creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

조 영 일 교수 지도

석사학위 청구논문

몬테카를로 시뮬레이션을 이용한 K-평균 군집분석과 잠재계층분석의 비교

: 군집회복(cluster recovery)을 중심으로

2017

성신여자대학교 대학원

심리학과

구 소 희

몬테카를로 시뮬레이션을 이용한 K-평균 군집분석과 잠재계층분석의 비교

: 군집회복(cluster recovery)을 중심으로

조 영 일 교수 지도

이 논문을 석사학위 논문으로 제출함

2017년 5월

성신여자대학교 대학원
심리학과
구 소 희

인 준 서

구소희의 석사학위 논문으로 인준함.

2017년 5월

심사위원장 _____(인)

심 사 위 원 _____ (인)

심 사 위 원 _____ (인)

성신여자대학교 대학원

논문개요

본 논문은 K-평균 군집분석과 잠재계층분석의 군집회복 정도를 비교해보는 시뮬레이션 연구이다. 이를 위하여 총 4개의 시뮬레이션 연구가 사용되었다. 4개의 연구에서는 다섯 개의 조작변수(즉, 사례수, 문항수, 군집수, 효과크기, & 잠재집단확률)가 사용되었다. 조작변수의 수준에 따라 3(N)×3(V)×3(K)×3(F)×3(γ)의 요인설계(factorial design)를 바탕으로 243개의 실험 조건을 설정하고 각 조건을 500번씩 반복하였다. 자료의 생성과 분석에는 통계프로그램 R이 사용되었다.

잠재계층분석의 분석 결과, 문항수 수준, 효과크기 수준, 잠재집단확률 수준에 따른 군집회복 정도에는 유의한 차이가 있었다. K-평균 군집분석의 결과, 군집수 수준, 효과크기 수준, 잠재집단확률 수준에 따른 군집회복 정도에는 유의한 차이가 있었다. 분석방법과 군집수 수준, 분석방법과 효과크기 수준, 그리고 분석방법과 잠재집단확률 수준이 군집회복에 미치는 상호작용효과는 통계적으로 유의하였다. 전반적으로 잠재계층분석이 K-평균 군집분석보다 높은 군집회복 지수를 보였다.

또한 추정의 효율성을 검증하기 위하여 잠재계층분석과 K-평균 군집분석을 실시하는 과정에서 지정한 반복(iteration) 횟수 이후에도 수렴하지 못하는 경우에 대해 살펴보았다. K-평균 군집분석에서는 NC(Not Converged)가 한 건도 발생하지 않은 반면에 잠재계층분석에서는 총 121,500의 자료 중에 약 0.1%인 141건의 NA가 발생하였다. 따라서 연구자에게 주어진 데이터의 형태와 조건에 따라 잠재계층분석과 K-평균 군집분석이 각각 가질 수 있는 장점과 단점이 있으므로, 상이한 연구문제에 따라 적합한 분석방법을 사용할 것에 대해 논의하였다. 마지막으로 후속연구를 위한 제언을 논의하였다.

주요어: 군집회복, ARI, 잠재계층분석, K-평균 군집분석, 시뮬레이션, 통계 프로그램 R

목 차

논문개요

Ⅰ. 서론 ····· 1
1. 연구의 필요성 및 목적 ····· 1

Ⅱ. 이론적 배경 ····· 10
1. 군집분석(Cluster Analysis) ····· 10
1) 군집화 방법 ····· 12
2) 군집의 종류 ····· 16
2. K-평균 군집분석(K-means Clustering) ····· 20
1) K-평균 군집분석의 알고리즘 ····· 21
2) K-평균 군집분석의 쟁점 ····· 26
3. 잠재계층분석(Latent Class Analysis) ····· 31
1) 잠재계층모형의 기본 ····· 33
2) 모수추정 ····· 39
3) 잠재계층분석의 쟁점 ····· 42

Ⅲ. 연구문제 및 가설 ····· 45
1. 연구문제1 ····· 45
2. 연구문제2 ····· 45
3. 연구문제3 ····· 46
4. 연구문제4 ····· 46

Ⅳ. 연구방법 ····· 47
1. 자료 생성(Data generation) ····· 47
1) 몬테카를로 시뮬레이션(Monte Carlo simulation) ····· 48
2) 군집회복지수(Adjusted Rand Index) ····· 50
2. 자료 분석 ····· 55

Ⅴ. 연구결과 ····· 56
1. 연구1: 잠재계층분석의 군집회복 정도 ····· 56
1) 사례수에 따른 군집회복 정도의 차이 ····· 56
2) 문항수에 따른 군집회복 정도의 차이 ····· 57
3) 군집수에 따른 군집회복 정도의 차이 ····· 59
4) 효과크기에 따른 군집회복 정도의 차이 ····· 60
5) 잠재집단확률에 따른 군집회복 정도의 차이 ····· 61
2. 연구2: K-평균군집분석의 군집회복 정도 ····· 63
1) 사례수에 따른 군집회복 정도의 차이 ····· 63
2) 문항수에 따른 군집회복 정도의 차이 ····· 64
3) 군집수에 따른 군집회복 정도의 차이 ····· 65
4) 효과크기에 따른 군집회복 정도의 차이 ····· 67
5) 잠재집단확률에 따른 군집회복 정도의 차이 ····· 68
3. 연구3: 각 조작변수와 각 분석방법의 상호작용 효과 ····· 71
1) 사례수와 분석방법의 상호작용 효과 ····· 71
2) 문항수와 분석방법의 상호작용 효과 ····· 76
3) 군집수와 분석방법의 상호작용 효과 ····· 82
4) 효과크기와 분석방법의 상호작용 효과 ····· 86
5) 잠재집단확률과 분석방법의 상호작용 효과 ····· 92

4. 연구4: 분석방법에 따른 군집회복 정도의 차이 ····· 97

Ⅵ. 논의 ····· 98

참고문헌

ABSTRACT

부록

표 목 차

<표 1> P와 Q 분할의 공통부분 (Steinley, 2004)

<표 2> 네 가지 다른 쌍(pairs) (Steinley, 2004)

<표 3-1> 사례수 수준에 따른 잠재계층분석 군집회복의 기술통계

<표 3-2> 사례수 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

<표 3-3> 문항수 수준에 따른 잠재계층분석 군집회복의 기술통계

<표 3-4> 문항수 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

<표 3-5> 문항수 수준에 따른 잠재계층분석 사후분석 결과

<표 3-6> 군집수 수준에 따른 잠재계층분석 군집회복의 기술통계

<표 3-7> 군집수 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

<표 3-8> 효과크기 수준에 따른 잠재계층분석 군집회복의 기술통계

<표 3-9> 효과크기 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

<표 3-10> 효과크기 수준에 따른 잠재계층분석 사후분석 결과

<표 3-11> 잠재집단확률 수준에 따른 잠재계층분석 군집회복의 기술통계

<표 3-12> 잠재집단확률 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

<표 3-13> 잠재집단확률 수준에 따른 잠재계층분석 사후분석 결과

<표 4-1> 사례수 수준에 따른 K-평균 군집분석 군집회복의 기술통계

<표 4-2> 사례수 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

<표 4-3> 문항수 수준에 따른 K-평균 군집분석 군집회복의 기술통계

<표 4-4> 문항수 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

<표 4-5> 군집수 수준에 따른 K-평균 군집분석 군집회복의 기술통계

<표 4-6> 군집수 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

<표 4-7> 군집수 수준에 따른 K-평균 군집분석 사후분석 결과

<표 4-8> 효과크기 수준에 따른 K-평균 군집분석 군집회복의 기술통계

<표 4-9> 효과크기 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

<표 4-10> 효과크기 수준에 따른 K-평균 군집분석 사후분석 결과

<표 4-11> 잠재집단확률 수준에 따른 K-평균 군집분석군집회복의 기술통계

<표 4-12> 잠재집단확률 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

<표 4-13> 잠재집단확률 수준에 따른 K-평균 군집분석 사후분석 결과

<표 5-1-1> 사례수 수준에 따른 각 분석방법 군집회복의 기술통계

<표 5-1-2> 사례수와 분석방법이 군집회복에 미치는 효과에 대한 이원분산분석 결과

<표 5-1-3> 사례수의 주효과 검증 결과

<표 5-1-4> 사례수 수준에 따른 군집회복 정도의 일원분산분석 결과

<표 5-1-5> 사례수 수준에 따른 K-평균 군집분석 군집회복의 기술통계

<표 5-1-6> 사례수 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

<표 5-1-7> 사례수 수준에 따른 잠재계층분석 군집회복의 기술통계

<표 5-1-8> 사례수 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

<표 5-2-1> 문항수 수준에 따른 각 분석방법 군집회복의 기술통계

<표 5-2-2> 문항수와 분석방법이 군집회복에 미치는 효과에 대한 이원분산분석 결과

<표 5-2-3> 문항수의 주효과 검증 결과

<표 5-2-4> 문항수 수준에 따른 군집회복 정도의 일원분산분석 결과

<표 5-2-5> 문항수 수준에 따른 K-평균 군집분석 군집회복의 기술통계

<표 5-2-6> 문항수 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

<표 5-2-7> 문항수 수준에 따른 잠재계층분석 군집회복의 기술통계

<표 5-2-8> 문항수 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

<표 5-2-9> 문항수 수준에 따른 잠재계층분석 사후분석 결과

<표 5-3-1> 군집수 수준에 따른 각 분석방법 군집회복의 기술통계

<표 5-3-2> 군집수와 분석방법이 군집회복에 미치는 효과에 대한 이원분산분석 결과

<표 5-3-3> 군집수 수준에 따른 K-평균 군집분석 군집회복의 기술통계

<표 5-3-4> 군집수 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

<표 5-3-5> 군집수 수준에 따른 K-평균 군집분석 사후분석 결과

<표 5-3-6> 군집수 수준에 따른 잠재계층분석 군집회복의 기술통계

<표 5-3-7> 군집수 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

<표 5-4-1> 효과크기 수준에 따른 각 분석방법 군집회복의 기술통계

<표 5-4-2> 효과크기와 분석방법이 군집회복에 미치는 효과에 대한 이원분

산분석 결과

<표 5-4-3> 효과크기 수준에 따른 K-평균 군집분석 군집회복의 기술통계

<표 5-4-4> 효과크기 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

<표 5-4-5> 효과크기 수준에 따른 K-평균 군집분석 사후분석 결과

<표 5-4-6> 효과크기 수준에 따른 잠재계층분석 군집회복의 기술통계

<표 5-4-7> 효과크기 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

<표 5-4-8> 효과크기 수준에 따른 잠재계층분석 사후분석 결과

<표 5-5-1> 잠재집단확률 수준에 따른 각 분석방법 군집회복의 기술통계

<표 5-5-2> 잠재집단확률과 분석방법이 군집회복에 미치는 효과에 대한 이원분산분석 결과

<표 5-5-3> 잠재집단확률 수준에 따른 K-평균 군집분석 군집회복의 기술통계

<표 5-5-4> 잠재집단확률 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

<표 5-5-5> 잠재집단확률 수준에 따른 K-평균 군집분석 사후분석 결과

<표 5-5-6> 잠재집단확률 수준에 따른 잠재계층분석 군집회복의 기술통계

<표 5-5-7> 잠재집단확률 수준에 따른 잠재계층분석 군집회복 정도의 일원

분산분석 결과

<표 5-5-8> 잠재집단확률 수준에 따른 잠재계층분석 사후분석 결과

<표 6> K-평균 군집분석과 잠재계층분석 간에 군집회복 정도의 차이

그 림 목 차

<그림 1> 병합적 군집화와 분할적 군집화 ····· 14

<그림 2> 군집의 종류 ····· 19

<그림 3> 그래프기반 군집의 예 ····· 19

<그림 4> K-평균 군집분석의 알고리즘을 사용한 예 ····· 23

<그림 5> 전역최소해와 지역최소해 ····· 27

<그림 6> 잠재계층모형 ····· 31

<그림 7> 확률모형 ····· 34

<그림 8> 지역독립성을 가정한 모형과 그렇지 않은 모형 ····· 36

<그림 9> EM 알고리즘 ····· 41

<그림 10> 몬테카를로 시뮬레이션 ····· 47

<그림 11> 분석방법에 따라 사례수가 군집회복 정도에 미치는 영향 ····· 76

<그림 12> 분석방법에 따라 문항수가 군집회복 정도에 미치는 영향 ·····81

<그림 13> 분석방법에 따라 군집수가 군집회복 정도에 미치는 영향 ·····86

<그림 14> 분석방법에 따라 효과크기가 군집회복 정도에 미치는 영향 ·····91

<그림 15> 분석방법에 따라 잠재집단확률이 군집회복 정도에 미치는 영향 ·
96

I. 서론

1. 연구의 필요성 및 목적

심리학은 인간의 마음과 행동에 대하여 연구하는 학문으로, 인간의 다양성을 이해하며 설명하고자 한다. 다양성(Diversity)이란 구성원들 간의 다름을 의미하며, 이들 간의 다름 또는 개인차는 변산성(Variation)으로 요약된다. 구성원들 간의 개인차를 이해하고 설명하기 위한 방법은 크게 변수중심 접근방법(Variable-centered approach)과 사람중심 접근방법(Person-centered approach)[1]으로 나뉜다(Meyer & Morin, 2016). 변수중심 접근방법은 표본에 속해있는 모든 객체[2]가 단일 모집단에서 표집되었다고 가정하는 반면에, 사람중심 접근방법은 표본의 객체가 각각 다른 모집단에서 표집 되어 객체들 끼리 여러 개의 독립적인 소집단으로 묶여 질 수 있다고 가정한다(Meyer & Morin, 2016). 변수중심 접근방법에서는 유사한 변수끼리 묶이는 것을 수렴타당도(convergent validity)[3], 상이한 변수끼리 나누어 구분하는 것을 변별타당도(discriminant validity)[4]의 개념에서

1) 사람중심 접근방법(Person approach) 이라는 용어는 Block(1971)이 사용했고 이후 전체론적-상호작용적 관점의 틀에서 Magnusson 과 Allen(1983) 과 Magnusson (1985)에 의해 논의 되었다(Magnusson, 2003). 이러한 관점은 개인을 전체로서 기능하고 발전하며 조직된 전체로서 간주하는 관점이다(Bergman & Magnusson, 1997).

2) 공상이 아닌 실세계(real world)에 존재하거나 생각할 수 있는 것을 일컫는 객체(Object)란 수학적, 과학적 이론과 경험적 관계를 통해 어떠한 현상을 다양한 방법으로 개념화하고 수식화 하는 과정에서 보면 명확한 의미를 담고 있는 대상 또는 개념이다. 또는 프로그래밍 관점에서 볼 때 객체는 데이터에서 생성된 변수이며 소프트웨어를 구성하는 한 요소를 의미한다. 즉 실세계에 존재하는 대상을 가상공간인 컴퓨터 세계에 적용한 것이다(김치수, 2015).

3) 수렴타당도(convergent validity)는 타당도를 검증하는 방법 중 하나로 동일한 특성을 상이한 방법으로 측정해서 얻은 점수를 비교하여 이들 간의 상관이 높은 경우 측정이 타당하다는 수렴의 근거가 된다(교육평가 용어사전, 1998).

4) 변별타당도(discriminant validity)는 타당도를 검증하는 방법 중 하나로 상이한 특성을 동일한 방법으로 측정해서 얻은 점수를 비교하여 이들 간의 상관이 낮은 경우 측정이 타당하다는 변별의 근거가 된

설명하게 된다. 반면에 사람중심 접근방법은 특정 심리적 속성에서의 유사한 수준을 보이는 사람끼리 묶는 것이 수렴타당도의 개념이 되며 상이한 수준을 보이는 사람끼리 구분하는 것이 변별타당도의 개념이 되는 것이다.

변수중심 접근방법은 집단의 구성원[5]들이 모두 하나의 동질적인 모집단을 구성하며, 변수 자체가 분석단위가 되고 연구 주제 및 문제에 대한 주요개념이 되는 방식이다(Magnusson, 2003). 즉 단일 모집단의 모수를 추정하기 위해 동질적인 모집단에서 구성원들이 표집 되었다는 가정 아래에서, 모집단에 동일하게 적용될 수 있는 예측변수와 결과변수 간의 관계를 설명하는 방식으로 연구문제에 대한 가설을 검증한다. 이에 따른 통계기법으로는 상관분석, 회귀분석, 분산분석, 요인분석, 경로분석 등이 해당되며, 변수중심 접근방법에 기초한 분석 모델은 예측변수가 결과변수의 분산을 설명할 때 예측변수의 상대적 기여도, 제 3변수의 매개 또는 조절 등의 관계를 추정하는 연구주제에 빈번하게 활용된다(Laursen & Hoff, 2006).

반면에 사람중심 접근방법은 집단의 구성원들이 이질적인 모집단에서 표집 되었다는 가정에 기초한다. 이에 따라서 변수 간의 설정된 관계성이 개인 간에 다름을 가정하고 이를 설명하고자 한다. 즉, 사람중심 접근방법은 한 개인을 다른 개인과 비교했을 때 비교의 대상은 하나의 개별 속성이 아닌 개인의 전체적인 속성의 집합이 된다. 그러므로 사람중심 접근방법은 유사하거나 상이한 속성, 또는 하나의 속성뿐만 아니라 다양한 속성들 간의 관계의 형태에 따라서 특정 속성을 공유하는 소집단을 식별하는 방식이다(Magnusson, 2003). 따라서 사람중심 접근방법은 변수중심 접근방법과 달리 변수간의 관련성이 주요한 연구문제이기 보다는, 이 변수들의 수준에 따라서 개인들이 어떻게 군집화 되는지를 분석하고 변수들의

다. 즉 수렴타당도와 대비되는 개념이다(교육평가 용어사전, 1998).

5) 객체의 속성과 기능을 가지고 표집(sampling)이라는 과정을 거쳐 집단에 속하게 된 하나의 객체를 집단의 구성원 이라고 표기했다. 이후 본 논문에서 객체와 구성원을 동의어로 이해해도 무방하다.

속성이 어떻게 각 개인별로 상이하게 조합되어 있는지를 규명한다(Morin, Morizot, Boudrias, & Madore, 2010). 이후 그 집단이 미치는 영향 또는 이러한 집단 유형의 원인이 되는 요인들에 대해 연구할 수 있도록 한다(김도환, 황상민, & 최윤식, 2011; 양경욱 외, 2016). 결과적으로 동일한 집단 내에 속한 구성원들이 공유하고 있는 속성의 수준 및 여러 속성들 간의 관계의 형태를 추정한다. 사람중심 접근방법에 기초한 통계분석 기법으로는 잠재프로파일 분석(latent profile analysis), 잠재계층 분석(latent class analysis), 군집 분석(cluster analysis), 잠재성장혼합모형(latent growth mixture model) 등이 해당되며, 이들은 변수 사이의 관계의 형태의 개인 간 차이에 관한 연구 문제에 적합하다(Laursen & Hoff, 2006).

인간 행동에 대한 원인이나 인간 발달에 영향을 미치는 요인을 찾아내고 이러한 요인들 간의 중요도를 판단하는 데 있어 변수중심 접근방법은 큰 공헌을 하고 있다(Meyer, Stanley, & Vandenberg, 2013). 다만 여러 가지 요인의 상호작용으로 인해 다양성을 보이는 인간을 연구하는 심리학에 있어서, 변수중심 접근방법은 하나의 변수가 다른 변수에 의해 설명되는 분산에 대한 정보를 제공하기 때문에 변수들 사이에서 작용하는 상호작용을 식별하기에는 제한적일 수 있다(Meyer et al., 2013). 따라서 인간의 속성이 표출되고 작용하는 다양한 방식을 가정한다면, 개인의 특성을 판단하기 위하여 하나의 요인만을 고려하기보다는, 개인 내의 다양한 요인들 간의 상호작용에 대한 고려를 통해서 이러한 특성들의 공통성(common characteristics)이 다르게 반영되는 하위집단을 파악하는 연구(Wang & Hanges, 2011)가 반드시 필요하다.

사람중심 접근방법은 심리학에서 인간의 행동과 마음에 대해 일반적인 규칙을 제공해주기보다 속성이 비슷한 사람끼리 그룹화 시켜 분석대상을 세분화하는 접근방법에 기초하여 개인차에 대한 다양한 시각을 제공해 줄 수

있다. 인간의 태도나 행동이 하나의 변수에 의해서만 움직이지 않는 현실을 반영하기 위하여(양경욱, 윤세준, & 채연주, 2016) 사람중심 접근방법과 변수중심 접근방법과의 상호보완적 사용을 통해 인간의 마음과 행동에 대한 보다 깊은 이해가 가능하게 된다(이목화, 문형구, & 권석균, 2016; Morin et al., 2010; Meyer et al., 2013). 또한 개인을 전체로서 기능하고 발전하며 조직된 전체로서 간주하는 전체론적(holistic) 이며 상호작용적(interactionistic)인 관점에 대한 수용이 증가하고 있는 오늘날, 사람중심 접근방법의 필요성이 더욱 증대될 것으로 보인다(Bergman & Magnusson, 1997).

예를 들어, 2016년 「한국심리학회지:상담 및 심리치료」에서 출판된 총 48개의 논문 중에서 3개의 논문이 사람중심 접근방법을 이용하였다. 특히, 2016년도 논문의 94%가 변수중심 접근방법을 활용하여 연구를 진행하였으며 사람중심 접근방법을 이용한 논문은 6%를 차지하였다. 이처럼 변수중심 접근방법은 통계분석과 거의 유사어처럼 사용되어 왔고 심리학 연구에서 지배적인 위치에 있어왔다(Laursen & Hoff, 2006). 구체적으로, 2016년「한국심리학회지:상담 및 심리치료」에서 출판된 총 48개의 논문 중에서 3개의 사람중심 접근방법을 사용한 논문을 살펴보면 아래와 같다. 박은선, 손은영, 임지숙(2016)은 맞벌이 남성의 일, 가정 양립 경험에 대한 개념도를 연구하기 위하여 위계적(계층적) 군집분석을 실시하여 총 6개의 군집을 산출하였다. 황선정, 조성호(2016)도 위계적(계층적) 군집분석을 실시하여 병리적 자기애의 하위유형을 구분하고 각 유형별 인지 위계 특성을 살펴보았다. 또한 이홍숙, 최한나(2016)는 부부상담의 치료적 개입요인에 대한 상담자의 인식에 대해 연구하기 위해 위계적(계층적) 군집분석을 실시하여 부부 상담의 치료적 개입 요인에 대한 개념도를 도출하였다. 3개의 논문 모두 위계적 군집분석의 한 가지 방법인 Ward 방식을 사용하였고 이홍숙,

최한나(2016)는 추가적으로 중심연결법과 평균연결법을 함께 실시하였다.

변수중심 접근방법에 비해 사람중심 접근방법이 활성화 되지 못한 이유로 크게 세 가지가 논의되었다(Laursen & Hoff, 2006). 첫째, 사람중심 접근방법을 사용한 분석결과를 해석하기 위한 지침이 부족하다 생각하는 인식이 있을 수 있다는 점이다. 둘째, 일부 학자들은 분산분석에서의 집단 간 요인 차이나 회귀분석에서의 상호작용효과가 이미 개인 차이를 설명하고 있기 때문에 사람중심 접근방법이 필요하지 않다고 생각하는 경우가 있다. 셋째, 사람중심 접근방법을 질적 분석과 동일시하는 경우가 있다. 물론 사람중심 접근방법은 여러 변수의 군집화를 기반으로 하여 질적으로 다른 집단을 식별하도록 설계되긴 했지만, 이러한 변수의 측정과 집단의 도출은 양적자료에 기초한 분석방법이다. 즉 사람중심 접근방법을 사용한 분석방법은 핵심적으로 유사한 특성을 공유하는 의미 있는 군집을 식별하거나 확인하기 위해 질적 분석을 통해 얻은 질적 자료가 아닌 양적자료에 통계기법을 적용하는 것을 포함한다. 하지만 인간의 심리적 특성의 발달 또는 현상을 연구하는 심리학 분야에서만큼은 현실적인 인간행동의 다양성을 반영하기 위하여 변수중심 접근방법 뿐 아니라 사람중심 접근방법의 관심도 필요하다(Laursen & Hoff, 2006).

하지만, 자료에 변수를 조합하는 시스템의 유연성(flexibility)은 사람중심 접근방법의 장점이자 단점일 수 있다. 가령 구성원의 소속집단은 분석에 사용되는 변수의 구성에 따라 달라질 수 있기 때문이다. 어떤 변수들을 사용하여 구성원들을 분류하는 지에 따라 구성원들의 분류화가 달라질 수 있다는 것이다. 즉 사람중심 접근방법에서는 변수의 구성에 따라 군집 구성(class-membership)의 추정이 불안정(unstable) 하게 추정될 수 있음을 말한다(Griffiths, Borkan, Byme, Crabtree, Dowrick, Gunn, & Reis, 2010; Steinley, 2006).

사람중심 접근방법의 주된 목표는 여러 집단의 구성원들에 대해서 수집된 경험 자료들이 혼재되어 있으며 이러한 경험 자료들을 생성한 구성원들의 소속집단을 모르는 경우에 변수들의 조합과 관련하여 유사한 속성을 갖는 객체들의 군집을 찾아내고, 각 군집끼리 차별화된 형태를 보여주는 잠재적인 하위집단을 식별하는 것으로(이정진, 2011) 인간의 다양성, 또는 이질성을 중요시한다. 수십 년 동안 이러한 목표를 달성하기 위한 많은 분석 방법이 개발되었다. 지금까지 대부분의 절차는 크게 두 가지로 나누어지는데 이로는 기존에 많이 사용되고 있는 경험적인(heuristic)[6] 군집분석과 유한혼합모형(Finite mixture model)[7] 군집분석이 있다. 경험적인 군집분석은 다변량 공간에서 객체의 유사성을 판단하기 위해서 거리측도(distance measure)를 사용하고 혼합모형 군집분석은 통계적 모델(statistical models)을 기반으로 한다(Steinly & Brusco, 2011).

본 연구에서는 관측 대상 간의 유사성 및 집단 내의 분산의 개념을 기반으로 한 경험적 군집분석 중 가장 대표적으로 사용되는 K-평균 군집분석(Kogan, 2007; Steinley, 2006)과 확률을 기반으로 한 통계적 모형을 사용하여 K-평균 군집분석을 확장하고 일반화 시킨(이정진, 2011) 유한혼합모형인 잠재집단분석[8] (Bauer & Shanahan, 2007) 중 잠재계층분석(Latent Class Analysis)을 비교하여 살펴보고자 한다. K-평균

6) Heuristic(경험적, 체험적, 발견적)이라고 부르는 이유는 분석방법이 확률에 근거하지 않고 시행착오적인 방법을 사용하여 연구하기 때문이다. 다시 말해 논리적인 최적의 알고리즘이 확립되지 않았을 때 가장 적당하다고 생각되는 방법으로 도형인식 등의 기능을 사용하여 최적을 답을 구하는 방법이다(Kozak, Scaman, 2008).

7) 유한혼합모형(Finite mixture model)은 관찰된 데이터에서 잠재된 군집을 찾아 주는 모형으로 전체 모집단 내의 하위집단의 존재를 나타내는 확률적 모형이라고 할 수 있다. 관찰할 수 없는 잠재변수를 분석해야 하는 상황에서 큰 도움이 되며 데이터를 군집화 하는데 용이하게 쓰일 수 있다. 단 혼합모형의 특유한 점이라면 범주형의 잠재변수를 분석한다는 점이다. 관찰변수가 연속형이냐 범주형이냐에 따라 혼합모형은 다른 이름을 가지게 된다(Oberski, 2016).

8) 잠재집단분석은 관찰된 자료에서 잠재되어 있는 군집을 찾아내는데 중점을 두는 분석방법이다. 군집분석과 흡사하지만 자료에 근거한 경험적인 모형을 기반으로 사용하기 때문에 더욱 융통성이 있다는 장점을 가지고 있다. 이는 측정변수의 형태에 따라 두 가지로 나뉘는데 측정변수가 연속형 변수인 경우 잠재 프로파일 분석을, 그리고 측정변수가 범주형 자료인 경우는 잠재계층 분석을 진행하도록 한다(Oberski, 2016).

군집분석과 잠재계층분석의 알고리즘을 기술하고 이들의 효율성을 비교 대조하여 어떠한 조건에서 어떤 분석방법이 더 효율적인지 알아보는 것이 본 연구의 목적이다.

식(1.1)은 정규분포를 가정하는 잠재계층모형의 함수이다.

$$f(x;\alpha,\mu_k,\Sigma_k)=\sum_{k=1}^{K}\alpha_k\frac{\exp-\frac{1}{2}(x_i-\mu_k)'\Sigma_k^{-1}(x_i-\mu_k)}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \quad \text{식 (1.1)}$$

μ_k는 p-차원에서의 평균 벡터를 의미하며 Σ_k는 k번째 군집의 $p\times p$ 공분산 행렬을 의미한다. α_k는 비율을 의미하고 모든 k에서 $\alpha_k \geq 0$이며 총 합은 1이 된다($\sum_{k=1}^{K}\alpha_k=1$). 그리고 식 (1.2)는 K-평균 군집분석에서 최소화 하고자 하는 함수이다.

$$\sum_{k=1}^{K}\sum_{i\in C_k}(x_i-\mu_k)'(x_i-\mu_k) \quad \text{식 (1.2)}$$

식 (1.1)과 식 (1.2)을 비교해보면 유사한 부분과 상이한 부분을 알 수 있다. 식 (1.1)의 분모는 상수이기 때문에 무시할 수 있으며 식 (1.2)의 시그마 부분 $\sum_{i\in C_k}$은 식 (1.1)의 α_k 부분과 관련이 있다. 다만 다른 부분이 있다면 잠재계층분석에서는 군집구성이 확률적이지만 K-평균 군집분석에서는 군집구성이 이분법적이라는 것이다. 또한 식 (1.1)의 분자에서 공분산 행렬이 삭제되면 상수 -1/2가 무시될 수 있으며, 이에 따라 분자가 식 (1.2)의 내적(inner product)와 같아진다. 그러므로 K-평균 군집분석의 목적함수(objective function)와 잠

재계층모형의 목적함수는 일치하게 된다(Steinley & Brusco, 2011).

