R을 활용한 데이터분석

호서대학교 빅데이터경영공학부 연규필 (kpyeon1@hoseo.edu)

비지도학습모형

- 군집분석, 연관성분석, 주성분분석, 인자분석 -

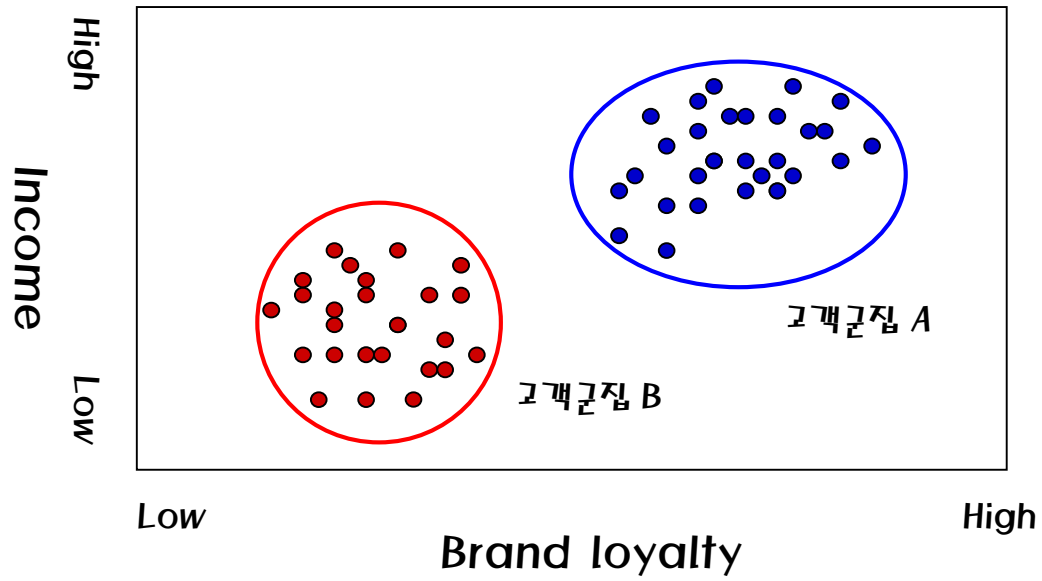
1. 군집분석
2. 연관성규칙
3. 주성분분석
4. 인자분석

■ 군집분석 (Cluster Analysis)

개체들을 유사한 것들끼리 그룹화하여 몇 개의 집단(cluster)을 구성

- 각 집단의 성격을 파악(profiling)
- 데이터 전체 구조에 대한 이해
- 데이터에 대한 의미 있는 정보 추출

≪예≫ 소득수준과 상표충성도 기준으로 고객 세분화(Segmentation)



■ 군집화 (clustering)

- **군집화의 기준**

  동일한 군집내 개체들은 여러 속성이 유사하도록,

  서로 다른 군집간 개체들은 상이한 속성을 갖도록 군집을 구성.

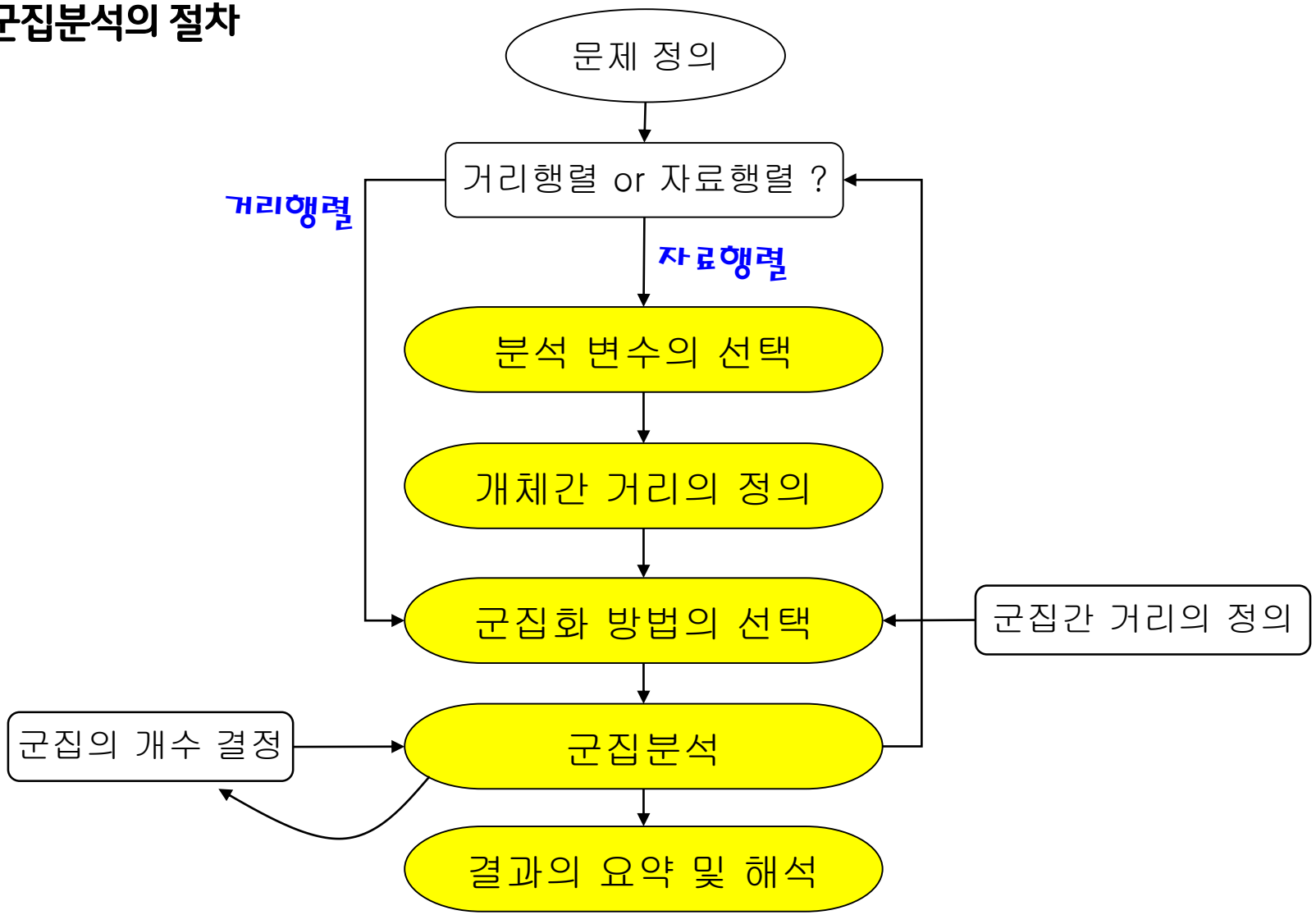
- **군집화를 위한 변수**

  전체 개체(개인)의 속성을 판단하기 위한 기준

  ≪예≫ 고객세분화

  - 인구통계적 변인 (성별, 나이, 거주지, 직업, 소득, 교육, 종교, …)
  - 구매패턴 변인 (상품, 주기, 거래액, …)
  - 생활패턴 변인 (라이프스타일, 성격, 취미, 가치관, …)

■ 군집분석의 절차

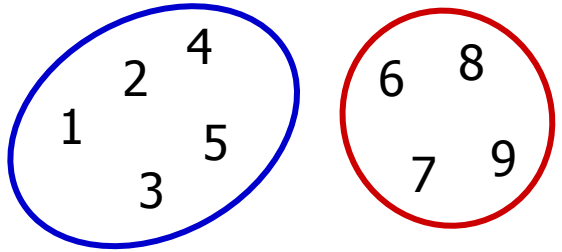

문제 정의
거리행렬 or 자료행렬 ?
거리행렬
자료행렬
분석 변수의 선택
개체간 거리의 정의
군집화 방법의 선택
군집간 거리의 정의
군집의 개수 결정
군집분석
결과의 요약 및 해석

■ 군집의 유형

• 상호배반적(disjoint) 군집

– 각 관찰치가 상호배반적인 여러 군집 중, 오직 하나에만 속함.

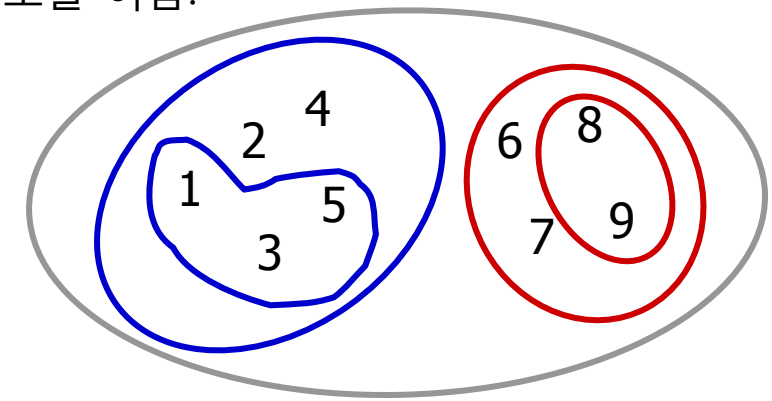
(예) 한국인, 중국인, 일본인



• 계보적(hierarchical) 군집

– 한 군집이 다른 군집의 내부에 포함되는 형태로 군집간의 중복은 없으며 군집들이 매단계 계층적인(나무) 구조를 이룸.

