응용통계연구 (2010)
23(4), 767–776

Broken-Stick 모형에 기초한 주성분 공헌도평가

강유정[1] · 변자현[2] · 김기영[3]

[1]롯데카드주식회사, [2]고려대학교 통계학과, [3]고려대학교 통계학과

(2010년 6월 접수, 2010년 6월 채택)

요 약

Broken-Stick 모형 (Barton과 David, 1956) 하에서 순서화된 분절구간의 기대길이를 기초로 유효차원의 개수를 결정하는 Frontier (1976)방법은 일관된 모의실험 결과를 제공하는 기준 중의 하나로 보고된 바 있다 (Jackson, 1993). 이 연구에서는 Broken-Stick 모형(BSM) 하에서 분절구간길이의 분포를 이용하여 주성분 상대공헌도의 크기를 확률적으로 평가하는 BSM 유의확률기준을 제안한다. 이에 부가하여 소득분포의 불균등성을 도식화한 로렌츠 곡선과 이에 대응하는 지니계수를 통해 주성분 공헌도의 포괄적 균등성을 탐구한다.

주요용어: 주성분 상대공헌도, Broken-Stick 모형, 로렌츠곡선, 지니계수.

1. 서론

주성분의 공헌도평가는 적절한 차원으로의 자료축약이라는 점에서 다차원 통계분석의 주요과제중의 하나이다. 이를 위해 여러 가지 대수적 (Hotelling, 1933; Bartlett, 1950; Jolliffe, 1972), 기하적 (Cattell, 1966; Box 등, 1973; Lambert 등, 1990) 기법들이 제안되었고, 이들 간 상대적 유용성은 Zwick과 Velicer (1986), Jackson (1991), Jackson (1993) 등에 의해 평가된 바 있다. 특히 Broken-Stick Model(BSM; Barton과 David, 1956) 하에서 개별 분절구간의 기대길이를 기준으로 해당 주성분의 보유여부를 결정하는 방법이 Frontier (1976)에 의해 고려되었는데, 이 방법은 고유값/고유벡터에 관한 붓스트랩 측도와 더불어 가장 일관된 모의실험 결과를 주는 방법 중의 하나로 보고된 바 있다 (Jackson, 1993). 이 연구에서는 주성분의 상대공헌도를 BSM에 대비하여 그의 확률적 유의미성을 평가하고자 한다. 이와 더불어 소득분포의 불균등정도를 나타내는 전통적 지표인 지니계수 (Gini, 1912)와 이에 관련된 로렌츠곡선 (Lorenz, 1905)을 통해 BSM 하에서 주성분 공헌도의 균등성을 지표화하고, 도식화하는 방법을 고려한다.

크기 $n\times p$ 자료행렬로부터 계산된 $p\times p$ 표본 공분산행렬 $S=\{s_{ij}\}$ [혹은 상관행렬 $R=\{r_{ij}\}$]의 k번째 고유값과 고유벡터를 각각 $(l_k, \underline{v_k})$, $k=1,\ldots,p$로 표기하자(단, $l_1>\cdots>l_k>\cdots>l_p$). 이때 k번째 주성분의 p-변량 총변이에 대한 상대공헌도 $c_k, k=1,\ldots,p$는 S-기초 주성분분석에서는 $c_k=l_k/\mathrm{tr}(S)$, R-기초 주성분분석의 경우 $c_k=l_k/p$로서 $\sum_{k=1}^{p} c_k=1$이 된다. 따라서 주성분 상대공헌도는 단위길이 상에서 만들어진 확률분절구간의 길이에 대응되어, 이 값의 임의성은 BSM을 통해 묘사할 수 있다.

2. Broken-Stick 모형(BSM)

종의 공평성이나 생태계 건강성을 묘사하기 위한 확률모형으로 널리 고려된 BSM (MacArthur, 1957; Pielou, 1975; Almorza와 Garcia, 2008; 등)은 변수들 간 상관관계가 클 경우 변수선택 (Jackson,

본 연구는 교신저자의 연구학기 기간(2009년 03월 01일-2009년 08년 31일)에 수행되었음.
[3]교신저자: (136-701) 서울 성북구 안암동 5-1, 고려대학교 통계학과, 교수. E-mail: kykim@korea.ac.kr

1993)에, 그리고 환경자료에 대해 변수선택 (King과 Jackson, 1999)이나, cDNA 자료에 대한 주성분 추정 (Richard와 Alain, 2007)에 이용되어왔다. 다음은 BSM에 대한 분포함수와 모멘트 함수들이다.