이렇게 K-평균 군집분석과 잠재계층모형은 수학적으로 유사한 형태를 하고 있어 이론적으로 비슷한 특징을 가지고 있다. 하지만 선행연구 조사 결과, K-평균 군집분석과 잠재계층모형이 수학적으로 유사하긴 하지만 다른 결과를 준다는 경험적인 연구결과들이 있다. Eshghi, Haughton, Legrand, Skaletsky & Woolford (2011)는 잠재계층분석과 K-평균 군집분석을 비교한 결과 군집의 개수와 군집에 속하는 구성(membership) 결과가 다르게 나온다고 주장하였다. 이는 각 방법마다 자료를 다르게 해석하게 될 수 있음을 말하며 가장 적합한 방법을 택해야 한다고 주장했다. 또한 Magidson & Vermunt (2002b)는 잠재계층분석이 K-평균 군집분석에 비해 몇 가지 이점을 제공한다는 것을 밝혔다. 또한 Magidson & Vermunt (2002a)의 시뮬레이션 연구에서도 잠재계층분석이 객체들을 분류하는 데 있어서 더 좋은 결과를 보였다고 말했다. 더하여 K-평균 군집분석은 변수들이 같은 분산을 가지고 있고 지역 독립성 가정을 유지하고 있을 때만 성과가 좋았다. 반면에 Steinley 와 Brusco (2011)는 잠재계층분석에 비해 K-평균 군집분석이 더 우수한 성과를 낸다는 결론을 이끌어 냈다. 하지만 이에 관하여 Vermunt(2011)는 Steinley 와 Brusco가 수행한 시뮬레이션에서는 K-평균 군집분석의 성능이 잠재계층분석과 동일할 수 있지만 실제 상황에서는 다를 것이라고 말했다. 왜냐하면 이론들은 잠재계층분석이 K-평균 군집분석을 능가할 것이라고 제안하기 때문이다(Vermunt, 2011). 이에 따라 Vermunt(2011)는 잠재계층분석이 K-평균 군집분석보다 더 좋은 모형임을 보여주는 추가적인 시뮬레이션 연구 결과를 제공했다(Vermunt, 2011).

따라서 본 연구에서는 유사한 수학식을 가지고 있지만 다양한 조건에 따라 상이한 결과를 준다는 선행연구들을 바탕으로 어떤 조건에서 어떤

분석방법이 더 효율적인지 검증해보고자 한다. 또한 상이한 연구문제에 따라 어떤 분석방법이 더 적합한지에 관하여 논의해보고자 한다. 이러한 연구문제를 검증하기 위하여 K-평균 군집분석과 잠재계층분석을 위한 모의자료의 생성과 모형의 평가는 통계 프로그램 R을 활용하여 이루어 질 것이다. 구체적으로, 프로그램 R에서 연구자가 설정한 모형에 기초하여 표본을 임의적이고 가상적으로 생성해주며, 시뮬레이션 표본들로부터 계산된 결과를 보고해 주는 몬테카를로 시뮬레이션 기능(이현숙, 김수진, 전수현, 2010)이 실행가능하다. 그러므로 본 연구에서는 몬테카를로 시뮬레이션 기능을 사용하여 임의 표본을 생성한 후, 이를 사용하여 잠재계층분석과 K-평균 군집분석 실행 후 각 분석방법의 성과를 비교분석 할 것이다.

사례수, 문항수, 군집수, 효과크기 그리고 잠재집단확률의 조작변수를 사용해 잠재계층분석과 K-평균 군집분석의 군집회복(cluster recovery)을 비교할 것이다. 군집회복이란 생성한 자료와 어떠한 분석을 거쳐 재 생성된 구조(recovered structure)를 비교했을 때 이것이 연구자가 초기에 생성한 자료와 얼마나 일치하는지를 보여주는 것이다. 즉 한 구성원이 원래의 군집으로 올바르게 할당되었는지의 여부를 알고자 하는 것이다. 실제 군집구성(true cluster membership) 과 조정된 군집구성(assigned cluster membership) 사이의 군집회복 정도를 비교하기 위해 ARI(Adjusted Rand Index)를 사용할 것이다.

II. 이론적 배경

1. 군집분석 (Cluster Analysis)

군집분석은 데이터의 구조나 내용을 잘 알지 못할 때 데이터에 대한 이해를 도모(clustering for understanding)하기도 하고, 유용하게 형성된 각각의 군집이 가지는 특성과 군집 간의 관계를 파악하여 다른 분석, 또는 다른 목적을 위해 유용한 출발점(clustering for utility)이 되는 분석 방법이기도 하다(이정진, 2011). 데이터를 이해하는 것이든 또는 또 다른 분석을 위해 유용하게 사용하려는 것이든 관계없이 군집분석은 심리학을 포함한 다양한 분야에서 오랫동안 중요한 역할을 수행해 왔다. 왜냐하면 본래 인간은 어떠한 대상을 집단 별로 분배하고(군집화 Clustering) 또 이러한 집단에 특정 대상을 할당(분류화 Classification)[9]시키는데 익숙하기 때문이다(Pang-Ning, Steinbach, & Kumar, 2006). 이러한 인간의 특성 덕분에 군집분석은 실용적인 문제에 대해서도 많이 응용되어져왔다. 따라서 군집분석을 통해 얻게 되는 집단, 또는 공통된 특징을 공유하는 개념적으로 의미 있는 군집은 사람들이 세상을 분석하고 설명하는데 에 중요한 역할을 한다(Pang-Ning et al., 2006).

데이터를 이해하고자 하는 맥락에서, 군집분석은 가지고 있는 데이터 안에서 두 개 이상의 소집단을 찾아내는 분석방법이다(Pang-Ning et al., 2006). 심리학에서도 군집분석이 다양한 역할을 할 수 있다. 가령 질병이나 심리건강상태를 측정하는 다양한 변수가 있을 수 있으며 이러한 변수들은 군집분석을 통해 각기 다른 소집단들을 식별해 낼 수 있다. 예를 들어, Paykel

9) 판별분석(discriminant analysis or classification analysis)은 군집분석(cluster analysis)와 관측치를 분류한다는 점에서 유사하지만 판별분석은 소속집단을 이미 알고 있는 상태에서 관측된 데이터를 사용하여 모형을 만든 후 이 모형을 통해 새로 들어온 자료, 즉 소속집단을 모르는 어떤 관측 데이터가 속하는 집단을 예측, 결정하고 분류하는 분석방법이다(노형진, 2014).

(1971)연구에서와 같이 군집화는 여러 유형의 우울증을 식별하는 데에도 사용되었다. 군집분석의 목표는 집단 내의 구성원이 그 집단 내의 다른 구성원과 유사하거나, 또는 다른 집단의 구성원과는 다르도록 분류하여 군집화 하는 것이다. 집단 내 유사성이 클수록(동질성) 또는 집단 간 차이가 클수록(이질성) 군집화가 더 우수하고 명확한 것이다(Pang-Ning et al., 2006).

두 번째로 또 다른 분석의 출발점이 되는 기술로서의 군집분석은 각각의 구성원으로부터 이런 구성원들이 속해있는 군집까지의 개념을 보여준다. 프로토타입(prototype)[10]을 이용하여 각 군집의 특징을 정하고, 프로토타입을 근거로 각각의 구성원이 같은 군집 안에서 다른 구성원들을 서로 대표할 수 있도록 한다. 이러한 프로토타입은 다른 분석을 위한 기초로도 사용될 수 있게 된다. 따라서 다른 분석의 출발점으로 사용하는 맥락에서의 군집분석은 즉 데이터를 가장 잘 대표하는 프로토타입을 찾는 기술에 대한 연구라고도 볼 수 있다(Pang-Ning et al., 2006).

군집의 프로토타입은 여러 장면에서 쓰일 수 있다. 첫 번째는 데이터 요약에 관련 한 것인데 흔히 사례수가 아주 큰 경우에는 표준오차가(standard error)가 작아져 효과크기(effect size)가 없는데도 효과가 있다는 결론을 얻을 수가 있다. 따라서 데이터 전체를 어떠한 알고리즘에 적용시키기 보다는 군집의 프로토타입 만을 포함하고 있는 보다 축소된 데이터를 사용하여 데이터를 요약하는 차원에서 군집분석이 이용될 수 있다. 군집의 프로토타입이 사용될 수 있는 또 다른 방법은 또는 벡터 양자화[11]를 이용하여 데이터를 압축하는 것이다. 벡터 양자화를 이용한 데이터 압축은 주로 데이터

10) 여기서 프로토타입은 군집의 특징 즉 중심점을 의미하며 군집에 속한 점들의 평균, 중앙값 또는 어떠한 대푯값이 될 수 있다. 주어진 알고리즘에 따라 여러 번의 반복을 거쳐(iteration) 적합한 프로토타입을 얻을 때까지 프로토타입을 재설정하게 된다.

11) 현실적인 실제 상황에서 데이터의 입력 신호는 연속적인 실수 값을 가지기 때문에 데이터의 양이 무한대로 늘어나게 된다. 따라서 입력 신호의 데이터 양을 이산적인 값(discrete)으로 표현하게 되어 여러 개의 입력 값을 가지고 있는 벡터를 간단한 형태의 멕터로 매핑 하는 방법이다(곽호완, 박창호, 이태연, 김문수, & 진영선, 2008).

간의 유사성이 매우 높거나, 정보 일부의 손실이 허용 가능하거나 또는 데이터 크기의 상당한 감소가 요구될 때 주로 사용된다. 마지막으로 군집의 프로토타입은 가장 가까운 대상을 좀 더 효율적으로 찾을 수 있도록 해준다. 원래 유사한 대상을 찾기 위해서는 데이터상 가장 가까운 점을 찾는 것은 모든 점 사이의 쌍으로의 거리를 계산을 하도록 한다. 하지만 군집과 군집의 프로토타입을 더 효율적으로 찾을 수 있는 방법이 있다. 어떤 대상이 군집의 프로토타입에 상대적으로 가까운 경우는 프로토타입을 사용해서 그 대상에서 가장 가까운 점을 찾는데 필요한 거리 계산의 수를 줄일 수 있는 것이다. 직관적으로 생각했을 때, 두 군집의 프로토타입이 멀리 떨어져 있다면, 각 군집의 대상은 서로 가까운 이웃이 될 수 없는 것이다. 결과적으로 어떤 대상의 가장 가까운 점을 찾기 위해서는 두 군집의 근접성이 포로토타입 간의 거리로 측정되는 인근 군집의 대상까지의 거리를 계산하기만 하면 되는 것이다.

1) 군집화 방법

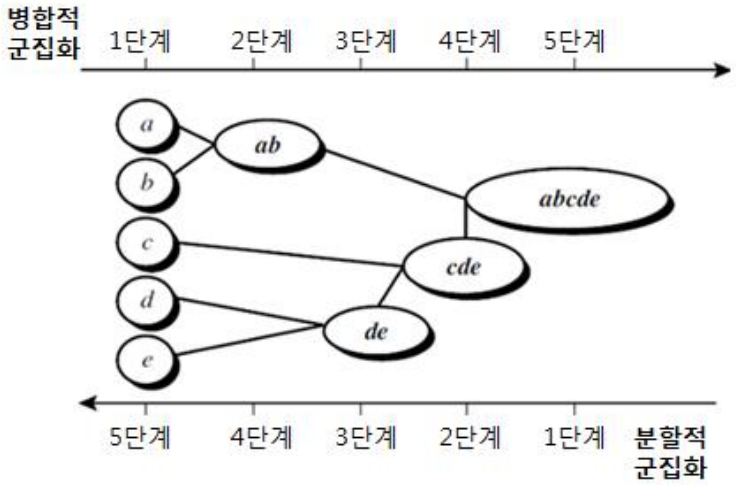
군집화를 하는 방법은 구분방법에 사용되는 기준에 따라 다양한 유형으로 분류될 수 있다. 상위군집과 하위군집의 개념여부에 따라 1)계층적(hierarchical) 군집화와 비계층적(non-hierarchical; partitional) 군집화, 분석 대상이 군집에 중복으로 속할 수 잇는지의 여부에 따라 2)독점식(exclusive) 군집화, 중복식(overlapping) 군집화 와 퍼지식(fuzzy) 군집화, 그리고 분석대상을 모두 이용하는지의 여부에 따라 3)완전(complete) 군집화와 부분(partial) 군집화로 나누어 볼 수 있다 (Pang-Ning et al., 2006).

첫 번째 방법은 상위 군집이 하위 군집을 포함하는지의 여부에 따라 계층적 군집분석과 비계층적 군집분석으로 나누어 질 수 있다. 계층적 군집분석은 각

군집이 하위군집을 갖도록 하는 것으로 개별대상 간의 거리에 의해 가장 가까이에 있는 대상으로부터 시작해서 결합해 감으로써 나무모양의 계층구조를 형성해 가는 방법이다. 덴드로그램을 그려 줌으로써 군집이 형성되는 과정을 정확히 파악 할 수 있다는 장점이 있다. 반면에 각 단계에서 지역적으로 최적인 군집을 찾기 때문에 전체 목적함수(objective function)[12]를 최적화 하는 일반적인 방법이 아니며 자료의 크기가 커지면 분석이 어려워진다는 단점도 가지고 있다(이정진, 2011).

계층적 군집분석은 다시 크게 두 가지 방식인 병합적 군집화와 분할적 군집화로 나누어진다. <그림 1>에서 보는 것처럼 병합적(Agglomerative) 계층 군집화는 상향식으로 가까운 객체끼리 군집을 형성해 나감으로써 더 큰 군집으로 만들고 군집을 통합해 가는 과정이다. 마지막에는 모든 객체가 하나의 군집으로 되고 연구자가 정해놓은 모든 조건을 만족하게 되면 군집화가 종료된다. 대부분의 군집화는 병합적으로 이루어지는 경우가 많다. 반면에 분할적(Divisive) 계층 군집화는 <그림 1>과 같이 병합적 군집화와 반대로 하향식으로 진행된다. 즉 모든 객체를 하나의 큰 군집에 속하는 것으로 가정하고 군집화가 시작되어 마지막으로는 각 객체가 하나의 군집으로 형성 되거나, 모든 조건을 만족하게 되면 군집화의 세분화가 종료된다(Han, Pei, & Kamber, 2011).

12) 일반적으로 최적화 문제를 다루며 함수의 최대값이나 최소값을 찾고자 할 때 사용되는 그 함수를 목적함수라 부른다(수학백과, 2015).



<그림 1> 병합적 군집화와 분할적 군집화

계층적 군집분석과는 달리 비계층적 군집분석 또는 분할 군집화[13]는 자료를 중복되지 않는 군집으로 나누는 것이다. 구하고자 하는 군집의 수를 정한 상태에서 설정된 군집의 중심에 가장 가까운 객체를 하나씩 포함해 나가는 방식으로 군집을 형성해 가는 방법이다. 이는 많은 자료를 빠르고 쉽게 분류할 수 있다는 장점을 지니고 있지만 동시에 군집의 수를 미리 정해주어야 하는 단점과 군집을 형성하기 위한 초기값에 따라 군집 결과가 달라진다는 단점을 가지고 있다(Pang-Ning et al., 2006).

따라서 계층적 군집분석을 통해 대략적인 군집의 수를 파악한 후 이를 초기 군집 수로 설정하는 방법을 이용하기도 한다(Pang-Ning et al., 2006). 일반적으로 군집분석에 쓰이는 데이터의 크기는 큰 경우가 많기 때문에 주어진 데이터에서 소규모 자료를 이용해 계층적 군집분석을 실시한다. 계층적 군집분석을 통해 가장 적적한 군집의 수를 결정한 후 다시 비계층적

13) 분할 군집화(Partitional clustering) 는 비계층적 군집분석(Non-hierarchical)과 같으며 데이터의 객체를 겹치지 않는 하위집단 즉 군집으로 나누어 각각의 객체가 정확히 하나의 하위집단에 포함되도록 하는 것이다.

군집분석을 실시하여 군집화하는 것이다(이훈영, 2013). 또한 군집분석을 통해 적절한 군집화가 이루어지면 교차분석, 분산분석 등을 이용하여 군집의 특성을 나타내는 변수를 추출할 수 도 있다(이훈영, 2013).

두 번째 방법으로는 분석 대상이 군집에 중복으로 속할 수 있는지의 여부에 따라 독점식(exclusive) 군집화, 중복식(overlapping) 군집화와 퍼지식(fuzzy) 군집화로 구분된다. 독점식 방법에서는 분석 대상이 하나의 군집에만 속할 수 있으며 중복식 방법에서는 분석 대상이 하나 이상의 군집에 속할 수 있도록 한다. 마지막으로, 퍼지식[14] 방법은 확률론적(probabilistic) 방법이라고도 불리는데 이는 어떠한 하나의 특정 군집이 아니라 다른 군집에도 속할 가능성을 표현해주기 때문이다. 하나의 객체가 모든 군집에 속할 가능성이 있다고 가정하고, 각 군집에 속할 가능성을 가중치로 표현한다(이정진, 2011). 이는 수학공식 즉 알고리즘을 사용하여 변수가 개인 간에 어떻게 결합되는지를 분석하는 방법이며(Pang-Ning et al., 2006) 측정모형 혹은 통계모형에 기반을 두어 변수 간의 관련성을 함수화 하는 것이다.

N개의 데이터를 K개의 군집으로 나누고자 할 때, 하나의 객체가 k번째 군집에 할당 될 가능성을 가중치 w_{ik}로 정한다면 $(0 \leq w_{ik} \leq 1)$, 그 객체가 각 군집에 속할 가중치들의 합은 1이 된다(이정진, 2011).

$$\sum_{k=1}^{K} w_{ik} = 1 \qquad \text{식 (2.1.1)}$$

그리고 각 군집에 속한 객체는 0이 아닌 가중치를 갖고 모든 가중치가 1이 될 수는 없다. 그러므로 데이터 전체에 해당하는 총 객체의 가중치들의 합은

14) 퍼지(fuzzy)의 사전적 의미는 애매한 또는 불문명함으로 1965년 Zadeh가 처음 제안한 이론이다. 애매하고 불분명한 개념들을 수치로 정량화하기 위한 시도로 퍼지이론을 창안했다(Zadeh, 1965).

0보다 크고 전체 데이터 객체의 수보다는 작다(이정진, 2011).

$$0 < \sum_{i=1}^{N} w_{ik} < N \qquad \text{식 (2.1.2)}$$

세 번째로는 분석 대상을 모두 군집에 속하도록 하는 완전 군집화(complete clustering)와 적적한 군집에 속하지 않는 대상은 포함하지 않고 군집화를 진행하는 부분 군집화(partial clustering)가 있다. 데이터가 극단치를 포함하는 경우를 쉽게 찾아볼 수 있기 때문에 부분 군집화는 어떠한 대상이 아무런 군집에도 속하지 않을 때 주로 사용된다(Pang-Ning et al., 2006). 또는 만약 각 군집에 대한 가중치가 모두 같은 경우 즉 주어진 모든 군집에 속하는 확률이 같은 경우가 있다. 이렇게 소속군집이 불분명한 경우는 완전 군집화보다 부분군집화를 실시하여 그 객체를 분석에서 제외하는 것이 더 효율적일 수 있다.

2) 군집의 종류

“잘 분리된 군집”(Well-separated clusters)은 군집 밖의 객체들 보다 같은 군집에 있는 객체들끼리 더 가까운 객체들의 집합을 말한다. 군집 내의 모든 객체가 서로 충분히 가깝도록 임계값을 지정해야 하는 경우도 종종 있다. 이러한 이상적인 군집의 정의는 데이터가 서로 멀리 떨어져 있는 자연적인(natural) 군집을 포함 할 때만 충족이 된다. <그림 2>를 보면 동그라미로 나타낸 잘 분리된 군집 두 개를 쉽게 찾아볼 수 있다. “잘 분리된 군집”은 꼭 구형(globular)일 필요가 없으며 어떠한 모양도 가능하다. 어떠한 모양도 가능하다는 것은 집단 간 공분산 행렬(covariance matrices)이 다를 수

있다는 것을 의미한다(Steinley, Brusco, 2011).

“프로토타입 기반 군집” (prototype-based clusters)은 각각의 대상이 다른 군집의 프로토타입 보다 자신이 속해 있는 군집을 대표하는 프로토타입에 거리 상 가장 가까운 대상끼리의 집합이다. <그림 2>를 보면 두 개의 포로토타입으로 인해 구분된 두 개의 군집을 볼 수 있다. 데이터의 형태가 연속형 변수 일 때 군집의 프로토타입은 주로 중심점[15]이 된다. 데이터가 범주형 자료라서 중심점이 더 이상 의미를 가지지 못할 때는 메도이드[16]를 사용하기도 한다. 프로토타입은 대개 대부분의 중심점으로 간주되기 때문에 “중심기반 군집” (center-based clusters)이라고도 부른다. 이러한 군집은 주로 구형을 띄게 된다.

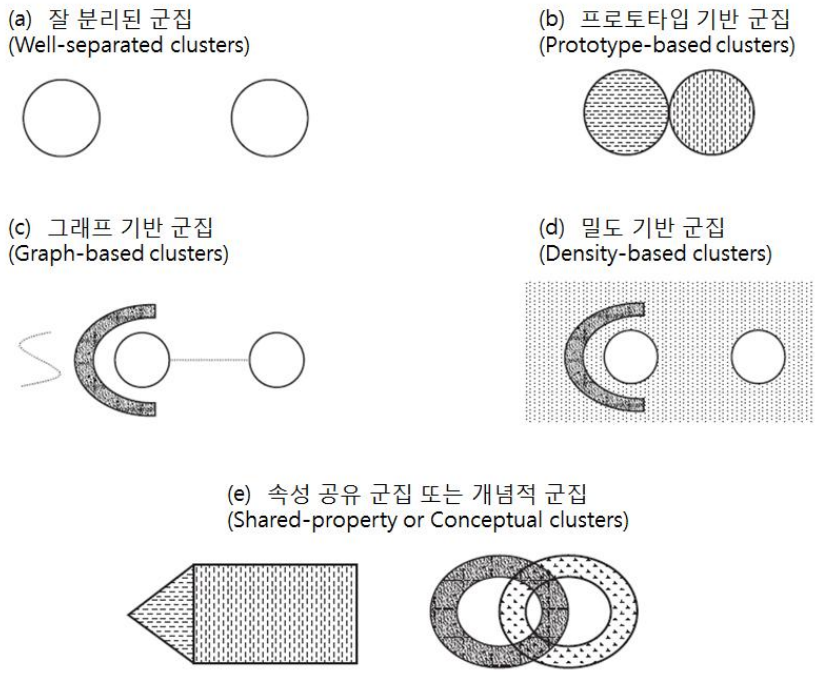
<그림 3>을 보면 데이터 전체를 그래프로 표현하고자 할 때, 각각의 점은 객체가 되고 객체끼리의 연결은 링크가 된다. 이때 “그래프 기반 군집 (graph-based clusters)은 군집 밖의 대상과는 연결되지 않고 군집 내의 대상과는 서로 연결된 객체들의 집합인 “연결된 구성요소(connected components)”로 정의될 수 있다. “그래프 기반 군집” (graph-based clusters)의 중요한 예로 두 객체가 서로 지정된 거리 내에 있는 경우에만 연결되는 “근접도 기반 군집” (contiguity-based clusters)이 있다. 이것이 말해 주는 것은, 근접도 기반 군집에 있는 각각의 대상은 다른 군집에 있는 대상보다 그 군집 안에 있는 대상들과 더 가깝다는 것이다. 이는 주로 군집이 불규칙 하거나 얽힌 경우 특히 유용하지만 <그림 2>에서 보는 것처럼 데이터에 노이즈(noise)[17]가 있는 경우는 두 개의 동그라미 모양의 군집을

15) 중심점 (centroid)는 해당 군집 안에 있는 대상들의 평균값을 의미한다(Pang-Ning, 2006).
16) 메도이드 (medoid)는 군집 내의 모든 객체들과 평균 비유사도가 가장 작은 값으로 전체 데이터를 대표하는 값이다. 중심점이나 중앙값과 비슷하지만 메도이드는 항상 데이터 안에 실제로 있는 객체라를 것이 다르다. 이산형 또는 연속형 데이터 모두에 적용할 수 있으며 실제 데이터중 이를 가장 잘 대표하는 대상 하나를 군집의 중심으로 쓰는 방법이다(이정진,2011).
17) 데이터 마이닝에서 노이즈(noise)란 측정변수의 랜덤 오류나 분산에 해당되며 데이터가 분석에 사용될 때 제거되거나 무시되어야 하는, 의미 없는 객체들을 의미한다(컴퓨터인터넷IT용어대사전, 2011).

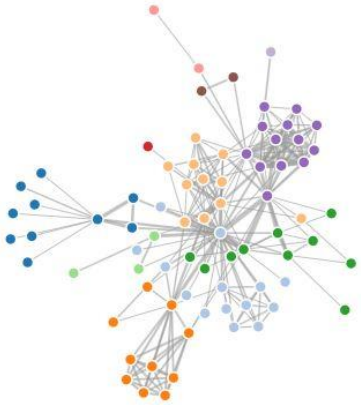
연결해주는 다리 때문에 두 개의 동그라미 군집을 병합시켜 버리는 문제가 발생 할 수 있다.

“밀도 기반 군집”(density-based clusters)은 객체끼리 뭉쳐진 밀도에 따라 군집을 형성하는 것을 말한다. <그림 2>에서는 C 모양의 낮은 밀도의 영역으로 둘러싸인 대상들의 고밀도 영역을 동그라미로 표현하고 있다. 이는 군집이 불규칙적이거나 얽혀있고 노이즈(noise)나 극단치가 존재하는 경우 종종 사용된다. <그림 2>의 (d)는 <그림 2>의 (c)그래프 기반 군집에서 노이즈를 추가하여 얻은 밀도기반 군집의 형태를 보여주고 있다. 따라서 (c)에서 보이는 S 모양의 굴곡과 두 동그라미 군집을 연결하는 다리 부분이 노이즈로 인식 된 것이다.

“속성 공유 군집 또는 개념적 군집” (shared-property or conceptual clusters)은 유사한 속성을 공유하는 대상끼리 군집화 하는 것을 말한다. 이것은 위에 언급한 군집의 정의들을 모두 포함하는 개념이다. <그림 2>에서 보는 것과 같이 세모 모양의 군집과 직사각형 모양의 군집은 서로 인접해 있고 두 개의 동그라미 모양의 군집은 서로 얽혀있다. 다양한 모양의 군집을 찾기 위해서는 특정 군집화 알고리즘이 필요하다. 이러한 과정을 개념적 군집이라고 일컫는다.



<그림 2> 군집의 종류 (Pang-Ning, Steinbach, & Kumar, 2006)



<그림 3> 그래프기반 군집의 예

2. K-평균 군집분석 (K-means Clustering)

데이터의 객체들을 분할화(Partitioning)하는 것은 통계학, 심리학 등 많은 분야에서 큰 관심거리이다. N개의 대상을 K개의 집합으로 나누는 경우의 수는 제 2종 스털링 수[18] 로 계산될 수 있다(Steinley, 2006)

$$\frac{1}{N!}\sum_{i=1}^{N}(-1)^{N-i}\binom{N}{i}N^k \qquad \text{식 (2.2.1)}$$

하지만 N이 커질수록 가능한 모든 분할(Partition)을 완벽하게 계산하는 것은 비현실적이다. 이러한 계산상의 어려움 때문에 적절한 시간 내에 최선의 분할방법 즉 군집화 방법을 제공하는 방법들이 연구되었다. 이에 따라 Cormack(1971)은 어느 정도의 군집 내 유사성과 군집 간 이질성을 나타내야 하며 즉 군집들은 군집 간에 외적으로 구분되어야 하며 내적으로는 결합력이 있어야 한다고 주장했다. 역사적으로 많은 연구자들이 군집 내 분산을 최소화함으로써 이러한 정의를 사용하려는 많은 시도가 있었다(Cox, 1957; Engelman & Hartigan, 1969; Fisher, 1958; Thorndike, 1953). 군집 내 유사성을 극대화시키는 이러한 초기 시도에 이어서 Sebestyen (1962)와 MacQueen(1967)은 최적의 분할방법을 찾기 위해서 K-평균군집(K-means)을 독자적으로 개발하였다(Steinley, 2006).

K-평균 군집분석 방법은 통계 프로그램 SPSS를 통해 실행 가능하다.

---

18) 제 2종 스털링 수(Stirling number of the second kind) $S(n,k)$는 n명의 사람을 k개의 그룹으로 분활하는 경우의 수를 말한다(Weisstein, 2002). 제2종 스털링수는 다음 식 $S(n,k)=S(n-1,k-1)+kS(n-1,k)$을 만족한다. 여기서 집합 S의 분할이란 서로 공통 원소를 갖지 않고 그것들의 합집합이 S가 되는 S의 부분집합들의 모임이다(김가온, 이동화, 안승민, 문혜성, 조완영, 2013). 예를 들어 집합 $S(4.2)$는 4개의 원소 {1,2,3,4}를 갖는 집합의 분할들 중에서 2개의 그룹으로 나누어지는 것들의 총 개수를 말한다. 그래서 제 2종 스털링 수 공식에 적용하면 $S(4.2)$는 7이 된다.

분석메뉴의 분류분석에서 계층적 군집을 실시하거나 또는 연구자의 주관적인 판단, 이론에 입각하여 최적의 군집 수를 결정한다. 그 후 다시 분류분석 중 K-평균 군집을 실시하여 최대반복계산 횟수와 소속군집, 그리고 군집의 수를 설정해준 후 최종군집중심 표를 보고 결과를 해석하면 된다.

