응용통계연구 제 18권 1호, 2005년, pp. 129-135

주성분의 자기일치성에 기초한 다변량 대표관찰치의 기하적 표현

김기영[1] 박용주[2]

요 약

일변량 자료의 경우 대표관찰치는 사분위수 등에 기초하여 자료의 분포와 변이를 함축적으로 표현하기 위한 목적으로 사용되는 소수 개의 관찰치이다. Jones와 Rice(1992)는 다변량 자료에 대한 대표관찰치를 선택하기 위해 주성분분석에 근거한 방법을 제시한 바 있다. 이 연구에서는 주성분의 자기일치성을 이용하여 대표관찰치를 선택하고, 이를 표현하는 방안을 고찰한다. 기존의 방법에 의한 대표관찰치가 자료의 표본변이에 민감한 한편, 여기에서 제안되는 방법의 결과는 자기일치성을 가진다.

주요용어: 주성분분석, 자기일치성, 대표관찰치

1. 서론

다변량 자료의 변이와 분포에 관한 정보는 일변량의 경우 관찰 자료의 분위수(quan-tiles) 등에 해당하는 다변량 대표관찰치(representative observation)들을 통해 요약할 수 있다. Flury(1990)는 관찰개체들과 대표관찰치 간의 평균제곱거리로 정의되는 손실함수를 최소화하는 해로서 주요점(principal point)의 개념을 도입하였다. 여기서 주요점은 주어진 분포 혹은 자료를 점방식 근사(pointwise approximation)하는 소수 개의 대표관찰치라고 할 수 있다. 따라서 다변량 대표관찰치를 구하기 위해서는 이와 같은 주요점을 추정하면 될 것이다. 그러나 Jones와 Rice(1992)는 주요점을 유도하는 과정에서 일어나는 국소 최적(local optimum)의 문제를 감안하여 이를 직접 추정하는 대신 직관적인 방법으로 주성분분석을 이용하여 대표관찰치를 선택하는 방법을 제시하였는데, 그 후 주요점이 주성분 부분공간에 있음이 밝혀짐에 따라(Tarpey, 1995) Jones와 Rice가 제시한 방법은 이론적 타당성을 가지게 되었다. 이 연구에서는 자기일치성(self-consistency)의 개념을 주성분분석에 적용하여 Jones와 Rice의 방법을 개선한 대안을 제안하고, 그 결과를 기존의 방법에 따른 것과 비교, 분석한다. 관련된 몇 가지 기본개념과 정리들을 약술하면 다음과 같다.

두 확률벡터 $\mathbf{X}$, $\mathbf{Z}$ 에 대해 다음 조건부 기대값 $\mathbf{E_x}(\mathbf{X}|\mathbf{Z}) = \mathbf{Z}$의 관계가 만족되면 $\mathbf{Z}$는 $\mathbf{X}$에 대해 자기일치성을 가진다고 한다(Tarpey와 Flury, 1996). 이 때 $\mathbf{Z}$가 연속형이라면 $\mathbf{Z}$는 $\mathbf{X}$의 자기일치곡선(self-consistent curve)이 되고, 이산형인 경우 자기일치점(self-consistent point)이 된다. 그리고 주요점은 자기일치점 중에서 $\mathbf{X}$와의 평균제곱거리가 최소인 것이다. 자기일치성과 주성분과의 관계는 다음과 같다.

1) (136-701) 서울시 성북구 안암동 5-1, 고려대학교 정경대학 통계학과 교수
E-mail: kykim@korea.ac.kr
2) (110-180) 서울시 중구 다동 39, 한미은행 소비자금융리스크관리부 과장
E-mail: oneofn@goodbank.com

정리 1.1 (Tarpey, 1999) 크기$p \times p$ 인 행렬 $\mathbf{P}$ 를 p 차원 유클리드 공간 $\mathbf{R^p}$에서 q 차원의 선형부분공간 $\mathbf{V}$ 으로의 직교투영행렬이라고 하자. 이 때 평균이 $\mathbf{0}$인 p 차원 벡터 $\mathbf{X}$에 대해 $\mathbf{PX}$가 $\mathbf{X}$에 대해 자기일치성을 가지면, $\mathbf{V}$는 $\mathbf{X}$의 공분산행렬의 고유벡터에 의해 생성된다. □