2.1. 분포와 모멘트

BSM은 단위길이 $[0,1]$상에 임의로 $(p-1)$개의 점을 택할 때 얻어지는 p개 분절구간의 길이(G)에 관한 모형이다. 이들을 내림차순 $g_1 > \cdots > g_k > \cdots > g_p$로 표기할 때(단, $\sum_{k=1}^{p} g_k = 1$, $0 < g_k < 1$, $k = 1, 2, \ldots, p$), 다음은 마지막 m개의 G들에 대한 결합확률함수와 몇 가지 모멘트이다.

$$f(g_{p-m+1}, \ldots, g_p) \propto [1 - g_p - \cdots - g_{p-m+2} - (p-m+1)g_{p-m+1}]^{p-m-1} \tag{2.1}$$

$$E(G_k) = \frac{1}{p}Q(k) \tag{2.2}$$

$$\mathrm{Cov}(G_k, G_j) = \frac{1}{p(p+1)}\left[Q(k^2) - \frac{1}{p}Q(k)Q(j)\right], \quad k \geq j$$

$$\text{단, } Q(k^a) = \sum_{i=k}^{p} i^{-a}.$$

$p = 4$일 때 식 (2.1)로부터 유도한 G_k, $k = 1, 2, 3, 4$의 주변분포는 식 (4.1)에 주어진다.

2.2. 점근분포

분절구간의 개수(p)가 증가함에 따라 식 (2.1)로부터 개별 주변확률함수 $f(g_k)$, $k = 1, 2, \ldots, p$를 구하는 과정은 점차 복잡해진다. 이 경우 G_k의 점근분포를 이용할 수 있는데, 이는 다음 3부분으로 나누어 표현된다.

(1) 상위 k번째 극단값

$$2\exp(-pG_k + \ln p) \sim \chi^2_{2k}. \tag{2.3}$$

(2) 중간 $100q\%$ 분위값: $k - 1 \sim qp$, $0 < q < 1$

$$p^{\frac{1}{2}}(pG_k + \ln q) \sim N\left(0, q^{-1} - 1 - \left(\ln q^{-1}\right)^2\right).$$

(3) 하위 k번째 극단값

$$2p^2 G_k \sim \chi^2_{2(p-k+1)}.$$

3. BSM에 기초한 주성분 공헌도평가

Horn (1965)은 독립적 정규난수자료에 대한 Scree플롯을 주성분 보유여부의 기준으로 하는 방법을 고려하였다. 이와 유사한 개념으로 k번째 주성분의 상대공헌도(c_k)는 적어도 BSM 하에서 k번째 분절구간의 기대길이$(E(G_k))$ 보다 커야 보유가치가 있는 주성분으로 간주하는 방법(BSM Frontier기준)이 Frontier (1976)에 의해 고려되었다. 이 연구에서는 k번째 주성분의 상대공헌도를 BSM 하에서 G_k의 확률분포에 근거하여 평가하는 방법을 제안한다. 더불어 주성분 공헌도의 포괄적 균등성을 점검하기 위해 로렌츠곡선과 지니계수의 도입을 고려한다. 로렌츠곡선과 지니계수는 주성분분석에서 관찰개체의 영향력을 분석하기 위해 사용된 바 있다 (Bénasséni, 2005).

3.1. BSM 하에서 주성분 공헌도의 유의확률($\mathbf{PROB}_k$)

관찰된 k번째 주성분의 상대공헌도(c_k)가 BSM 하에서의 임의성을 인정할 수 없을 정도로 클 경우, 다음 확률 $P[G_k > c_k] = \text{PROB}_k$은 충분히 작은 값을 가질 것으로 기대된다. 이와 같은 PROB_k는 k번째 주성분의 상대공헌도가 유의적이지 않다는 명제에 반하는 증거의 강도를 나타내는 지표로, 해당 주성분의 보유여부를 결정하는데 이용할 수 있어 "BSM 유의확률"로 표기한다. 따라서 BSM 유의확률은 통상적 가설검정에서의 유의확률과 비슷하게 그 값이 작을수록 그 주성분의 공헌도가 크게 됨을 의미하며, 사용자에 의해 주어진 임계수준 α에 대해 보유하게 될 주성분은 다음 조건을 만족하는 첫 m개가 된다.