(예) 전자제품 → 주방용 → 냉장고



■ 군집의 유형

- 중복(overlapping) 군집
  - 두개 이상의 군집에 한 관찰치가 동시에 소속되는 것을 허용

- 퍼지(fuzzy) 군집

  - 관찰치가 소속되는 특정한 군집을 표현하는 것이 아니라 각 군집에 속할 가능성을 표현

$$\text{Prob}(\text{개체 } 1 \in \text{군집 A}) = 0.7$$
$$\text{Prob}(\text{개체 } 1 \in \text{군집 B}) = 0.3$$

■ 계층적 군집분석

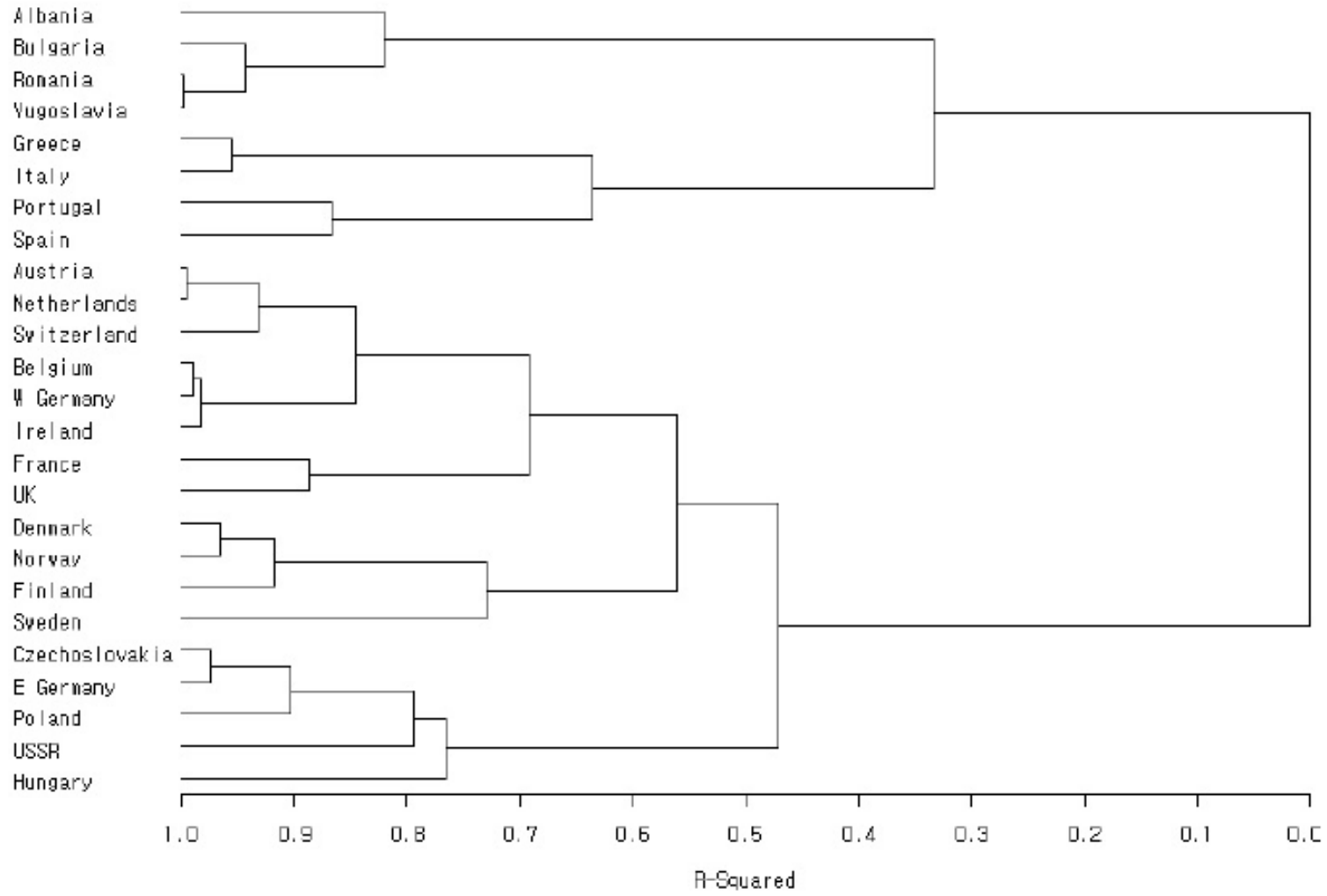
- 가까운 관측값들 끼리 묶는 병합(agglomeration)방법과
  먼 관측값들을 나누어가는 분할(division)방법으로 나눌 수 있다.

- 계층적 군집분석에서는 주로 병합 방법이 주로 사용된다.

- 계층적 군집분석의 결과는 나무구조인 덴드로그램(dendrogram)을 통해 간단하게 나타낼 수 있고, 이를 이용하여 전체 군집들간의 구조적 관계를 쉽게 살펴볼 수 있다.

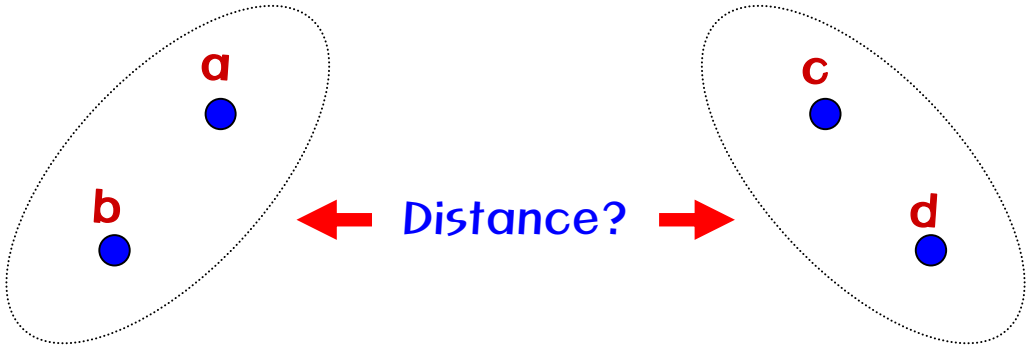
1. 군집분석

■ 덴드로그램


Albania
Bulgaria
Romania
Yugoslavia
Greece
Italy
Portugal
Spain
Austria
Netherlands
Switzerland
Belgium
W Germany
Ireland
France
UK
Denmark
Norway
Finland
Sweden
Czechoslovakia
E Germany
Poland
USSR
Hungary
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
R-Squared

■ 계층적 군집분석시 군집간 거리의 정의

- 최단 연결법 (Single linkage, 가장 가까운 항목)
- 최장 연결법 (complete linkage, 가장 먼 항목)
- 평균 연결법 (average linkage)
- 중심 연결법 (centroid method)
- 중위수 연결법 (median method)
- Ward의 방법 (Ward's minimum variance method)