1) K-평균 군집분석의 알고리즘

K-평균군집은 각 집단의 프로토타입을 중심점 (centroid)으로 특징지어진다. 중심점은 주로 점들의 평균값을 뜻하며 연속형 n차원의 공간에서의 대상에 적용이 된다. C_k 군집에 있는 j번째 변수의 중심점은 다음과 같이 계산된다.

$$\overline{x}_j^{(k)} = \frac{1}{n_k} \sum_{i \in C_k} x_{ij} \qquad \text{식 (2.2.2)}$$

그리고 이에 따른 C_k 군집의 완성된 중심점의 벡터는 다음과 같다.

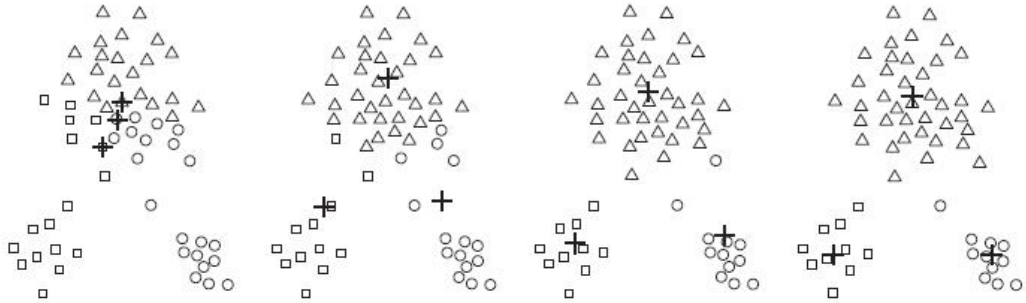
$$\overline{X}^{(k)} = (\overline{x}_1^{(k)}, \overline{x}_2^{(k)}, \cdots, \overline{x}_P^{(k)})' \qquad \text{식 (2.2.3)}$$

K-평균 군집분석의 첫 번째 단계는 K개의 초기 중심점의 개수를 정하는 것으로 이는 연구자가 지정하는 데로 설정 가능하며 K는 연구자가 원하는 군집의 수와 같다. K 초기 종자(initial seed)[19]는 P-차원 벡터($s_1^{(k)},...,s_p^{(k)}$)들로 정의되며($1 \le k \le K$) i번째 대상과 k번째 종자 벡터 사이의 거리를 다음과 같이 얻게 된다.

19) 초기종자(initial seed)는 난수(random number) 등을 생성하기 위한 처음의 기초 값을 말한다.

$$d^2(i,k) = \sum_{j=1}^{P} (x_{ij} - s_j^{(k)})^2 \qquad \text{식 (2.2.4)}$$

각 대상들은 식 (2.2.4) 값이 가장 작은 군집으로 배정이 된다. 그 후 식 (2.2.3)을 이용하여 각 군집의 중심점을 만들어 낸다. 다시 $d^2(i,k)$를 사용하여 각각의 대상을 해당 중심점과 비교한 후 각각의 점, 즉 데이터의 대상들에서 각자 가장 가까운 중심점에 배정이 된다. 각 중심점에 배정이 된 대상들의 집합이 바로 군집이 되는 것이다. 각 군집의 중심점은 그 군집에 배정된 대상들을 기반으로 다시 계산되어 업데이트가 된다. 그러면 각각의 대상들은 업데이트 된 중심점에 따라 가장 가까운 중심점으로 다시 배정이 된다. 이러한 배정과 업데이트 과정은 중심점이 지정한 오차의 범위 안에 속하여 유지될 때까지 반복(iteration)된다. 즉 중심점을 계산하고 그 중심점에 가까운 대상들을 배정한 후 배정된 대상들을 기반으로 중심점이 다시 계산되는 반복의 과정을 거치는 것이다. 이러한 과정은 <그림 4>를 통해 쉽게 알 수 있는데 세 개의 십자가 모양이 중심점을 말하며 구성원들이 세 개의 군집으로 나누어지는 절차를 보여주고 있다.



<그림 4> K-평균 군집분석의 알고리즘을 사용한 예
(Pang-Ning, Steinbach, & Kumar, 2006)

따라서 반복의 과정은 대상들이 한 군집에서 다른 군집으로 더 이상 옮겨가지 않는 시점이 종료시점이 되는 것이다. 이렇게 K-평균 군집분석 알고리즘이 종료되었다는 것은 중심점이 구성원들의 집단을 식별했다는 의미가 된다. 이것이 K-평균 군집분석의 기본 알고리즘이다(Pang-Ning et al., 2006; Steinley, 2006).

K-평균 군집분석의 기본 알고리즘에서 좀 더 세부적으로 알아보자면, 각 대상들을 그들과 가장 가까운 중심점으로 배정을 하는데 "가장 가까운(the closest)" 이라는 개념을 정량화하기 위해 근접성 함수[20] 을 통한 근접도 평가[21]를 해야 한다. 여러 가지 근접도 평가가 있지만 주어진 데이터의 형태(가령, 비연속자료 또는 연속형 자료)에 따라 적합한 방법을 선택하게

20) 근접성 함수(Proximity function)은 특정 거리나 위치 내에 존재하는 대상물들 간의 관계를 분석하는 것을 말한다(이강원, 송호웅, 2016).
21) 근접도 평가(Proximity measure)는 근접성 기능을 통해 대상 간의 가까운 정도 또는 먼 정도를 수량화 하는 것으로 방법으로는 유클리드 거리 측정법(Euclidean Distance Measure), 제곱 유클리드(Squared Euclidean) 거리법, 맨하탄(Manhattan) 거리 측정법, 코사인(Cosine) 거리 측정법, 타니모토(Tanimoto) 거리 측정법(또는 확장 자카드(Jaccard) 계수) 등이 있다(Owen, Abil, Dunning, & Friedman, 2012).

된다. 흔히 유클리드 거리[22] (L_2) 는 유클리드 공간에서의 각 측정값과 중심점 간의 거리를 측정하는데 주로 쓰이며 가장 일반적인 방법이다. 또한 군집화의 목표는 일반적으로 각 점들끼리의 근접도나 중심점으로의 근접도에 따른 목적함수[23]로 표현된다. 가령, 각각의 점과 그에 가장 가까운 중심점의 거리를 제곱한 값을 최소화 시키는 것이다.

$$SSE = \sum_{j=1}^{P} \sum_{k=1}^{K} \sum_{i \in C_k} (x_{ij} - \bar{x}_j^{(k)})^2 \qquad \text{식 (2.2.5)}$$

즉, 식(2.2.5)의 목적함수는 오차제곱합(sum of the squared error;SSE)을 사용하여 군집화의 질을 측정하게 된다. 다시 말해, 각 관찰값과 중심점 간의 차이인 오차를 계산하여, 가장 가까운 중심점과의 유클리드 거리를 계산, 전체 오차 제곱합을 계산하는 것이다. 오차제곱합이 작다는 것 다시 말해, 군집 내 분산이 작다는 것은 군집의 중심점이 그 군집 안에 있는 측정값들을 더 잘 대표하고 있다는 뜻이 되기 때문에 오차제곱합이 더 작은 군집을 더욱 선호하게 된다.

K-평균의 재배치 알고리즘은 X와 두 개의 추가적인 행렬을 사용해서 공식화 될 수 있다. 하나는 군집구성(membership) 행렬이며 또 하나는 군집표현(cluster representation) 행렬이다. 군집구성 행렬 M은 $\{m_{ik}\}_{N \times K}$ 으로 나타내며 한 객체가 군집에 속하면 m_{ik}는 1과 같고 아니면 0과 같다.

---

22) 유클리드 거리(Euclidean Distance)는 가장 직관적이고 일반적으로 생각 되는 거리의 개념과 비슷하다. 이는 a 와 b가 각각의 데이터 포인트 일 때 $d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$ 으로 계산된다(Owen, et al., 2012).

23) 목적함수(Objective Function)은 어떤 목적을 위해 사용하는 함수로 일반적으로 함수의 최대값 또는 최소값을 구하여 최적화 문제를 다룰 때 사용되는 함수이다(이강원, 송호웅, 2016).

군집표현 행렬 $R_{K\times P}=\begin{bmatrix} r'_1 \\ r'_2 \\ \vdots \\ r'_K \end{bmatrix}$의 k번째 행, $r_k = \{r_{k1},\cdots,r_{kP}\}$, 은 k 군집의 객체들을 가장 잘 대표하는 벡터이다. 그리고 각각의 행은 K 군집의 P변수에 해당하는 중심점 벡터의 평균을 말한다. 그러면 식 (2.2.5)는 M과 R행렬의 함수로 다시 쓰여 질수 있다(Steinley, 2006).

$$F(R,M) = tr[(X-MR)'(X-MR)] \qquad \text{식 (2.2.6)}$$

R의 추정치가 주어지면 M에 대한 식 (2.2.6)을 최소화 한 후 주어진 군집구성 추정치를 기반으로 하여 M에 대한 값을 고정하고 R을 다시 추정하는 것을 반복한다. 즉, K-평균군집분석의 목표는 μ_k가 군집의 평균 벡터일 때,

$$\sum_{k=1}^{K}\sum_{i\in C_k}(x_i-\mu_k)'(x_i-\mu_k) \qquad \text{식 (2.2.7)}$$

를 최소화 시키는 것이 된다. 더하여 (X-MR)′(X-MR)는 k번째 군집의 군집 내 다변량제곱합 행렬인 $\sum_{k=1}^{K} W_k$와 같다는 것에 주목할 필요가 있다. 이러한 등가성은 W_k를 $X_{N\times P}=\begin{bmatrix} X'_1 \\ X'_2 \\ \vdots \\ X'_N \end{bmatrix}$ 이며, $X_i=\{x_{i1},\cdots,x_{ip}\}$인

$$W_k = \frac{1}{2n_k} \sum_{i \in C_k} \sum_{i^* \in C_k} (x_i - x_{i^*})(x_i - x_{i^*})' \qquad \text{식 (2.2.8)}$$
$$\forall\, , i^* = 1, \ldots, n_k$$

으로 다시 나타낼 수 있다. 이로 인해 W의 대각합과 W_k 대각합의 관계로 식 (2.2.5)와 식 (2.2.6)이 같다는 것을 보일 수 있다.

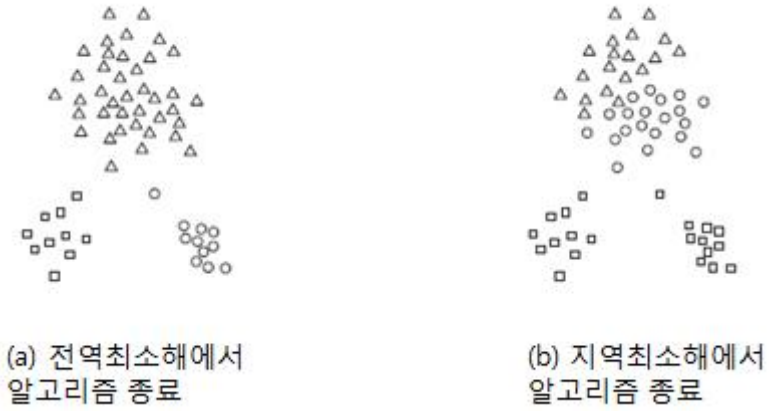
$$tr(W) = \sum_{k=1}^{K} tr(W_k) \qquad \text{식 (2.2.9)}$$

2) K-평균 군집분석의 쟁점

K-평균 알고리즘에서 중심점을 정할 때 무작위로 정하는 방법이 가장 많이 쓰이는 방법이다(Peña, Loxano, & Larrañaga, 1999). 무작위 방법은 각각의 객체들을 임의의 군집에 배당한 후, 각 군집에 배당된 객체들의 평균을 초기 중심점으로 설정한다. 초기 군집이 각 객체들에 의해서 고르게 분포되어 있기 때문에 각 초기 군집들의 중심점은 데이터 전체의 중심에 가깝게 위치하는 경향을 보인다. 하지만 이러한 알고리즘이 반드시 전역최소해[24]를 제공해주지 못하며, 사용된 초기값에 따라 전역최소해로 검증가능하지 않은 지역최소해[25]에서 알고리즘이 종료되어 문제가 되는 경우가 있다(MacQueen, 1967; Hartigan & Wong, 1979). <그림 5>는 전역최소해에서 알고리즘이 종료된 경

24) 전역최소해(global minimum)은 특정 지점의 함수 값이 그 지점 주변의 다른 모든 함수 값과 비교했을 때 정의역에서 가장 작은 경우의 값을 말한다.
25) 지역최소해(local minimum)은 특정 지점의 함수 값이 그 지점 주변의 다른 함수 값과 비교했을 때 특정 범위 안에서만 가장 작은 경우의 값을 말한다.

우와 지역최소해에서 알고리즘이 종료되어 군집화가 마무리 된 경우를 보여주고 있다.



<그림 5> 전역최소해와 지역최소해
(Pang-Ning, Steinbach, & Kumar, 2006)

지역최소해에서 알고리즘이 종료되는 것을 피하기 위해서 몇몇의 학자들은 상이한 초기값을 사용한 K-평균 분석방법을 여러 번 실행해보고 각각 상이한 총 SSE값 중 가장 작은 값을 보이는 결과를 선택할 것을 제안하기도 한다(Falkenauer & Marchand, 2001). 하지만 어느 정도 적당한 크기의 데이터에 대한 지역최소해의 개수는 수천 개로 증가할 수 있기 때문에 (Steinley, 2003), 알고리즘의 수렴이 전역최소해가 아닌 지역최소해로 수렴하여 결과적으로 좋지 못한 결과를 보여줄 가능성이 있다(Makarenkov & Legendre, 2001; Pang-Ning et al., 2006). 따라서 올바른 K-평균 알고리즘 초기값을 선택하는 것은 K-평균 알고리즘의 과정에서 중요한 단계이다(Steinley, 2006). 이에 따라 지역최소해를 이용한 해결책을 피하고 전역최소해를 찾기 위한 몇 가지 대안이 있다(Steinley, 2006).

대표적으로 Forgy(1965)가 고안한 Forgy 방법은 데이터로부터 임의의 k개

의 데이터를 선택한 후 상호 배타적인 군집으로 나누어 그 그룹의 평균을 해당 군집의 초기 중심값으로 설정한다. 초기 군집이 임의의 k개의 점들에 의해서 설정되기 때문에 각 초기 군집들의 중심값은 데이터 전체의 중심에서 펴져 있는 경향을 보인다. 또한 MacQueen(1967)이 고안한 MacQueen 방법은 데이터로부터 임의의 k개의 데이터를 선택한 후 각 군집의 초기 중심점으로 설정한다. 그 후 지정되지 않은 각 객체들을 그로부터 가장 가까운 군집에 배당하게 된다. 이로 인해 모든 객체가 각자의 군집에 배당되면 각 군집의 중심값을 다시 계산하여 초기 중심점으로 재설정하게 된다. MacQueen 방법은 최종 수렴에 가까운 군집을 찾는 것은 비교적 빨리 진행되지만 최종 수렴에 해당되는 군집을 찾는 것은 매우 느리다는 단점이 있다(Du, Wong, 2002).

이밖에도 초기값을 선택하는 여러 가지 방법이 있다(Celebi, Kingravi, & Vela, 2013). Jancey 방법(1966), Ball & Hall's 방법(1967), González, & Tou (1974)가 고안한 간단한 군집 찾기 방법(The Simple Cluster Seeking method), späth 방법(1977), Astrahan 방법 (1970), Milligan 방법(1980), Bradley & Fayyad 방법 (1998), Hajnal & Loosveldt 방법 (2000), Kaufman & Rousseeuw (1990)가 고안한 Kaufman 방법 등이 있다.

K-평균 군집분석에서 가장 어려운 문제 중 하나는 군집의 수를 결정하는 것이다. Milligan 와 Cooper(1985)는 올바른 군집의 수를 결정하기 위해 계층적 군집분석에 사용되는 30가지 방법에 대해 몬테카를로 시뮬레이션을 실시했다. 하지만 이런 방법 중 상당수는 군집이 계층적으로 중첩되는 계층적 군집분석의 경우에 사용가능 하기 때문에 비계층적 군집분석인 K-평균 군집분석에는 적합하지 않은 방식들이다(Steinley, 2006). 계층적 군집분석을 실시한 후 정하는 군집의 수는 통계적인 수치가 아니라 연구자의 주관적인 판단 또는 이론에 입각하여 최적의 집단 수를 규명하는 것이기 때문에 신뢰도가 떨어진다는 단점이 있다(Magidson & Vermunt, 2002b). 이에 따라 알고리즘 방법

(Algorithmic methods), 그래프 방법(graphical method), 그리고 수식화 방법(formulatic method)을 사용하여 군집의 수를 정하기도 한다(Steinley, 2006). 알고리즘 방법은 알고리즘 자체 내에서 의사결정을 통해 군집의 수를 결정하는 것이다. K-평균 군집분석 알고리즘과 같이 사용자가 군집의 수를 결정하지만 알고리즘이 사용자의 제공값을 수정 할 수 있도록 하는 것이다. 그래프 방법은 곡선이 평탄화(flattening)되는 것이 K의 정확한 값을 나타낸다는 다양한 K값에 대한 SSE의 객관적 기준을 지지하는 방법이다. 수식화 방법은 K의 범위에 걸쳐 방정식을 계산하고 기준을 최소화 하거나 최대화 하는 값을 선택하게 된다.

오차제곱합의 기준을 사용하기 때문에 극단치는 과도하게 영향을 미칠 수 있다(Pang-Ning et al., 2006). 중심점을 구할 때 모든 데이터가 적용이 되어 평균 중심점 계산 시 극단치가 영향을 미치게 된다. 그래서 극단치가 존재하는 경우 군집의 프로토타입 즉 중심점은 그 군집의 구성원들을 더 이상 잘 대표하지 못하게 되며 오차제곱합 또한 높아질 수밖에 없다. 이 때문에 극단치는 미리 찾아서 제거하는 것이 도움이 된다.

구형(spherical)이 아닌 군집을 찾는데 에는 K-평균 군집분석이 적절하지 않을 수 있다. 즉 공분산 행렬이 같은 경우 용이하다는 것이다. 왜냐하면 중심점과 각 객체 간의 유클리드 거리를 계산하기 때문에 알고리즘 수행 시 중심점으로부터 구형으로 군집화가 이루어지게 된다. 따라서 만약 데이터의 분포가 구형이 아닌 경우는 군집화 결과가 예상과 다를 수 있다. 이 같은 경우는 비슷한 밀도를 가진 객체들을 군집으로 묶는 DBSCAN(Density-Based Apatial Clustering of Allication with Noise)[26] 방법이나 mean-shift[27]

---

26) DBSCAN은 밀도에 기초한 군집화 방법 중 하나로 두 가지 모수를 가진다. 주어진 객체들의 반경인 ϵ(엡실론)과 p(객체의 최소 개수)이다. 임의의 점에서 시작하여 ϵ반경 안에 최소의 개수를 만족하는 충분한 객체들이 포함되면 군집화를 시키고 그렇지 않으면 노이즈(noise)로 식별한다. 그 다음 ϵ 반경 안에 있는 객체들에서 같은 방법으로 군집을 확장해 나간다. (Sander, Ester, Kriegel, & Xu, 1998).

27) 평균점 이동 군집화(Mean shift clustering)은 비모수적 군집 방법으로 데이터의 분포에서 밀도가 가

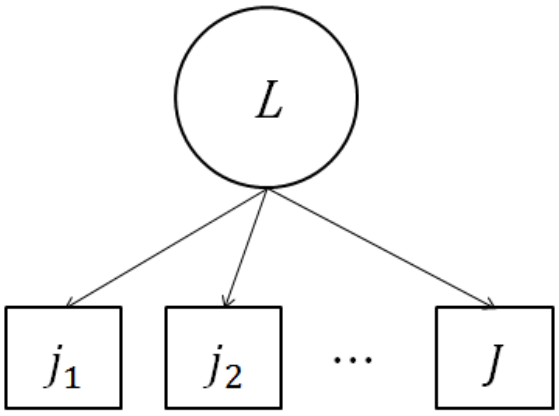
군집화를 사용하는 것이 낫다(Sander, Ester, Kriegel, & Xu, 1998).

또한 특정 구성원은 오직 한 군집에만 속하는 것으로 나타나며 그 구성원은 그 군집의 특징만을 반영하는 것으로 해석이 된다. 즉 다른 군집에도 미미하게나마 속할 확률을 전혀 고려하지 않고 이분법적으로 구성원들을 군집으로 할당하게 되는 것이다. 확률이 아닌 이분법적 결정을 하게 되면 측정에서의 오차를 고려하지 못하고 오차를 모형 안에 포함하지 못하게 된다. 또한 이는 현실적인 인간행동의 다양성을 반영하고 있지 못하기 때문에 문제점으로 인식되고 있다 (Steinley & Bruco, 2011).

장 높은 곳을 찾는 방식으로 군집화를 실시한다. 이 알고리즘은 각 객체에서 미리 지정된 창(window) 안의 객체 들을 찾고 그 창 내의 평균값을 구한다. 그리고 다시 그 평균값에서 같은 크기의 창 이내의 객체들을 찾고, 반경 내의 평균값을 구하는 방식을 계속하면 객체들의 밀도가 최대인 곳에서 이 창의 움직임이 멈추게 된다. 이런 방식으로 모든 데이터에 대해 지역 최적해를 알아낸다. 이때 같은 최빈값으로 수렴한 객체들은 같은 군집으로 할당 되는 것이다(Little & Jones, 2011)

3. 잠재계층분석(Latent Class Analysis)

사회과학에서 연구되는 많은 개념들은 직접 눈으로 관찰 할 수 없다. 예를 들어 우울, 행복, 사랑, 인종차별 같은 것들은 직접 관찰 할 수 없는 것들이다. 사실 이론적으로 흥미로운 개념들은 수백 가지가 있지만 오차 없이 이러한 개념들을 완전한 측정하기에는 어려움이 있다(McCutcheon, 1987). 가령 우울을 직접 관찰 할 수는 없지만, 우울이라는 것이 증가하면 사람은 식욕, 성욕, 수면욕 등이 감소할 수 있으며 일상 활동에 대한 흥미나 즐거움이 저하될 것이고, 사고력이나 집중력이 감소하는 등 의 증상을 보일 것이라 가정할 수 있다. 왜냐하면 각각의 관찰된 지표들은 눈에 보이지 않는 또는 잠재된(latent) 관심 변수에 의해 일어났다고 생각할 수 있기 때문이다.



<그림 6> 잠재계층모형

<그림 6>은 일반적인 잠재계층모형을 그린 것이다. 네모 안에 있는 변수($j_1 \cdots J$)는 관찰변수를 말하며 동그라미 안에 있는 변수는 잠재변수(L)를 말한다. 잠재변수를 이해하고 특성화하기 위해 관찰변수들 사이의 공분산을 구

하고 이들 간의 상호관계의 패턴을 연구한다. 잠재변수 연구의 기본적인 전제는 관찰변수 사이에서 실제로 추정된 공분산이 관찰변수들 관계를 설명하는 잠재변수와 각 관찰변수의 관계에 의해서 발생된다는 것이다. 결과적으로 잠재변수는 원래 관찰된 공분산의 진정한 원인이라고 보는 것이다. 최근 많은 방법론적 발달은 연구자가 이러한 관계를 체계적으로 분석할 수 있게 하여 잠재변수를 사용한 모형의 장점이 더욱 부각되고 있다(McCutcheon, 1987).

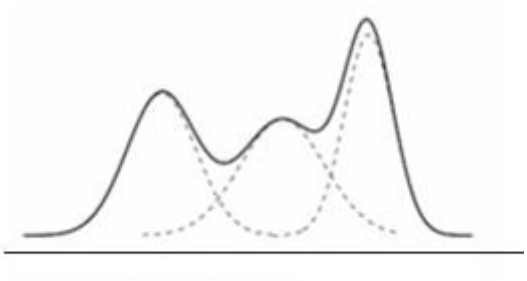
특히 여러 관찰변수들에서 추출한 공분산이 잠재변수로 사용되기 때문에 측정오차(measurement error)가 통제 가능하다(김진호, 홍세희, 추병대, 2007). 따라서 하나의 측정변수보다 여러 개의 측정변수로 생성된 잠재변수가 더욱 신뢰도와 타당도가 높기 때문에 구성개념의 다양한 측면을 반영 할 수 있다(이현숙, 김수진, 전수현, 2010).

다양한 잠재변수 모형들 중에서 Lazarsfelds는 연구자가 두 개 이상의 이산형 관찰변수에서 이산형 잠재변수를 경험적으로 식별할 수 있게 해주는 잠재계층분석(Latent Class Analysis)를 처음 소개했다 (Lazarsfeld, 1950). 이후 Goodman(1974)이 모수의 최대 가능성 추정치를 얻기 위한 알고리즘을 개발함으로써 실제 적용 가능한 모델을 만들었다. 그리고 이산형 잠재변수를 이용한 범주형 자료 분석을 위한 일반적인 틀은 Hagenaars(1990)에 의해 제안되었고 Vermunt(1997)에 의해 확장되었다. 지난 10여 년간 군집분석 방법으로 잠재계층분석을 적용하는데 많은 관심이 있었다(Magidson & Vermunt, 2002b). 잠재계층분석을 일컫는 동의한 용어들은 다음과 같다: mixture likelihood approach to clustering (McLachlan & Basford 1988; Everitt, Landau, & Leese, 1993), model-based clustering (Banfield & Raftery 1993; Bensmail, Celeux, Rafterym, & Robert, 1997; Fraley & Raftery 1998), mixture-model clustering (Jorgensen & Hunt 1996; McLachlan, Peel, Basford, & Adams, 1999), Bayesian classification

(Cheeseman & Stutz 1995), unsupervised learning (McLachlan & Peel 1996), 그리고 latent class cluster analysis (Vermunt & Magidson 2002).

1) 잠재계층모형의 기본

잠재계층모형은 여러 관찰변수에서의 응답으로부터 측정된 범주형 잠재변수를 식별한다. 잠재계층분석은 범주형 자료의 분류(Classification)를 위한 혼합 또는 확률모형(mixture or probability models)의 한 형태로 데이터를 생성하는 기본 확률함수를 가정하고 이러한 확률함수는 일반적으로 다항식 범주형 자료인 경우가 많다(Hickendorff, Putten, Heiser, & Verhelst, 2005). 확률모형(probability model)이란 <그림 7>과 같이 데이터가 어떻게 분포되어 있는지를 알기 위해 적절한 확률밀도함수(probability density function)를 가정하고 데이터의 분포에 대한 모형을 만드는 것이다(Bishop, 2006). 이의 목표는 관찰변수를 이용하여 객체를 집단으로 분류하고 집단을 가장 잘 구분하는 예측변수를 식별하는 것이다(Nylund, Asparouhov, & Muthén, 2007). 그러므로 잠재계층분석의 목적은 서로 배타적인 잠재집단을 식별하는 것이기 때문에 분할(partitioning) 방법에 해당 된다.



<그림 7> 확률모형

잠재계층모형은 여러 관찰변수와 이로부터 얻어낸 잠재변수로 구성이 된다. 만약, j=1,…,J의 관찰변수가 있고, 관찰변수 j는 $r_j = 1,\cdots,R_j$ 의 응답의 범주를 가진다고 가정하자. 그러면 변수 J는 $W = \prod_{j=1}^{J} R_j$ 의 셀, 즉 응답에 대한 경우의 수를 갖게 된다. 각각의 셀은 완전한 응답 패턴이 되며 이것은 $y = (r_{1,} \cdots, r_j)$로 쓰고 즉 변수 J에 대한 응답의 벡터가 된다. 그리고 Y는 이러한 응답 패턴의 총집합이며 W 행과 J 열을 갖게 된다. 각각의 응답패턴 y는 $P(Y = y)$와 $\sum P(Y = y) = 1$ 과 관련이 있다(Lanza, Flaherty, & Collins, 2003).

잠재계층모형은 관측변수 사이의 공분산을 설명할 수 있는 범주형 잠재변수를 추정하는 것이 목적이다(McCutcheon, 1987; Goodman, 2002). 따라서 잠재계층모형은 추정된 잠재집단 확률 모수(Latent class probability parameter)와 문항반응 확률 모수(Item response probability parameter)로 구성되어 있는데 (Nylund et al., 2007), 잠재집단 확률 모수는 주로 $\gamma's$ (gamma's)로 표기하고 문항반응 확률 모수는 $\rho's$ (rho's)로 표기한다(Lanza, Flaherty, & Collins, 2003).

L은 잠재집단 c=1,…,C를 가진 범주형 잠재변수를 말한다. 그리고 잠재집단

확률 γ_c는 잠재집단 c의 우위성(prevalence)를 말하며 다시 말해 잠재변수 L의 잠재집단 c에 속할 확률을 뜻한다. 잠재집단들은 서로 상호 배타적이기 때문에 각각의 개인은 사후 확률 변수에 따라서 유일한 하나의 잠재집단에 할당이 된다. 그러므로 다음과 같은 식이 만족된다.

$$\sum_{c=1}^{C} \gamma_c = 1 \qquad \text{식 (2.3.1)}$$

문항반응 확률 $\rho_{j,r_j|c}$는 잠재집단 c에 속할 조건 하에 관찰변수 j에 대한 응답 r_j 의 확률을 말한다. 모수 ρ의 집합은 각각의 관찰변수와 각 잠재집단 사이의 관계를 보여준다. 그러므로 관찰변수가 주어졌을 때 개인이 잠재집단으로 얼마나 잘 분류될 수 있는지를 말해주는 것이다. 모수 ρ는 또한 잠재변수의 측정과 잠재집단의 해석의 기초를 제공해주기 때문에 측정모수(measurement parameter) 라고도 불린다.