$$m = \max_{1 \le k \le p} \{k : \text{PROB}_k \le \alpha\} \tag{3.1}$$

이와 같은 BSM 유의확률기준은 평균(상대)공헌도(Guttman-Kaiser) 기준($l_k > \text{tr}(S)/p$일 때 보유)이나, 이를 보정한 Jolliffe 기준($l_k > 0.7 \times \text{tr}(S)/p$일 때 보유), 90% 이상의 누적상대공헌도기준 등 기존에 통상적으로 이용되어온 직관적인 기준에 대해 주성분 상대공헌도의 확률적 크기를 기준으로 한다는 특징을 지닌다. 또한 이는 BSM Frontier기준에서처럼 각 k에서 기준점을 한 값($E(G_k)$)에 고정하지 않고 연구자의 판단에 따라 임계수준을 정함으로서 보유주성분의 수에 대해 신축적인 의사결정을 내릴 수 있다. 물론 임계수준을 작게 택하면 Frontier 기준에 따른 결과에서 보다 통상적으로 작은 개수의 주성분을 보유하는 경향을 보이겠지만, 이런 현상은 주성분분석이 가지는 자료탐색적 입장에서 수용할 수 있는 특성이라고 할 수 있다.

3.2. 주성분 공헌도에 대한 로렌츠곡선

주성분 공헌도의 포괄적 균등성을 도식화한 로렌츠곡선은 다음과 같이 구축된다. 마지막 m개 주성분들의 누적상대공헌도를 $C_m = \sum_{k=p-m+1}^{p} c_k$로 표시하자(단, $C_0 = 0$). 이때 주성분 공헌도에 대한 로렌츠곡선은 가로축을 주성분개수의 누적백분비($x_m = m/p$)를, 그리고 세로축을 주성분의 누적상대공헌비($y_m = C_m$)로 하는 점(x_m, y_m), $m = 0, 1, \ldots, p$들을 선분으로 잇는 연속조각선형함수(continuous piecewise linear function)이다(단, $x_0 = 0$, $y_0 = 0$) (Anand, 1983).

이에 따라 모든 주성분이 다변량 총변이에 균등한 공헌을 할 경우 로렌츠곡선은 기울기 45°인 직선의 형태를 취하게 되고, 첫 주성분을 제외한 나머지 주성분들이 전혀 설명력을 가지지 못하면 ($x < 1.0$)일 때 $y = 0$이고, $x = 1.0$일 때 $y = 1$(완전불균등선)이 된다. 그러므로 주성분들의 공헌도가 균등상태로부터 이탈한 정도는 로렌츠곡선과 완전균등선 사이의 (불균등)면적의 크기를 통해 표현되며, 이는 Scree(혹은 LEV)플롯 등에 대한 하나의 대안으로 생각할 수 있다.

또한, 공헌도의 불균등성을 평가할 때 비교기준을 완전균등선 대신 BSM 하에서의 기대 로렌츠곡선인 점 $(m/p, \sum_{k=p-m+1}^{p} E(G_k))$을 연결한 연속조각선형함수를 고려할 수 있을 것이다. 이때 관찰된 로렌츠곡선이 기대로렌츠곡선 보다 상위에 있으면 주성분 공헌도들이 BSM 하에서의 경우에 비해 더 균등하다고 할 수 있으나, 그 반대의 경우 공헌도들이 평균해서 BSM 하에서의 임의적 균등성을 가지지 못하는 것으로 판단된다.

3.3. 주성분 공헌도에 대한 지니계수

로렌츠곡선에 나타나는 균등상태로부터의 이탈정도를 대수적으로 지표화한 고전적 지니계수는 3.2절에서의 불균등면적을 2배한 값으로 주성분 공헌도의 경우 다음과 같이 표현할 수 있다.

$$\text{gini} = 1 - \sum_{k=1}^{p} \left(\frac{1}{p}\right)(C_{k-1} + C_k) = 1 - \sum_{k=1}^{p} \left(\frac{1}{p}\right)(2k-1)c_k. \tag{3.2}$$

통용되는 기준(OECD)에 의하면 지니계수의 값이 0.2 보다 작으면 매우 균등하고, $[0.2, 0.5]$면 보통, 그리고 0.5 보다 크면 매우 불균등함을 뜻한다.