■ K-평균 군집분석

- **특징**

  각 관찰치를 상호배반적인 K개의 군집을 형성

  - ✓ 초기에 부적절한 병합(분리)이 일어났을 때 회복 가능
  - ✓ 군집의 수 K를 사전에 정의
  - ✓ 대용량 자료의 경우 유용

- **알고리즘**

  [단계 0] 군집 수 K를 사전에 결정하고 각 군집 중심을 임의로 설정

  [단계 1] 각 개체를 그 중심과 가장 가까운 거리에 있는 군집에 할당

  [단계 2] 각 군집별로 [단계 1]을 통해 할당된 개체를 이용해 군집중심 재산출

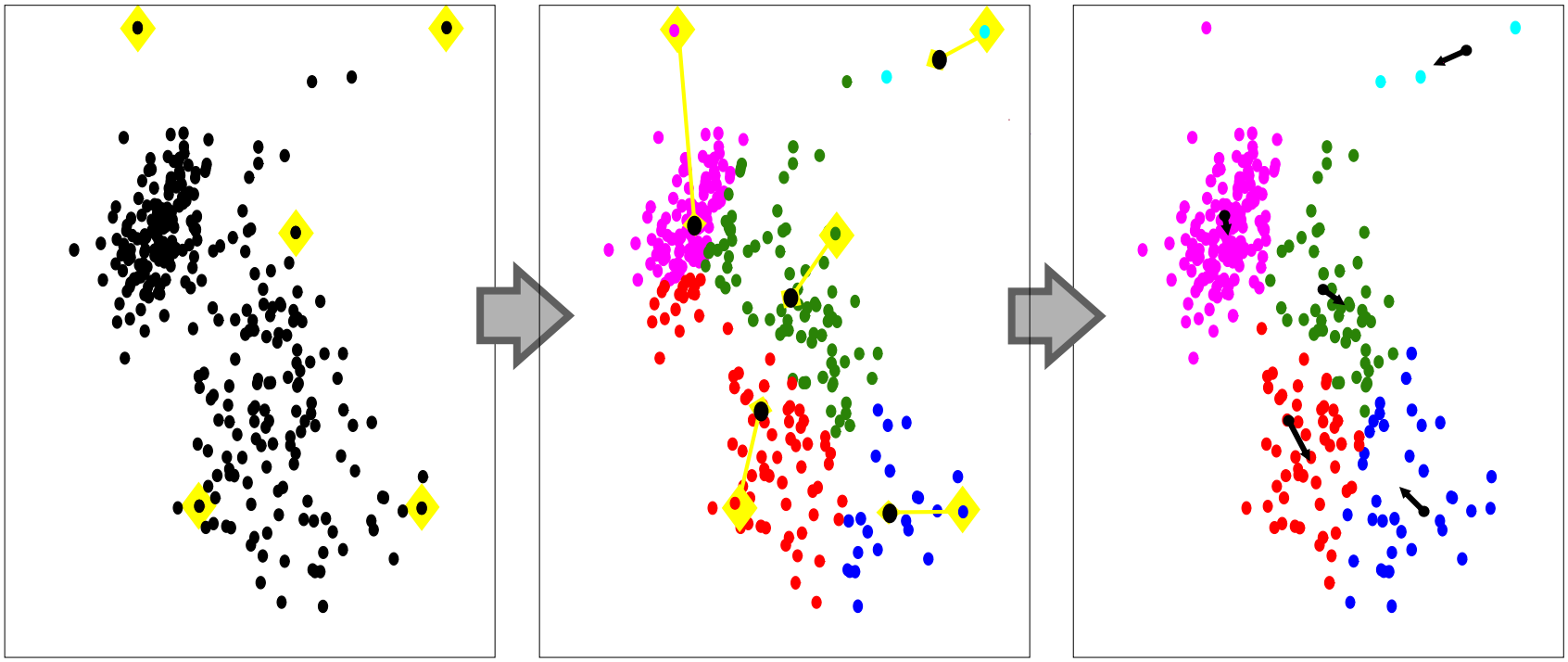
  [단계 3] [단계 1]과 [단계 2]의 과정을 기존 중심과 새로운 중심의 차이가 없을 때까지 반복

■ K-평균 군집분석의 절차

군집의 수 K 결정 : K=5
최초 군집기준값 결정

개체의 할당
군집중심 재 산출

개체의 할당
군집중심 재 산출 (반복)



(단계 0)

(단계 1, 2)

(단계 3)

■ 초기 군집수의 결정

방법 1

- 일부 표본을 이용하여 계층적 군집분석을 수행한 후
  적절한 초기 군집수를 결정한 후

- 선택된 군집수를 이용하여 전체 자료에 대하여 k-평균 군집화를 수행

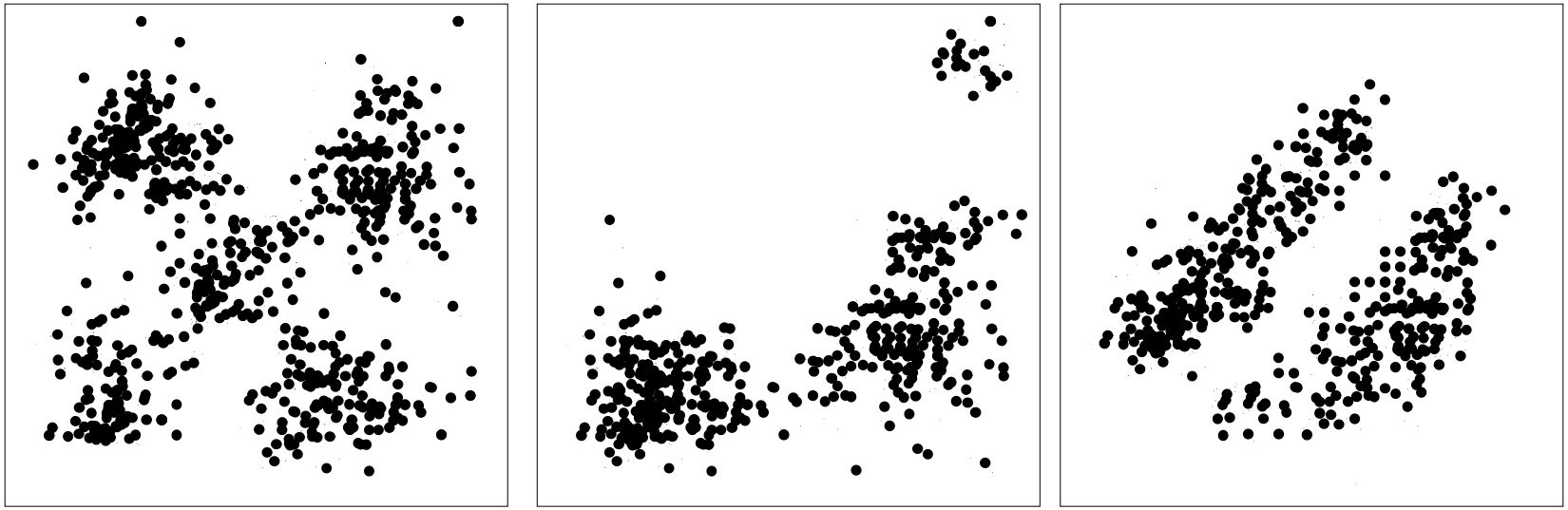
방법 2

- 여러 군집수 (k=2,3,4,…)에 대하여 군집분석을 수행한 후,
  적절한 지표에 따른 적정 군집수를 선택

  - 군집수 선택 지표: Silhouette width, PG(Pearson Gamma) 통계량 등

■ K-평균 군집화의 주의점

- 군집 수 K의 사전 결정
- 초기 군집중심의 설정
- 특이점 또는 자료가 내포한 특이한 군집구조
- 변수의 표준화



■ K-평균 군집화의 특징

- 장점
  - 탐색적인 기법
  - 다양한 형태의 데이터에 적용 가능
  - 분석방법의 적용 용이성

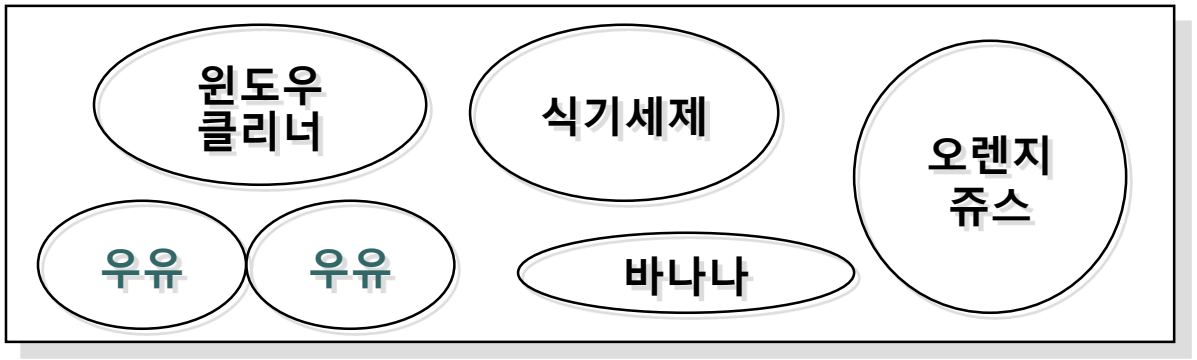
- 단점
  - 가중치와 거리의 정의
  - 초기 군집 수의 설정
  - 결과 해석의 어려움

2. 연관성 규칙

■ 연관성 규칙 발견의 개념

둘 이상의 거래나 사건에 포함되어 있는 항목들의 관련성을 파악하는 탐색적 데이터 분석 기법

≪ex≫ Products in Shop Cart (One trip, Together)



1) '오렌지 주스와 식기세제' 구입시 '윈도우 클리너'를 같이 구입하는가?

2) '우유'를 '바나나' 구입시 함께 구입하는가?

또한 구입 할 때 특정 브랜드를 구입하는가?

3) '식기세제'를 어는 곳에 위치시켜야지만 판매고를 최대화하는가?

2. 연관성 규칙

■ 의미있는 연관성 규칙

어떤 항목(들)의 존재가 다른 항목(들)의 존재를 암시하는 것을 의미

(항목 집합 A) ⟶ (항목집합 B)

(if A then B : 만일 A 가 일어나면 B 가 일어난다.)

- 연관성 규칙의 예
  - 신발 구매 → 양말 구입
  - 최근에 구좌정리와 이율상담을 요구 → 이후 1달 내에 거래 중단
- 해석
  - 원인과 결과의 직접적인 인과관계가 아님
  - 둘 또는 그 이상 품목들 사이의 상호 관련성

2. 연관성 규칙

■ 연관성 규칙의 결과 유형

- Useful Result

  - 마케팅 전략상 유용한 결과가 나온 경우

    ex) 주말을 위해, 목요일 소매점에 기저귀를 사러 온 아빠들은 맥주도 함께 사간다.

- Trivial Result

  - 기존의 마케팅 전략에 의해 연관성이 높게 나온 경우

    ex) 정비계약을 맺은 소비자들은 많은 설비를 구매 할 것 같다.

