Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

K-평균 군집분석을 활용한 다중대응분석의 재해석

김경희[1] 최용석[2]

요 약

다원분할표에서 범주들의 대응관계를 그래프적으로 보여주는 다중대응분석(multiple correspondence analysis)은 주결여성(principal inertia)이 총결여성(total inertia)에서 차지하는 비율이 전반적으로 낮아 설명력(goodness-of-fit)이 낮은 2차원의 대응분석그림을 얻게 된다. 이를 극복하기 위해 Benzécri의 공식을 사용하면 낮은 주결여성을 높이고 새로운 2차원 대응분석그림을 얻을 수 있다. 그러나 이 새로운 대응분석그림도 범주들의 대응관계를 명확히 보여주지는 못한다(Greenacre and Blasius, 1994, chapter 10). 앤드류 플롯(Andrews plot)을 이용하여 범주들의 군집화(clustering)로 다중대응분석을 재해석 하고자 하나 범주의 수가 많은 경우 해석상 어려움이 따른다. 본 소고에서 이와 같은 경우 K-평균 군집분석을 활용하여 다중대응분석의 해석을 용이하게 하고자 한다.

주요용어 : 다중대응분석, 앤드류 플롯, K-평균 군집법

1. 서론

대응분석은 비정칙치분해(singular value decomposition)를 이용한 차원 축소와 함께 2차원의 그래프적 표현을 통해 분할표 자료의 행과 열 범주들간의 대응관계를 탐구하려는 것을 목적으로 한다(Greenacre, 1994, pp. 3-8; 최용석, 2001, 1장). 대응분석이 다변량 자료분석으로 널리 알려진 때는 1980년대이다. 1960년대 Jean-Paul Benzécri에 의해 대응분석의 기하적인 면이 발전되어 졌으며 Greenacre(1984)와 Lebart et al. (1984)에서 자세히 소개하고 있다.

대응분석은 이원분할표에 적용되는 단순대응분석(simple correspondence analysis)과 다원분할표에 적용되는 다중대응분석으로 나눌 수 있다. 단순대응분석 대개 처음 두 개의 주결여성에 대응하는 제 1축과 제 2축에 의한 2차원의 대응분석그림을 통하여 원자료의 행과 열의 대응관계를 충분히 보여줄 수 있지만 다중대응분석은 비정칙치(singular value)의 제곱인 주결여성이 총결여성에서 차지하는 비율이 낮기 때문에 2차원의 대응분석그림으로는 범주들 간의 대응관계를 정확하게 파악할 수 없는 경우가 많다.

Greenacre and Blasius(1994, chapter 10)는 Benzécri의 공식을 사용하여 주결여성의 차원도 줄이면서 주결여성의 비율을 높게 하는 수정된 주결여성으로 다중대응분석을 하고자 한다. 그러나 이 새로운 대응분석그림도 범주들의 대응관계를 명확히 보여주지 못한다. 이들은 추가적으로 앤드류 플롯을 통해 재해석하고 있다. 그러나 앤드류 플롯에서는 자료의 모든 정보를 시각적으로 표현할 수 있는 장점은 가지고 있지만 이 방법은 탐색적인 방법으로 객관적이지 못하고 범주가 많은 경우 자료의 구조를 파악하기 힘든 문제점이 있다.

이런 문제점을 해결하기 하기 위해 군집분석을 생각할 수 있다. 그 중에서도 K-평균 군집분석

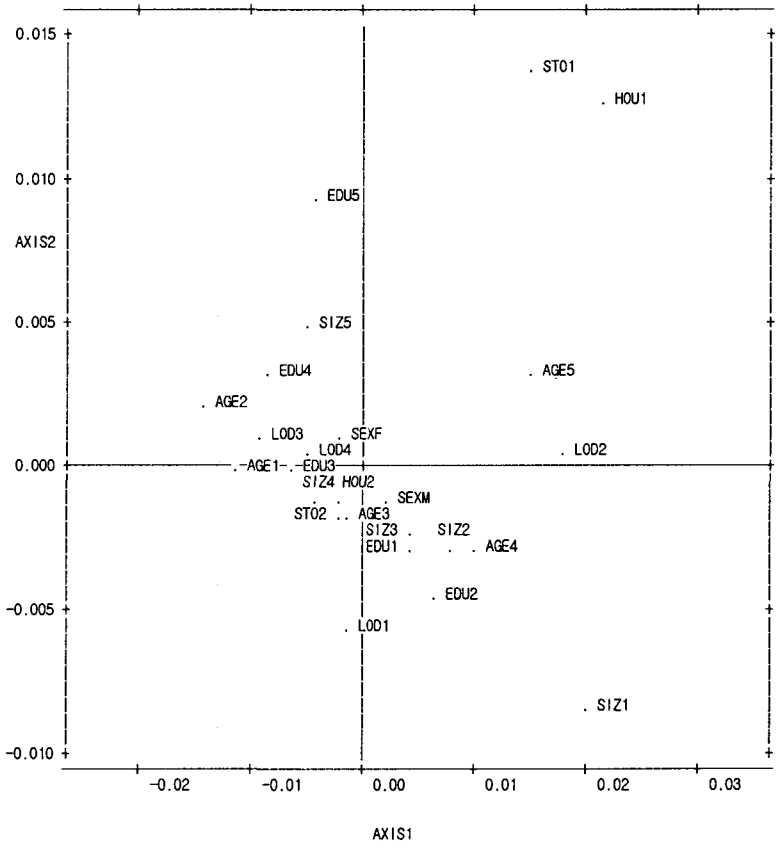
1) 부산시 금정구 장전동, 부산대학교 자연과학대학 통계학과, 석사과정
2) 부산시 금정구 장전동, 부산대학교 자연과학대학 통계학과, 부교수

은 군집분석 중 계산이 비교적 간단하고 여러 실험적 상황의 수행평가에서 상당히 좋은 결과를 내는 것으로 알려져 있다(김미경, 2000).