한편 식 (3.2)와 같은 통상적 gini 대신, 식 (2.2)를 이용하여 구한 BSM 하에서의 모멘트를 기준으로 평가한 다음 형식의 표준화 지니계수(Z_{gini})를 고려할 수 있을 것이다.

$$Z_{gini} = \frac{\text{gini} - E_{\text{BSM}}(\text{Gini})}{\sqrt{\text{VAR}_{BSM}(\text{Gini})}} \tag{3.3}$$

$$\text{단, } E_{BSM}(\text{Gini}) = \frac{p-1}{2p}$$

$$\text{Var}_{BSM}(\text{Gini}) = \frac{1}{p^2}\left[\sum_{k=1}^{p}(2k-1)^2\text{Var}(G_k) + 2\sum_{k>j}^{p}(2k-1)(2j-1)\text{Cov}(G_k, G_j)\right].$$

4. 평가기준의 적용

여기에서는 3장에서 고려한 기준들의 이용가능성을 점검하고, 기존에 제안된 방법들과 비교하기 위해 실제 응용예로서 탄도미사일자료에 여러 기준들을 적용한 결과를 평가한다. 특히 p가 클 때 BSM 극한분포를 이용하는 예는 음소자료를 통해 고찰한다.

4.1. 탄도미사일자료 (Jackson, 1959, 1960)

탄도미사일(ballistic missile)자료($n = 40$, $p = 4$)에 대한 변수내역과 표본공분산행렬(S), 그리고 고유값 및 고유벡터(l_k, $\underline{v_k}$, $k = 1, 2, 3, 4$)는 각각 다음과 같다.

$$X_1 = \text{integrator reading(gauge♯1)}, \qquad X_2 = \text{planimeter measurement(gauge♯1)}$$

$$X_3 = \text{integrator reading(gauge♯2)}, \qquad X_4 = \text{planimeter measurement(gauge♯2)}$$

$$S = \begin{bmatrix} 102.74 & & & \\ 88.67 & 142.74 & & \\ 67.04 & 86.56 & 84.57 & \\ 54.06 & 80.03 & 69.42 & 99.06 \end{bmatrix}$$

$$(l_1, l_2, l_3, l_4) = (335.34, 48.03, 29.33, 16.41), \qquad \text{tr}(S) = 429.11$$

$$\underline{v_1}^T = (0.468, 0.608, 0.459, 0.448), \qquad \underline{v_2}^T = (-0.622, -0.179, 0.139, 0.750).$$

이 자료($p = 4$)에서 $c_k = l_k/\text{tr}(S)$, $E(G_k) = 1/4\sum_{i=k}^{4} 1/i$와 BSM 유의확률 [PROB$_k$] 및 첫 k개 주성분의 누적공헌도($\sum_{i=1}^{k} c_i$) 등이 표 4.1에 주어진다.

PROB$_k$을 구하기 위해 사용된 다음과 같은 G_k의 주변분포(식 (4.1))은 식 (2.1)로부터 유도되었다.

$$f_{234}(g_2, g_3, g_4) = 144, \quad 0 < g_4 < \frac{1}{4},\ g_4 < g_3 < \frac{1-g_4}{3},\ g_3 < g_2 < \frac{1-g_4-g_3}{2}. \tag{4.1}$$

$$f(g_1) = \begin{cases} 12(4g_1 - 1)^2, & \frac{1}{4} < g_1 < \frac{1}{3}, \\ 12\left(10g_1 - 2 - 11g_1^2\right), & \frac{1}{3} < g_1 < \frac{1}{2}, \\ 12(1 - g_1)^2, & \frac{1}{2} < g_1 < 1, \\ 0, & \text{otherwise}, \end{cases}$$

표 **4.1.** 탄도미사일자료($p = 4$)

k	c_k	$E(G_k)$	PROB_k	$\sum_{i=1}^{k} c_i$
1	**0.7815**	**0.5208**	**0.1488**	0.7815
2	0.1119	0.2708	0.9601	**0.8934**
3	0.0684	0.1458	0.8936	**0.9618**
4	0.0382	0.0625	0.5457	1

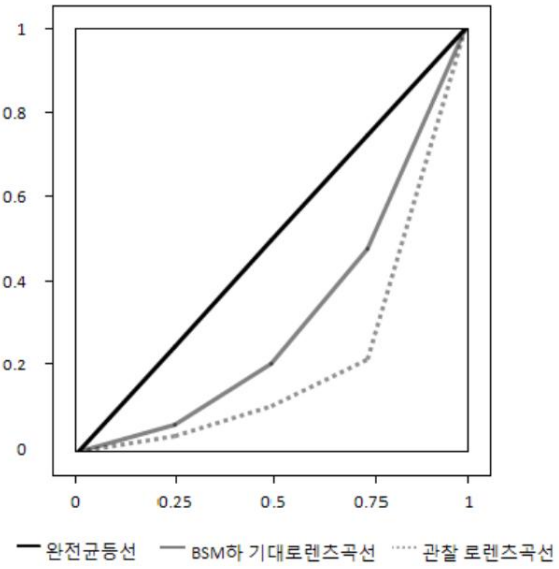

그림 **4.1.** 로렌츠곡선-탄도미사일자료

$$f(g_2) = \begin{cases} 72g_2^2, & 0 < g_2 < \frac{1}{4}, \\ 36(8g_2 - 1 - 14g_2^2), & \frac{1}{4} < g_2 < \frac{1}{3}, \\ 36(1 - 2g_2)^2, & \frac{1}{3} < g_2 < \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases}$$