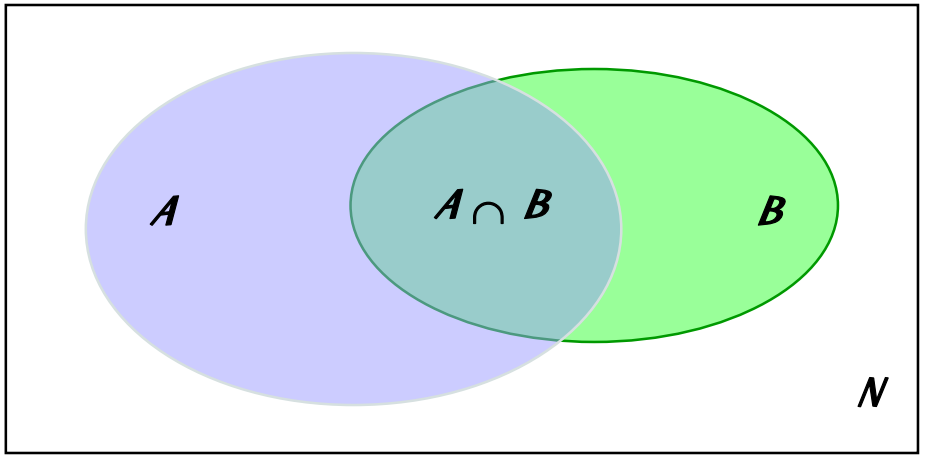
- Inexplicable Result

  - 의미를 발견하기 위해 많은 고민이 필요한 경우

    ex) 새로 철물점을 개업하면, 대개 화장실 문고리를 많이 사 간다.

■ 지지도 (Support)

전체 거래 중 항목 A 와 항목 B 를 동시에 포함하는 거래의 비율



$$\text{Support}(A \Rightarrow B) = \Pr(A \cap B)$$

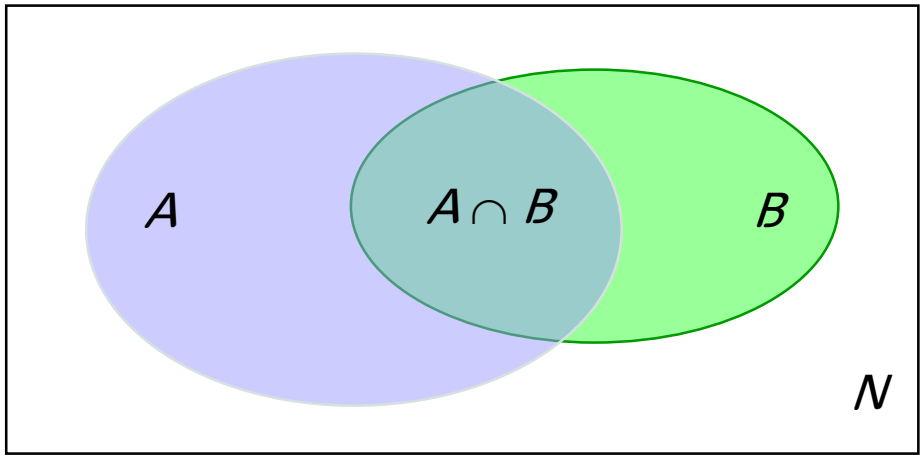
$$= \frac{\text{항목 } A \text{와 항목 } B \text{가 동시에 포함된 거래수}}{\text{전체 거래수 } (N)}$$

$$= \text{Support}(B \Rightarrow A)$$

2. 연관성 규칙

■ 신뢰도 (Confidence)

항목 A 를 포함하는 거래 중에서 항목 B 가 포함된 거래의 비율



$$\text{Confidence}(A \Rightarrow B) = \Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

$$= \frac{\text{항목 } A \text{와 항목 } B \text{가 동시에 포함된 거래수}}{\text{항목 } A \text{가 포함된 거래수}}$$

$$\text{Coverage}(A) = \Pr(A) = \frac{\text{항목 } A \text{가 포함된 거래수}}{\text{전체 거래수}(N)}$$

2. 연관성 규칙

■ 연관성 규칙 발견의 예

(a) 요약된 거래 데이터

항목	거래의 수
버섯	100
페페로니	150
치즈	200
버섯 + 페페로니	400
버섯 + 치즈	300
페페로니 + 치즈	200
버섯 + 페페로니 + 치즈	100
추가토핑 안함	550
합 계	2,000

(b) 재구성된 데이터

항목	항목이 포함된 거래의 수	포함률
버섯	100+400+300+100=900	45.0%
페페로니	150+400+200+100=850	42.5%
치즈	200+300+200+100=800	40.0%
버섯 + 페페로니	400+100=500	25.0%
버섯 + 치즈	300+100=400	20.0%
페페로니 + 치즈	200+100=300	15.0%
버섯 + 페페로니 + 치즈	100	5.0%

(c) 지지도와 신뢰도의 계산

규칙 $(A \Rightarrow B)$	지지도, $Pr(A \cap B)$	신뢰도, $Pr(B\|A)$
버섯 ⇒ 페페로니	25%	25/45=55.6%
(버섯 + 페페로니) ⇒ 치즈	5%	5/25=20.0%
(버섯 + 치즈) ⇒ 페페로니	5%	5/20=25.0%
(페페로니 + 치즈) ⇒ 버섯	5%	5/15=33.3%
...	...	...

■ 향상도 (Lift)

항목 A 를 구매한 경우, 그 거래가 '항목 B 를 포함하는 경우'와 '항목 B에 구매여부와 관계없이 구매되는 경우'의 비

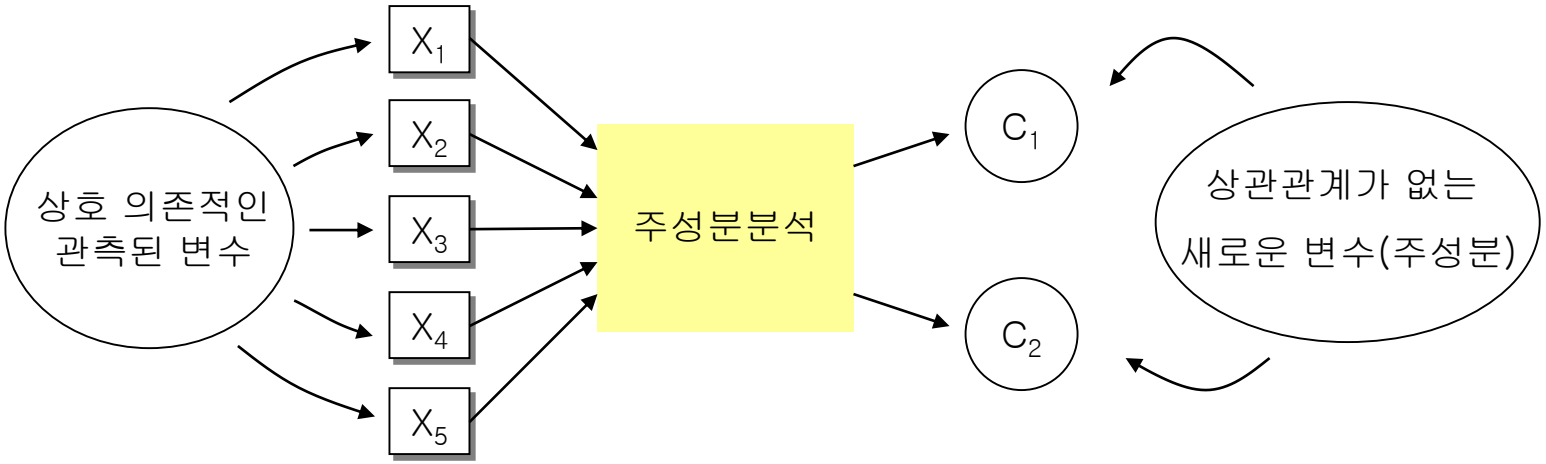
$$\text{Lift}(A \Rightarrow B) = \frac{\Pr(B \mid A)}{\Pr(B)} = \frac{\Pr(A \cap B)}{\Pr(A) \cdot \Pr(B)}$$

규칙 $(A \Rightarrow B)$	$Pr(A \cap B)$	$Pr(A)$	$Pr(B)$	향상도
버섯 ⇒ 페페로니	25%	45.0%	42.5%	1.31
(버섯 + 페페로니) ⇒ 치즈	5%	25.0%	40.0%	0.50
(버섯 + 치즈) ⇒ 페페로니	5%	20.0%	42.5%	0.59
(페페로니 + 치즈) ⇒ 버섯	5%	15.0%	45.0%	0.74

Lift	의미	예
1	두 품목이 서로 독립적인 관계	과자와 후추
> 1	두 품목이 서로 양의 상관 관계	빵과 버터
< 1	두 품목이 서로 음의 상관 관계	지사제와 변비약