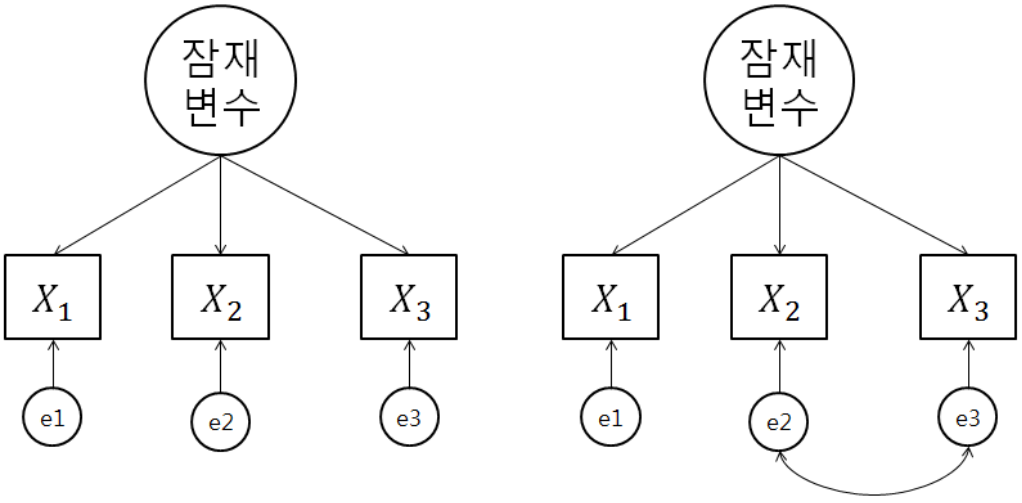
각 개인은 변수 j에 대해 하나의 응답을 하기 때문에 특정 잠재집단에 속한다는 조건 하에 특정 변수 대한 문항반응 확률의 벡터는 항상 1이 된다. 그러므로 모든 j에 대해 다음과 같은 식이 성립된다.

$$\sum_{r_j=1}^{R_j} \rho_{J,r_j|c} = 1 \qquad \text{식 (2.3.2)}$$

문항반응 확률 모수는 요인분석에서 요인계수와 비슷한 조건부 확률이며 잠재집단 확률 모수는 각 집단의 상대적 크기에 상응하는 잠재집단 확률을 말한다. 문항반응 모수란 조건부 문항 확률인데 이는 해당 집단의 구성원이 해

당 문항을 지지할 확률에 대한 정보를 제공한다. 즉 한 구성원이 속해 있는 어떤 한 잠재집단이 있을 때 특정 변수에 대해 특정 범주에 속할 확률을 말한다. 이는 잠재집단의 특성을 식별하는데 도움이 될 수 있다(Hickendorff, Putten, Heiser, & Verhelst, 2005).

잠재계층분석의 이론적 근거는 지역 독립성이며 이는 관찰변수들의 확률함수가 잠재집단 변수(c개의 범주로 나뉘는 변수 L로 표기)의 수준에 조건부로 서로 통계적으로 독립적이라는 것을 의미한다(McCutcheon, 1987). 다시 말해서, 잠재변수의 효과를 통제한다면 관찰변수들 사이의 관계가 서로



(a) 지역독립성을 가정한 모형 (b) 지역독립성을 가정하지 않은 모형

<그림 8> 지역독립성을 가정한 모형과 그렇지 않은 모형
(Collins & Lanza, 2013)

독립적이라는 것이다. <그림 8>을 보면 지역독립성을 가정한 모형에서는 각 관찰변수와 잠재변수 간의 화살표는 있지만 관찰변수 간에 관계성을 보여주는 화살표는 없다. 반면에 지역독립성을 가정하지 않은 모형에는 관찰변수 x_2와 x_3간에 잠재변수 만을 통한 관련성이 아닌 이들의 오차 간의 관련성을

보여주고 있다. 따라서 지역 독립성 가정으로 인하여 잠재계층모형이 잠재집단 확률과 문항반응 확률의 곱으로 표현될 수 있음을 말한다(Hickendorff, Putten, Heiser, & Verhelst, 2005).

$$P(Y=y)=\sum_{c=1}^{C}\gamma_c\prod_{j=1}^{J}\prod_{r_j=1}^{R_j}\rho_{j,r_j|c}^{I(y_i=r_j)} \qquad \text{식 (2.3.3)}$$

y_j를 응답패턴 y의 구성요소 j라고 가정하고, 변수에 대한 응답이 $j=r_j$일 때 1이 되고 아닌 경우 0이 되는 표시함수 $I(y_i=r_j)$를 가정한다. 이 표시함수는 단지 서로 곱할 적절한 모수 ρ를 골라내기 위함이다. 식 (2.3.3)은 반응에 대한 특정 벡터가 관찰 될 확률이 잠재집단 확률과 문항반응 확률의 함수라는 것을 말해주고 있다. 즉 각 잠재집단에 속할 확률과 어떠한 잠재집단에 속한다는 조건 하에 변수에 대해 각각의 응답을 할 확률의 함수이다.

식(2.3.3) 의 두 번째 부분은, $\prod_{j=1}^{J}\prod_{r_j=1}^{R_j}\rho_{j,r_j|c}^{I(y_i=r_j)}$, 잠재집단 c에 속한다는 조건하에 특정관찰변수에 대한 응답 패턴 y의 확률을 말하고 있다.

$$P(Y=y|L=c)=\prod_{j=1}^{J}\prod_{r_j=1}^{R_j}\rho_{j,r_j|c}^{I(y_i=r_j)} \qquad \text{식 (2.3.4)}$$

따라서 식 (2.3.4)는 모두 잠재집단 c에 속한다는 전제하에, 관찰변수 1에 대한 반응 r_1의 확률, 관찰변수 2에 대한 반응 r_2의 확률, 즉 관찰변수 J에 대한 반응 r_J의 확률을 말한다.

각 잠재집단 c의 $P(Y=y|L=c)$를 계산 한 뒤 $P(Y=y)$를 계산하는 데, 이것은 각 잠재집단 C에 대한 잠재집단 c와 응답 패턴 y의 무조건 동시 확률(unconditional joint probability)이다. 일반적으로 동시확률은 식 (2.3.5)에 기초하여 계산할 수 있다.

$$P(AB) = P(A)P(B|A) \qquad \text{식 (2.3.5)}$$

그러므로, $\gamma_c = P(L=c)$ 이기 때문에 동시확률은 다음과 같다.

$$P(Y=y, L=c) = P(L=c)P(Y=y|L=c) = \gamma_c \prod_{j=1}^{J} \prod_{r_j=1}^{R_j} \rho_{j,r_j|c}^{I(y_j=r_j)}$$

식 (2.3.6)

동시확률이 계산된 뒤에는 잠재집단에 대한 동시확률의 합으로 $P(Y=y)$의 주변화를 할 수 있다.

$$P(Y=y) = \sum_{c=1}^{C} P(Y=y, L=c) \qquad \text{식 (2.3.7)}$$

그러므로 식 (2.3.6)을 식 (2.3.7)에 대입하면 식 (2.3.3)을 얻을 수 있다.

2) 모수 추정

잠재계층모형에 주어진 데이터를 적용시키기 위해서는 자료를 기반으로 위에서 언급한 잠재집단 확률 모수와 문항반응 확률 모수를 추정해야한다(Lanza et al., 2003). 일반적으로 사용되는 평균 μ이나 표준편차 σ 같은 통계적 모수는 간단하게 방정식을 풀어서 얻어낼 수 있다. 이런 것을 흔히 폐쇄형 해(closed-form solution)라고 한다. 하지만 잠재계층모형처럼 좀 더 복잡한 통계적 모형일 경우 모형 모수의 폐쇄형 해는 구할 수가 없기 때문에 모수 추정을 위해 반복적인 접근방법을 사용해야 한다(Lanza, et al., 2003). 따라서 잠재계층모형은 최대가능성(maximum likelihood) 모수 추정치를 찾기 위해서 또는 가능성 함수를 최대화시키기 위하여 반복적인 접근방법 중 Dempster, Laird, & Rubin (1977)이 제안한 기대치-최대화(Expectation-Maximization; EM)알고리즘 절차를 이용해 모수를 추정한다((Hickendorff et al., 2005; Lanza et al., 2003). EM 알고리즘은 잠재변수를 사용한 모형에서 최대가능성 추정치를 찾는 일반적인 기법이다(Bishop, 2006). 주어진 데이터와 모형에 대해서 가능성 함수는 모형 모수 추정치에 조건하에 경험적으로 관측된 데이터의 가능성을 나타낸다(Agresti, 2010). 즉 최대가능성 추정은 데이터가 가장 많이 관찰 될 수 있는 모수 값을 나타내는 것이다.

실제로 가능성 함수의 로그(log)를 사용하여 최소화 하는 것이 일반적으로 더 편리하다(Lanza, et al., 2003). 그러므로 로그 가능성(log likelihood)의 기대치를 계산하는 기대치 단계(Expectation stage; E)와 이 기대치를 최대로 만들어주는 변수 값을 구하는 최대화 단계(Maximization stage; M)를 번갈아가며 반복하는 것이다. 최대화 단계에서 얻은 변수의 값은 다시 다음 기대치 단계의 추정치로 쓰이며 일정한 값에 수렴할 때까지 이 단계가

반복된다(Mooijaart & Heijden, 1992).

관찰변수를 J, 잠재변수를 L, 그리고 알고자하는 모수의 벡터를 θ라 할 때 (J,L)에 대한 확률분포는

$$L(\theta;J,L) = p(J,L|\theta) \qquad \text{식 (2.3.8)}$$

로 주어진다. θ에 대한 최대 가능성 추정치(Maximum likelihood estimation; MLE)는 관찰변수의 주변화 가능성(marginal likelihood)을 사용하여 정의할 수 있다.

$$L(\theta;J) = p(X|\theta) = \sum_{L} p(J,L|\theta) \qquad \text{식 (2.3.9)}$$

하지만 잠재변수의 차원이 증가할수록 취할 수 있는 값의 수는 기하급수적으로 증가하기 때문에 이 수식을 계산하기는 어려워진다. 따라서 최대 가능성 추정치를 찾기 위해 EM 알고리즘을 사용하여 기대치 단계와 최대화 단계를 반복적으로 계산하는 것이다. 기대치(E) 단계에서는 모수 $\theta^{(t)}$의 θ를 추정할 때 가능성의 기대치 Q를 정의한다. 이때 잠재변수 L의 조건부 분포에 관하여 로그 가능성 함수의 기대치를 계산한다.

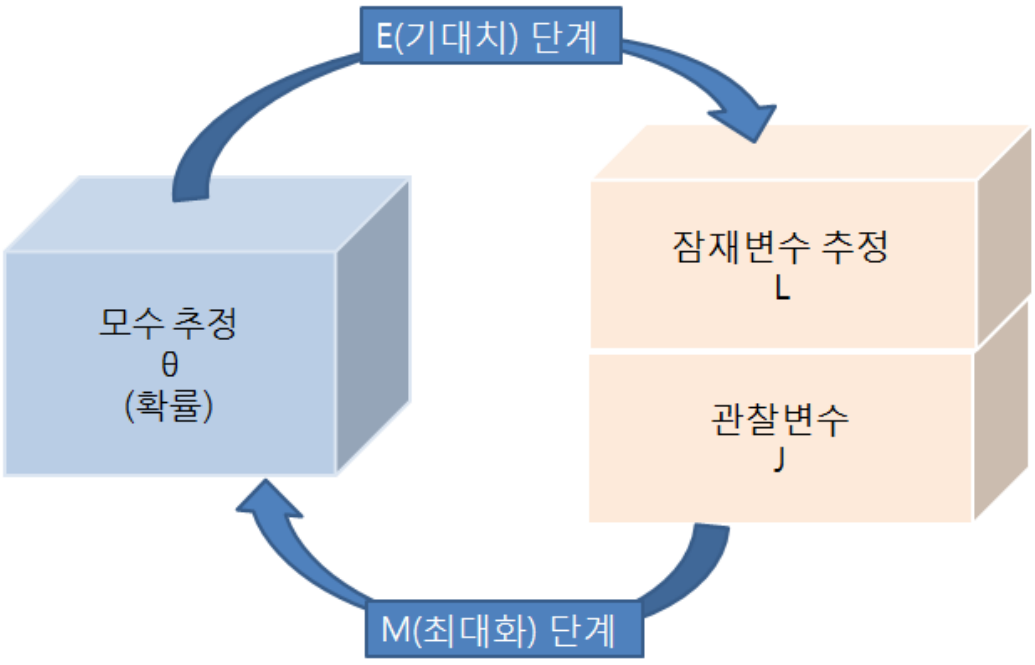
$$Q(\theta|\theta^{(t)}) = E_{L|J,\theta^{(t)}}[\log Likelihood(\theta;J,L)] \qquad \text{식 (2.3.10)}$$

그리고 최대화(M) 단계에서는 위에서 구한 Q를 최대화 하는 새로운 모수

$\theta^{(t+1)}$을 계산한다.

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}) \qquad \text{식 (2.3.11)}$$

그 후 식 (2.3.10)과 식(2.3.11)을 번갈아가며 반복적으로 계산하여 연구자가 정한 값에 수렴하게 되면 EM 알고리즘은 종료하게 된다(Zhu, 2005). <그림 9>을 참고하면 기대치(E)단계와 최대화(M)단계를 반복하는 것을 알 수 있다.



<그림 9> EM 알고리즘

하지만 잠재계층분석에서 모형의 모수를 추정하는 것만이 관심사는 아니다. 구성원이 어떤 군집으로 분류(classification)가 되어져야 하는지도 큰 관심분

야이다. 따라서 잠재계층분석에서 객체들의 분류가 직접적으로 추정되지는 않지만 추정된 모수들(잠재집단 확률과 문항반응 확률)로부터 이러한 분할(partitioning)을 도출해 낼 수 있다. 관찰변수에 대한 응답패턴이 주어졌을 때 이러한 추정치들로부터 어떠한 한 객체가 잠재집단 c에 속할(class-membership) 사후확률을 알아낼 수 있다(Vermunt & Magidson, 2002).

$$\pi_{c|r_1,r_2,r_3}^{L|j_1j_2j_3} = \frac{\pi_c^L \times \pi_{r_1|c}^{j_1|L} \times \pi_{r_2|c}^{j_2|L} \times \pi_{r_3|c}^{j_3|L}}{\sum_c \pi_c^L \times \pi_{r_1|c}^{j_1|L} \times \pi_{r_2|c}^{j_2|L} \times \pi_{r_3|c}^{j_3|L}} \qquad \text{식 (2.3.12)}$$

그 후에는 구성원이 가장 높은 사후 확률을 갖는 잠재집단에 할당되는 것이다. 이로 인해서 확률적인 분류가 이루어지기 때문에 불확실 정도를 평가할 수도 있다.

3) 잠재계층분석의 쟁점

위에서 언급한 바와 같이 잠재계층모형에서는 잠재집단 확률 모수와 문항반응 확률 모수를 추정해야 한다. 모형은 EM 알고리즘이라는 반복적인 최대 가능성 절차에 의해 추정이 되는데 이는 지역 최적해로 인해 어려움을 겪을 수 있다.

또한 모형에 변수 또는 변수의 범주가 너무 많이 투입되는 경우 모형의 식별(identification of the model)이 문제가 될 수 있다. 주어진 데이터에 대해 너무 많은 모수를 추정하게 되면서 모형을 식별할 수 없게 되는 것이다(Collins, Fidler, & Wugalter, 1996). 또한 이런 경우의 가능성 통계치는 χ^2 분포를 따르지 않아서 모형 합치도 평가를 할 때 어려움이 있을 수 있다

(Collins et al., 1996). 이러한 문제를 해결하기 위해 동일성 제약(equality constraints)이나 특정 값에 제약을 거는 등 모수에 제약을 거는 방법이 있다.

모델의 식별 및 선택은 잠재계층분석에 관한 주요한 연구문제 중 하나이다 (Vermunt & Magidson, 20002). 잠재계층분석은 K-평균 군집분석에 비해서 통계적 지표를 활용하기 때문에 분류의 질(quality)를 평가할 수 있다는 장점이 있다. 하지만 잠재계층분석이 주는 이점에도 불구하고 군집의 수를 결정(class enumeration)할 때 일반적으로 받아들여지는 통계적 지표(statistical indicator)가 아직 없다는 쟁점이 있다. 따라서 모델을 평가하고 선택하기 위해, 즉 군집의 수를 결정하기 위해 몇 가지 기준들을 조합하여 주로 사용하고 있다(Nylund et al., 2007). 추정하는 모수의 개수가 늘어날수록 패널티를 더 부과하는 Akaike's Information Criterion (AIC; Akaike, 1987)이나 Bayesian Information Criterion (BIC; Schwartz, 1978)같은 통계적 정보기준(information criteria)이 주된 예이다. 이밖에도 가능성 통계치 χ^2, 모델이 통계적으로 데이터에 적합한지의 여부에 대해 검증할 수 있는 L_2, 군집의 개수와 그 형태를 결정하기 위해 부트 스트래핑(McLachlan et al., 1999)과 같은 방법을 쓰는 경우도 있고 Markov chain Monte Carlo (Bensmail et al., 1997) 방법을 사용하기도 한다. 이러한 통계적 기준들은 상이한 모형을 비교할 수 있고 분류의 질을 평가할 수 있다는 장점을 가진다(Hickendorff et al., 2005).

하지만 지금까지의 여러 가지 제안에도 불구하고 잠재계층모형에서 집단의 수를 결정하기 위한 최상의 기준은 공통적으로 받아들여지지 않고 있다 (Nylund et al., 2007). 결과를 해석하고 추론하는 데 군집이 사용되기 때문에 잠재계층모형을 실제 데이터에 적용하기 위해서는 이러한 통계적 지수의 올바른 선택과 사용이 중요한 문제이다(Nylund et al., 2007).

또한 한 구성원이 주어진 각 잠재집단에 속할 사후확률이 모두 같다면 어떤 집단에 할당 되어야 하는 것이 옳은지에 대한 연구는 아직 진행된 바

없다. 또는 각 잠재집단에 속할 사후확률이 서로 미세하게 차이가 나는 경우, 그 미세한 차이 때문에 어느 한 집단에 할당이 되는 것이 적절한 선택인지에 관한 연구 또한 미비한 실정이다. 그에 관련하여 어느 정도의 차이가 미세한지 또는 적절한 차이인지에 관한 연구도 많이 이루어지지 않았다.

이에 따라 Nylund, Asparouhov, 그리고 Muthén(2007)은 잠재계층모형에서 군집 수를 결정하는 데 사용되는 통계적 정보기준(Information Criterion; ICs)의 성능을 검사하는 시뮬레이션 연구를 진행하였다. 사례 수, 문항 수, 군집의 개수, 잠재집단 확률 그리고 문항반응 확률을 조작변수로 사용하여 어떤 정보기준이 군집의 개수를 가장 잘 식별하는지에 대해 연구하였다. 이밖에도 Steinley 와 Brusco(2011)도 시뮬레이션 연구를 통해 잠재집단분석과 K-평균군집분석을 비교분석하였다.

III. 연구문제 및 가설

연구문제1: 잠재계층에서 조작변수 수준에 따라 군집회복(cluster recovery) 정도의 차이가 있는가?

[가설 1-1] 사례수의 수준에 따라서 잠재계층분석 결과의 군집회복 정도의 차이가 있다.

[가설 1-2] 문항수의 수준에 따라서 잠재계층분석 결과의 군집회복 정도의 차이가 있다.

[가설 1-3] 군집수의 수준에 따라서 잠재계층분석 결과의 군집회복 정도의 차이가 있다.

[가설 1-4] 효과크기의 수준에 따라서 잠재계층분석 결과의 군집회복 정도의 차이가 있다.

[가설 1-5] 잠재집단확률의 수준에 따라서 잠재계층분석 결과의 군집회복 정도의 차이가 있다.

연구문제2: K-평균군집분석에서 조작변수 수준에 따라 군집회복(cluster recovery) 정도의 차이가 있는가?

[가설 1-1] 사례수의 수준에 따라서 K-평균군집분석 결과의 군집회복 정도의 차이가 있다.

[가설 1-2] 문항수의 수준에 따라서 K-평균군집분석 결과의 군집회복 정도의 차이가 있다.

[가설 1-3] 군집수의 수준에 따라서 K-평균군집분석 결과의 군집회복 정도의 차이가 있다.

[가설 1-4] 효과크기의 수준에 따라서 K-평균군집분석 결과의

군집회복 정도의 차이가 있다.

[가설 1-5] 잠재집단확률의 수준에 따라서 K-평균군집분석 결과의 군집회복 정도의 차이가 있다.

연구문제3: 각 조작변수가 군집회복 정도에 미치는 영향이 어떤 분석방법을 쓰느냐에 따라 달라지는가?

[가설 3-1] 사례수 수준에 따른 군집회복의 차이는 분석방법에 따라 차이가 있다.

[가설 3-2] 문항수 수준에 따른 군집회복의 차이는 분석방법에 따라 차이가 있다.

[가설 3-3] 군집수 수준에 따른 군집회복의 차이는 분석방법에 따라 차이가 있다.

[가설 3-4] 효과크기 수준에 따른 군집회복의 차이는 분석방법에 따라 차이가 있다.

[가설 3-5] 잠재집단확률 수준에 따른 군집회복의 차이는 분석방법에 따라 차이가 있다.

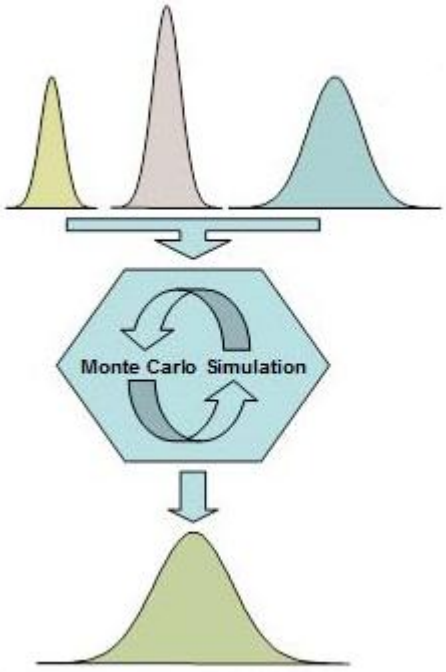
연구문제4: 잠재계층분석을 실시한 경우와 K-평균 군집분석을 실시한 경우, 군집회복 정도에 차이가 있는가?

[가설 4-1] 잠재계층분석을 실시한 경우와 K-평균군집분석을 실시한 경우, 군집회복 정도에 차이가 있다.

IV. 연구방법

1. 자료 생성 (data generation)

프로그램 R에서 몬테카를로 시뮬레이션(Monte Carlo Simulation)[28] 기능을 사용하여 자료를 생성한 후 K-평균 군집분석과 잠재계층분석을 실시하였다.



<그림 10> 몬테카를로 시뮬레이션
(Carvalho, Stubstad, Briggs,
Selezneva, Mustafa, &
Ramachandran, 2012)

28) 몬테카를로 시뮬레이션(Monte Carlo Simulation)은 모의실험의 한 형태로서 불확실한 상황에서 의사결정을 하기 위해 확률적 분포를 따르는 자료를 무작위(random)로 생성하는 방법이다(Stephenson, Holbert, 2003).

확정모형(deterministic model)은 변수들 간의 관계가 확실하여 방정식을 사용하여 모형을 만들고 수학적으로 변수에 대한 해를 찾아 예측치를 정확하게 찾을 수 있는 모형이다. 반면에 확률모형(probability model)은 결과를 정확하게 예측할 수 없는 모형이기 때문에 확률모형의 모수(parameter)에 대해 반복적으로 여러 수치적인 방법(numerical method)을 시도하여 확률변수의 분포를 얻어내게 된다. 이러한 분포를 생성하기 위해 반복적으로 이용하는 수치를 일련의 난수(random number)로부터 얻을 때 몬테카를로 시뮬레이션을 이용하게 된다(Stephenson, Holbert, 2003). <그림 10>과 같이 여러 확률분포 함수로부터 대응되는 확률변수를 추출하여 모델에 적용하고 반복 실행된 결과 값들이 분포를 지어 보이는 것이다.

몬테카를로 시뮬레이션 연구에서는 두 가지 유형의 사양을 다르게 지정할 수 있는데 첫 번째는 몬테카를로 변수(Monte Carlo Variable)이며 두 번째는 모집단의 사양(population specification), 즉 조작요인(manipulated factors)이다. 몬테카를로 변수는 표본의 사례와 반복횟수(replication)를 말하는데 본 연구에서 표본의 사례 수는 n=200, 500, 그리고 1,000으로 지정하며 주어진 조건을 각각 500번 씩 반복하였다. 조작요인으로는 문항의 수(number of items), 군집의 개수(number of clusters), 효과크기 (effect size), 잠재집단확률(latent class probability)을 포함하였다. 그리고 문항에 대하여 이분법 적인 응답(예 또는 아니오)을 포함한 모형에서 교차부하(cross loading)가 없는 문항반응 확률을 사용하여 집단을 구별하였다.

1) 몬테카를로 시뮬레이션 (Monte Carlo simulation)

본 연구에서는 네 가지 조작요인이 표본 데이터 생성을 결정하게 된다. 이를 위해 200, 500, 그리고 1,000의 사례수, 문항수와 군집수의 비율이 각각 1:1,

2:1, 3:1인 경우, 2, 4, 6개의 군집의 수, 군집간의 평균의 차이 즉 효과크기가 0.2, 0.5, 0.8로 각각 다르게 설정한 세 가지 상황과, 잠재집단확률이 10%, 균등, 60%로 분배되는 세 가지 상황을 가지는 3(N)×3(V)×3(K)×3(F)×3(γ)의 요인설계(factorial design)를 바탕으로 자료를 생성하였다. 따라서 243개의 조건은 각 500번씩 반복 될 것이므로 모두 121,500개의 자료를 생성하였다.

지정한 문항반응 확률은 교차부하(cross loading)가 일어나지 않은 단순한 구조로 하나의 문항반응 확률이 하나의 잠재집단에서만 높거나 또는 낮게 나타나는 경우이다. 따라서 이 문항들이 집단들 끼리를 구별하게 되는 것이다. 요인분석(factor analysis)의 관점에서 보았을 때 교차부하가 없이 각 요인을 식별하는 고유한 문항이 있는 것이다(Nylund et al., 2007). 그리고 사례 수는 Nylund, Asparouhov과 Muthén(2007)의 연구에서와 같이 작은 사례수, 중간 사례수, 많은 사례수를 구분하기 위하여 200, 500 그리고 1000으로 조작하였다.

첫 번째 조작변수인 문항의 수는 V=4, 6, 8, 그리고 10이 관측치들을 군집화하고자 하는 연구에서 일반적으로 쓰이는 범위(Milligan, 1996)이지만, 군집의 수에 따라 문항의 수를 다르게 지정하기 위하여 이들의 비율을 사용하였다. 따라서 문항수와 군집수의 비율을 1:1, 2:1, 3:1 로 지정하였다. 두 번째 조작변수인 군집의 수 2, 4, 6개로 조작하였다. 이러한 범위의 상한값을 6로 정한 것은 8개 이상의 군집은 관련한 문헌연구가 거의 없기 때문이고, 하한값을 2로 정한 것은 군집 2개의 관한 연구가 진행되어 반복 검증해 보는 것의 필요성이 제기되었기 때문이다(Knotkova, Clark, Keonha, Kuhl, Winer, & Wharton, 2006; Pastor, Baarron Miller, & Davis, 2007; Richardson, 2007; Velicer, Redding, Anatchkova, Fava, & Prochaska, 2007). 세 번째 조작변수인 효과크기는 집단 간의 평균 차이를 일컬으며, 즉 집단이 서로 얼마나 다른지를 말해준다. 이는 0.2, 0.5, 0.8 세 가지 수준으로 조작하였다. 네 번째 조작변수인 잠

재집단 확률이란 군집의 상대적 크기를 말하며 즉 전체 집단에서 특정 군집이 발생할 확률을 일컫는다(Steinley & Brusco, 2011). 이 요소는 세 가지 수준으로 나뉘며 (a) 모든 군집이 같은 개수의 관측치를 갖는 경우, (b) 하나의 군집이 전체 관측치 60%를 갖고 나머지 관측치들은 나머지 군집에 똑같이 분배된 경우로 하나의 크고 우세한 군집 하나와 나머지 작은 소집단들로 이루어진 경우와 (c) 하나의 군집이 전체 관측치 10%를 갖고 나머지 관측치들은 나머지 군집에 똑같이 분배된 경우로 군집 하나의 기본 구성 비율이 다른 군집들보다 훨씬 작은 경우이다(Milligan, 1980; Steinly, 2003; Steinly, 2006).

2) 군집회복지수(Adjusted Rand Index)

근래에 전산 기술이 도입된 이후로 다양한 군집화 절차와 종류가 기하급수적으로 증가하였다(Steinley, 2004). 또한 어떤 군집화 방법이 가장 적합한지 결정하기 위하여 다양한 절차들끼리의 비교분석도 많이 진행되었다(Steinley, 2004). 이러한 비교 연구는 재 생성된 구조(recovered structure)와 연구자가 원하는 특성을 가진 생성된 구조(generated structure)를 비교하는 몬테카를로 연구를 통해 이루어졌다. 그 후 군집회복지표(cluster recovery index)를 사용하여 둘 사이의 유사성을 알아보았다(Brusco & Cradit, 2001; Milligan, 1980; Milligan & Copper, 1985). 연구에서 가장 많이 사용되는 군집회복지표 5가지는 Rand index(Rand, 1971), Morey & Agresti(1984)의 adjusted Rand index, Hubert & Arabie(1985)의 adjusted Rand index, Jaccard index (Downton & Brennan, 1980), 그리고 Fowlkes & Mallows (1983)의 지표이다. 즉 Rand Index는 두 가지의 군집화의 유사성을 측정하는 지표로 두 가지 군집화가 완벽하게 일치하면 그 값이 1이 되며 그 반대는 0이 된다.