$$f(g_3) = \begin{cases} 72\left(g_3 - \frac{7}{2}g_3^2\right), & 0 < g_3 < \frac{1}{4}, \\ 36(1 - 3g_3)^2, & \frac{1}{4} < g_3 < \frac{1}{3}, \\ 0, & \text{otherwise.} \end{cases}$$

$$f(g_4)) = \begin{cases} 12(1 - 4g_4)^2, & 0 < g_4 < \frac{1}{4}, \\ 0, & \text{otherwise.} \end{cases}$$

표 4.1을 기초로 그린 로렌츠곡선이 그림 4.1이다.

(1) 로렌츠곡선과 지니계수: 그림 4.1에서 관찰로렌츠곡선이 기대로렌츠곡선 보다 상당히 아래에 위치하여 주성분 공헌도는 BSM 하에서의 기대되는 것보다 불균등한 상황이다. 한편 gini = 0.5684로 주성분 공헌도의 포괄적 균등성을 OECD 기준으로 파악할 때 "매우 불균등한" 상태를 나타내고 있다. 이런 상황은 BSM 하 표준화 지니계수에서도 비슷하게 나타나고 있는데, 이 경우 $E_{BSM}(\text{Gini}) = 0.375$, $\text{Var}_{BSM}(\text{Gini}) \doteq 0.0156$이므로 $Z_{gini} = 1.5472$가 된다. 즉, 관찰된 지니계수는 BSM 하에서 그의 평균으로부터 1.55 표준편차 거리에 있음을 알 수 있다. 전체적으로 4개 주성분 공헌도는 불균등하여 따라서 차원축약의 가능성을 보이고 있다.

보유주성분의 개수를 결정하기 위해 표 4.1에 주어진 값들을 기초로 몇 가지 기준들을 적용해본 결과는 다음과 같다.

(2) BSM에 따른 기준: 표 4.1에서 BSM 유의확률기준을 살펴보면, $\text{PROB}_1 = 0.1488$로 15% 정도의 임계수준에서 첫 주성분은 유의적인 상대공헌도를 가지고 있다고 하겠고, 나머지 모든 주성분에 대한 PROB 값은 상당히 크므로 첫 주성분만을 보유하는 것이 적절한 것으로 판단된다. BSM Frontier기준을 적용한 경우도 $k = 1$일 때만 $c_1 > E(G_1)$의 부등관계를 만족하므로 BSM 유의확률 기준에서와 동일한 결과를 가진다.

(3) 기타:

- Guttman-Kaiser 기준($c_k > 1/p = 0.25$일 때): 첫 주성분 보유
- Jolliffe 기준($c_k > 0.175$일 때): 첫 주성분 보유
- 누적상대공헌도: 80% 이상의 경우에는 첫2개의 주성분을, 통상 기계적으로 이용되는 90% 이상의 경우는 첫 3개 주성분을 보유
- Bartlett 카이제곱검정 (Anderson, 1963): 마지막 m개 주성분공헌도의 균등성에 대한 X^2검정은 다음과 같이 첫 주성분을 제외한 나머지 주성분들의 공헌도에 유의한 차이가 없음을 나타내고 있다.

$$m = 2 : X^2 = 3.23, \quad df = 2 : p\text{-값} > 0.05$$
$$m = 3 : X^2 = 10.85, \quad df = 5 : p\text{-값} > 0.05$$
$$m = 4 : X^2 = 110.67, \quad df = 9 : p\text{-값} < 0.05$$

- 편상관기준 (Velicer, 1976): 유의미한 첫 m개 주성분들을 제거한 후 잔차들 간 제곱편상관계수들의 평균(f_m)은 각각 $f_0 = 0.50$, $f_1 = 0.14$, $f_2 = 0.38$, $f_3 = 1.00$로 $m = 1$에서 최소값을 가지므로 첫 주성분만을 보유하게 된다.

전체적으로 탄도미사일자료의 경우 누적상대공헌도의 경우를 제외한 모든 기준들이 첫 주성분만을 보유하는 것이 타당하다는 일치된 결과를 제공하고 있다.