■ 주성분 분석 (PCA: Principal Component Analysis

주성분분석의 이해



- **주성분**
  - ✓ 주성분은 측정된 변수들의 선형결합(linear combination)으로 구성됨

$$C_1 = w_1X_1 + w_2X_2 + w_3X_3 + w_4X_4 + w_5X_5$$

  - ✓ 변수를 가장 잘 설명할 수 있는 선형결합의 계수 값을 찾는 것이 주성분분석

■ 선형변환과 주성분

- 주성분분석은 차원의 단순화를 통해 서로 상관되어 있는 변수들 간의 복잡한 구조를 분석하는 데 그 목적을 두고 있으며, 이를 위하여 관찰변수들을 선형변환시켜 '주성분'(principal component)이라고 불리는 서로 상관되어 있지 않은(혹은 독립적인) 새로운 인공변수들을 유도한다.

obs	x_1	x_2	x_3	x_4	x_5		평점	주성분
1	3	33	73	8	12		25.8	16.0
2	3	30	59	28	20		28.0	21.4
3	35	83	91	32	34		55.0	45.4
4	35	83	85	33	32	$\longrightarrow$	53.6	44.8
5	15	40	55	68	52	차원축소	46.0	43.5
6	3	53	76	10	8		30.0	18.8
7	68	83	85	48	50		66.8	62.4
8	15	47	77	76	76		56.2	53.6
9	46	60	83	83	68		67.8	65.1
10	98	83	91	80	72		84.8	84.9
평균	32.1	59.5	77.5	46.6	42.4		51.4	45.6
표준편차	31.6	22.0	12.4	28.6	25.0		19.3	22.2

■ 주성분 점수

- 선형변환 (Linear Transformation)
  - ✓ $Y = a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5$
  - ✓ 주성분점수는 동일한 제약조건을 가지는 모든 가능한 선형결합 중 가장 변이가 크다는 (즉, 개체들의 상대 위치를 멀리 떨어뜨려 놓는다는) 점에서 최적의 성질을 가진다.
  - ✓ 주성분점수는 관찰변수들과 최대 다중상관계수를 가진다.

(상관계수)	국어	영어	제2외국어	수학	과학
국어	1.000				
영어	0.784	1.000			
제2외국어	0.683	0.860	1.000		
수학	0.559	0.212	0.138	1.000	
과학	0.610	0.309	0.279	0.973	1.000
(표준편차)	31.6	22.0	12.4	28.6	25.0

  - ✓ 평균점수 = $0.20\ x_1 + 0.20\ x_2 + 0.20\ x_3 + 0.20\ x_4 + 0.20\ x_5$
  - ✓ 주성분점수 = $0.30\ x_1 + 0.15\ x_2 + 0.07\ x_3 + 0.25\ x_4 + 0.23\ x_5$

3. 주성분 분석

■ 주성분의 성질

- Σ : **공분산행렬** ($p \times p$)
- $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_k \geq \cdots \geq \delta_{p-1} \geq \delta_p$: **고유값**

  $\mathbf{e}_k = (e_{1k}, e_{2k}, \cdots, e_{pk})'$: 고유벡터, $k=1, \cdots, p$

- **선형결합** $y = \mathbf{a}'\mathbf{x} = a_1x_1 + a_2x_2 + \cdots + a_px_p$ 에 대하여

① $\mathrm{Var}(y)=\mathrm{Var}(\mathbf{a}'\mathbf{x})$를 최대로 하는 $\mathbf{a}$를 구하면?

→ $\mathbf{a}=\mathbf{e}_1$, $\mathrm{Var}(\mathbf{e}_1'\mathbf{x})=\mathrm{Var}(y_1)=\delta_1$, y_1: 첫번째 주성분

② $\mathrm{Cov}(\mathbf{e}_1'\mathbf{x}, \mathbf{a}'\mathbf{x})=0$이고, $\mathrm{Var}(y)=\mathrm{Var}(\mathbf{a}'\mathbf{x})$를 최대로 하는 $\mathbf{a}$를 구하면?

→ $\mathbf{a}=\mathbf{e}_2$, $\mathrm{Var}(\mathbf{e}_2'\mathbf{x})=\mathrm{Var}(y_2)=\delta_2$, y_2: 두번째 주성분

③ $\mathrm{Cov}(\mathbf{e}_1'\mathbf{x}, \mathbf{a}'\mathbf{x})=\mathrm{Cov}(\mathbf{e}_2'\mathbf{x}, \mathbf{a}'\mathbf{x})=0$이고, $\mathrm{Var}(y)=\mathrm{Var}(\mathbf{a}'\mathbf{x})$를 최대로 하는 $\mathbf{a}$를 구하면?

→ $\mathbf{a}=\mathbf{e}_3$, $\mathrm{Var}(\mathbf{e}_3'\mathbf{x})=\mathrm{Var}(y_3)=\delta_3$, y_3: 세번째 주성분

…

■ 고객만족 자료 사례

obs	x_1	x_2	x_3	x_4	x_5	성별	연령
1	1	2	4	1	1	여자	10대
2	1	2	3	2	1	여자	10대
3	2	5	5	2	2	여자	20대
4	2	5	5	2	2	여자	20대
5	1	2	3	4	3	여자	30대
6	1	3	4	1	1	남자	30대
7	4	5	5	3	3	남자	40대
8	1	3	4	4	4	남자	40대
9	3	3	5	5	4	남자	50대
10	5	5	5	4	4	남자	50대
평균	2.10	3.50	4.30	2.80	2.50		

- 변수:

  x_1(가격), x_2(성능), x_3(편리성), x_4(디자인), x_5(색상)

- 변수값:

  1='매우 만족하지 않는다.', 2='만족하지 않는다.', 3='보통이다.',
  4='만족한다.', 5='매우 만족한다.'

■ 기초통계량과 상관계수행렬

① Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
x1	10	2.10000	1.44914	21.00000	1.00000	5.00000
x2	10	3.50000	1.35401	35.00000	2.00000	5.00000
x3	10	4.30000	0.82327	43.00000	3.00000	5.00000
x4	10	2.80000	1.39841	28.00000	1.00000	5.00000
x5	10	2.50000	1.26930	25.00000	1.00000	4.00000

② Pearson Correlation Coefficients, N = 10

Prob > |r| under H0: Rho=0

	x1	x2	x3	x4	x5
x1	1.00000	0.70784	0.71713	0.44960	0.57386
		0.0220	0.0196	0.1924	0.0828
x2	0.70784	1.00000	0.84725	0.05868	0.29093
	0.0220		0.0020	0.8721	0.4148
x3	0.71713	0.84725	1.00000	0.15442	0.37215
	0.0196	0.0020		0.6702	0.2896
x4	0.44960	0.05868	0.15442	1.00000	0.93897
	0.1924	0.8721	0.6702		¡.0001
x5	0.57386	0.29093	0.37215	0.93897	1.00000
	0.0828	0.4148	0.2896	<.0001	

3. 주성분 분석

■ 고유값과 고유벡터

① Covariance Matrix

	x1	x2	x3	x4	x5
x1	2.100000000	1.388888889	0.855555556	0.911111111	1.055555556
x2	1.388888889	1.833333333	0.944444444	0.111111111	0.500000000
x3	0.855555556	0.944444444	0.677777778	0.177777778	0.388888889
x4	0.911111111	0.111111111	0.177777778	1.955555556	1.666666667
x5	1.055555556	0.500000000	0.388888889	1.666666667	1.611111111

Total Variance 8.1777777778

S

λ_1

$S\,\underline{e}_1=\lambda_1\underline{e}_1$

$\underline{e}_1$

② Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	5.08214264	2.62396015	0.6215	0.6215
2	2.45818249	2.01407206	0.3006	0.9221
3	0.44411043	0.30213695	0.0543	0.9764
4	0.14197348	0.09060473	0.0174	0.9937
5	0.05136875		0.0063	1.0000

③ Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
x1	0.573726	0.255875	-.773457	-.042393	-.073031
x2	0.410198	0.581434	0.509999	-.465832	0.128735
x3	0.260041	0.285983	0.223020	0.877978	0.173233
x4	0.452373	-.601742	0.080055	-.086101	0.647644
x5	0.479910	-.390617	0.292435	0.054270	-.727078

3. 주성분 분석

■ 해석

- 제1 주성분은 일종의 '전반적인 만족도'를 나타낸다고 할 수 있다.

  P1 = (0.574 x1 + 0.410 x2 + 0.260 x3 + 0.452 x4 + 0.480 x5)

  ≈ (0.574 x1 + 0.410 x2 + 0.260 x3 + 0.452 x4 + 0.480 x5) / 2.176

  = 0.264 x1 + 0.188 x2 + 0.119 x3 + 0.208 x4 + 0.221 x5

- 제2 주성분은 가격, 성능, 편리성 등 제품의 '내형적 요인'과 디자인, 색상 등 '외형적 요인'의 차이를 나타낸다고 해석할 수 있다.