Milligan (1996)에 의하면 군집화 기법의 검증을 위해서는 연구자가 데이터

의 정확한 구조를 사전적으로 알아야하기 때문에 데이터를 직접 생성하고 몬테카를로 시뮬레이션을 통해 검증을 하게 된다. 자료 생성과 자료 분석이 끝나면 재생산된 군집을 이미 알고 있는 데이터와 비교하게 된다(Steinley, 2004). N이 사례 수이고 V가 변수의 수 일 때, 데이터 행렬은 $X=\{x_{ij}\}_{N\times V}$로 주어진다. 그리고 N개의 객체를 R개의 소집단(subset)즉 군집으로, $P=\{p_1,\cdots,p_R\}$, 분할하게 된다. 모든 군집을 통합하면 모든 객체의 집합이 되고 P 분할에 있는 어느 두 개의 군집의 교차점은 비어있게 된다.

R 군집을 가진 분할 P와 C 군집을 가진 분할 Q 두 개가 있다고 가정하자. 그러면 P와 Q가 겹치는 부분을 나타내기 위해 <표 1>처럼 T를 만들 수 있다.

<표 1> P와 Q 분할의 공통부분 (Steinley, 2004)

		Q				
분할	군집	q_1	q_2	$\cdots$	q_C	Total
P	p_1	t_{11}	t_{12}	$\cdots$	t_{1C}	t_{1+}
	p_2	t_{21}	t_{22}	$\cdots$	t_{2C}	t_{2+}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	p_R	t_{R1}	t_{R2}	$\cdots$	t_{RC}	t_{R+}
Total		t_{+1}	t_{+2}	$\cdots$	t_{+C}	$t_{++}=N$

<표 1>은 P와 Q의 겹치는 군집을 보여주는 T를 말하며 <표 1>에서 포괄적인 t_{rc}는 분할 R의 r번째 군집과 분할 C의 c번째 군집에 할당된 객체의 수를 말한다(Steinley, 2004). ARI는 객체의 쌍들이 T로 어떻게 할당되는지에 기초하여 P와 Q 사이의 관련성을 측정하는 것이다.

$\binom{N}{2}$가 네 가지의 다른 쌍(pairs)으로 나타나는 모든 쌍의 개수를 나타낸다고

하자. 첫 번째는 짝지어진 객체 둘이 P 에서도 같은 군집에 있고 Q에서도 같은 군집에 있는 경우, 두 번째는 짝지어진 객체 둘이 P에서는 같은 군집에 있지만 Q에서는 다른 군집에 있는 경우, 세 번째는 짝지어진 객체 둘이 P에서는 다른 군집에 있지만 Q에서는 같은 군집에 있는 경우, 그리고 마지막은 짝지어진 객체 둘이 P에서도 다른 군집에 있고 Q에서도 다른 군집에 있는 경우이다. 그렇다면 이러한 네 가지 경우를 기반으로 T를 2×2 표로 다음 <표 2> 처럼 나타낼 수 있다.

<표 2> 네 가지 다른 쌍(pairs) (Steinley, 2004)

	Q	
P	Pair in same group	Pair in different groups
Pair in same group	a	b
Pair in different groups	c	d

그리고 T를 R×C 행렬로 나타낸다면, $T=\{t_{rc}\}$, 네 가지 경우는 다음과 같이 계산 될 수 있다.

$$a = \frac{\sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2 - N}{2} \qquad \text{식 (4.2.1)}$$

$$b = \frac{\sum_{r=1}^{R} t_{r+}^2 - \sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2}{2} \qquad \text{식 (4.2.2)}$$

$$c = \frac{\sum_{c=1}^{C} t_{+c}^2 - \sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2}{2} \qquad \text{식 (4.2.3)}$$

$$d = \frac{\sum_{r=1}^{R}\sum_{c=1}^{C} t_{rc}^2 + N^2 - \sum_{r=1}^{R} t_{r+}^2 - \sum_{c=1}^{C} t_{+c}^2}{2} \qquad \text{식 (4.2.4)}$$

Rand index (Rand, 1971)는 군집을 확인(validation)하는 몇몇 연구에서 사용된 지표다(Dreger, Fuller, & Lemine, 1988; Dubes & Jain, 1976; Milligan, 1980; Milligan & Isaac, 1980; Milligan, Soon & Sokol, 1983). 그리고 다음과 같이 계산된다.

$$Rand = \frac{a+d}{a+b+c+d} \qquad \text{식 (4.2.5)}$$

식 (4.2.5)는 P와 Q에 동시에 함께 있거나 또는 따로 분할된 객체들에게 가중치를 주는 것이다. 하지만 불행하게도 Rand는 군집의 수가 증가함에 따라 Rand의 통계치가 상한선인 1에 가까워진다는 것을 밝혔다. 이러한 한계를 극복하기 위한 시도로 몇 가지 다른 측정방법이 소개되었다(Steinley, 2004).

그 중 Hubert & Arabie(1985)가 소개한 Adjusted Rand idex는 식 (4.2.5)를 다음과 같이 수정하였다.

$$ARI_{HA} = \frac{\binom{N}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{N}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]}$$

식 (4.2.6)

Hubert 와 Arabie(1985)의 Adjusted Rand Idex의 범위는 ±1이 되며 이것은 우연에 의한 오차를 수정한 영가설 아래에서 추정되기 때문에 확률적 해석의 이점을 가진다. Hubert 와 Arabie(1985)의 ARI에서 1은 완벽한 분류의 정확성을 뜻하며 -1은 완벽한 분류의 부정확성을 뜻한다. 따라서 양수의 ARI는 분류방법이 통계적, 실질적 이득을 제공하여 분류의 정확성이 높아짐을 의미하며 음수의 ARI는 분류의 부정확성이 높아짐을 말한다. 또한 0은 영가설 하에서의 기댓값(expected value)로서 우연에 의해서 발생되는 분류의 정확성을 뜻한다.

본 연구에서는 Hubert 와 Arabie(1985)의 Adjusted Rand idex를 사용하여 각 조작변수에 따라 군집의 회복 정도를 비교 분석하였다. 하지만 Hubert 와 Arabie 가 소개한 ARI의 표준화된 절대적 기준이 제공되지 않기 때문에 그 값이 1에 가까울수록 군집화 간의 높은 유사성을 보이며, -1에 가까울수록 그 반대가 된다는 것을 기반으로, 각각의 조작변수에 따라 ARI값을 상대적으로 비교하였다. 그 후 잠재집단분석과 K-평균분석 간에 군집 회복의 정도에서 차이가 있는지를 검증하였다.

2. 자료 분석

연구 1에서는 생성된 자료로 잠재계층분석을 실시한 후, 각 조작변수의 수준에 따라서 군집회복의 정도에 차이가 있는지 검증해보기 위해 SPSS에서 일원분산분석을 실시하였다. 또한 연구 2의 K-평균군집분석 결과 각 조작변수의 수준에 따라 군집회복 정도에 차이가 있는지 검증해보기 위해 SPSS에서 일원분산분석을 실시하였다. 연구 3에서는 각 조작변수와 분석방법 간에 상호작용 효과가 있는지 확인해보기 위해 SPSS에서 이원분산분석을 실시하여 검증하였다. 마지막으로 군집회복 정도에서 잠재계층분석과 K-평균군집분석 간에 차이가 있는지 검증하기 위해 SPSS를 사용하여 대응표본 t검증을 실시하였다.

V. 연구결과

1. 연구1: 잠재계층분석의 군집회복 정도

연구1은 잠재계층분석을 실시한 결과, 각 조작변수의 수준에 따라서 군집회복 정도의 차이가 있는지 검증해 보기 위한 목적으로 진행되었다. R의 몬테카를로 시뮬레이션을 사용하여 자료를 생성하고 생성자료를 대상으로 R의 e1071 패키지를 사용하여 잠재계층분석을 실시하였다. 조건별로 시행된 잠재계층분석의 군집회복 ARI 지수는 SPSS에서 일원분산분석으로 분석하였다.

각 조작변수의 수준에 따른 군집회복의 정도를 살펴보기 위하여, 사례수는 200, 500, 1000으로, 문항수와 군집수의 비율은 1:1, 2:1, 3:1으로, 군집수는 2, 4, 6으로, 효과크기는 0.2, 0.5, 0.8로 마지막으로 잠재집단확률은 .1, .균등, .6으로 조작하였다. 즉, 연구 1은 3(사례수) x 3(문항수) x 3(군집수) x 3(효과크기) x 3(잠재집단확률) 요인설계로 모두 243개의 조건으로 구성되어 있으며, 설정된 각 조건은 500번씩 반복되었다. 그 결과 산출된 군집회복 정도의 지수인 ARI 지수의 수행을 분석하였다.

1) 사례수에 따른 군집회복

사례수의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 3-1>에 제시하였다. 사례수가 200인 경우 군집회복 정도의 평균은 .0384, 표준편차는 .0373이고, 사례수가 500인 경우 군집회복 정도의 평균은 .0405, 표준편차는 .0369이며, 사례수가 1000인 경우 군집회복 정도의

평균은 .0428, 표준편차는 .0377로 관찰되었다. 기술통계 결과 사례수가 커짐에 따라 군집회복 정도가 커지는 것으로 관찰되었다.

<표 3-1> 사례수 수준에 따른 잠재계층분석 군집회복의 기술통계

	200	500	1000	합계
사례수	81	81	81	243
평균	.0384	.0405	.0428	.0406
표준편차	.0373	.0369	.0377	.0372

사례수 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 3-2>에 제시하였다. 그 결과, 사례수 수준에 따른 군집회복의 정도에는 유의한 차이가 없었다, $F(2,240)=.277$, $p=.759$.

<표 3-2> 사례수 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
사례수	.001	2	.000	.277	.759
오차	.334	240	.001		
합계	.335	242			

2) 문항수에 따른 군집회복

문항수의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 3-3>에 제시하였다. 문항수와 군집수의 비율이 1:1인 경우 군집회복 정도의 평균은 .0309, 표준편차는 .0361이고, 문항수와 군집수의 비율이 2:1인 경우 군집회복 정도의 평균은 .0426, 표준편차는 .0364이며, 문항수와 군집수의 비율이 3:1인 경우 군집회복 정도의 평균은 .0481, 표준편차는 .0375으로 관찰되었다. 기술통계 결과 문항수가 늘어남에 따라 군집회복

정도가 커지는 것으로 관찰되었다.

<표 3-3> 문항수 수준에 따른 잠재계층분석 군집회복의 기술통계

	1:1	2:1	3:1	합계
사례수	81	81	81	243
평균	.0309	.0426	.0481	.0406
표준편차	.0361	.0364	.0375	.0372

문항수 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 3-4>에 제시하였다. 그 결과, 문항수 수준에 따른 군집회복의 정도에는 유의한 차이가 있었다, $F(2,240)=4.638$, $p=.011$.

<표 3-4> 문항수 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
문항수	.012	2	.006	4.638	.011
오차	.323	240	.001		
합계	.335	242			

사후분석인 Scheffe검증 결과는 <표 3-5>에 제시하였다. 문항수와 군집수의 비율이 1:1인 경우 군집회복 정도(평균=.0609, 표준편차=.0361)와 문항수와 군집수의 비율이 2:1인 경우 군집회복 정도(평균=.0426, 표준편차=.0364)와의 차이가 없었다. 또한 문항수와 군집수의 비율이 2:1인 경우 군집회복 정도와 문항수와 군집수의 비율이 3:1인 경우 군집회복 정도(평균=.0481, 표준편차=.0375)의 차이가 없었다. 하지만, 문항수와 군집수의 비율이 1:1인 경우는 문항수와 군집수의 비율이 3:1인 경우보다 군집회복 정도가 낮았다.

<표 3-5> 문항수 수준에 따른 잠재계층분석 사후분석 결과

문항비율	평균차	표준오차	유의확률
1:1 vs 2:1	−.0116	.0058	.132
1:1 vs 3:1	−.0172	.0058	.013
2:1 vs 3:1	−.0055	.0058	.630

3) 군집수에 따른 군집회복

군집수의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 3-6>에 제시하였다. 군집수가 2개인 경우 군집회복 정도의 평균은 .0267, 표준편차는 .0252이고, 군집수가 4개인 경우 군집회복 정도의 평균은 .0243, 표준편차는 .0471이며, 군집수가 6개인 경우 군집회복 정도의 평균은 .0069, 표준편차는 .0636으로 관찰되었다. 기술통계 결과 군집수가 4개인 경우, 군집회복 정도가 가장 높은 것으로 관찰되었다.

<표 3-6> 군집수 수준에 따른 잠재계층분석 군집회복의 기술통계

	2	4	6	합계
사례수	81	81	81	243
평균	.0337	.0457	.0423	.0406
표준편차	.0292	.0365	.0439	.0372

군집수 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 3-7>에 제시하였다. 그 결과, 군집수 수준에 따른 군집회복의 정도에는 유의한 차이가 없었다, $F(2,240)=2.271$, $p=.105$.

<표 3-7> 군집수 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
군집수	.006	2	.003	2.271	.105
오차	.329	240	.001		
합계	.335	242			

4) 효과크기에 따른 군집회복

효과크기의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 3-8>에 제시하였다. 효과크기가 0.2씩 차이가 나는 경우 군집회복 정도의 평균은 .0198, 표준편차는 .0268이고, 효과크기가 0.5씩 차이가 나는 경우 군집회복 정도의 평균은 .0369, 표준편차는 .0273이며, 효과크기가 0.8씩 차이가 나는 경우 군집회복 정도의 평균은 .0649, 표준편차는 .0409으로 관찰되었다. 기술통계 결과 효과크기가 커짐에 따라 군집회복 정도가 커지는 것으로 관찰되었다.

<표 3-8> 효과크기 수준에 따른 잠재계층분석 군집회복의 기술통계

	0.2	0.5	0.8	합계
사례수	81	81	81	243
평균	.0198	.0369	.0649	.0406
표준편차	.0268	.0273	.0409	.0372

효과크기 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 3-9>에 제시하였다. 그 결과, 효과크기 수준에 따른 군집회복의 정도에는 유의한 차이가 있었다, $F(2,240)=40.227$, $p<.001$.

<표 3-9> 효과크기 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
효과크기	.084	2	.042	40.227	<.001
오차	.251	240	.001		
합계	.335	242			

사후분석인 Scheffe검증 결과는 <표 3-10>에 제시하였다. 효과크기가 0.2인 경우 군집회복 정도(평균=.0198 표준편차=.0268)가 가장 낮았으며 효과크기가 0.5인 경우(평균=.0369 표준편차=.0273)가 중간, 또한 효과크기가 0.8인 경우 군집회복 정도(평균=.0649, 표준편차=.0409)가 가장 높았다.

<표 3-10> 효과크기 수준에 따른 잠재계층분석 사후분석 결과

효과크기	평균차	표준오차	유의확률
0.2 vs 0.5	−.0171	.0051	.004
0.5 vs 0.8	−.0280	.0051	<.001
0.2 vs 0.8	−.0451	.0051	<.001

5) 잠재집단확률에 따른 군집회복

잠재집단확률의 수준에 따른 군집회복정도에 대한 사례수, 평균, 그리고 표준편차를 <표 3-11>에 제시하였다. 잠재집단확률이 한 집단에 10%가 모여 있는 경우 군집회복 정도의 평균은 .0380, 표준편차는 .0346이고, 잠재집단확률이 고르게 분배된 경우 군집회복 정도의 평균은 .0548, 표준편차는 .0366이며, 잠재집단확률이 한 집단에 60%가 모여 있는 경우 군집회복 정도의 평균은 .0289 표준편차는 .0361로 관찰되었다. 기술통계 결과 잠재집단확률이 균등하게 분배된 경우의 군집회복 정도가 가장 큰 것으로 관찰되었다.

<표 3-11> 잠재집단확률 수준에 따른 잠재계층분석 군집회복의 기술통계

	10%	even	60%	합계
사례수	81	81	81	243
평균	.0380	.0548	.0289	.0406
표준편차	.0346	.0366	.0361	.0372

잠재집단확률 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 3-12>에 제시하였다. 그 결과, 잠재집단확률 수준에 따른 군집회복의 정도에는 유의한 차이가 있었다, $F(2,240)=10.958$, $p<.001$.

<표 3-12> 잠재집단확률 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
잠재집단확률	.028	2	.014	10.958	<.001
오차	.307	240	.011		
합계	.335	242			

사후분석인 Scheffe검증 결과는 <표 3-13>에 제시하였다. 잠재집단확률이 10%인 경우 군집회복 정도와(평균=.0380, 표준편차=.0346) 잠재집단확률이 60%로 분배된 경우 군집회복 정도(평균=.0289, 표준편차=.0361)는 동일하였다. 하지만 잠재집단확률이 고르게 분배된 경우 군집회복 정도가 가장 높았다.

<표 3-13> 잠재집단확률 수준에 따른 잠재계층분석 사후분석 결과

잠재집단확률	평균차	표준오차	유의확률
10% vs even	−.0168	.0056	.012
even vs 60%	.0259	.0056	<.001
10% vs 60%	.0090	.0056	.272

2. 연구2: K-평균 군집분석의 군집회복 정도

연구2는 K-평균 군집분석을 실시한 결과, 각 조작변수의 수준에 따라서 군집회복 정도의 차이가 있는지 검증해 보기 위한 목적으로 진행되었다. R의 몬테카를로 시뮬레이션을 사용하여 자료를 생성하고 생성자료를 대상으로 R의 cluster 패키지를 사용하여 잠재계층분석을 실시하였다. 조건별로 시행된 K-평균군집분석의 군집회복은 ARI 지수는 SPSS에서 일원분산분석으로 분석하였다.

각 조작변수의 수준에 따른 군집회복의 정도를 살펴보기 위하여, 사례수는 200, 500, 1000으로, 문항수와 군집수의 비율은 1:1, 2:1, 3:1으로, 군집수는 2, 4, 6으로, 효과크기는 0.2, 0.5, 0.8로 마지막으로 잠재집단확률은 .1, 균등, .6으로 조작하였다. 즉, 연구 1은 3(사례수) x 3(문항수) x 3(군집수) x 3(효과크기) x 3(잠재집단확률) 요인설계로 모두 243개의 조건으로 구성되어 있으며, 설정된 각 조건은 500번씩 반복되었다. 그 결과 산출된 군집회복 정도의 지수인 ARI 지수의 수행을 분석하였다.

1) 사례수에 따른 군집회복

사례수의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 4-1>에 제시하였다. 사례수가 200인 경우 군집회복 정도의 평균은 .0189, 표준편차는 .0498이고, 사례수가 500인 경우 군집회복 정도의 평균은 .0192, 표준편차는 .0487이며, 사례수가 1000인 경우 군집회복 정도의 평균은 .0198, 표준편차는 .0478로 관찰되었다. 기술통계 결과 사례수가 커짐에 따라 군집회복 정도가 커지는 것으로 관찰되었다.

<표 4-1> 사례수 수준에 따른 K-평균 군집분석 군집회복의 기술통계

	200	500	1000	합계
사례수	81	81	81	243
평균	.0189	.0192	.0198	.0193
표준편차	.0498	.0487	.0478	.0486

사례수 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 4-2>에 제시하였다. 그 결과, 사례수 수준에 따른 군집회복의 정도에는 유의한 차이가 없었다, $F(2,240)=.009$, $p=.992$.

<표 4-2> 사례수 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
사례수	.000	2	.000	.009	.992
오차	.571	240	.002		
합계	.571	242			

2) 문항수에 따른 군집회복

문항수의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 4-3>에 제시하였다. 문항수와 군집수의 비율이 1:1인 경우 군집회복 정도의 평균은 .0170, 표준편차는 .0492이고, 문항수와 군집수의 비율이 2:1인 경우 군집회복 정도의 평균은 .0190, 표준편차는 .0477이며, 문항수와 군집수의 비율이 3:1인 경우 군집회복 정도의 평균은 .0219, 표준편차는 .0493으로 관찰되었다. 기술통계 결과 문항수가 늘어남에 따라 군집회복 정도가 커지는 것으로 관찰되었다.

<표 4-3> 문항수 수준에 따른 *K*-평균 군집분석 군집회복의 기술통계

	1:1	2:1	3:1	합계
사례수	81	81	81	243
평균	.0170	.0190	.0219	.0193
표준편차	.0492	.0477	.0493	.0486

문항수 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 4-4>에 제시하였다. 그 결과, 문항수 수준에 따른 군집회복의 정도에는 유의한 차이가 없었다, $F(2,240)=.209$, $p=.811$.

<표 4-4> 문항수 수준에 따른 *K*-평균 군집분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	*F*	유의확률
문항수	.001	2	.000	.209	.811
오차	.570	240	.002		
합계	.571	242			

3) 군집수에 따른 군집회복

군집수의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 4-5>에 제시하였다. 군집수가 2개인 경우 군집회복 정도의 평균은 .0267, 표준편차는 .0252이고, 군집수가 4개인 경우 군집회복 정도의 평균은 .0243, 표준편차는 .0471이며, 군집수가 6개인 경우 군집회복 정도의 평균은 .0069, 표준편차는 .0636으로 관찰되었다. 기술통계 결과 군집수가 늘어남에 따라 군집회복 정도가 작아지는 것으로 관찰되었다.

<표 4-5> 군집수 수준에 따른 K-평균 군집분석 군집회복의 기술통계

	2	4	6	합계
사례수	81	81	81	243
평균	.0267	.0243	.0069	.0193
표준편차	.0252	.0471	.0636	.0486

군집수 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 4-6>에 제시하였다. 그 결과, 군집수 수준에 따른 군집회복의 정도에는 유의한 차이가 있었다, $F(2,240)=4.130$, $p=.017$.

<표 4-6> 군집수 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
군집수	.019	2	.009	4.130	.017
오차	.552	240	.002		
합계	.571	242			

사후분석인 Scheffe검증 결과는 <표 4-7>에 제시하였다. 군집의 수가 2개인 경우 군집회복 정도와(평균=.0267, 표준편차=.0252) 군집의 수가 4개인 경우 군집회복 정도(평균=.0243, 표준편차=.0471)의 차이가 없었다. 또한 군집의 수가 4개인 경우 군집회복 정도와 군집의 수가 6개인 경우 군집회복 정도(평균=.0069, 표준편차=.0636)의 차이가 없었다. 하지만, 군집의 수가 2개인 경우는 군집의 수가 6개인 경우보다 군집회복 정도가 높았다.

<표 4-7> 군집수 수준에 따른 K-평균 군집분석 사후분석 결과

군집수	평균차	표준오차	유의확률
2 vs 4	.0023	.0075	.953
4 vs 6	.0175	.0075	.070
2 vs 6	.0198	.0075	.033

4) 효과크기에 따른 군집회복

효과크기의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 4-8>에 제시하였다. 효과크기가 0.2씩 차이가 나는 경우 군집회복 정도의 평균은 .0069, 표준편차는 .0184이고, 효과크기가 0.5씩 차이가 나는 경우 군집회복 정도의 평균은 .0207, 표준편차는 .0466이며, 효과크기가 0.8씩 차이가 나는 경우 군집회복 정도의 평균은 .0303, 표준편차는 .0659으로 관찰되었다. 기술통계 결과 효과크기가 커짐에 따라 군집회복 정도가 커지는 것으로 관찰되었다.

<표 4-8> 효과크기 수준에 따른 K-평균 군집분석 군집회복의 기술통계

	0.2	0.5	0.8	합계
사례수	81	81	81	243
평균	.0069	.0207	.0303	.0193
표준편차	.0184	.0466	.0659	.0486

효과크기 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 4-9>에 제시하였다. 그 결과, 효과크기 수준에 따른 군집회복의 정도에는 유의한 차이가 있었다, $F(2,240)=4.919$, $p=.008$.

<표 4-9> 효과크기 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
효과크기	.022	2	.011	4.919	.008
오차	.549	240	.002		
합계	.571	242			

사후분석인 Scheffe검증 결과는 <표 4-10>에 제시하였다. 효과크기가 0.2인 경우 군집회복 정도와(평균=.0069, 표준편차=.0184) 효과크기가 0.5인 경우 군집회복 정도(평균=.0207, 표준편차=.0466)의 차이가 없었다. 또한 효과크기가 0.5인 경우 군집회복 정도와 효과크기가 0.8인 경우 군집회복 정도(평균=.0303, 표준편차=.0659)의 차이가 없었다. 하지만, 효과크기가 0.2인 경우는 효과크기가 0.8인 경우보다 군집회복 정도가 낮았다.

<표 4-10> 효과크기 수준에 따른 K-평균 군집분석 사후분석 결과

효과크기	평균차	표준오차	유의확률
0.2 vs 0.5	−.0138	.0075	.188
0.5 vs 0.8	−.0097	.0075	.439
0.2 vs 0.8	−.0234	.0075	.008

5) 잠재집단확률에 따른 군집회복

잠재집단확률의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 4-11>에 제시하였다. 잠재집단확률이 한 집단에 10%가 모여 있는 경우 군집회복 정도의 평균은 .0400, 표준편차는 .0283이고, 잠재집단확률이 고르게 분배된 경우 군집회복 정도의 평균은 .0471, 표준편

차는 .0250이며, 잠재집단확률이 한 집단에 60%가 모여 있는 경우 군집회복 정도의 평균은 -.0293, 표준편차는 .0459로 관찰되었다. 기술통계 결과 잠재집단확률이 균등할 때 군집회복 정도가 가장 큰 것으로 관찰되었다.

<표 4-11> 잠재집단확률 수준에 따른 *K*-평균 군집분석군집회복의 기술통계

	10%	even	60%	합계
사례수	81	81	81	243
평균	.0400	.0471	−.0293	.0193
표준편차	.0283	.0250	.0459	.0486

잠재집단확률 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 4-12>에 제시하였다. 그 결과, 잠재집단확률 수준에 따른 군집회복의 정도에는 유의한 차이가 있었다, $F(2,240)=122.679$, $p<.001$.

<표 4-12> 잠재집단확률 수준에 따른 *K*-평균 군집분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	*F*	유의확률
잠재집단확률	.289	2	.144	122.679	<.001
오차	.282	240	.001		
합계	.571	242			

사후분석인 Scheffe검증 결과는 <표 4-13>에 제시하였다. 잠재집단확률이 10%인 경우 군집회복 정도와(평균=.0400, 표준편차=.0283) 잠재집단확률이 50%로 고르게 분배된 경우 군집회복 정도(평균=.0471, 표준편차=.0250)는 잠재집단확률이 60%인 경우(평균=-.0293, 표준편차=.0459)보다 군집회복 정도가 높았다. 하지만 잠재집단확률이 10%인 경우 군집회복 정도와 잠재집단확률이 50%로 고르게 분배된 경우 군집회복 정도에서는 차이가 없었다.

<표 4-13> 잠재집단확률 수준에 따른 K-평균 군집분석 사후분석 결과

잠재집단확률	평균차	표준오차	유의확률
10% vs even	−.0071	.0054	.419
even vs 60%	.0764	.0054	<.001
10% vs 60%	.0693	.0054	<.001

3. 연구3: 각 조작변수와 각 분석방법의 상호작용효과

연구3은 각 조작변수가 군집회복 정도에 미치는 영향이 K-평균 군집분석을 실시했을 때와 잠재계층분석을 실시했을 때 달라지는지 알아보기 위한 목적으로 진행되었다. R의 몬테카를로 시뮬레이션 기능을 사용하여 자료를 생성하였고, 생성한 자료는 SPSS에서 이원분산분석으로 분석하였다.

1) 사례수에 따른 군집회복

사례수 수준이 200, 500, 1000인 경우와 이에 따라 K-평균 군집분석과 잠재계층분석을 각각 실시한 결과, 군집회복 정도의 평균과 표준편차는 <표 5-1-1>에 제시하였다. 잠재계층분석을 실시한 결과의 군집회복 정도는 (평균=.0406, 표준편차=.0372) K-평균 군집분석을 실시한 결과의 군집회복 정도(평균=.0193 표준편차=.0486)에 비해서 상대적으로 높았다. 덧붙여서, 사례수의 수준에 따른 군집회복 정도에서 사례수가 1000으로 가장 높은 경우 군집회복 정도(평균=.0313, 표준편차=.0445)가 가장 높았다.

<표 5-1-1> 사례수 수준에 따른 각 분석방법 군집회복의 기술통계

		200	500	1000	합계
K-평균 군집분석	평균	.0189	.0192	.0198	.0193
	표준편차	.0498	.0487	.0478	.0486
	사례수	81	81	81	243
잠재계층 분석	평균	.0384	.0405	.0428	.0406
	표준편차	.0373	.0369	.0377	.0372
	사례수	81	81	81	243
합계	평균	.0286	.0298	.0313	.0299

	표준편차	.0449	.0444	.0445	.0445
	사례수	162	162	162	486

분석방법과 사례수의 수준이 군집회복 정도에 미치는 효과에 대하여 2(*K*-평균 군집분석 vs 잠재계층분석) X 3(200 vs 500 vs 1000) 분산분석을 실시하였다. 그 결과는 <표 5-1-2>에 제시하였다. 분석방법과 사례수 수준이 군집회복 정도에 미치는 상호작용효과는 통계적으로 유의하지 않았다, $F(2,480)=0.061$, $p=.94$.

<표 5-1-2> 사례수와 분석방법이 군집회복에 미치는 효과에 대한 이원분산분석 결과

분산원	제곱합	자유도	평균제곱	*F*	*p*
사례수	.001	2	.000	.154	.858
분석방법	.055	1	.055	29.154	<.001
사례수 x 분석방법	.000	2	.000	.061	.940
오차	.905	480	.002		
합계	.961	485			

따라서 주효과를 검증하기 위하여 일원변량분석을 실시하여 군집회복 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 5-1-3>에 제시하였다. 사례수가 200인 경우 군집회복 정도의 평균은 .0286, 표준편차는 .0449이고, 사례수가 500인 경우 군집회복 정도의 평균은 .0298, 표준편차는 .0444이며, 사례수가 1000인 경우 군집회복 정도의 평균은 .0313, 표준편차는 .0445로 관찰되었다.