위의 [**정리 1.1**]에서 $q = 1$인 경우 $\mathbf{e_k}(k = 1, 2, \cdots, p)$를 $\mathbf{X}$의 공분산행렬의 k 번째 고유벡터라 할 때 $\mathbf{P} = \mathbf{e_k e_k^T}$를 만족하는 $\mathbf{e_k}$가 존재하고, 자기일치곡선의 패턴은 선형성을 만족하여 자기일치선(self-consistent line)이 되며, 이 때 자기일치선은 곧 k 번째 주성분 축과 일치한다.

한편, 모든 주성분의 자기일치성이 성립하는 분포는 강대칭분포(strongly symmetric distribution)인데, 이는 모든 주성분이 서로 독립이 되는 분포로서 타원형분포(*elliptical distribution*)는 그의 특별한 경우라고 할 수 있다(Tarpey, 1995). 또한 주성분의 자기일치성이 만족되면 자기일치성의 패턴이 선형성을 가지게 되고, 자료의 분포중심이 주성분 축 위에 있게 된다. 따라서 주성분의 자기일치성이란 자료의 분포가 선형근사를 통해 왜곡없이 요약됨을 의미하므로, 주성분분석의 타당성을 평가하는 기준이 될 수도 있다. 이에 따라 주성분의 자기일치성이 만족되지 않는다는 것은 자료의 근사에 있어 비선형적 방법을 도입해야 함을 나타낸다.

2. 대표관찰치의 선택과 표현

그림 2.1에서의 각 곡선은 표준정규분포에 따르는 크기 50의 확률표본으로부터 추정한 확률밀도함수 $\hat{f}(\cdot)$ 로서, 이런 과정을 독립적으로 100회 반복하여 얻은 결과가 중첩되어 그려져 있다. 이 예는 Jones와 Rice(1992)에 의해 고려된 바 있는데, 실제 연구에서는 수평좌표의 [-4, +4] 구간을 101개의 등간격점으로 나누고, 이들 각 점에서 계산한 $\hat{f}(\cdot)$들을 원소로 하는 $n(= 100)$ 개의 $p(= 101)$ 차원 벡터들을 자료로 하고 있다. 즉, i 번째 곡선을 나타내는 벡터는 다음과 같이 주어지게 된다.

$$\mathbf{x_i^T} = (x_{i1}, \cdots, x_{i51}, \cdots, x_{i101}) = [\hat{f}_i(-4), \cdots, \hat{f}_i(0), \cdots, \hat{f}_i(4)], i = 1, 2, \cdots, n(= 100) \tag{2.1}$$

다변량 자료를 시각적으로 표현하기 위해 그림 2.1과 같은 프로화일 곡선을 생각할 수 있는데, 개체수가 많아질수록 서로 겹쳐지는 곡선들로 인해 통계적 정보도출의 유용성이 떨어진다. 이런 경우 이 프로파일 곡선들이 가지는 전반적 변이에 관한 주요 정보를 요약할 수 있는 대표관찰치(곡선)를 선택하여 이를 해석하고 표현하는 것이 하나의 해결방안이 될 수 있을 것이다.

대표관찰치의 선택이라는 문제를 고려하기에 앞서 다변량 변이의 요약이라는 관점에서 원자료 그림 2.1에 대해서 공분산행렬을 사용하여 얻은 주성분분석의 주요 결과를 기하적으로 표현한 것을 보면 다음과 같다. 그림 2.2는 처음 두 개의 주성분점수 (PRIN1, PRIN2)에 대한 산점도인데 주성분의 설명비율은 각각 45% , 32% 로서 누적설명비율은 77%가 된다. 또한 그림 2.3은 처음 두 개의 주성분에 대한 주성분적재(PC loading)의 프로파일을 나타