  P2 = (0.256 x1 + 0.581 x2 + 0.286 x3) – (0.602 x4 + 0.391 x5)

3. 주성분 분석

■ 개별 관찰치에 대한 주성분 점수 산출

subject	gender	age	x_1	x_2	x_3	x_4	x_5	Prin1	Prin2
1	F	10	1	2	4	1	1	-2.859	0.430
2	F	10	1	2	3	2	1	-2.666	-0.458
3	F	20	2	5	5	2	2	0.138	1.723
4	F	20	2	5	5	2	2	0.138	1.723
5	F	30	1	2	3	4	3	-0.802	-2.443
6	M	30	1	3	4	1	1	-2.448	1.011
7	M	40	4	5	5	3	3	2.218	1.243
8	M	40	1	3	4	4	4	0.349	-1.966
9	M	50	3	3	5	5	4	2.208	-1.770

$$y_{i1} = 0.574c_{i1} + 0.410c_{i2} + 0.260c_{i3} + 0.452c_{i4} + 0.480c_{i5}$$
$$y_{i2} = 0.256c_{i1} + 0.581c_{i2} + 0.286c_{i3} - 0.602c_{i4} - 0.391c_{i5}$$

■ 공분산 행렬에 기초한 주성분 분석

① Covariance Matrix

	x1	x2	x3	x4	x5
x1	2.100000000	1.388888889	0.855555556	0.911111111	1.055555556
x2	1.388888889	1.833333333	0.944444444	0.111111111	0.500000000
x3	0.855555556	0.944444444	0.677777778	0.177777778	0.388888889
x4	0.911111111	0.111111111	0.177777778	1.955555556	1.666666667
x5	1.055555556	0.500000000	0.388888889	1.666666667	1.611111111

Total Variance 8.1777777778

② Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	5.08214264	2.62396015	0.6215	0.6215
2	2.45818249	2.01407206	0.3006	0.9221
3	0.44411043	0.30213695	0.0543	0.9764
4	0.14197348	0.09060473	0.0174	0.9937
5	0.05136875		0.0063	1.0000

③ Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
x1	0.573726	0.255875	-.773457	-.042393	-.073031
x2	0.410198	0.581434	0.509999	-.465832	0.128735
x3	0.260041	0.285983	0.223020	0.877978	0.173233
x4	0.452373	-.601742	0.080055	-.086101	0.647644
x5	0.479910	-.390617	0.292435	0.054270	-.727078

$$
\begin{aligned}
\hat{y}_1 &= \hat{e}_{11}c_1 + \hat{e}_{21}c_2 + \cdots + \hat{e}_{p1}c_p \\
&= \hat{e}_{11}(x_1 - \bar{x}_1) + \hat{e}_{21}(x_2 - \bar{x}_2) + \cdots + \hat{e}_{p1}(x_p - \bar{x}_p) \\
&= 0.574(x_1 - 2.1) + 0.410(x_2 - 3.5) + \cdots + 0.480(x_5 - 2.5) \\
&= 0.574x_1 + 0.410x_2 + 0.260x_3 + 0.452x_4 + 0.480x_5 - 6.225
\end{aligned}
$$

- 전체 분산의 합계 : $\mathrm{tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^{p} \sigma_i^2 = \sum_{i=1}^{p} \delta_i$ (= 8.178)
- k 번째 주성분 y_k가 전체 씨스템 변이를 설명하는 부분 : $\delta_k / (\delta_1 + \cdots + \delta_p)$
- 첫 m개의 주성분 $y_1, y_2, \cdots, y_m$에 의해 설명되는 부분 : $(\delta_1 + \cdots + \delta_m)/\mathrm{tr}(\boldsymbol{\Sigma})$

3. 주성분 분석

■ 개별 관찰치에 대한 주성분 점수 산출

① Correlation Matrix

	x1	x2	x3	x4	x5
x1	1.0000	0.7078	0.7171	0.4496	0.5739
x2	0.7078	1.0000	0.8473	0.0587	0.2909
x3	0.7171	0.8473	1.0000	0.1544	0.3722
x4	0.4496	0.0587	0.1544	1.0000	0.9390
x5	0.5739	0.2909	0.3722	0.9390	1.0000

② Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.07431211	1.57142527	0.6149	0.6149
2	1.50288684	1.25573414	0.3006	0.9154
3	0.24715269	0.10099234	0.0494	0.9649
4	0.14616036	0.11667236	0.0292	0.9941
5	0.02948799		0.0059	1.0000

③ Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
x1	0.511879	-.109228	-.839489	-.121834	-.080395
x2	0.437061	-.464649	0.206940	0.725773	0.153326
x3	0.463022	-.392368	0.422350	-.668075	0.083396
x4	0.361259	0.623080	0.070902	0.015384	0.689927
x5	0.449511	0.479572	0.262729	0.108859	-.697906

$$
\begin{aligned}
\hat{y}_1 &= \hat{e}_{11} z_1 + \hat{e}_{21} z_2 + \cdots + \hat{e}_{p1} z_p \\
&= 0.512 \times [(x_1 - 2.1)/1.449] + \cdots + 0.450 \times [(x_5 - 2.5)/1.269] \\
&= 0.353x_1 + 0.323x_2 + 0.562x_3 + 0.258x_4 + 0.354x_5 - 5.899
\end{aligned}
$$

■ 보유 주성분 개수에 대한 판정

- 전체변이에 대한 공헌도

✓ $$C(m) = \begin{cases} 100(\hat{\delta}_1 + \hat{\delta}_2 + \cdots + \hat{\delta}_m)/\mathrm{tr}(\mathbf{S}), & (\mathbf{S}\text{를 사용하는 경우}) \\ 100(\hat{\delta}_1 + \hat{\delta}_2 + \cdots + \hat{\delta}_m)/p, & (\mathbf{R}\text{을 사용하는 경우}) \end{cases}$$

✓ $C(m) > c^*$를 만족하는 최소 정수값 m

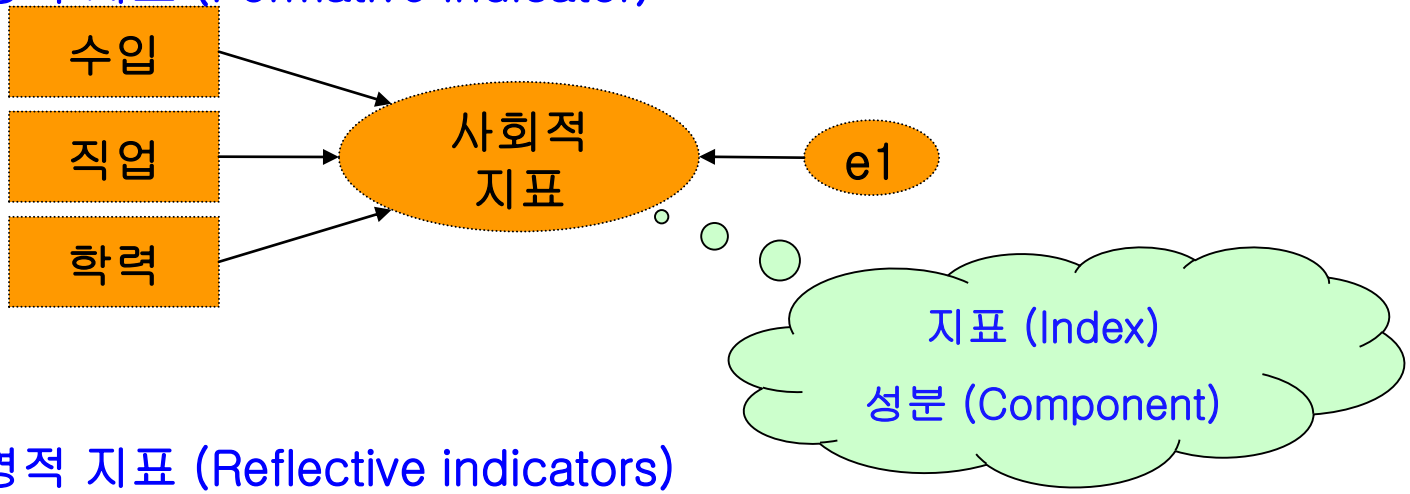
- 고유값의 크기 (Kaiser의 규칙)

✓ 주성분이 상관행렬에 기초하고 있다면 상관행렬의 대각원소가 1이므로 모든 주성분의 분산은 1이 된다. 따라서 1보다 작은 고유값을 가지는 주성분은 원래 반응변수 중의 어느 하나 보다 작은 정보를 가지므로 보유할 가치가 없다고 하겠다.