<표 5-1-3> 사례수의 주효과 검증 결과

	200	500	1000	합계
사례수	162	162	162	486
평균	.0286	.0298	.0313	.0299
표준편차	.0449	.0444	.0445	.0445

그리고 사례수 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 5-1-4>에 제시하였다. 그 결과, 사례수 수준에 따른 군집회복의 정도에는 유의한 차이가 없었다, $F(2,483)=.146$, $p=.864$.

<표 5-1-4> 사례수 수준에 따른 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
사례수	.001	2	.000	.146	.864
오차	.961	483	.002		
합계	.961	485			

또한 각 독립변인의 단순주효과를 검증하기 위해 분석방법 별로 두 개의 집단을 만든 후에 각각의 집단에서 사례수의 수준이 군집회복정도에 미치는 효과를 검증하였다. 먼저 K-평균 군집분석결과, 사례수의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 5-1-5>에 제시하였다. 사례수가 200인 경우 군집회복 정도의 평균은 .0189, 표준편차는 .0498이고, 사례수가 500인 경우 군집회복 정도의 평균은 .0192, 표준편차는 .0487이며, 사례수가 1000인 경우 군집회복 정도의 평균은 .0198, 표준편차는 .0478로 관찰되었다.

<표 5-1-5> 사례수 수준에 따른 *K*-평균 군집분석 군집회복의 기술통계

	200	500	1000	합계
사례수	81	81	81	243
평균	.0189	.0192	.0198	.0193
표준편차	.0498	.0487	.0478	.0486

사례수 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 5-1-6>에 제시하였다. 그 결과, 사례수 수준에 따른 군집회복의 정도에는 유의한 차이가 없었다, $F(2,240)=.009$, $p=.992$.

<표 5-1-6> 사례수 수준에 따른 *K*-평균 군집분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	*F*	유의확률
사례수	.000	2	.000	.009	.992
오차	.571	240	.002		
합계	.571	242			

잠재계층분석 결과, 사례수의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 5-1-7>에 제시하였다. 사례수가 200인 경우 군집회복 정도의 평균은 .0384, 표준편차는 .0373이고, 사례수가 500인 경우 군집회복 정도의 평균은 .0405, 표준편차는 .0369이며, 사례수가 1000인 경우 군집회복 정도의 평균은 .0428, 표준편차는 .0377로 관찰되었다.

<표 5-1-7> 사례수 수준에 따른 잠재계층분석 군집회복의 기술통계

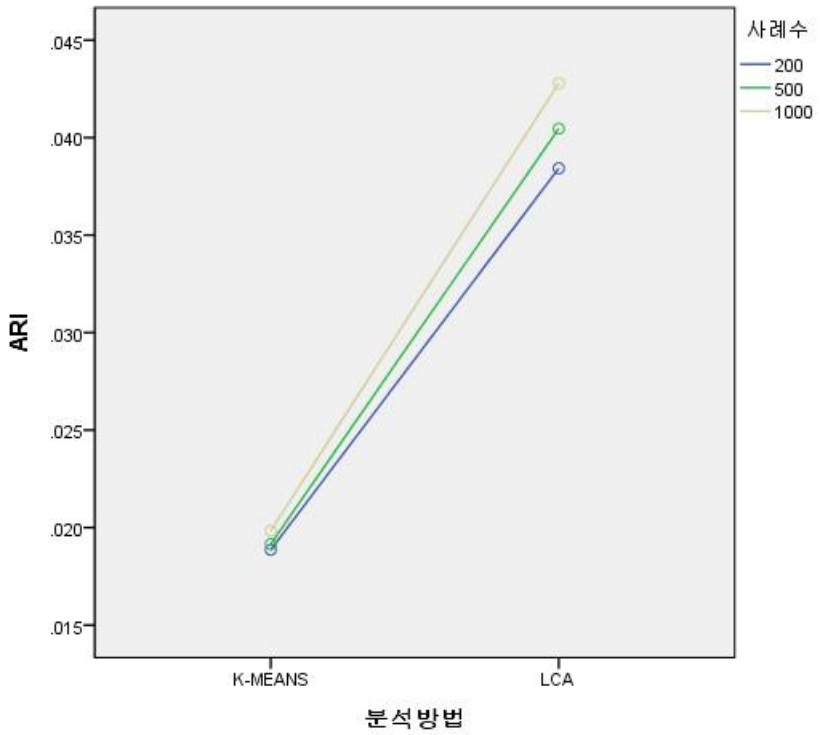
	200	500	1000	합계
사례수	81	81	81	243
평균	.0384	.0405	.0428	.0406
표준편차	.0373	.0369	.0377	.0372

사례수 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 5-1-8>에 제시하였다. 그 결과, 사례수 수준에 따른 군집회복의 정도에는 유의한 차이가 없었다, $F(2,240)=.277$, $p=.759$.

<표 5-1-8> 사례수 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
사례수	.001	2	.000	.277	.759
오차	.334	240	.001		
합계	.335	242			

사례수가 군집회복 정도에 미치는 영향이 K-평균 군집분석을 실시했을 때와 잠재계층분석을 실시했을 때 어떠한 차이를 보이는지 <그림 11>에 제시하였다.



<그림 11> 분석방법에 따라 사례수가 군집회복 정도에 미치는 영향

2) 문항수에 따른 군집회복

문항과 군집수의 비율이 1:1, 1:1, 3:1인 경우와 이에 따라 K-평균 군집분석과 잠재계층분석을 각각 실시한 결과, 군집회복 정도의 평균과 표준편차는 <표 5-2-1>에 제시하였다. 잠재계층분석을 실시한 결과의 군집회복 정도는 (평균=.0406, 표준편차=.0372) K-평균 군집분석을 실시한 결과의 군집회복 정도(평균=.0193 표준편차=.0486)에 비해서 상대적으로 높았다. 덧붙

여서, 문항수의 수준에 따른 군집회복 정도에서 문항비율이 3:1로 가장 높은 경우 군집회복 정도(평균=.0350, 표준편차=.0456)가 가장 높았다.

<표 5-2-1> 문항수 수준에 따른 각 분석방법 군집회복의 기술통계

		1:1	2:1	3:1	합계
K-평균 군집분석	평균	.0170	.0190	.0219	.0193
	표준편차	.0492	.0477	.0493	.0486
	사례수	81	81	81	243
잠재계층 분석	평균	.0309	.0426	.0481	.0406
	표준편차	.0361	.0364	.0375	.0372
	사례수	81	81	81	243
합계	평균	.0240	.0308	.0350	.0299
	표준편차	.0436	.0440	.0456	.0445
	사례수	162	162	162	486

분석방법과 문항수의 수준이 군집회복 정도에 미치는 효과에 대하여 2(K-평균 군집분석 vs 잠재계층분석) X 3(1:1 vs 2:1 vs 3:1) 분산분석을 실시하였다. 그 결과는 <표 5-2-2>에 제시하였다. 분석방법과 문항수 수준이 군집회복 정도에 미치는 상호작용효과는 통계적으로 유의하지 않았다, $F(2,480)=0.910$, $p=.403$.

<표 5-2-2> 문항수와 분석방법이 군집회복에 미치는 효과에 대한 이원분산분석 결과

분산원	제곱합	자유도	평균제곱	F	p
문항수	.010	2	.005	2.710	.068
분석방법	.055	1	.055	29.568	<.001
문항수 x 분석방법	.003	2	.002	.910	.403
오차	.893	480	.002		
합계	.961	485			

따라서 주효과를 검증하기 위하여 일원변량분석을 실시하여 군집회복 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 5-2-3>에 제시하였다. 문항비율이 1:1인 경우 군집회복 정도의 평균은 .0240, 표준편차는 .0436이고, 문항비율이 2:1인 경우 군집회복 정도의 평균은 .0308, 표준편차는 .0440이며, 문항비율이 3:1인 경우 군집회복 정도의 평균은 .0350, 표준편차는 .0456으로 관찰되었다.

<표 5-2-3> 문항수의 주효과 검증 결과

	1:1	2:1	3:1	합계
사례수	162	162	162	243
평균	.0240	.0308	.0350	.0299
표준편차	.0436	.0440	.0456	.0445

그리고 문항수 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 5-2-4>에 제시하였다. 그 결과, 사례수 수준에 따른 군집회복의 정도에는 유의한 차이가 없었다, $F(2,483)=2.560$ $p=.078$.

<표 5-2-4> 문항수 수준에 따른 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
문항수	.010	2	.005	2.560	.078
오차	.951	483	.002		
합계	.961	485			

또한 각각 요인의 주효과별로 평균의 차이를 검증하기 위하여 일원변량분석을 실시하였다. K-평균 군집분석결과, 문항수의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 5-2-5>에 제시하였다. 문항수와 군집수의 비율이 1:1인 경우 군집회복 정도의 평균은 .0170, 표준편차는 .0492이고, 문항수와 군집수의 비율이 2:1인 경우 군집회복 정도

의 평균은 .0190, 표준편차는 .0477이며, 문항수와 군집수의 비율이 3:1인 경우 군집회복 정도의 평균은 .0219, 표준편차는 .0493으로 관찰되었다.

<표 5-2-5> 문항수 수준에 따른 *K*-평균 군집분석 군집회복의 기술통계

	1:1	2:1	3:1	합계
사례수	81	81	81	243
평균	.0170	.0190	.0219	.0193
표준편차	.0492	.0477	.0493	.0486

문항수 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 5-2-6>에 제시하였다. 그 결과, 문항수 수준에 따른 군집회복의 정도에는 유의한 차이가 없었다, $F(2,240)=.209$, $p=.811$.

<표 5-2-6> 문항수 수준에 따른 *K*-평균 군집분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	*F*	유의확률
문항수	.001	2	.000	.209	.811
오차	.570	240	.002		
합계	.571	242			

잠재계층분석 결과, 문항수의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 5-2-7>에 제시하였다. 문항수와 군집수의 비율이 1:1인 경우 군집회복 정도의 평균은 .0309, 표준편차는 .0361이고, 문항수와 군집수의 비율이 2:1인 경우 군집회복 정도의 평균은 .0426, 표준편차는 .0364이며, 문항수와 군집수의 비율이 3:1인 경우 군집회복 정도의 평균은 .0481, 표준편차는 .0375으로 관찰되었다.

<표 5-2-7> 문항수 수준에 따른 잠재계층분석 군집회복의 기술통계

	1:1	2:1	3:1	합계
사례수	81	81	81	243
평균	.0309	.0426	.0481	.0406
표준편차	.0361	.0364	.0375	.0372

문항 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 5-2-8>에 제시하였다. 그 결과, 문항수 수준에 따른 군집회복의 정도에는 유의한 차이가 있었다, $F(2,240)=4.638$, $p=.011$.

<표 5-2-8> 문항수 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

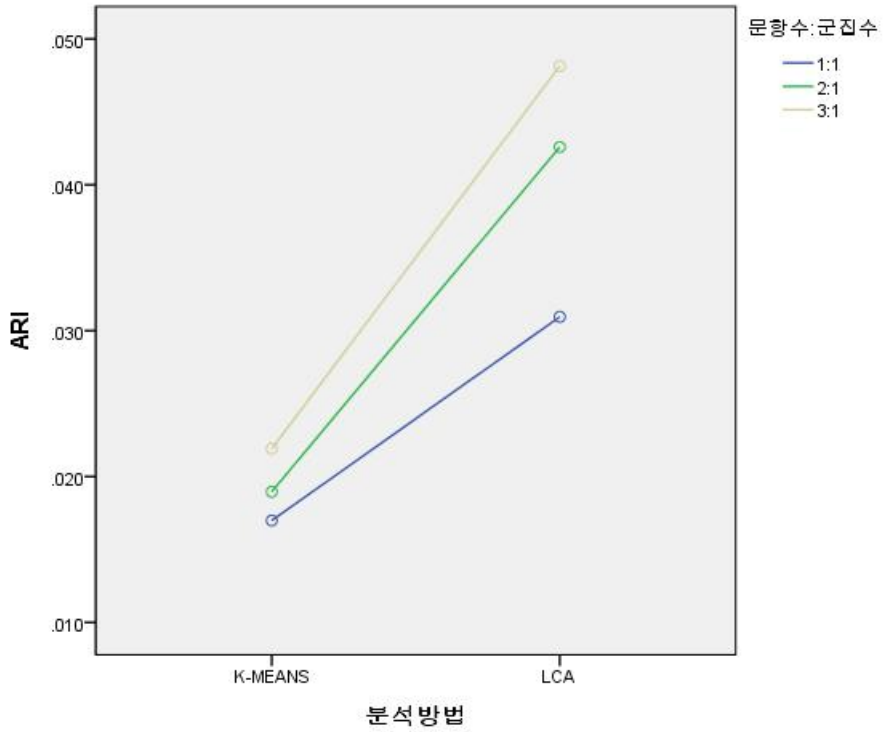
	제곱합	자유도	평균제곱	F	유의확률
문항수	.012	2	.006	4.638	.011
오차	.323	240	.001		
합계	.335	242			

사후분석인 Scheffe검증 결과는 <표 5-2-9>에 제시하였다. 문항수와 군집수의 비율이 1:1인 경우 군집회복 정도(평균=.0609, 표준편차=.0361)와 문항수와 군집수의 비율이 2:1인 경우 군집회복 정도(평균=.0426, 표준편차=.0364)와의 차이가 없었다. 또한 문항수와 군집수의 비율이 2:1인 경우 군집회복 정도와 문항수와 군집수의 비율이 3:1인 경우 군집회복 정도(평균=.0481, 표준편차=.0375)의 차이가 없었다. 하지만, 문항수와 군집수의 비율이 1:1인 경우는 문항수와 군집수의 비율이 3:1인 경우보다 군집회복 정도가 낮았다.

<표 5-2-9> 문항수 수준에 따른 잠재계층분석 사후분석 결과

문항비율	평균차	표준오차	유의확률
1:1 vs 2:1	−.0116	.0058	.132
1:1 vs 3:1	−.0172	.0058	.013
2:1 vs 3:1	−.0055	.0058	.630

문항수가 군집회복 정도에 미치는 영향이 K-평균 군집분석을 실시했을 때와 잠재계층분석을 실시했을 때 어떠한 차이를 보이는지 <그림 12>에 제시하였다.



<그림 12> 분석방법에 따라 문항수가 군집회복 정도에 미치는 영향

3) 군집수에 따른 군집회복

군집수가 2,4,6인 경우와 이에 따라 K-평균 군집분석과 잠재계층분석을 각각 실시한 결과, 군집회복 정도의 평균과 표준편차는 <표 5-3-1>에 제시하였다. 잠재계층분석을 실시한 결과의 군집회복 정도는 (평균=.0406, 표준편차=.0372) K-평균 군집분석을 실시한 결과의 군집회복 정도(평균=.0193 표준편차=.0486)에 비해서 상대적으로 높았다. 덧붙여서, 군집수의 수준에 따른 군집회복 정도에서 군집수가 4개일 때 가장 높은 군집회복 정도(평균=.0350, 표준편차=.0434)를 보였다.

<표 5-3-1> 군집수 수준에 따른 각 분석방법 군집회복의 기술통계

		2	4	6	합계
K-평균 군집분석	평균	.0267	.0243	.0069	.0193
	표준편차	.0252	.0471	.0636	.0486
	사례수	81	81	81	243
잠재계층분석	평균	.0337	.0457	.0423	.0406
	표준편차	.0292	.0365	.0439	.0372
	사례수	81	81	81	243
합계	평균	.0302	.0350	.0246	.0299
	표준편차	.0274	.0434	.0573	.0445
	사례수	162	162	162	486

분석방법과 군집수의 수준이 군집회복 정도에 미치는 효과에 대하여 2(K-평균 군집분석 vs 잠재계층분석) X 3(2 vs 4 vs 6) 분산분석을 실시하였다. 그 결과는 <표 5-3-2>에 제시하였다. 분석방법과 군집수 수준이 군집회복 정도에 미치는 상호작용효과는 통계적으로 유의하였다, $F(2,480)=4.464$, $p=.012$.

<표 5-3-2> 군집수와 분석방법이 군집회복에 미치는 효과에 대한 이원분산분석 결과

분산원	제곱합	자유도	평균제곱	F	p
군집수	.009	2	.004	2.408	.091
분석방법	.055	1	.055	29.962	<.001
군집수 x 분석방법	.016	2	.008	4.464	.012
오차	.881	480	.002		
합계	.961	485			

따라서 각 독립변인의 주효과를 검증하는 대신에 단순주효과를 검증하였다. 보다 구체적으로 분석방법 별로 두 개의 집단을 만든 후에 각각의 집단에서 군집수의 수준이 군집회복정도에 미치는 효과를 검증하였다. 먼저 K-평균 군집분석을 실시한 경우, 군집수의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 5-3-3>에 제시하였다. 군집의 수가 2개인 경우 군집회복 정도의 평균은 .0267, 표준편차는 .0252이고, 군집의 수가 4개인 경우 군집회복 정도의 평균은 .0243, 표준편차는 .0636이며, 군집의 수가 6개인 경우 군집회복 정도의 평균은 .0069, 표준편차는 .0636으로 관찰되었다.

<표 5-3-3> 군집수 수준에 따른 K-평균 군집분석 군집회복의 기술통계

	2	4	6	합계
사례수	81	81	81	243
평균	.0267	.0243	.0069	.0193
표준편차	.0252	.0471	.0636	.0486

K-평균 군집분석을 실시한 경우, 군집수 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 5-3-4>에 제시하였다. 그 결과, 군집수 수준에

따른 군집회복의 정도에는 .05 수준에서 통계적으로 유의한 차이가 있었다, $F(2,240)=4.130$, $p=.017$.

<표 5-3-4> 군집수 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
군집수	.019	2	.009	4.130	.017
오차	.552	240	.002		
합계	.571	242			

집단 간의 차이를 보다 구체적으로 살펴보기 위해서 사후검증인 Scheffe 검정을 실시하여 그 결과는 <표 5-3-5>에 제시하였다. 군집의 수가 2개인 경우 군집회복 정도와(평균=.0267, 표준편차=.0252) 군집의 수가 4개인 경우 군집회복 정도(평균=.0243, 표준편차=.0471)의 차이가 없었다. 또한 군집의 수가 4개인 경우 군집회복 정도와 군집의 수가 6개인 경우 군집회복 정도(평균=.0069, 표준편차=.0636)의 차이가 없었다. 하지만, 군집의 수가 2개인 경우는 군집의 수가 6개인 경우보다 군집회복 정도가 높았다.

<표 5-3-5> 군집수 수준에 따른 K-평균 군집분석 사후분석 결과

군집수	평균차	표준오차	유의확률
2 vs 4	.0023	.0075	.953
4 vs 6	.0175	.0075	.070
2 vs 6	.0198	.0075	.033

다음으로 잠재계층분석을 실시한 경우, 군집수의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 5-3-6>에 제시하였다. 군집수가 2개인 경우 군집회복 정도의 평균은 .0267, 표준편차는 .0252이고, 군집수가 4개인 경우 군집회복 정도의 평균은 .0243, 표준편차는 .0471

이며, 군집수가 6개인 경우 군집회복 정도의 평균은 .0069, 표준편차는 .0636으로 관찰되었다.

<표 5-3-6> 군집수 수준에 따른 잠재계층분석 군집회복의 기술통계

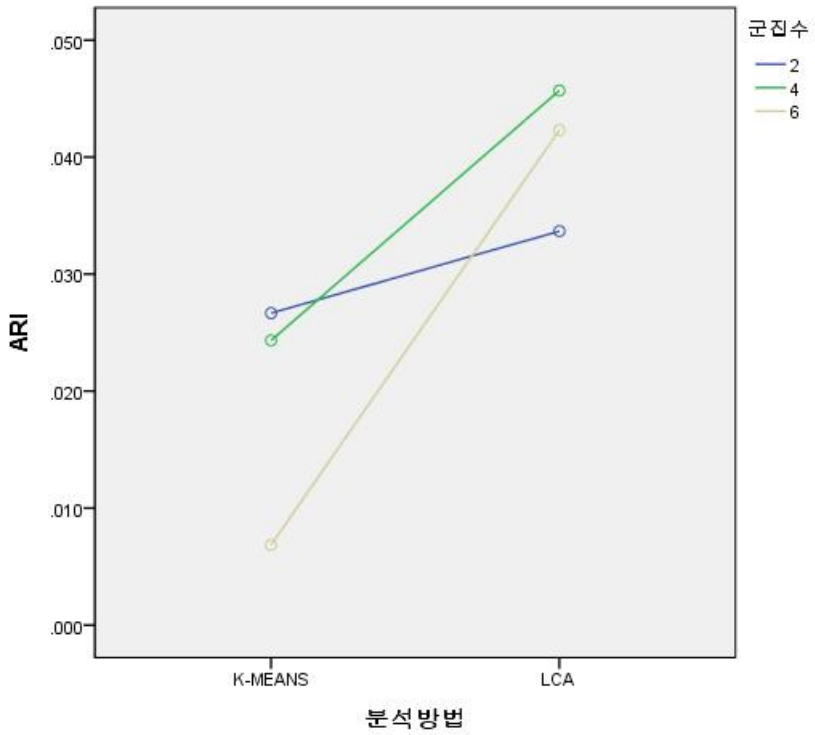
	2	4	6	합계
사례수	81	81	81	243
평균	.0337	.0457	.0423	.0406
표준편차	.0292	.0365	.0439	.0372

잠재계층분석을 실시한 경우, 군집수 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 5-3-7>에 제시하였다. 그 결과, 군집수 수준에 따른 군집회복의 정도에는 .05 수준에서 통계적으로 유의한 차이가 없었다, $F(2,240)=2.271$, $p=.105$.

<표 5-3-7> 군집수 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
군집수	.006	2	.003	2.271	.105
오차	.329	240	.001		
합계	.335	242			

군집수가 군집회복 정도에 미치는 영향이 K-평균 군집분석을 실시했을 때와 잠재계층분석을 실시했을 때 어떤 차이를 보이는지 <그림 13>에 제시하였다.



<그림 13> 분석방법에 따라 군집수가 군집회복 정도에 미치는 영향

4) 효과크기에 따른 군집회복

효과크기가 수준이 0.2, 0.5, 0.8인 경우와 이에 따라 K-평균 군집분석과 잠재계층분석을 각각 실시한 결과, 군집회복 정도의 평균과 표준편차는 <표 5-4-1>에 제시하였다. 잠재계층분석을 실시한 결과의 군집회복 정도는 (평균=.0406, 표준편차=.0372) K-평균 군집분석을 실시한 결과의 군집회복 정

도(평균=.0193 표준편차=.0486)에 비해서 상대적으로 높았다. 덧붙여서, 효과크기의 수준에 따른 군집회복 정도에서 효과크기가 0.8로 가장 높은 경우 군집회복 정도(평균=.0476, 표준편차=.0574)가 가장 높았다.

<표 5-4-1> 효과크기 수준에 따른 각 분석방법 군집회복의 기술통계

		0.2	0.5	0.8	합계
K-평균 군집분석	평균	.0069	.0206	.0303	.0193
	표준편차	.0184	.0466	.0659	.0486
	사례수	81	81	81	243
잠재계층 분석	평균	.0198	.0369	.0649	.0406
	표준편차	.0268	.0273	.0409	.0372
	사례수	81	81	81	243
합계	평균	.0133	.0288	.0476	.0299
	표준편차	.0239	.0390	.0574	.0445
	사례수	162	162	162	486

분석방법과 효과크기의 수준이 군집회복 정도에 미치는 효과에 대하여 2(*K*-평균 군집분석 vs 잠재계층분석) X 3(0.2 vs 0.5 vs 0.8) 분산분석을 실시하였다. 그 결과는 <표 5-4-2>에 제시하였다. 분석방법과 효과크기 수준이 군집회복 정도에 미치는 상호작용효과는 통계적으로 유의하였다, $F(2,480)=3.318$, $p=.037$.

<표 5-4-2> 효과크기와 분석방법이 군집회복에 미치는 효과에 대한 이원분산분석 결과

분산원	제곱합	자유도	평균제곱	*F*	*p*
효과크기	.096	2	.048	28.684	<.001
분석방법	.055	1	.055	33.012	<.001
효과크기 x 분석방법	.011	2	.006	3.318	.037
오차	.800	480	.002		

합계	.961	485

따라서 각 독립변인의 주효과를 검증하는 대신에 단순주효과를 검증하였다. 보다 구체적으로 분석방법 별로 두 개의 집단을 만든 후에 각각의 집단에서 효과크기의 수준이 군집회복정도에 미치는 효과를 검증하였다. 먼저 *K*-평균 군집분석을 실시한 경우, 효과크기의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 5-4-3>에 제시하였다. 효과크기가 0.2씩 차이가 나는 경우 군집회복 정도의 평균은 .0069, 표준편차는 .0184이고, 효과크기가 0.5씩 차이가 나는 경우 군집회복 정도의 평균은 .0207, 표준편차는 .0466이며, 효과크기가 0.8씩 차이가 나는 경우 군집회복 정도의 평균은 .0303, 표준편차는 .0659으로 관찰되었다.

<표 5-4-3> 효과크기 수준에 따른 *K*-평균 군집분석 군집회복의 기술통계

	0.2	0.5	0.8	합계
사례수	81	81	81	243
평균	.0069	.0207	.0303	.0193
표준편차	.0184	.0466	.0659	.0486

K-평균 군집분석을 실시한 경우, 효과크기 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 5-4-4>에 제시하였다. 그 결과, 효과크기 수준에 따른 군집회복의 정도에는 .05 수준에서 통계적으로 유의한 차이가 있었다, $F(2,240)=4.919$, $p=.008$.

<표 5-4-4> 효과크기 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
효과크기	.022	2	.011	4.919	.008
오차	.549	240	.002		
합계	.571	242			

집단 간의 차이를 보다 구체적으로 살펴보기 위해서 사후검증인 Scheffe 검정을 실시하여 그 결과는 <표 5-4-5>에 제시하였다. 효과크기가 0.2인 경우 군집회복 정도와(평균=.0069, 표준편차=.0184) 효과크기가 0.5인 경우 군집회복 정도(평균=.0207, 표준편차=.0466)의 차이가 없었다. 또한 효과크기가 0.5인 경우 군집회복 정도와 효과크기가 0.8인 경우 군집회복 정도(평균=.0303, 표준편차=.0659)의 차이가 없었다. 하지만, 효과크기가 0.2인 경우는 효과크기가 0.8인 경우보다 군집회복 정도가 낮았다.

<표 5-4-5> 효과크기 수준에 따른 K-평균 군집분석 사후분석 결과

효과크기	평균차	표준오차	유의확률
0.2 vs 0.5	−.0138	.0075	.188
0.5 vs 0.8	−.0097	.0075	.439
0.2 vs 0.8	−.0234	.0075	.008

다음으로 잠재계층분석을 실시한 경우, 효과크기의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 5-4-6>에 제시하였다. 효과크기가 0.2씩 차이가 나는 경우 군집회복 정도의 평균은 .0198, 표준편차는 .0268이고, 효과크기가 0.5씩 차이가 나는 경우 군집회복 정도의 평균은 .0369, 표준편차는 .0273이며, 효과크기가 0.8씩 차이가 나는 경우 군집회복 정도의 평균은 .0649, 표준편차는 .0409으로 관찰되었다.

<표 5-4-6> 효과크기 수준에 따른 잠재계층분석 군집회복의 기술통계

	0.2	0.5	0.8	합계
사례수	81	81	81	243
평균	.0198	.0369	.0649	.0406
표준편차	.0268	.0273	.0409	.0372

잠재계층분석을 실시한 경우, 효과크기 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 5-4-7>에 제시하였다. 그 결과, 효과크기 수준에 따른 군집회복의 정도에는 .05 수준에서 통계적으로 유의한 차이가 있었다, $F(2,240)=40.227$, $p<.001$.

<표 5-4-7> 효과크기 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
효과크기	.084	2	.042	40.227	<.001
오차	.251	240	.001		
합계	.335	242			

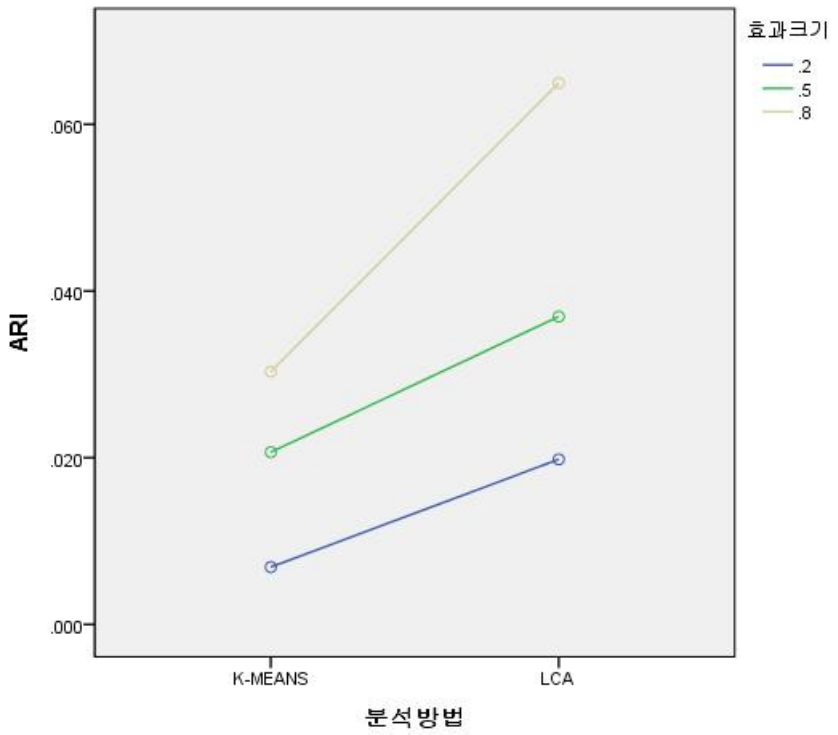
집단 간의 차이를 보다 구체적으로 살펴보기 위해서 사후검증인 Scheffe 검정을 실시하여 그 결과는 <표 5-4-8>에 제시하였다. 효과크기가 0.2인 경우 군집회복 정도(평균=.0198 표준편차=.0268)가 가장 낮았으며 효과크기가 0.5인 경우(평균=.0369 표준편차=.0273)가 중간, 또한 효과크기가 0.8인 경우 군집회복 정도(평균=.0649, 표준편차=.0409)가 가장 높았다.