4.2. 음소(音素)자료 (Hastie 등, 1995)

p가 매우 클 때 BSM 유의확률은 (2.3)에 주어진 G_k의 점근분포를 통해 구할 수 있다. 이 경우에 대한 응용예로서 음성인지에 관한 연구에서 크기 $n = 4509$, $p = 256$의 음소(phoneme)자료에 대한 상관행렬을 고려한다. 우선 이 행렬의 고유값으로부터 얻은 그림 4.2의 로렌츠곡선을 보면, 관찰로렌츠곡선이 BSM하 기대로렌츠곡선 보다 훨씬 아래에 위치하고 있음을 알 수 있다. 여기서 gini = 0.8862이고, $E_{BSM}(\text{Gini}) \doteq 0.498$, $\text{Var}_{BSM}(\text{Gini}) \doteq 0.00032$로서 $Z_{Gini} \doteq 21.7$가 된다. 즉, 음소자료의 경우 주

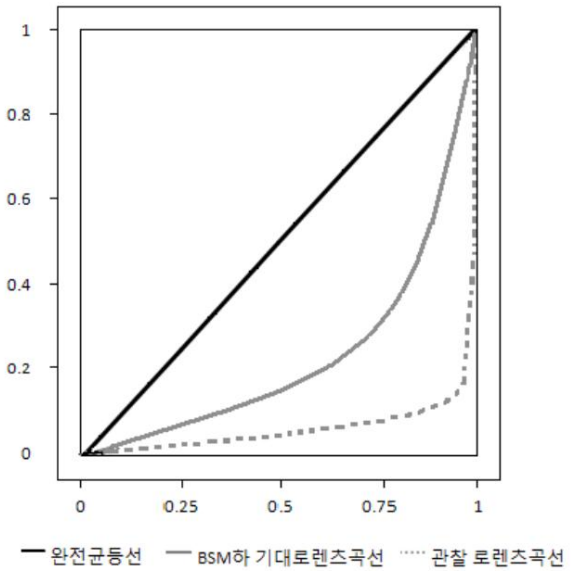

그림 4.2. 로렌츠곡선-음소자료

표 4.2. 음소자료($p = 256$)

k	R의 고유값 l_k	c_k	$\sum_{i=1}^{k} c_i$	$E(G_k)$	PROB_k
1	141.86538	0.55416	0.55416	**0.02392**	**0.0000**
2	38.87824	0.15187	0.70603	**0.02002**	**0.0000**
3	12.36727	0.04831	0.75434	**0.01806**	**0.0000**
4	5.15807	0.02015	0.77449	**0.01676**	**0.0622**
5	3.75896	0.01468	0.78917	0.01579	0.7112
6	3.27519	0.01279	**0.80197**	0.01500	0.9202
7	2.35208	0.00919	0.81115	0.01435	1.0000
8	2.20591	0.00862	0.81977	0.01379	1.0000
9	1.64267	0.00642	0.82619	0.01331	1.0000
10	1.40550	0.00549	0.83168	0.01287	1.0000
11	1.34284	0.00525	0.83692	0.01248	1.0000
12	1.15816	0.00452	0.84145	0.01213	1.0000
13	**1.12700**	0.00440	0.84585	0.01180	1.0000
14	0.99636	0.00389	0.84974	0.01150	1.0000
15	0.89348	0.00349	0.85323	0.01122	1.0000
16	0.85279	0.00333	0.85656	0.01096	1.0000
17	0.77508	0.00303	0.85959	0.01072	1.0000
18	**0.71254**	**0.00278**	0.86237	0.01049	1.0000
19	0.68028	0.00266	0.86503	0.01027	1.0000
20	0.64945	0.00254	**0.86757**	0.01006	1.0000
⋮	⋮	⋮	⋮	⋮	⋮

성분 공헌도는 전체적으로 매우 불균등하며, 유의적 차이가 있는 것으로 판단할 수 있다. 이제 표 4.2을 통해 개별 주성분 공헌도를 평가하고, 이를 바탕으로 주성분의 축약된 차원수를 판단해본다.

우선 Guttman-Kaiser 기준($l_k \geq 1$)에 따르면 첫 (13개) 주성분을 그리고 Jolliffe 기준($l_k \geq 0.7$)에서는 (18개)를 보유하게 된다. 그러나 누적공헌도기준에서 '80% 이상'을 택하면 (6개)를 보유해야 하고, '90% 이상'을 택하는 경우에는 (20개) 보다 더 많은 주성분을 보유해야 함을 나타내고 있다. 이에 대해 BSM Frontier기준이나 BSM 유의확률기준에서는 모두 첫 (4개) 주성분을 보유하는 것이 적절한 것으로 판단된다. 이런 결과는 Guttman-Kaiser, Jolliffe, 혹은 누적공헌도 90% 이상을 기준으로 한 경우와 현격한 차이를 보이고 있다.