따라서 본 소고에서는 K-평균 군집분석을 활용하여 수정된 주결여성을 사용한 다중대응분석을 재해석 하고자 한다.

2. 군집분석을 활용한 다중대응분석의 재해석

다중대응분석을 위해 Lebart et al. (1984)의 프랑스의 사회 경제적 문제에 대한 자료를 한 예로 생각해보자. SAS/PROC CORRESP에서 채택하고 있는 Greenacre(1984, p. 141)의 다중대응분석 알고리즘을 이용하면 2차원을 구성하는 처음 두 개 주결여성이 각각 9.72%와 8.49%로 이들은 총 18.21%로 낮아 Benzécri의 공식을 사용하면 수정된 주결여성의 비율이 각각 55.65%와 27.63%로 높일 수 있다. 이 때 수정된 주결여성에 대응하는 제 1축과 제 2축에 의한 2차원의 대응분석그림 <그림1>을 보면 2차원 보다는 더 높은 차원에서 해석되어야 할 LOD4, AGE1 등의 범주들이 있다. 이런 범주들의 대응관계를 자세히 알아보기 위해 Greenacre and Blasius (1994, chapter 10)는 앤드류 플롯을 이용하고 있다. 그들은 해석하기 복잡한 <그림 2>를 몇 개의 범주로 나누어 그들에 대한 여러 개의 앤드류 플롯의 비교를 통해 크게 (1) STO1, HOU1



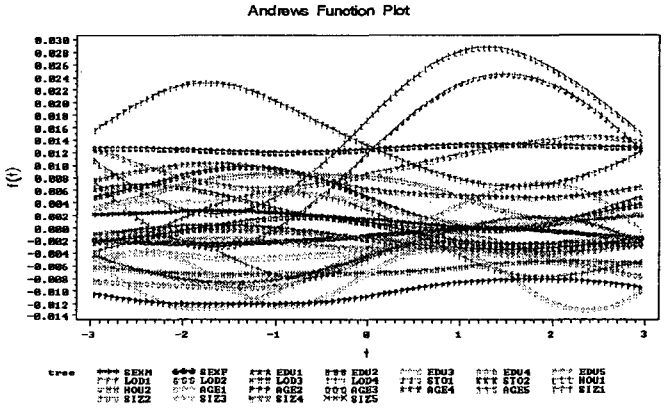
<그림 1> 제 1축과 제 2축의 대응분석그림

(2) AGE2, LOD3 (3) AGE1, LOD4 (4) SEXM, SEXF, STO2, HOU2 (5) EDU5, SIZ5 5개의 군집으로 묶어 주고 있다. 이와 같이 여러 개의 앤드류 플롯을 비교해야 하는 번거로움이 많다.

특히, Sharama(1996)는 계층적 군집분석(hierarchical clustering)에서 결정한 군집의 수를 K-평균 군집분석에서의 K개 초기 군집의 수로 두고 초기 시드점으로는 계층적 군집법의 군집 중심을 사용하고 있다.

이 방법을 따라 본 예제에서는 계층적 군집법에 의한 결과, 군집의 수를 4개로 결정하여 K-평균 군집을 하게 되면 <표 1>과 같은 결과를 얻을 수 있다. 이 결과 군집 1에서는 교육수준이 높을수록 (EDU3-EDU5) 나이가 젊고(AGE1-AGE3) 거주지역 규모도 크다(SIZ4-SIZ5). 하지만 주식과 동산은 소유하지 않고 있음을 알 수 있다.

이와 같이 다중대응분석에서 K-평균 군집법을 활용하면 주결여성의 비율을 높였음에도 불구하고 해석되지 못했던 범주들 간의 대응관계를 잘 파악할 수 있다.



<그림 2> 앤드류 플롯

군집	대 응 범 주
군집1	SEXF, EDU3, EDU4, EDU5, LOD1, LOD3, LOD4 STO2, HOU2, AGE1, AGE2, AGE3, SIZ4, SIZ5
군집2	SEXM, EDU1, EDU2, AGE4, SIZ2, SIZ3
군집3	LOD2, AGE5, SIZ1
군집4	STO1, HOU1

<표 1> centroid 군집중심을 이용한 K-평균 군집법

3. 결론

지금까지 다중대응분석에서 Benzécri의 공식을 활용하여 낮은 주결여성을 높이는 수정된 다중대응분석을 소개하고 범주들의 대응관계를 앤드류 플롯을 활용한 Greenacre and Blasius(1994, chapter 10)의 방법을 소개하고 그것의 문제점을 지적하였다. 본 소고에서는 다중대응분석에서 K-평균 군집분석을 활용하여 범주들 간의 대응관계를 보다 더 쉽게 해석하였다.

참고문헌

[1] 최용석 (2001). <SAS 대응분석의 이해와 응용>, 자유아카데미, 서울.
[2] 허명회 (1999). <다변량 수량화>, 자유아카데미, 서울.
[3] 김미경 (2000). 저차원 K-평균 군집화, 고려대학교 박사학위논문.
[4] Greenacre, M. J. (1984). *Theory and Application of Correspondence Analysis*, Academic Press, London.
[5] Greenacre, M., and Blasius, J. (1994). *Correspondence Analysis in the Social Sciences*, Academic Press, London.
[6] Lebart, L., Morineau, A., and Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*, Wiley, New York.
[7] SAS Institute Inc. (1999). *Applied Multivariate Statistics with SAS Software*, Second Edition, SAS Institute Inc., NC.
[8] Sharma, S. (1996). *Applied Multivariate Techniques*, Wiley, New York.