<표 5-4-8> 효과크기 수준에 따른 잠재계층분석 사후분석 결과

효과크기	평균차	표준오차	유의확률

0.2 vs 0.5	−.0171	.0051	.004
0.5 vs 0.8	−.0280	.0051	<.001
0.2 vs 0.8	−.0451	.0051	<.001

효과크기가 군집회복 정도에 미치는 영향이 K-평균 군집분석을 실시했을 때와 잠재계층분석을 실시했을 때 어떠한 차이를 보이는지 <그림 14>에 제시하였다.



<그림 14> 분석방법에 따라 효과크기가 군집회복 정도에 미치는 영향

5) 잠재집단확률에 따른 군집회복

잠재집단확률의 수준이 10%, 균등, 60%인 경우와 이에 따라 K-평균 군집분석과 잠재계층분석을 각각 실시한 결과, 군집회복 정도의 평균과 표준편차는 <표 5-5-1>에 제시하였다. 잠재계층분석을 실시한 결과의 군집회복 정도는 (평균=.0406, 표준편차=.0372) K-평균 군집분석을 실시한 결과의 군집회복 정도(평균=.0193 표준편차=.0486)에 비해서 상대적으로 높았다. 덧붙여서, 잠재집단확률의 수준에 따른 군집회복 정도에서 균등하게 집단이 나누어진 경우 군집회복 정도(평균=.0510, 표준편차=.0315)가 가장 높았다.

<표 5-5-1> 잠재집단확률 수준에 따른 각 분석방법 군집회복의 기술통계

		10%	even	60%	합계
K-평균 군집분석	평균	.0400	.0471	-.0293	.0193
	표준편차	.0283	.0250	.0459	.0486
	사례수	81	81	81	243
잠재계층 분석	평균	.0380	.0548	.0289	.0406
	표준편차	.0346	.0366	.0361	.0372
	사례수	81	81	81	243
합계	평균	.0390	.0510	-.0002	.0299
	표준편차	.0315	.0315	.0504	.0445
	사례수	162	162	162	486

분석방법과 잠재집단확률의 수준이 군집회복 정도에 미치는 효과에 대하여 2(K-평균 군집분석 vs 잠재계층분석) X 3(0.1 vs 균등 vs 0.6) 분산분석을 실시하였다. 그 결과는 <표 5-5-2>에 제시하였다. 분석방법과 잠재집단확률 수준이 군집회복 정도에 미치는 상호작용효과는 통계적으로 유의하였다, $F(2,480)=34.457$, $p<.001$.

<표 5-5-2> 잠재집단확률과 분석방법이 군집회복에 미치는 효과에 대한 이원분산분석 결과

분산원	제곱합	자유도	평균제곱	F	p
잠재집단확률	.232	2	.116	94.508	<.001
분석방법	.055	1	.055	44.781	<.001
잠재집단확률 x 분석방법	.085	2	.042	34.457	<.001
오차	.589	480	.001		
합계	.961	485			

따라서 각 독립변인의 주효과를 검증하는 대신에 단순주효과를 검증하였다. 보다 구체적으로 분석방법 별로 두 개의 집단을 만든 후에 각각의 집단에서 잠재집단확률 수준이 군집회복정도에 미치는 효과를 검증하였다. 먼저 K-평균 군집분석을 실시한 경우, 잠재집단확률의 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 5-5-3>에 제시하였다. 잠재집단확률이 한 집단에 10%가 모여 있는 경우 군집회복 정도의 평균은 .0400, 표준편차는 .0283이고, 잠재집단확률이 고르게 분배된 경우 군집회복 정도의 평균은 .0471, 표준편차는 .0250이며, 잠재집단확률이 한 집단에 60%가 모여 있는 경우 군집회복 정도의 평균은 -.0293, 표준편차는 .0459로 관찰되었다.

<표 5-5-3> 잠재집단확률 수준에 따른 K-평균 군집분석 군집회복의 기술통계

	10%	even	60%	합계
사례수	81	81	81	243
평균	.0400	.0471	−.0293	.0193
표준편차	.0283	.0250	.0459	.0486

K-평균 군집분석을 실시한 경우, 잠재집단확률 수준에 따른 군집회복 정

도에 대한 일원분산분석 결과는 <표 5-5-4>에 제시하였다. 그 결과, 잠재집단확률 수준에 따른 군집회복의 정도에는 유의한 차이가 있었다, $F(2,240)=122.679$, $p<.001$.

<표 5-5-4> 잠재집단확률 수준에 따른 K-평균 군집분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
잠재집단확률	.289	2	.144	122.679	<.001
오차	.282	240	.001		
합계	.571	242			

집단 간의 차이를 보다 구체적으로 살펴보기 위해서 사후검증인 Scheffe 검정을 실시하여 그 결과는 <표 5-5-5>에 제시하였다. 잠재집단확률이 10%인 경우 군집회복 정도와(평균=.0400, 표준편차=.0283) 잠재집단확률이 50%로 고르게 분배된 경우 군집회복 정도(평균=.0471, 표준편차=.0250)는 잠재집단확률이 60%인 경우(평균=-.0293, 표준편차=.0459)보다 군집회복 정도가 높았다. 하지만 잠재집단확률이 10%인 경우 군집회복 정도와 잠재집단확률이 50%로 고르게 분배된 경우 군집회복 정도에서는 차이가 없었다.

<표 5-5-5> 잠재집단확률 수준에 따른 K-평균 군집분석 사후분석 결과

잠재집단확률	평균차	표준오차	유의확률
10% vs even	−.0071	.0054	.419
even vs 60%	.0764	.0054	<.001
10% vs 60%	.0693	.0054	<.001

다음으로 잠재계층분석을 실시한 경우, 잠재집단확률 수준에 따른 군집회복(ARI) 정도에 대한 사례수, 평균, 그리고 표준편차를 <표 5-5-6>에 제시하였다. 잠재집단확률이 한 집단에 10%가 모여 있는 경우 군집회복 정도의

<표 5-5-6> 잠재집단확률 수준에 따른 잠재계층분석 군집회복의 기술통계

	10%	even	60%	합계
사례수	81	81	81	243
평균	.0380	.0548	.0289	.0406
표준편차	.0346	.0366	.0361	.0372

평균은 .0380, 표준편차는 .0346이고, 잠재집단확률이 고르게 분배된 경우 군집회복 정도의 평균은 .0548, 표준편차는 .0366이며, 잠재집단확률이 한 집단에 60%가 모여 있는 경우 군집회복 정도의 평균은 .0289 표준편차는 .0361로 관찰되었다.

잠재계층분석을 실시한 경우, 잠재집단확률 수준에 따른 군집회복 정도에 대한 일원분산분석 결과는 <표 5-5-7>에 제시하였다. 그 결과, 잠재집단확률 수준에 따른 군집회복의 정도에는 유의한 차이가 있었다, $F(2,240)=10.958$, $p<.001$.

<표 5-5-7> 잠재집단확률 수준에 따른 잠재계층분석 군집회복 정도의 일원분산분석 결과

	제곱합	자유도	평균제곱	F	유의확률
잠재집단확률	.028	2	.014	10.958	<.001
오차	.307	240	.011		
합계	.335	242			

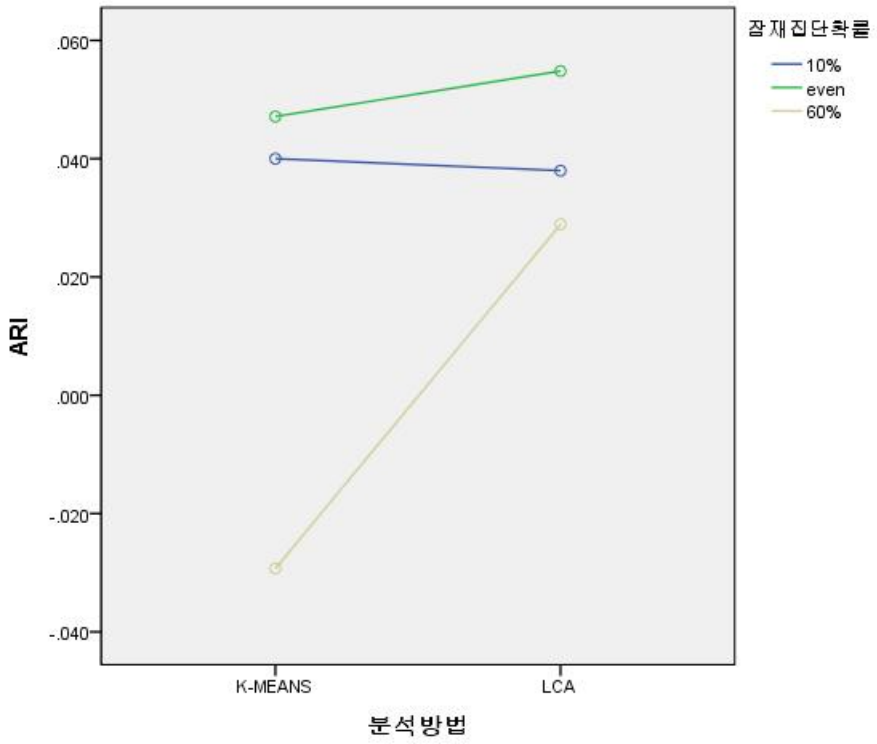
집단 간의 차이를 보다 구체적으로 살펴보기 위해서 사후검증인 Scheffe 검정을 실시하여 그 결과는 <표 5-5-8>에 제시하였다. 잠재집단확률이 10%인 경우 군집회복 정도와(평균=.0380, 표준편차=.0346) 잠재집단확률이 60%로 고르게 분배된 경우 군집회복 정도(평균=.0289, 표준편차=.0361)는 동일하였다. 하지만 잠재집단확률이 50%로 고르게 분배된 경우 군집회복 정도가 가장 높

았다.

<표 5-5-8> 잠재집단확률 수준에 따른 잠재계층분석 사후분석 결과

잠재집단확률	평균차	표준오차	유의확률
10% vs even	−.0168	.0056	.012
even vs 60%	.0259	.0056	<.001
10% vs 60%	.0090	.0056	.272

잠재집단확률이 군집회복 정도에 미치는 영향이 K-평균 군집분석을 실시했을 때와 잠재계층분석을 실시했을 때 어떠한 차이를 보이는지 <그림 15>에 제시하였다.



<그림 15> 분석방법에 따라 잠재집단확률이 군집회복 정도에 미치는 영향

4. 연구4: 분석방법에 따른 군집회복 정도의 차이

연구4는 K-평균 군집분석을 실시한 결과와 잠재계층분석을 실시한 결과에서 군집회복 정도에 차이가 있는지 검증해 보기 위한 목적으로 진행되었다. R의 몬테카를로 시뮬레이션 기능을 사용하여 자료를 생성하였고, 생성한 자료는 SPSS에서 대응표본 t검증으로 분석하여 결과는 <표 6>에 제시하였다.

K-평균 군집분석과 잠재계층분석의 군집회복 정도를 비교한 결과, 잠재계층분석이 K-평균 군집분석보다 높은 군집회복 지수를 보이며 통계적으로 유의하게 더 나은 결과를 준다고 말할 수 있다.

<표 6> K-평균 군집분석과 잠재계층분석 간에 군집회복 정도의 차이

	사례수	평균	표준편차	평균차이	t	df	유의수준
K-means	243	.0193	.0486	-.0213	-6.778	242	<.001
LCA	243	.0406	.0372				

VI. 논의

본 연구는 Eshghi, Haughton, Legrand, Skaletsky 와 Woolford (2011)가 잠재계층분석과 K-평균 군집분석을 비교한 결과 군집의 개수와 군집에 속하는 구성(membership) 결과가 다르게 나온다고 주장한 것을 바탕으로 어떠한 조작변수들이 이에 영향을 미치는지 알아보고자 하였다. 따라서 상이한 연구문제에 따라 K-평균 군집분석과 잠재계층분석 간에 어떤 분석방법이 더 효율적인지 검증해보고자 하였다.

연구문제1에서 조작변수 수준에 따라 잠재계층분석 결과의 군집회복 정도의 차이를 관찰한 결과 가설은 아래와 같이 지지 또는 기각되었다. 가설 1-1인 '사례수의 수준에 따라서 잠재계층분석 결과의 군집회복 정도의 차이가 있을 것이다' 는 기각되었다. 가설 1-2인 '문항수의 수준에 따라서 잠재계층분석 결과의 군집회복 정도의 차이가 있을 것이다' 는 지지되었다. 가설 1-3인 '군집수의 수준에 따라서 잠재계층분석 결과의 군집회복 정도의 차이가 있을 것이다' 는 기각되었다. 가설 1-4인 '효과크기의 수준에 따라서 잠재계층분석 결과의 군집회복 정도의 차이가 있을 것이다' 는 지지되었다. 가설 1-5인 '잠재집단확률의 수준에 따라서 잠재계층분석 결과의 군집회복 정도의 차이가 있을 것이다' 는 지지되었다.

연구 1의 결과가 의미하는 것은 사례수가 100, 500, 1000의 수준으로 달라짐에 따라 군집회복의 정도에는 유의한 차이가 없었다고 해석할 수 있다. 또한 문항수와 군집수의 비율이 1:1, 2:1, 3:1의 수준으로 달라짐에 따라 군집회복의 정도에는 유의한 차이가 있었다. 즉 문항의 수가 많을수록 잠재계층분석 결과가 더 양호하다고 해석할 수 있다. 군집수가 2개, 4개, 6개의 수준으로 달라짐에 따라 군집회복의 정도에는 유의한 차이가 없었다고 해석할 수 있다. 효과크기가 0.2, 0.5, 0.8의 수준으로 달라짐에 따라 군집회복의

정도에는 유의한 차이가 있었다. 즉 효과크기가 클수록 잠재계층분석 결과가 더 양호하다고 해석할 수 있다. 잠재집단확률이 10%, 균등, 60%의 수준으로 달라짐에 따라 군집회복의 정도에는 유의한 차이가 있었다. 즉 잠재집단확률이 고르게 분배될수록 잠재계층분석 결과가 더 양호하다고 해석할 수 있다.

연구문제2에서 조작변수 수준에 따라 K-평균 군집분석 결과의 군집회복 정도의 차이를 관찰한 결과 가설은 다음과 같이 지지 또는 기각되었다. 가설 2-1인 '사례수의 수준에 따라서 K-평균 군집분석 결과의 군집회복 정도의 차이가 있을 것이다' 는 기각되었다. 가설 2-2인 '문항수의 수준에 따라서 K-평균 군집분석 결과의 군집회복 정도의 차이가 있을 것이다' 는 기각되었다. 가설 2-3인 '군집수의 수준에 따라서 K-평균 군집분석 결과의 군집회복 정도의 차이가 있을 것이다' 는 지지되었다. 가설 2-4인 '효과크기의 수준에 따라서 K-평균 군집분석 결과의 군집회복 정도의 차이가 있을 것이다' 는 지지되었다. 가설 2-5인 '잠재집단확률의 수준에 따라서 K-평균 군집분석 결과의 군집회복 정도의 차이가 있을 것이다' 는 지지되었다.

연구 2의 결과가 의미하는 것은 사례수가 100, 500, 1000의 수준으로 달라짐에 따라 군집회복의 정도에는 유의한 차이가 없었다고 해석할 수 있다. 문항수와 군집수의 비율이 1:1, 2:1, 3:1의 수준으로 달라짐에 따라 군집회복의 정도에는 유의한 차이가 없었다고 해석할 수 있다. 군집수가 2개, 4개, 6개의 수준으로 달라짐에 따라 군집회복의 정도에는 유의한 차이가 있었다. 즉 군집수가 많아질수록 K-평균 군집분석 결과가 더 양호하다고 해석할 수 있다. 효과크기가 0.2, 0.5, 0.8의 수준으로 달라짐에 따라 군집회복의 정도에는 유의한 차이가 있었다. 즉 효과크기가 클수록 K-평균 군집분석 결과가 더 양호하다고 해석할 수 있다. 잠재집단확률이 10%, 균등, 60%의 수준으로 달라짐에 따라 군집회복의 정도에는 유의한 차이가 있었다. 즉 잠재집

단확률이 고르게 분배될수록 잠재계층분석 결과가 더 양호하였고 특히 60%가 한 군집에 분배된 경우, 군집회복률이 가장 좋지 않다고 해석할 수 있다.

연구문제3에서는 각 조작변수가 군집회복 정도에 미치는 영향이 어떤 분석방법을 쓰느냐에 따라 달라지는지를 관찰한 결과 가설은 다음과 같이 지지 또는 기각되었다. 가설 3-1인 '사례수의 수준에 따른 군집회복의 차이는 분석방법에 따라 차이가 있을 것이다' 는 기각되었다. 가설 3-2인 '문항수의 수준에 따른 군집회복의 차이는 분석방법에 따라 차이가 있을 것이다' 는 기각되었다. 가설 3-3인 '군집수의 수준에 따른 군집회복의 차이는 분석방법에 따라 차이가 있을 것이다' 는 지지되었다. 가설 3-4인 '효과크기의 수준에 따른 군집회복의 차이는 분석방법에 따라 차이가 있을 것이다' 는 지지되었다. 가설 3-5인 '잠재집단확률의 수준에 따른 군집회복의 차이는 분석방법에 따라 차이가 있을 것이다' 는 지지되었다.

연구 3의 결과가 의미하는 것은 분석방법과 사례수는 상호작용효과를 가지지 않기 때문에, 사례수가 군집회복에 미치는 영향이 각 분석방법에서 차이가 없다고 해석할 수 있다. 또한 분석방법과 문항수도 상호작용효과를 가지지 않기 때문에, 문항수가 군집회복에 미치는 영향이 각 분석방법에서 차이가 없다고 해석할 수 있다. 반면에 분석방법과 군집수는 상호작용효과가 존재하기 때문에 군집수가 군집회복에 미치는 영향이 각 분석방법에서 차이가 있다고 해석할 수 있다. 다만 잠재계층분석에서는 K-평균 군집분석과 달리 문항수 수준에 따른 군집회복정도에 유의한 차이가 없었다. 즉, K-평균 군집분석에서만 군집수가 많을수록 더 좋은 결과를 준다고 해석할 수 있다. 분석방법과 효과크기는 상호작용효과를 가진다고 해석할 수 있고, 나아가 잠재계층분석과 K-평균 군집분석 모두에서 효과크기 수준에 따른 군집회복정도에 유의한 차이가 있었다. 또한 모든 분석방법에서 효과크기가 클수록 군집회복 정도에서 더 좋은 결과를 보이는 것으로 나타났다. 분석방법

과 잠재집단확률은 상호작용효과를 가진다고 해석할 수 있고, 나아가 잠재계층분석과 K-평균 군집분석 모두에서 잠재집단확률 수준에 따른 군집회복 정도에 유의한 차이가 있었다. 또한 모든 분석방법에서 잠재집단확률이 균등하게 분배된 경우 군집회복 정도에서 가장 좋은 결과를 보이는 것으로 나타났다.

연구문제4에서는 잠재계층분석을 실시한 경우와 K-평균군집분석을 실시한 경우, 군집회복 정도에 차이가 있는가를 관찰하였다. 이에 따라 가설 4-1인 '잠재계층분석을 실시한 경우와 K-평균군집분석을 실시한 경우, 군집회복 정도에 차이가 있을 것이다'는 지지되었다. 이는 조작변수와 관계없이 전반적으로 잠재계층분석이 K-평균군집분석보다 더 나은 결과를 주는 분석방법이라고 해석할 수 있다.

또한 잠재계층분석과 K-평균 군집분석을 비교함에 있어서 추정의 효율성도 고려되어야 했다. 특히, 잠재계층모형은 확률에 기초한 통계모형이기 때문에 모형의 추정에서 EM(Expectation-Maximization) 알고리즘에 기초한 최대우도법(Maximum likelihood)을 활용하여 반복(iteration)에 기초한 수렴(convergence)으로 모수의 추정이 이루어진다. 따라서 질적으로 구분되는 하위집단의 결합의 분리에 초점을 맞춘 잠재계층모형은 지역해(local maxima)의 문제점을 가지고 있다. 이에 대한 해결책으로 시작점(starting value)을 무선적으로 선택된 여러 개의 세트를 활용하여 지역해의 문제를 점검한다. 이에 따라서, R에서 실행된 K-평균 군집분석과 잠재계층분석에서 K-평균 군집분석의 함수는 nstart=10, iter.max=k*10 로 지정하여 10개의 랜덤 세트를 고르도록 하였고 군집의 개수의 10을 곱한 만큼의 반복을 최대로 하게끔 지정하였다. 이러한 함수의 조건에 충족되지 못하거나 수렴하지 못하면 NC(Not Converged) 값을 내도록 지정하였다. 그리고 잠재계층분석에서 함수는 이의 단점을 보완하기 위해 niter=100으로 지정하여 K-평균 군집분석보다 수렴할

수 있는 반복수를 늘려주었다. 하지만 그럼에도 불구하고 K-평균 군집분석에서는 총 121,500의 결과 중에 NA가 하나도 나타나지 않은 반면 잠재계층분석에서는 총 121,500의 결과 중에 약 0.1%인 141개의 NC를 얻었다.

따라서 잠재계층분석에서 얻은 141개 NA의 빈도를 각 조작변수별로 분석해본 결과는 다음과 같았다 첫째, 사례수의 수준이 각각 200, 500, 1000인 경우 사례수가 가장 작은 200의 경우에서 총 141개의 NA중 약 74%인 104건이 발생했다. 사례수 500에서는 37건, 사례수 1000에서는 0건이 발생하였다. 이러한 결과에 기초하여 사례수가 커질수록 군집회복이 더 양호하게 나타나며, 동시에 더 좋은 수렴결과를 나타낸다고 해석할 수 있다.

둘째, 문항비율의 수준이 각각 1:1, 2:1, 3:1인 경우 문항수가 군집수와 같은 경우는 NA가 40건, 2:1인 경우는 29건, 그리고 3:1로 문항수가 가장 많은 경우는 72건이 발생하였다. 이러한 결과로 미루어보아 문항수가 많아질수록 군집회복의 정도는 증가하지만 동시에 수렴하지 못하는 케이스도 함께 늘어난다는 것을 알 수 있다. 또한 2:1인 경우 NC 발생이 가장 낮은 것으로 보아 군집수에 비해 문항수가 2배 정도로 적당한 경우가 가장 잘 수렴한다고 해석할 수 있다.

셋째, 군집수의 수준이 각각 2,4,5개인 경우, 군집수가 가장 많은 6개인 경우에서 총 141개의 NA중 약 73%인 103건이 발생했다. 그리고 군집수가 2개인 경우는 37건, 군집수가 4개인 경우는 1건으로 가장 낮게 발생하였다. 이러한 결과로 미루어보아 군집수가 많아질수록 군집회복의 정도는 증가하지만 동시에 수렴하지 못하는 케이스도 함께 늘어난다는 것을 알 수 있다. 또한 군집수가 4개인 경우 NC 발생이 가장 낮은 것으로 보아 군집수가 너무 적거나 큰 경우는 수렴이 어려워진다고 해석할 수 있다.

넷째, 효과크기의 수준이 각각 0.2, 0.5, 0.8인 경우, 효과크기가 0.8로 가장 큰 경우에서 총 141개의 NA중 약 67%인 94건이 발생했다. 그리고 효과크

기가 0.2인 경우는 37건, 효과크기가 0.5인 경우는 10건으로 가장 낮게 발생하였다. 이러한 결과로 미루어보아 효과크기가 클수록 군집회복의 정도는 증가하지만 동시에 수렴하지 못하는 케이스도 함께 늘어난다는 것을 알 수 있다. 또한 효과크기가 0.5인 경우 NC 발생이 가장 낮은 것으로 나타나, 군집간의 평균차이 즉 효과크기가 너무 적게 차이나거나 크게 차이나는 경우는 수렴이 어려워진다고 해석할 수 있다.

마지막으로, 잠재집단확률의 수준이 각각 10%, 균등, 60%인 경우, 잠재집단확률이 10%인 경우가 75건, 균등한 경우가 65건 그리고 60%인 경우가 1건이 발생했다. 이러한 결과로 미루어보아 잠재집단확률이 균등한 경우 군집회복 정도는 가장 높지만, 수렴이 양호하게 이루어진다고 볼수는 없다. 반면에 60%인 경우 군집회복의 정도는 가장 저조했지만, 수렴은 가장 잘 이루어졌다고 해석할 수 있다.

시뮬레이션을 이용한 연구결과, 잠재계층분석이 K-평균군집분석 보다 나은 군집회복 정도를 보이기는 하지만 더 관대한 조건으로 반복(iteration)을 늘려주었음에도 불구하고 수렴하지 못하는 단점을 보였다. 이것은 잠재계층분석이 EM(Expectation-Maximization) 알고리즘을 사용하여 반복적인 최대가능성 절차를 통해 모수 추정이 이루어지는 과정에서 지역 최적해로 인한 어려움을 겪을 수 있다는 쟁점을 뒷받침 해주는 결과이다, 또한 본 연구의 결과는 Magidson 과 Vermunt (2002a)가 주장한 바와 같이 잠재계층분석에 객체들을 분류하는데 에 있어서 더 좋은 결과를 보였다는 부분과 일치한다. 다만 본 연구에서 변수들이 동일한 분산을 가지고 있고 지역 독립성 가정을 유지했음에도 불구하고 K-평균군집분석보다 잠재계층분석의 결과가 더 좋았다는 것은 Magidson 과 Vermunt (2002a)와 일치하지 않았다. 따라서 연구자가 가지고 있는 자료의 형태나 연구가설에 따라 가장 적합한 분석방법을 택하는 것이 필요하다고 생각된다.

사례수와 문항수가 군집회복에 미치는 영향은 분석방법에 따라 다르지 않았다. 사례수가 많을수록, 그리고 문항수가 많을수록 군집회복은 더 양호해지지만 분석방법 간에는 크게 차이가 없기 때문에 K-평균군집분석과 잠재계층분석 중 어느 방법을 사용하여도 무관하다. 다만 군집수, 효과크기 그리고 잠재집단확률이 군집회복에 미치는 효과는 분석방법마다 다르게 나타났다. 특히 군집수가 많아질수록 K-평균군집분석 결과가 잠재계층분석보다 좋지 않기 때문에 군집수가 6개 이상이라면 K-평균군집분석보다 잠재계층분석을 사용하는 것을 권고한다. 또한 효과크기가 작을수록 군집회복 정도가 나빠지며 K-평균군집분석 결과가 잠재계층분석보다 좋지 않기 때문에 효과크기가 작을수록 잠재계층분석을 사용하는 것을 권고한다. 잠재집단확률이 균등하게 분배되어 있거나 군집 하나가 상대적으로 매우 작은 경우는 어느 분석방법을 사용하여도 무관하다. 하지만 군집 하나가 상대적으로 매우 큰 경우는 K-평균군집분석 결과, 군집회복 정도가 음수의 값으로 아주 낮게 나타나기 때문에 K-평균군집분석 보다 잠재계층분석을 실시할 것을 권한다.

후속 연구를 위한 제언은 다음과 같다. 시뮬레이션이 아닌 실제상황에서 실제 데이터를 사용하여 잠재계층분석과 K-평균군집분석의 군집회복 정도를 비교해보는 것이 필요하다. 이에 따라 변수 간의 분산이 다른 경우와 문항반응확률이 본연구의 시뮬레이션 조건과 다른 경우에도 본 연구 결과가 일반화 될 수 있는지 확인하는 것이 필요하다.

참 고 문 헌

곽호완, 박창호, 이태연, 김문수, & 진영선 (2008). 실험심리학 용어사전. *서울: 시그마프레스.*

김가온, 이동화, 안승민, 문혜성, & 조완영. (2013). Stirling (스털링) 수와 생성함수에 대한 연구. *한국수학교육학회 학술발표논문집, 2013*(단일호), 333-337.

김도환・황상민・최윤식 (2011). 정치 현상에 대한 한국인의 인식 유형과 심리적 기제. 한국심리학회지: 사회 및 성격, 25(2), 183-208.

김범구, 조아미 (2013). 잠재계층분석(LCA)을 이용한 학업중단 청소년의 유형과 특성. 한국청소년연구, 24(3), 5-31.

김진호, 홍세희, 추병대 (2007). 경영학 연구에서의 구조방정식 모형의 적용: 문헌 연구와 비판. 경영학연구, 36(4), 897-923.

김치수 (2015). 쉽게 배우는 소프트웨어 공학. *한빛아카데미.*

노형진 (2014). SPSS를 활용한 조사방법 및 통계분석. *학현사.*

수학백과 (2015). *대한수학회*

양경욱, 윤세준, & 채연주 (2016). 감정노동의 사람 중심적 접근. *대한경영학회지, 29*, 1435-1459.

이강원, 송호웅 (2016). 지형 공간정보체계 용어사전. *구미서관.*

이목화, 문형구, & 권석균 (2016). 조직몰입 프로파일 유형과 선행요인: 조직몰입에 대한 사람중심 (Person-Centered) 의 접근. *조직과 인사관리연구, 40*, 93-131.

