 관측값 내의 분산 역시 설명하여야 한다. 본 논문에서는 구간형 자료의 세 가지 주성분 분석법을 소개하고자 한다. 또한 기존의 분위수 방법에서 균일분포를 사용하는 것이 아니라 구간의 중심점 부근이 좀 더 많은 정보를 가지고 있는 것으로 보고 절단정규분포를 사용하는 방법을 제안하였다. 모의실험과 OECD 관련 실제 통계 자료를 통하여 각 방법의 결과를 비교해 보았다. 마지막으로 분위수 방법의 경우 화살표 표현법을 통해 주성분 산점도를 그리고 분위수들의 위치와 분포를 확인하였다.

주요용어: 꼭짓점 방법, 분위수 방법, 심볼릭 자료, 절단정규분포, 중심점 방법

이 연구는 2019학년도 한국외국어대학교 교내학술연구비 지원과 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2017R1D1A1B03035026).

[1]교신저자: (17035) 경기도 용인시 처인구 모현면 외대로 81, 한국외국어대학교 통계학과.
E-mail: khkang@hufs.ac.kr