낸 것으로, 첫 번째 고유벡터($\mathbf{e_1}$)는 실선으로, 두 번째 고유벡터($\mathbf{e_2}$)는 점선으로 표시하였다. 우선 첫 번째 고유벡터는 프로화일의 양쪽 부분에 서로 반대 부호의 큰 적재가 있으므로, 첫 번째 주성분이 가지는 변이는 분포의 왜도(skewness)를 설명하고 있다고 해석할 수 있다. 즉, 첫 번째 주성분의 값이 클수록 오른쪽으로 치우친 곡선을 나타낸다는 의미이다. 이에 대해 두 번째 고유벡터는 양쪽 부분에 음(-)의 적재, 그리고 중심부분에 이보다 상대적으로 큰 양(+)의 적재가 있으므로, 둘째 주성분이 설명하는 변이는 분포의 첨도 정도를 나타낸다고 볼 수 있다. 따라서 두 번째 주성분가 클수록 첨도가 높은 곡선이 될 것으로 기대된다.

2.1. Jones와 Rice(1992)의 방법

Jones와 Rice의 방법은 보유한 $q(=2)$ 개의 주성분 각각에 대해서 최대값, 최소값, 사분위수 등과 같은 특정 분위수에 대응되는 원래 관찰치를 대표관찰치로 선택하는 것이다. 즉, $k(=1,2,\cdots,q)$ 번째 주성분에 대해 i 번째 관찰개체 $\mathbf{x_i}$의 주성분점수 y_{ik}가 특정 분위수에 해당된다면 $\mathbf{x_i}$를 대표관찰치로 선택하는 방법으로 $\mathbf{x_i}$와 주성분점수와 다음과 같은 관계를 가진다.

$$\mathbf{x_i} = (\mathbf{e_1},\cdots,\mathbf{e_p})(y_{i1},\cdots,y_{ik},\cdots,y_{ip})^T + \overline{x} = \sum_{k=1}^{p} y_{ik}\mathbf{e_k} + \overline{x} \tag{2.2}$$

이 관계식에 따르면 k 번째 주성분에서 대표관찰치로 선택된 $\mathbf{x_i}$ 에는 현재 고려하는 k 번째 주성분을 포함한 모든 나머지 주성분들의 효과가 내포되어 있어, 표현하고자 하는 특정 주성분에 대해 잡음으로 작용하게 됨을 알 수 있다.

아래 그림 2.4, 그림 2.5, 그림 2.6은 Jones와 Rice의 방법에 따라 첫 두 주성분 각각에 대해 대표관찰치를 선택하여 그린 것이다. 그림 2.4에서 숫자(1, 2)는 해당 주성분의 번호로서 첫 두 주성분 각각의 3개 분위수(최소값, 중위값, 최대값)의 위치를 나타낸 것이고, 주성분별 분위수에 해당하는 대표관찰치의 프로화일로 나타낸 것이 그림 2.5와 그림 2.6이다. 여기서 실선은 최대값, 점선은 중위값, 그리고 가는 점선은 최소값에 해당하는 대표관찰치를 나타낸다.

2.2. 주성분의 자기일치성을 이용한 방법

이 자료의 경우 첫 두 주성분 각각에 대해 주성분의 자기일치성 여부를 Tarpey(1999)의 적합결여검정을 이용하여 검정한 결과 모두 자기일치성을 만족하는 것으로 나타났다. 이와 같이 주어진 자료로부터 유도된 주성분들이 자기일치성을 가진다면 주성분들이 서로 독립이므로 k 번째 주성분에 대한 대표관찰치를 선택하여 표현할 때 k 번째 주성분이 가지는 변이의 내용만을 반영하도록 그의 주변분포만 고려하는 것이 타당할 것이다. 이런 점에서 여기에서 제안하는 주성분의 자기일치성을 이용한 방법은 k 번째 주성분에 대한 대표관찰치의 선택에서 k 번째 주성분을 제외한 나머지 모든 주성분점수를 균일하게 0으로 설정하는 것이다. 따라서 이 방법을 통해 선택된 관찰개체 $\mathbf{x_i^*}$ 는 주성분의 자기일치성에 의해 Jones와 Rice의 방법에 의해 선택된 대표관찰치 $\mathbf{x_i}$ 의 조건부 기대값이라는 성질을