이정진 (2011). R, SAS, MS-SQL을 활용한 데이터마이닝. *자유아카데미.*

이현숙, 김수진, & 전수현 (2010). 구조방정식모형 원리와 적용. *서울: 학지사.*

이훈영 (2013). SPSS를 이용한 데이터분석. *청람.*

컴퓨터인터넷IT용어대사전 (2011). *일진사*

한국성인교육학회 (1998). 교육평가 용어사전. *학지사.*

Agresti, A. (2010). *Analysis of ordinal categorical data* (Vol. 656). John Wiley & Sons.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*(3), 317-332.

Astrahan, M. M. (1970). *Speech analysis by clustering, or the hyperphoneme method* (No. AIM-124). STANFORD UNIV CALIF DEPT OF COMPUTER SCIENCE.

Ball, G. H., & Hall, D. J. (1967). A clustering technique for summarizing multivariate data. *Systems Research and Behavioral Science*, *12*(2), 153-155.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 803-821.

Bauer, D. J., & Shanahan, M. J. (2007). Modeling complex interactions: Person-centered and variable-centered approaches. *Modeling contextual effects in longitudinal studies*, 255-283.

Bensmail, H., Celeux, G., Raftery, A. E., & Robert, C. P. (1997). Inference in model-based cluster analysis. *Statistics and Computing*, *7*(1), 1-10.

Bergman, L. R., & Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Development and psychopathology*, *9*(02), 291-319.

Bishop, C. M. (2006). Pattern recognition. *Machine Learning*, *128*, 1-58.

Block, J.(1971). *Lives Through Time*. Berkely, Calif: Bancroft Books

Bradley, P. S., & Fayyad, U. M. (1998, July). Refining Initial Points for K-Means Clustering. In *ICML* (Vol. 98, pp. 91-99).

Brusco, M. J., & Cradit, J. D. (2001). A variable-selection heuristic for K-means clustering. Psychometrika, 66(2), 249-270.

Carvalho, R., Stubstad, R., Briggs, R., Selezneva, O., Mustafa, E., & Ramachandran, A. (2012). Simplified techniques for evaluation and interpretation of pavement deflections for network-level analysis (No. FHWA-HRT-12-023).

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, *40*(1), 200-210.

Cheeseman, P., and Stutz, J. (1995). "Bayesian classification (Autoclass): Theory and Results" *Advances in knowledge discovery and data mining* (Vol. 21). U. M. Fayyad, P. Smyth, & R. Uthurusamy (Eds.). Menlo Park: AAAI press.

Collins, L. M., & Lanza, S. T. (2013). Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences (Vol. 718). John Wiley & Sons.

Collins, L. M., Wugalter, S. E., & Fidler, P. L. (1996). Some Practical Issues Related to the Estimation of Latent Class and Latent Transition Parameters-6.

Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society. Series A (General)*, 321-367.

Cox, D. R. (1957). Note on grouping. *Journal of the American Statistical Association*, *52*(280), 543-547.

Deb, P. (2008). Finite mixture models. *Summer north American stata users' group, Stata Users Group*.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.

Downton, M., & Brennan, T. (1980). Comparing classifications: an evaluation of several coefficients of partition agreement. Class. Soc. Bull, 4(4), 53-54.

Dreger, R. M., Fuller, J., & Lemoine, R. L. (1988). Clustering seven data sets by means of some or all of seven clustering methods. Multivariate behavioral research, 23(2), 203-230.

Du, Q., & Wong, T. W. (2002). Numerical studies of MacQueen's k-means algorithm for computing the centroidal Voronoi tessellations. *Computers & Mathematics with Applications*, *44*(3), 511-523.

Dubes, R., & Jain, A. K. (1976). Clustering techniques: the user's dilemma. Pattern Recognition, 8(4), 247-260.

Engelman, L., & Hartigan, J. A. (1969). Percentage points of a test for clusters. *Journal of the American Statistical Association*, *64*(328), 1647-1648.

Eshghi, A., Haughton, D., Legrand, P., Skaletsky, M., & Woolford, S. (2011). Identifying groups: A comparison of methodologies. Journal of Data Science, 9(2), 271-292.

Everitt, B. S., Landau, S., & Leese, M. (1993). Cluster Analysis (Edward Arnold, London). *Google Scholar*.

Falkenauer, E., & Marchand, A. (2001, June). Using k-means? consider arrayminer. In *International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences, Las Vegas*.

Fisher,W. D. (1958). On grouping for maximum homogeneity. Journal of the American Statistical Association, 53, 789 - 798.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, *21*, 768-769.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. Journal of the American statistical association, 78(383), 553-569.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The computer journal*, *41*(8), 578-588.

González, R. C., & Tou, J. T. (1974). Pattern recognition principles. *Applied Mathematics and Computation. Reading (MA): Addison-Wesley*.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part IA modified latent structure approach. *American Journal of Sociology*, 1179-1259.

Griffiths, F., Borkan, J., Byrne, D., Crabtree, B. F., Dowrick, C., Gunn, J., ... & Reis, S. (2010). Developing evidence for how to tailor medical interventions for the individual patient. *Qualitative Health Research*, *20*(12), 1629-1641.

Hagenaars, J. A. (1990). *Categorical longitudinal data: Log-linear panel, trend, and cohort analysis*. Sage Publications, Inc.

Hajnal, I., & Loosveldt, G. (2000). The effects of initial values and the covariance structure on the recovery of some clustering methods. In *Data analysis, classification, and related methods* (pp. 47-52). Springer Berlin Heidelberg.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100-108.

Harvey, D. (1989). *The condition of postmodernity* (Vol. 14). Oxford: Blackwell.

Hickendorff, M., van Putten, C. M., Heiser, W. J., & Verhelst, N. D. (2005). Clustering Nominal Data with Equivalent Categories: a Simulation Study Comparing Restricted GROUPALS and Restricted Latent Class Analysis.

Höppner, F. (1999). *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*. John Wiley & Sons.

Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of classification, 2(1), 193-218.

Jancey, R. C. (1966). Multidimensional group analysis. *Australian Journal of Botany*, *14*(1), 127-130.

Jorgensen, M., & Hunt, L. (1996, August). Mixture model clustering of data sets with categorical and continuous variables. In *Proceedings of the Conference ISIS* (Vol. 96, pp. 375-384).

Kogan, J. (2007). *Introduction to clustering large and high-dimensional data*. Cambridge University Press.

Kozak, M., & Scaman, C. H. (2008). Unsupervised classification methods in food sciences: discussion and outlook. *Journal of the Science of Food and Agriculture*, *88*(7), 1115-1127.

Lanza, S. T., Flaherty, B. P., & Collins, L. M. (2003). Latent class and latent transition analysis. Handbook of psychology.

Laursen, B. P., & Hoff, E. (2006). Person-centered and variable-centered approaches to longitudinal data. *Merrill-Palmer Quarterly*, *52*(3), 377-389.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis Measurement and Prediction, SA Stouffer et al.

Little, M. A., & Jones, N. S. (2011, November). Generalized methods and solvers for noise removal from piecewise constant signals. I. Background theory. In Proc. R. Soc. A (Vol. 467, No. 2135, pp. 3088-3114). The Royal Society.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

Magidson, J., & Vermunt, J. K. (2002a). Latent class modeling as a probabilistic extension of K-means clustering. Quirk's Marketing Research Review, 20, 77-80.

Magidson, J., & Vermunt, J. (2002b). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing Research*, *20*(1), 36-43.

Magnusson, D. (2003). The person approach: Concepts, measurement models, and research strategy. *New directions for child and adolescent development*, *2003*(101), 3-23.

Magnusson, D., & Stattin, H. (2006). The person in context: A holistic interactionistic approach. *Handbook of child psychology*.

Makarenkov, V., & Legendre, P. (2001). Optimal variable weighting for ultrametric and additive trees and K-means partitioning: Methods and software. *Journal of Classification*, *18*(2), 245-271.

McCutcheon, A. L. (1987). *Latent class analysis* (No. 64). Sage.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering* (Vol. 84). Marcel Dekker.

McLachlan, G. J., & Peel, D. (1996). An algorithm for unsupervised learning via normal mixture models. *ISIS: Information, Statistics and Induction in Science, DL Dowe, KB Korb, and JJ Oliver (Eds.)*, 354-363.

McLachlan, G. J., Peel, D., Basford, K. E., & Adams, P. (1999). The EMMIX software for the fitting of mixtures of normal and t-components. *Journal of Statistical Software*, *4*(2), 1-14.

Meyer, J. P., & Morin, A. J. (2016). A person centered approach to commitment research: Theory, research, and methodology. *Journal of Organizational Behavior*.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, *45*(3), 325-342.

Milligan, G. W. (1996). Clustering validation: results and implications for applied analyses. In *Clustering and classification* (pp. 341-375).

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159-179.

Milligan, G. W., & Isaac, P. D. (1980). The validation of four ultrametric clustering algorithms. Pattern Recognition, 12(2), 41-50.

Milligan, G. W., Soon, S. C., & Sokol, L. M. (1983). The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure. IEEE Transactions on Pattern Analysis and Machine Intelligence, (1), 40-47.

Mooijaart, A., & Van der Heijden, P. G. (1992). The EM algorithm for latent class analysis with equality constraints. *Psychometrika*, *57*(2), 261-269.

Morey, L. C., & Agresti, A. (1984). The measurement of classification agreement: An adjustment to the Rand statistic for chance agreement. Educational and Psychological Measurement, 44(1), 33-37.

Morin, A. J., Morizot, J., Boudrias, J. S., & Madore, I. (2010). A multifoci person-centered perspective on workplace affective commitment: A

latent profile/factor mixture analysis. *Organizational Research Methods*.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. Structual Equation Modeling, 14(4), 535-569.

Oberski, D. (2016). Mixture models: Latent profile and latent class analysis. In *Modern Statistical Methods for HCI* (pp. 275-287). Springer International Publishing.

Owen, S., Anil, R., Dunning, T., & Friedman, E. (2012). Mahout in action.

Pang-Ning, T., Steinbach, M., & Kumar, V. (2006). Introduction to data mining. In Library of congress (Vol. 74).

Paykel, E. S. (1971). Classification of depressed patients: a cluster analysis derived grouping. *The British Journal of Psychiatry*, *118*(544), 275-288.

Pena, J. M., Lozano, J. A., & Larranaga, P. (1999). An empirical comparison of four initialization methods for the k-means algorithm. *Pattern recognition letters*, *20*(10), 1027-1040.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336), 846-850.

Rousseeuw, P. J., & Kaufman, L. (1990). *Finding Groups in Data*. Wiley Online Library.

Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm gdbscan and its applications. Data mining and knowledge discovery, 2(2), 169-194.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Sebestyen, G. S. (1962). Decision-making processes in pattern recognition (ACM monograph series).

Späth, H. (1977). Computational experiences with the exchange method: Applied to four commonly used partitioning cluster analysis criteria. *European Journal of Operational Research*, *1*(1), 23-31.

Steinley, D. (2003). Local optima in K-means clustering: what you don't know may hurt you. *Psychological methods*, *8*(3), 294.

Steinley, D. (2004). Properties of the Hubert-Arable Adjusted Rand Index. Psychological methods, 9(3), 386.

Steinley, D. (2006). K means clustering: a half century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 1-34.

Steinley, D., & Brusco, M. J. (2011). Evaluating mixture modeling for clustering: recommendations and cautions. *Psychological Methods*, *16*(1), 63.

Stephenson, M. T., & Holbert, R. L. (2003). A Monte Carlo simulation of observable versus latent variable structural equation modeling techniques. *Communication Research, 30*(3), 332-354.

Thorndike, R. L. (1953). Who belongs in the family?. *Psychometrika*, *18*(4), 267-276.

Tonidandel, S., & Overall, J. E. (2004). Determining the number of clusters by sampling with replacement. *Psychological Methods*, *9*(2), 238.

Vermunt, J. K. (1997). Log-linear models for event histories, advanced quantitative techniques in the social sciences series.

Vermunt, J. K. (2011). K-means may perform as well as mixture model clustering but may also be much worse: comment on Steinley and Brusco (2011).

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. *Applied latent class analysis*, *11*, 89-106.

Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. *The sage encyclopedia of social sciences research methods*, 549-553.

Wang, M., & Hanges, P. J. (2011). Latent class procedures: Applications to organizational research. *Organizational Research Methods, 14*(1), 24-31.

Weisstein, E. W. (2002). Stirling number of the second kind. *triangle*, *7*, 8.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, *8*(3), 338-353.

Zhu, X. (2005). Semi-supervised learning literature survey.

ABSTRACT

Comparison of *K*-means Clustering and Latent Class Analysis Using Monte Carlo Simulation : Evaluation of Cluster Recovery using ARI

Sohee Koo
Department of Psychology
The Graduate school of
Sungshin University

This is a comparative study of *K*-means clustering and Latent class analysis in research methodology. It is consisted of 4 research questions and conducted through simulation study to compare the result of cluster recovery of *K*-means clustering and Latent class analysis using ARI(Adjusted Rand Index). Study 1 conducted latent class analysis in order to test whether there is a difference in cluster recovery depending on the level of manipulated variable. And study 2 conducted *K*-means clustering in order to test whether there is a difference in cluster recovery depending on the level of manipulated variable. Following study 3 observed whether there is an interaction effect between the manipulated variables and the analysis methods. Lastly, in study 4, the

differences in the degree of cluster recovery according to the analysis method was verified.

Sample size, number of items, number of clusters, effect size, and latent class probability were used as manipulated variables and item response probability was specified with no cross loading. Based on the factorial design of 3(N)×3(V)×3(K)×3(F)×3(γ) depending on the level of each manipulated variables, each 243 experimental conditions were repeated 500 times. Then, 121,500 data were generated using Monte Carlo simulation in the statistical program R.

As a result of verifying the study 1 after analyzing the generated data with latent class analysis, it showed that there was no significant difference in the degree of cluster recovery according to the sample size and number of clusters. On the other hand there was a significant difference in the degree of cluster recovery according to the number of items, effect size, and latent class probability. The result of study 2 according to K-means clustering, it showed that there was no significant difference in the degree of cluster recovery according to the sample size and number of items. On the other hand there was a significant difference in the degree of cluster recovery according to the number of clusters, effect size, and latent class probability. In order to observe results from both latent class analysis and K-means clustering, the interaction effect between analysis methods and sample size, analysis methods and number of items on cluster recovery were not statistically significant. But the interaction effect between analysis methods and number of clusters, analysis methods and effect size, analysis methods

and latent class probability on cluster recovery were statistically significant. Finally the result of study 4 showed that latent class analysis showed higher cluster recovery than K-means clustering overall.

Besides these four studies, this dissertation study examined the cases where the convergence did not occur even after the specified number of iterations in the process of performing latent class analysis and K-means clustering to verify the efficiency of estimation. While no NC occurred in K-means clustering, 141 NC which is about 0.1% of the total of 121,500 data, were occurred in latent class analysis. This result suggested that convergence rate can be lower even though the degree of cluster recovery is high enough. It also includes the suggest of using appropriate analytical methods for different research problems as there are pros and cons that each methods can have depending on the type and condition of the data given to the researchers. Finally, a proposal for further research was discussed.

Key Words: Cluster recovery, ARI, Latent Class Analysis, K-means Clustering, Monte Carlo Simulation, Program R

부 록

<부록 1> 자료 생성을 위한 R 코드

<부록 2> K-평균군집분석 시뮬레이션을 위한 R 코드

<부록 3> 잠재계층분석 시뮬레이션을 위한 R 코드

<부록 4> 문항반응확률 표

<부록 1> 자료 생성을 위한 R 코드

```
#-------- Condition --------#
# n: 사례수 (200 , 500 , 1000)
# v: 문항비율 (1 , 2 , 3)
# k: 군집수 (2 , 4 , 6)
# f: 효과크기 (0.2 , 0.5 , 0.8)
# r: 잠재집단확률 (even , 60% , 10%)
#
# Number of Simulation : 500
#--------------------------#

# 필요한 패키지 로딩
library(cluster)
library(e1071)
library(dplyr)

#필요한 자료 로딩
table_2 <- read.csv('k2_table.csv' , header = T , sep = ',' )
table_4 <- read.csv('k4_table.csv' , header = T , sep = ',' )
table_6 <- read.csv('k6_table.csv' , header = T , sep = ',' )

# 잠재집단 계산을 위한 함수 생성
p_even <- function(n , k)  c(rep(round(n/k) , (k-1)) , (n -
(round(n/k)*(k-1))))
```

```
p_uneven <- function(n , k , p)  c(round(n * p) , rep( round((n - round(n * p))/(k-1)) , (k-2)) , (n - round(n * p) - round((n - round(n * p))/(k-1))*(k-2)))

# theta를 p_theta로 변환하는 함수 생성
p_theta <- function(theta , b) (1 / (1 + exp(-1.7 * (theta + b))))
p_bound <- function(theta , table , n , v , k) {
  res_mat <- matrix(0 , nrow = n , ncol = (v*k))
  for (z in 1 : n) {
    for (y in 1 : (v*k)) {
      res_mat[z,y] <- p_theta(thetas[z] , table[true_cluster[z], y])
    }}
  return(res_mat)
}

# result matrix의 생성
gr_data <- matrix(NA , nrow = 243 , ncol = 505)
colnames(gr_data) <- c('n' , 'v' , 'k' , 'f' , 'r' , c(1:500))
gr_data[,'n'] <- rep(c(200 , 500 , 1000) , each = 81 , times = 1)
gr_data[,'v'] <- rep(c(1 , 2 , 3) , each = 27 , times = 3)
gr_data[,'k'] <- rep(c(2 , 4 , 6) , each = 9 , times = 9)
gr_data[,'f'] <- rep(c(0.2 , 0.5 , 0.8) , each = 3 , times = 27)
gr_data[,'r'] <- rep(c(0.5 , 0.6 , 0.1) , 81)
gr_data <- as.data.frame(gr_data)
```

```
# Make process bar (: optional)
p <- progress_estimated(243)

# Start Simulation

for (i in 1:243) {
  p$tick()$print() # optional

  for (j in 1:500) {

    # define variables
    n <- gr_data[i , 1]
    v <- gr_data[i , 2]
    k <- gr_data[i , 3]
    f <- gr_data[i , 4]
    r <- gr_data[i , 5]
    n_each = c()

    if(r == 0.5) {
      n_each = p_even(n , k) } else {
        n_each = p_uneven(n , k , r) }

    if(k == 2) {
      table_k <- table_2} else if (k == 4) {
        table_k <- table_4} else {table_k <- table_6}
```

```
cluster_mean <- 0 + f * (0 : (k-1))

# design answer for this condition

true_cluster <- c(rep(1:k , n_each))

# Simulate data

# theta's for each person in each cluster
thetas <- rnorm(n_each[1] , cluster_mean[1] , 1)
for(each in 2:k) {
  thetas <- c(thetas , rnorm(n_each[each] , cluster_mean[each] , 1))
}

# Sample from uniform distribution
Sim_data_temp <- matrix( runif(n*v*k) , ncol = (v*k) , nrow = n)

# Decide 0 or 1

# Cut off boundary: p_theta
mat_ind <- (1 : (k*v))
if(v != 1) { mat_ind <- ((k * sum(1 : (v-1)) + 1) : (k * sum(1 : v))) }
```

```
    table_use <- table_k[ , mat_ind]

    Threshold <- p_bound(thetas , table_use , n , v , k)

    # In each Cluster, decide 0 or 1

    mymy <- Sim_data_temp < Threshold
    mymy <- ifelse(mymy == T , 1 , 0)
    write.csv(mymy , paste(i , "_" , j , ".csv")  , row.names = F)
}}
```

<부록 2> K-평균군집분석 시뮬레이션을 위한 R 코드

```
# Title : ARI for KMEANS

#-------- Condition --------#
# n: 사례수 (200 , 500 , 1000)
# v: 문항비율 (1 , 2 , 3)
# k: 군집수 (2 , 4 , 6)
# f: 효과크기 (0.2 , 0.5 , 0.8)
# r: 잠재집단확률 (even , 60% , 10%)
#
# Number of Simulation : 500
#--------------------------#

library(cluster)
library(e1071)
library(dplyr)

# ARI 계산을 위한 함수
my_ARI <- function(mat , n_combi){

  n_combi <- as.numeric(n_combi)
  mat <- as.matrix(mat)

  if( dim(mat)[2] == 0) {
```

```
        return(NA) } else{

          a <- as.numeric(mat[2 , 2])
          b <- as.numeric(mat[2 , 1])
          c <- as.numeric(mat[1 , 2])
          d <- as.numeric(mat[1 , 1])
          res <- (n_combi * (a+d) - ((a+b)*(a+c) + (c+d)*(b+d))) / (
n_combi^2 - ((a+b)*(a+c) + (c+d)*(b+d)))
          return(res)
        }
  }

  # result matrix의 생성
  result_KMEANS <- matrix(NA , nrow = 243 , ncol = 505)
  colnames(result_KMEANS) <- c('n' , 'v' , 'k' , 'f' , 'r' , c(1:500))
  result_KMEANS[,'n'] <- rep(c(200 , 500 , 1000) , each = 81 , times = 1)
  result_KMEANS[,'v'] <- rep(c(1 , 2 , 3) , each = 27 , times = 3)
  result_KMEANS[,'k'] <- rep(c(2 , 4 , 6) , each = 9 , times = 9)
  result_KMEANS[,'f'] <- rep(c(0.2 , 0.5 , 0.8) , each = 3 , times = 27)
  result_KMEANS[,'r'] <- rep(c(0.5 , 0.6 , 0.1) , 81)

  result_KMEANS <- as.data.frame(result_KMEANS)

  table_2 <- read.csv('k2_table.csv' , header = T , sep = ',' )
  table_4 <- read.csv('k4_table.csv' , header = T , sep = ',' )
```

```
table_6 <- read.csv('k6_table.csv' , header = T , sep = ',' )

# KMEANS

for (i in 1 : 243){
  for (j in 1 : 500) {

    # 데이터 로딩
    Sim_data <-  read.csv(paste(i , "_" , j , ".csv")  , header = T , sep
= ',')

    # define variables
    n <- result_KMEANS[i , 1]
    v <- result_KMEANS[i , 2]
    k <- result_KMEANS[i , 3]
    f <- result_KMEANS[i , 4]
    r <- result_KMEANS[i , 5]
    n_each = c()

    if(r == 0.5) {
      n_each = p_even(n , k) } else {
        n_each = p_uneven(n , k , r) }

    if(k == 2) {
      table_k <- table_2} else if (k == 4) {
```

```
        table_k <- table_4} else {table_k <- table_6}

    cluster_mean <- 0 + f * (0 : (k-1))

    # design answer for this condition

    true_cluster <- c(rep(1:k , n_each))

    # K-means Clustering

    res_kmeans <- kmeans(Sim_data , k , nstart=10, iter.max=k*10)$cluster

    # Compute ARI

    # all pairs of combination

    my_combi <- combn(1 : n , 2)

    # True Cluster Combi; whether two are in same group or not
    True_YN <- true_cluster[my_combi[1,]] == true_cluster[my_combi[2,]]
    KMEANS_YN <- res_kmeans[my_combi[1,]] == res_kmeans[my_combi[2,]]

    n_combi <- ncol(my_combi)
```

```
        KMEANS_tb <- as.matrix(table(True_YN , KMEANS_YN))

        result_KMEANS[i , (j+5)] <- my_ARI(KMEANS_tb , n_combi)

        }
}

write.csv(result_KMEANS , 'result_KMEANS.csv' , row.names = F)
```

<부록 3> 잠재계층분석 시뮬레이션을 위한 R 코드

```
# Title : ARI for LCA

#-------- Condition --------#
# n: 사례수 (200 , 500 , 1000)
# v: 문항비율 (1 , 2 , 3)
# k: 군집수 (2 , 4 , 6)
# f: 효과크기 (0.2 , 0.5 , 0.8)
# r: 잠재집단확률 (even , 60% , 10%)
#
# Number of Simulation : 500
#--------------------------#

library(cluster)
library(e1071)
library(dplyr)

# 잠재집단 계산을 위한 함수 생성
p_even <- function(n , k)  c(rep(round(n/k) , (k-1)) , (n - (round(n/k)*(k-1))))
p_uneven <- function(n , k , p)  c(round(n * p) , rep( round((n - round(n * p))/(k-1)) , (k-2)) , (n - round(n * p) - round((n - round(n * p))/(k-1))*(k-2)))
```

```
# ARI 계산을 위한 함수
my_ARI <- function(mat , n_combi){

  n_combi <- as.numeric(n_combi)
  mat <- as.matrix(mat)

  if( dim(mat)[2] == 0) {
    return(NA) } else{

      a <- as.numeric(mat[2 , 2])
      b <- as.numeric(mat[2 , 1])
      c <- as.numeric(mat[1 , 2])
      d <- as.numeric(mat[1 , 1])
      res <- (n_combi * (a+d) - ((a+b)*(a+c) + (c+d)*(b+d))) / (
n_combi^2 - ((a+b)*(a+c) + (c+d)*(b+d)))
      return(res)
    }
}

# result matrix의 생성
result_LCA <- matrix(NA , nrow = 243 , ncol = 505)
colnames(result_LCA) <- c('n' , 'v' , 'k' , 'f' , 'r' , c(1:500))
result_LCA[,'n'] <- rep(c(200 , 500 , 1000) , each = 81 , times = 1)
result_LCA[,'v'] <- rep(c(1 , 2 , 3) , each = 27 , times = 3)
result_LCA[,'k'] <- rep(c(2 , 4 , 6) , each = 9 , times = 9)
```

```
result_LCA[,'f'] <- rep(c(0.2 , 0.5 , 0.8) , each = 3 , times = 27)
result_LCA[,'r'] <- rep(c(0.5 , 0.6 , 0.1) , 81)

result_LCA <- as.data.frame(result_LCA)

table_2 <- read.csv('k2_table.csv' , header = T , sep = ',' )
table_4 <- read.csv('k4_table.csv' , header = T , sep = ',' )
table_6 <- read.csv('k6_table.csv' , header = T , sep = ',' )



# Make process bar (: optional)
p <- progress_estimated(243)



# LCA

for (i in 1 : 243){
  p$tick()$print() # optional

  for (j in 1 : 500) {

    #  데이터 로딩
    Sim_data <-  read.csv(paste(i , "_" , j , ".csv")  , header = T , sep
= ',')
```

```
# define variables
n <- result_LCA[i , 1]
v <- result_LCA[i , 2]
k <- result_LCA[i , 3]
f <- result_LCA[i , 4]
r <- result_LCA[i , 5]
n_each = c()

if(r == 0.5) {
  n_each = p_even(n , k) } else {
    n_each = p_uneven(n , k , r) }

if(k == 2) {
  table_k <- table_2} else if (k == 4) {
    table_k <- table_4} else {table_k <- table_6}

cluster_mean <- 0 + f * (0 : (k-1))

# design answer for this condition

true_cluster <- c(rep(1:k , n_each))

# LCA Clustering

model <- lca(x = as.matrix(Sim_data) , k = k , niter = 100)
```

```
      names(model)
      res_LCA <- predict(model , Sim_data)

      # Compute ARI

      # all pairs of combination

      my_combi <- combn(1 : n , 2)

      # True Cluster Combi; whether two are in same group or not
      True_YN <- true_cluster[my_combi[1,]] == true_cluster[my_combi[2,]]
      LCA_YN <- res_LCA[my_combi[1,]] == res_LCA[my_combi[2,]]

      n_combi <- ncol(my_combi)

      LCA_tb <- as.matrix(table(True_YN , LCA_YN))

      result_LCA[i , (j+5)] <- my_ARI(LCA_tb , n_combi)

    }
  }

write.csv(result_LCA , 'result_LCA.csv' , row.names = F)
```

<부록 4> 문항반응확률 표

군집수 2개, 문항수 2개

0.85	0.2
0.2	0.85

군집수 2개, 문항수 4개

0.85	0.85	0.1	0.2
0.1	0.2	0.85	0.85

군집수 2개, 문항수 6개

0.85	0.85	0.85	0.1	0.2	0.3
0.1	0.2	0.3	0.85	0.85	0.85

군집수 4개, 문항수 4개

0.85	0.1	0.2	0.3
0.1	0.85	0.2	0.3
0.1	0.2	0.85	0.3
0.1	0.2	0.3	0.85

군집수 4개, 문항수 8개

0.85	0.85	0.1	0.2	0.1	0.2	0.1	0.2
0.1	0.2	0.85	0.85	0.1	0.2	0.1	0.2
0.1	0.2	0.1	0.2	0.85	0.85	0.1	0.2
0.1	0.2	0.1	0.2	0.1	0.2	0.85	0.85

군집수 4개, 문항수 12개

0.85	0.85	0.85	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3
0.1	0.2	0.3	0.85	0.85	0.85	0.1	0.2	0.3	0.1	0.2	0.3
0.1	0.2	0.3	0.1	0.2	0.3	0.85	0.85	0.85	0.1	0.2	0.3
0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3	0.85	0.85	0.85

군집수 6개, 문항수 6개

0.85	0.2	0.2	0.2	0.2	0.2
0.2	0.85	0.2	0.2	0.2	0.2
0.2	0.2	0.85	0.2	0.2	0.2
0.2	0.2	0.2	0.85	0.2	0.2
0.2	0.2	0.2	0.2	0.85	0.2
0.2	0.2	0.2	0.2	0.2	0.85

군집수 6개, 문항수 12개

0.85	0.85	0.85	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3
0.1	0.2	0.3	0.85	0.85	0.85	0.1	0.2	0.3	0.1	0.2	0.3
0.1	0.2	0.3	0.1	0.2	0.3	0.85	0.85	0.85	0.1	0.2	0.3
0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3	0.85	0.85	0.85
0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3
0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3

군집수 6개, 문항수 18개

0.85	0.85	0.85	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3
0.1	0.2	0.3	0.85	0.85	0.85	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3
0.1	0.2	0.3	0.1	0.2	0.3	0.85	0.85	0.85	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3
0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3	0.85	0.85	0.85	0.1	0.2	0.3	0.1	0.2	0.3
0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3				0.85	0.85	0.85	0.1	0.2	0.3
0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3	0.85	0.85	0.85