그러나 첫 4개의 주성분이 가지는 누적공헌도가 전체변이의 약 78% 정도인 한편, 5, 6번째 주성분의 개별공헌도가 1%를 약간 상회하고 있으며, 특히 7번째 이후의 모든 주성분은 개별적으로 1% 미만의 미미한 공헌도를 가지고 있음을 주시할 필요가 있다. 이런 결과는 특정 자료에 대한 것이어서 제한적이기는 하나 BSM Frontier방법이 일관성 있는 결과를 제공한다는 기존의 모의실험보고 (Jackson, 1993)와 더불어 BSM 유의확률기준의 유용성에 대한 하나의 예증적 근거로 볼 수 있다. 특히 음소자료와 같이 원래변수차원의 수가 크나 미소크기의 공헌도를 가지는 많은 개수의 주성분이 있는 경우 Guttman-Kaiser, 누적공헌도, 혹은 Bartlett 검정 등의 방법이 사소한 잔여변이를 설명하기 위해 실제보다 지나치게 많은 개수의 의미 없는 주성분을 보유하게 되는 경향이 있음을 나타낼 가능성이 높다.

5. 결론

이 연구는 유의한 공헌도를 가지는 주성분을 식별하기 위한 기준으로 BSM 하에서 주성분 공헌도의 확률적 크기를 나타내는 BSM 유의확률기준의 이용가능성을 탐구한다. 이 기준에 따른 결과는 기존의 BSM Frontier기준이나 Guttman-Kaiser-Jolliffe기준, 누적공헌도 및 Bartlett 검정이나 Velicer 편상관기준 등의 결과와 비교평가 되었다. BSM에 기초한 방법들이 다른 방법의 결과에 비해 주성분의 개수를 상대적으로 작게 보유하는 경향이 있었다. 그러나 BSM Frontier 기준에 대한 Jackson의 모의실험결과를 감안하거나 여기서 고려한 두 가지 실제 예가 제공하는 결과를 볼 때 BSM에 기초한 기준은 타당한 수준의 보유주성분수를 추정하고 있는 것으로 판단된다.

BSM 유의확률기준은 통상적으로 설정되는 임계값의 수준을 고려하면 BSM Frontier기준보다 작은 개수의 주성분을 보유하게 되지만 주성분분석이 가지는 자료의 탐색과 간명성 추구하는 입장에서 이런 경향은 긍정적인 측면으로 간주할 수 있다. 구분될만한 차이는 전자가 BSM 하 분절구간의 확률분포에 기초한다면 후자는 단순히 확률분포의 한 특성인 기댓값에만 연관하고 있어 BSM 유의확률기준에서 의사결정의 신축성이 제고되고 있다는 점에 있다.

이에 추가하여 이 연구에서는 보유주성분수를 결정하는 차원축약과정에 앞서 모든 주성분 공헌도의 전반적 균등성 여부를 시각적으로 가늠해보는 장치로서 몇 가지 유형의 로렌츠곡선을 고려하며, 이 곡선의 불균등상태를 객관적인 수치로 지표화한 지니계수를 평가하는 방법을 제시하고 있다. 여기에서도 BSM 하에서의 BSM 로렌츠곡선과 BSM을 중심으로한 표준화 지니계수를 고려함으로서 개별적 주성분 공헌도의 평가 이전에 총괄적 차원축약가능성을 타진할 수 있다는 유용성을 가진다.

제안된 이들 기법들은 특히 원자료의 정규성 등에 관한 분포적 제한에서 해지되어 있으며, 특히 $n < p$와 같은 경우에도 제한 없이 적용할 수 있다는 점도 있다. 이 경우 물론 p가 크면 분절구간길이의 확률분포를 유도하는데 어려움이 예상되나 변수의 개수가 충분히 클 때 점근분포나 정규근사를 활용하므로 이 점이 절대적인 제약이라고 할 수는 없다고 판단된다. 또한. 이 연구는 BSM 관련 기준이 제공하는 보유주성분수의 추정에서의 일관성 (Jackson, 1993)을 배경으로 하고 있으나, BSM에 기초한 판정기준이 가지는 유용성이 일반화되기 위해서는 원자료의 분포적 특성과 반응변수들 간의 다양한 상관구조 및 표본크기 등을 감안한 광범위한 모의연구가 요구된다고 하겠다.