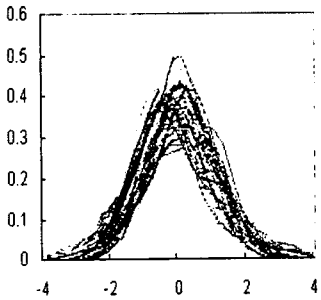

그림 2.1: 원자료

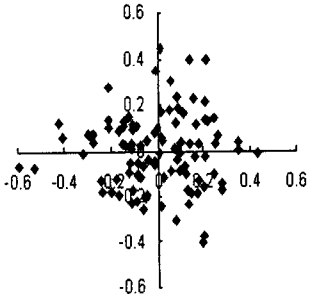

그림 2.2: 주성분점수

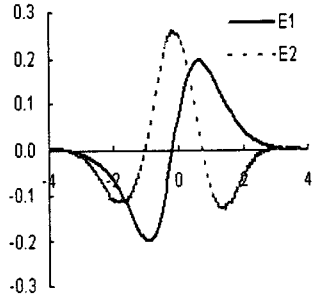

그림 2.3: 고유벡터

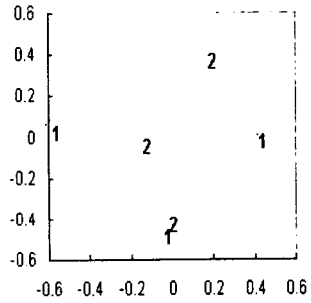

그림 2.4: 선택된 대표 관찰치

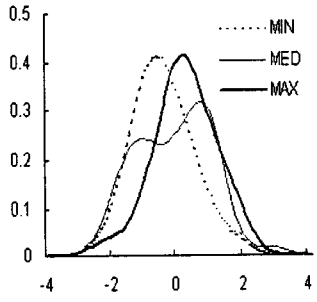

그림 2.5: PRIN1의 대표 관찰치

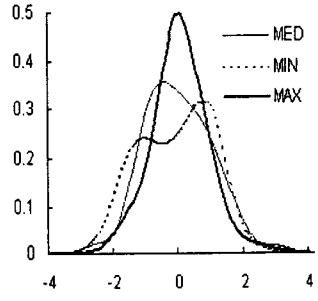

그림 2.6: PRIN2의 대표 관찰치

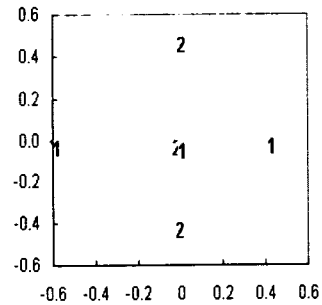

그림 2.7: 선택된 대표 관찰치

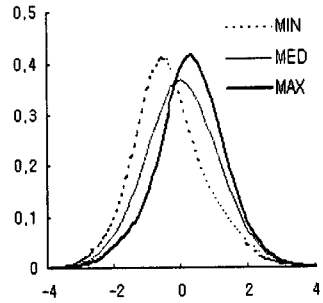

그림 2.8: PRIN1의 대표 관찰치

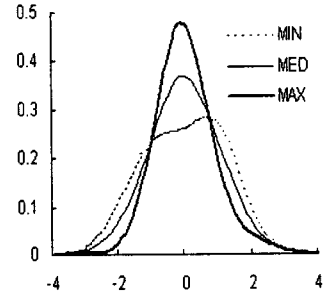

그림 2.9: PRIN2의 대표 관찰치

가진다. 다시 말해서, 관심대상이 되는 특정 분위수에 대해 두 가지 대표관찰치 $\mathbf{x_i}$와 $\mathbf{x_i^*}$는 $\mathbf{E(x_i|x_i^*)} = \mathbf{x_i^*}$ 의 관계를 가진다. 구체적으로 이 방법은 i 번째 관찰개체 $\mathbf{x_i}$ 의 k 번째 주성분점수 y_{ik} 가 특정 분위수에 해당된다면, $\mathbf{x_i^*} = y_{ik}\mathbf{e_k} + \bar{\mathbf{x}}$ 를 대표관찰치로 선택할 것을 제안하고 있다.