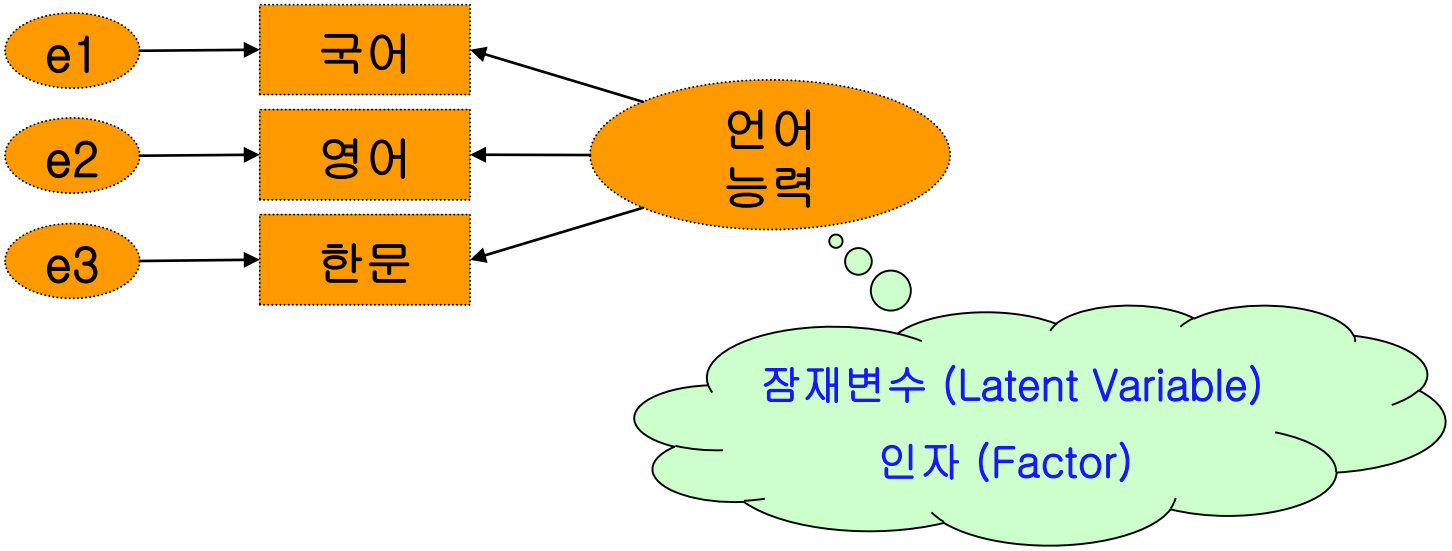
✓ 상관행렬에 기초하여 분석을 수행하는 경우, 고유값이 1이상인 주성분을 보유하는 'Kaiser(1960)의 규칙'을 기준으로 사용할 수 있다.

■ 인자분석의 개념

- 형성적 지표 (Formative indicator)



- 반영적 지표 (Reflective indicators)



■ 인자분석의 개념

(1) 서로 상관관계를 맺고 있어서,
(2) 직접적으로 해석하기 어려운,
(3) 여러 관찰(측정)변수들 간의 구조적 연관관계를,

(a) 상대적으로 독립적이면서,
(b) 변수들의 저변구조를 이해하기 위해 개념상 의미를 부여할 수 있는,
(c) 원래 변수들의 개수보다 훨씬 적은 개수의 공통인자들을 상정하여,
이들을 통해 분석하고자 하는 통계적 방법.

- 공통인자 (Common Factor)
  - ✓ 변수들이 그들의 구조적 측면에서 서로 공유하고 있는 확률적 인자로서, 변수들 간의 상관관계를 생성시키는 가설적인, 관찰할 수 없는, 혹은 저변에 깔려 있는 인자를 의미.

- 인자분석은 상관(혹은 공분산)행렬의 구조에 관한 통계적 모형을 구축하고, 그와 같은 구조를 생성시키는 소수 몇 개의 인자를 유도하여, 변수들 간의 구조적 관계를 해석하는 공분산 내지 상관중심의 기법.

■ 직교 인자 모형

$$\begin{cases} x_1 - \mu_1 = \lambda_{11}F_1 + \lambda_{12}F_2 + \cdots + \lambda_{1m}F_m + \varepsilon_1 \\ x_2 - \mu_2 = \lambda_{21}F_1 + \lambda_{22}F_2 + \cdots + \lambda_{2m}F_m + \varepsilon_2 \\ \cdots \\ x_p - \mu_p = \lambda_{p1}F_1 + \lambda_{p2}F_2 + \cdots + \lambda_{pm}F_m + \varepsilon_p \end{cases}$$

(1) $E(F_i)=0,\ Var(F_i)=1,\ Cov(F_i,F_j)=0$ for $i \neq j$

(2) $E(\varepsilon_i)=0,\ Var(\varepsilon_i)=\varphi_i,\ Cov(\varepsilon_i,\varepsilon_j)=0$ for $i \neq j$

(3) $Cov(F_i,\varepsilon_j)=0$ for all i, j

$F_1, F_2, \cdots, F_m$: 공통인자 (common factor)

$\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p$: 특수인자 (specific factor)

- $Var(x_i) = \{\lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{ip}^2\} + \varphi_i$ = 공통성 + 특수성
- $\lambda_{ik} = Cov(x_i, F_k)$, i.e. $\lambda_{ik} = Corr(z_i, F_k)$

■ 인자적재와 특수분산의 추정

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pm} \end{pmatrix} \qquad \Psi = \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{pmatrix}$$