참고문헌

Almorza, D. A. and Garcia, M. H. (2008). Results of exploratory data analysis in the broken Stick model, *Journal of Applied Statistics*, **35**, 979–983.

Anand, S. (1983). *Inequality and Poverty in Malaysia*, Oxford University Press, New York.

Anderson, T. W. (1963). An asymptotic expansion for the distribution of the latent roots of an estimated covariance matrix, *The Annals of Mathematical Statistics*, **36**, 1153–1173.

Bartlett, M. S. (1950). Tests of significance in factor analysis, *British Journal of Psychology(Statistical section)*, **3**, 77–85.

Barton, D. E. and David, F. N. (1956). Some notes on ordered random intervals, *Journal of the Royal Statistical Society, Series B*, **18**, 79–94.

Bénasséni, J. (2005). A concentration study of principal components, *Journal of Applied Statistics*, **32**, 947–957.

Box, G. E. P., Hunter, W. G., MacGregor, J. F. and Erjavaz, J. (1973). Some problems associated with the analysis of multiresponse data, *Technometrics*, **15**, 33–51.

Cattell, R. B. (1966). The scree test for the number of factors, *Multivariate Behavioral Research*, **1**, 245–276.

Frontier, S. (1976). Étude de la decroissance des valeurs propres dans une analyse en composantes principales: Comparison avec le modéle du baton brisé, *Journal of Experimental Marine Biology and Ecology*, **25**, 67–75.

Gini, C. (1912). Variabilità e mutabilità, Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). *Libreria Eredi Virgilio Veschi*, 1955, Rome.

Hastie, T., Buja, A. and Tibshirani, R. (1995). Penalized discriminant analysis, *The Annals of Statistics*, **23**, 73–102.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis, *Psychometrika*, **30**, 179–185.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, **24**, 417–441.

Jackson, D. A. (1993). Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches, *Ecological Society of America*, **74**, 2204–2214.

Jackson, J. E. (1959). Quality control methods for several related variables, *Technometrics*, **1**, 359–377.

Jackson, J. E. (1960). Multivariate analysis illustrated by Nike-Hercules, In *Proceedings of the Thirtieth Conference on the Design of Experiments in Army Research, Development, and Testing, U.S. Army Research Office*, Durham, N.C. 307–327.

Jackson, J. E. (1991). *A User's Guide to Principal Components*, John Wiley & Sons, INC.

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis I: Artificial data, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **21**, 160–173.

King, R. J. and Jackson, D. A. (1999). Variable selection in large environmental data sets using principal components analysis, *Environmetrics*, **10**, 67–77.

Lambert, Z. V., Wildt, A. R. and Durand, R. M. (1990). Assessing sampling variation relative to number-of-factors criteria, *Educational and Psychological Measurement*, **50**, 33–48.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth, *American Statistical Association*, **9**, 209–219.

MacArthur, R. H. (1957). On the relative abundance of bird species, In *Proceedings of the National Academy of Sciences USA*, **43**, 293–295.

Pielou, E. C. (1975). *Ecological Diversity*, Willy, New York.

Richard, C. and Alain, G. (2007). Component retention in principal component analysis with application to cDNA microarray data, *Biology Direct*, **2**, 2.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations, *Psychometrika*, **41**, 321–327.

Zwick, W. R. and Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain, *Psychological Bulletin*, **99**, 432–442.

Contribution of Principal Components Based on the Broken-Stick Model

Y. J. Kang[1] · J. H. Byun[2] · K. Y. Kim[3]

[1]Lotte Card Co., Ltd.; [2]Department of Statistics, Korea University
[3]Department of Statistics, Korea University

(Received June 2010; accepted June 2010)

Abstract

Frontier (1976) suggested a criterion based on the expected length of ordered random intervals under the Broken-stick model (Barton and David, 1956) to determine the optimal number of principal components retained. It is considered to be one of the methods that provide the most consistent simulation results (Jackson, 1993). This study is aimed to propose a method using the distribution of ordered random intervals to evaluate the contribution of principal components. We also examine several types of Gini indices along with the corresponding Lorenz curves to visualize the overall equivalence of those contributions.

Keywords: Contribution of principal components, Broken-stick model, Lorenz curve, Gini index.

[3]Corresponding author: Professor, Department of Statistics, Hankuk University, 86-1 Hankang-Dong, Dong-Gu, Seoul 122-807, Korea. E-mail: pghwang@hu.ac.kr