이 결과를 기하적으로 표현한 그림 2.7, 그림 2.8, 그림 2.9를 살펴보면 다음과 같다. 우선 그림 2.8, 그림 2.9가 이에 대응되는 그림 2.5, 그림 2.6보다 첫 두 주성분의 효과인 왜도와 첨도에 대한 정보를 더 명확히 묘사하고 있음을 볼 수 있다. 이는 그림 2.8, 그림 2.9를 생성시킨 첫 두 주성분에 대한 대표관찰치가 원자료의 주요변이를 두 개의 주성분으로 분해한 수정된 자료에서 선택된 것으로써 이 대표관찰치들은 각 주성분에서 해당 주성분점수와 평균벡터만의 합으로 구성되었기 때문이다.

3. 맺음말

다변량자료에 대한 대표관찰치를 선택함에 있어 기존의 Jones와 Rice가 제안한 방법은 각 주성분의 변이내용을 반영하는 대표관찰치에 해당 주성분은 물론 나머지 모든 주성분들의 효과가 포함됨으로써 해당 주성분에 대해 선택된 관찰치가 가지는 대표성의 역할이 모호해질 가능성이 있다. 이에 대해 주성분의 자기일치성을 이용한 방법은 주성분의 자기일치성이 만족될 때 각 주성분이 독립이 되는 성질을 이용하여 그의 대표관찰치가 해당 주성분의 변이만을 포함하도록 조정함으로써 대표성을 향상시킨다는 특징을 가진다. 또한, 각 주성분에 대한 대표관찰치를 선택할 때 해당 주성분의 주변분포만을 고려하는 것으로써 자연히 관련 문제를 단순화시키는 장점이 있다. 그러나, 만약 주성분의 자기일치성이 만족되지 않을 경우 통상적인 주성분분석과 같은 선형근사방법을 적용하는 것과 이에 근거해서 대표관찰치를 선택하는 방법은 왜곡된 결과를 초래할 수 있다.

대표관찰치를 주성분축 상에서 선택하는 것은 그 주성분축이 바로 자기일치선이기 때문임을 고려한다면, 주성분이 자기일치성을 만족하지 않는 경우 자기일치곡선(self-consistent curve)를 추정하여 관찰개체를 그 곡선으로 투영시킨 후 개체를 선택하는 방법을 생각할 수 있을 것이다.

참고문헌

Flury, B. (1990). Principal points, *Biometrika*, **77**, 33-41.

Hastie, T. and Stuetzle, W. (1989). Principal curves, *Journal of the American Statistical Association*, **84**, 502-516.

Jones, M. C. and Rice, J. A. (1992). Displaying the important feature of large collections of similar curves, *The American Statistician*, **46**, 140-145.

Tarpey, T. (1995). Principal points and self-consistent points of symmetric multivariate distributions, *Journal of Multivariate Analysis*, **53**, 39-51.

Tarpey, T. (1999). Self-consistency and principal component analysis, *Journal of the American Statistical Association*, **94**, 456-467.

Tarpey, T. and Flury, B. (1996). Self-consistency : a fundamental concept in statistics, *Statistical Science*, **11**, 229-243.

[2004년 4월 접수, 2004년 10월 채택]

A Method of Expressing Multivariate Representative Observations Based on the Self-Consistency of Principal Components

KeeYoung Kim[1] YongJu Park[2]

ABSTRACT

Representative observations are useful to express explicitly the distributional variation of the data by a few selected observations corresponding to the quantiles in the univariate situation. Jones and Rice(1992) extended it to the multidimensional case by the principal component based method. This study introduces a modified version of Jones and Rice exploiting the self-consistency of principal components in expressing the chosen observation vectors. Compared to that of Jones and Rice, the suggested method tends to provide with less susceptible representative observations to the sampling variation of the data and the resulted vectors benefits from the self-consistency.

Keywords: Principal component analysis, Self-consistency, Representative observation

1) Professor, Dept. of Statistics, Korea University. Anam-Dong 4-1, Seoul 136-701, Korea.
E-mail: kykim@korea.ac.kr
2) Manager, Dept. of Consumer Credit Risk Management, Koram Bank. Da-Dong 39, Seoul 110-180, Korea
E-mail: oneofn@goodbank.com