인자패턴(적재)행렬 　　특수분산행렬

- 주성분인자법 : 주성분분석 이용

$$F_1 = P_1 / \sqrt{l_1}, \quad F_2 = P_2 / \sqrt{l_2}, \quad \ldots$$

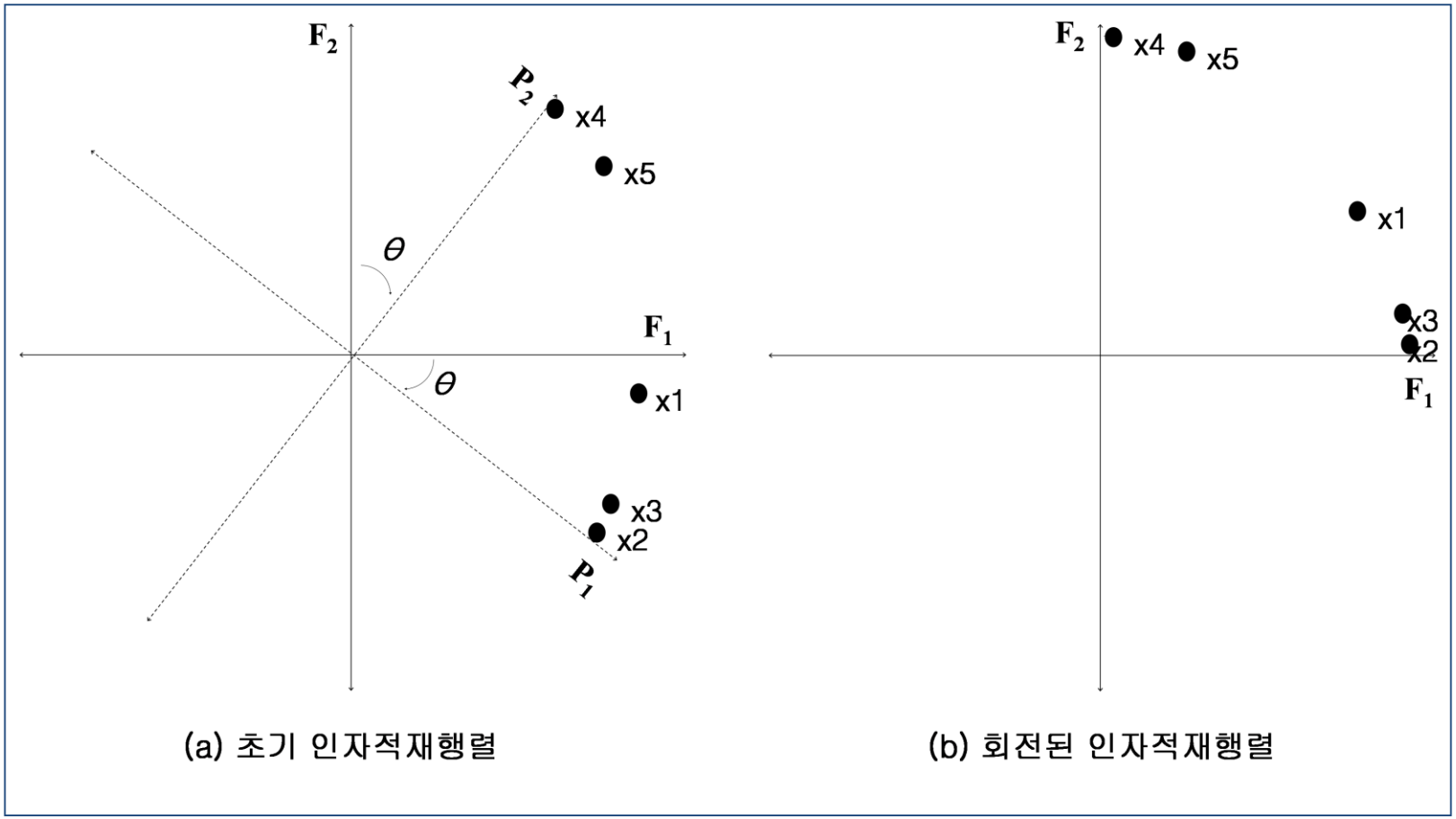
- 주축인자법 : 특수분산의 초기값을 정하고 반복적으로 계산

$$\mathrm{S} = \mathrm{LL}' + \mathrm{Y}$$

- 최대우도법 : 다변량정규분포의 가정 하에 우도함수를 최대화

■ 인자회전

- 인자의 해석이 보다 더 용이하도록 인자회전을 시행



■ 인자의 개수

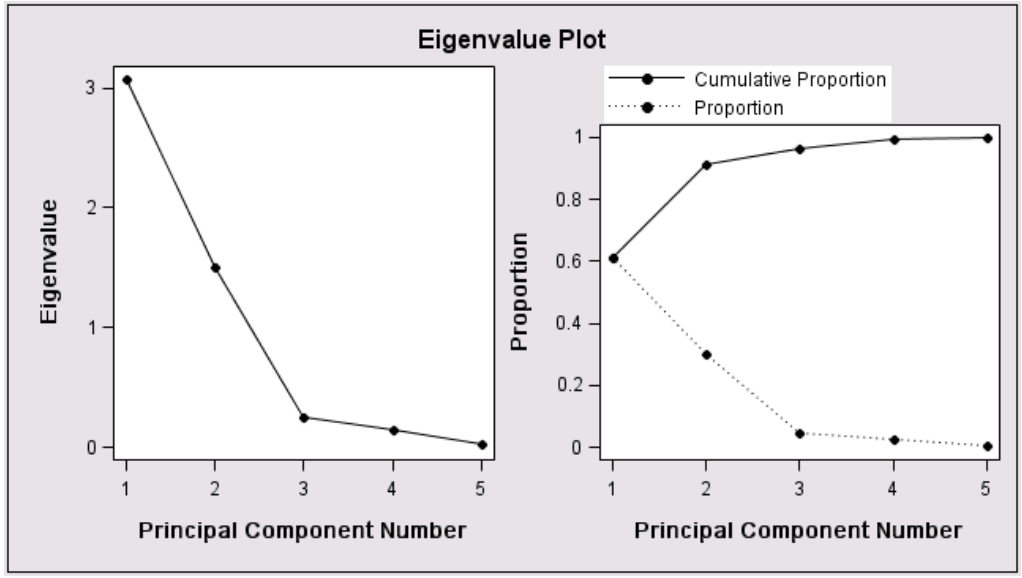
- 고유값의 크기

  ✓ 표본상관행렬 R에 기초하여 분석을 수행하는 경우, 표본상관행렬의 고유값 중 1보다 큰 개수만큼 인자를 보유. (Kaiser의 규칙)

  ✓ 일반적으로 연구자가 기대하는 인자의 개수와 일치하는 경우가 많음.

  ✓ 보유되어야 할 것보다 적은 개수의 인자를 결정하게 하는 경향이 있음.

- Scree plot



■ 인자의 공헌도

- 관찰변수의 공통성: $c_i = \sum_{j=1}^{m} \lambda_{ij}^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2$
  - ✓ m 개의 인자들에 의해 설명되는 변수 x_i의 분산
- 인자의 공헌도 (설명분산) : $v_k = \sum_{i=1}^{p} \lambda_{ij}^2 = \lambda_{1k}^2 + \lambda_{2k}^2 + \cdots + \lambda_{pk}^2$
  - ✓ 원래변수의 전체분산 중 k 번째 인자에 의해서 설명되는 분산의 양

	f_1	f_2	공통성 (c_i)
x_1	0.79130	0.44420	0.823
x_2	0.95465	0.01949	0.912
x_3	0.93632	0.11741	0.890
x_4	0.03244	0.99178	0.985
x_5	0.26254	0.94759	0.967
설명분산 (v_k)	2.484	2.093	4.577
설명비율 (%)	49.7	41.9	91.6

■ 인자의 점수화 (인자점수)

- 각 관측치에 대하여 인자점수를 추정(참고: indeterminacy)
- 원래 변수보다 적은 개수의 인자를 이용하여 회귀분석 등의 입력변수로 활용 (인자점수 회귀)

obs	z_1	z_2	z_3	z_4	z_5	f_1°	f_2°
1	-0.76	-1.11	-0.36	-1.29	-1.18	-0.60	-1.12
2	-0.76	-1.11	-1.58	-0.57	-1.18	-1.20	-0.63
3	-0.07	1.11	0.85	-0.57	-0.39	0.90	-0.73
4	-0.07	1.11	0.85	-0.57	-0.39	0.90	-0.73
5	-0.76	-1.11	-1.58	0.86	0.39	-1.47	0.86
6	-0.76	-0.37	-0.36	-1.29	-1.18	-0.29	-1.23
7	1.31	1.11	0.85	0.14	0.39	1.16	0.16
8	-0.76	-0.37	-0.36	0.86	1.18	-0.69	1.01
9	0.62	-0.37	0.85	1.57	1.18	0.09	1.43
10	2.00	1.11	0.85	0.86	1.18	1.22	0.98

4. 인자 분석

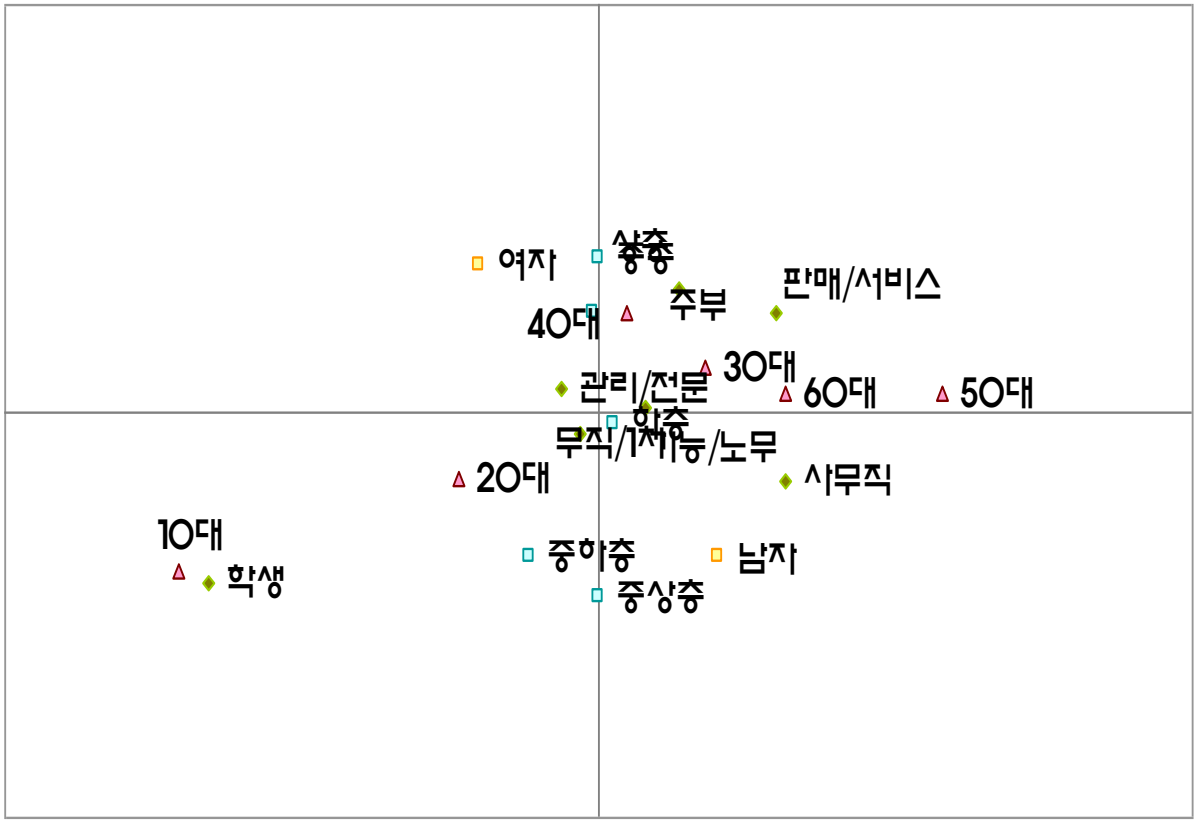
■ 사례

	Factor 1	Factor 2	Factor 3
유아들이나 어린이들이 등장하는 광고	0.793	0.014	0.010
일반 소비자가 등장하는 광고	0.712	0.143	0.054
전통이나 인간적 유대감을 강조하는 광고	0.701	0.148	-0.107
동물이 등장하는 광고	0.646	-0.008	0.219
유머가 있어서 재미있는 광고	0.531	-0.017	0.194
제품의 이미지를 강조하는 광고	0.061	0.885	0.122
제품의 내용을 강조하는 광고	0.111	0.860	0.04
경쟁제품과 자기제품을 비교하는 광고	0.046	0.734	0.283
외국의 탤런트/스포츠 스타가 나오는 광고	0.109	0.102	0.840
유명 연예인이 나오는 광고	0.196	0.060	0.750
성적인 느낌을 강조하는 광고	-0.095	0.356	0.624
사회 저명인사가 나오는 광고	0.374	0.362	0.399

- Factor 1 : 모델 및 구성의 친숙성
- Factor 2 : 메인 포커스
- Factor 3 : 모델의 주목성

4. 인자 분석

■ 사례



- 여자는 남자에 비해 '모델 및 구성의 친숙성'이 강한 광고를 선호함.
- 남자는 여자에 비해 '메인 포커스'와 '모델의 주목성'이 강한 광고를 산호함.
- 10대, 20대, 학생은 '모델의 주목성'이 강한 광고를 선호함.
- 50대, 60대, 주부는 제품 중심의 '메인 포커스'가 강한 실용적인 광고를 선호함.