

Latent Class Analysis(LCA) Methods

Sa-Hyun Kim

Department of Social Welfare, Daegu University

Latent Class Analysis?

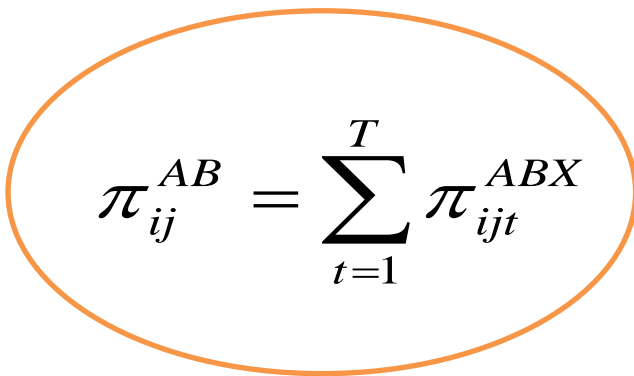
- 일종의 요인분석
 - 관찰(측정)변수들간 관계를 유발하는 잠재적 공통요인을 규명.
- 다만, LCA는 관찰변수와 잠재변수 모두 범주형.
- 계층분류방법
 - 요인분석: 요인별 부하량에 기초해 각 사례별 요인점수산출.
 - LCA : 각 사례별 잠재변수의 범주(잠재계층)에 속할 확률 도출.
확률의 상대적 크기에 기초해 계층분류.
잠재범주의 계층 수는 모형검증을 통해 도출.
- 잠재계층의 의미
 - 파악된 계층은 연구자의 판단에 기초해 해석.

LCA 기초 이론

- Fundamental equation

$$\pi_{ijt}^{ABX} = \pi_t^X \pi_{it}^{\bar{A}X} \pi_{jt}^{\bar{B}X}$$

for $i = 1, \dots, I; j = 1, \dots, J; t = 1, \dots, T$


$$\pi_{ij}^{AB} = \sum_{t=1}^T \pi_{ijt}^{ABX}$$

- π_{ijt}^{ABX} : 관찰값이 관찰변수 A의 i범주, 관찰변수 B의 j범주, & 잠재변수 X의 t범주에 속할 결합확률.
- π_t^X : 관찰값이 잠재변수 X의 t범주에 속할 확률.
- $\pi_{it}^{\bar{A}X}$: 잠재변수 X의 t범주에 속한 사례들이 관찰변수 A에서 i인 조건확률.
- $\pi_{jt}^{\bar{B}X}$: 잠재변수 X의 t범주에 속한 사례들이 관찰변수 B에서 j인 조건확률.

LCA 기초 이론

- Local independence

- A와 B는 잠재변수 X의 각 범주에서 각각 조건부 독립. 즉,

$$\pi_{ijt}^{\overline{A}\overline{B}X} = \pi_{ijt}^{ABX} / \pi_t^X = \pi_{it}^{\overline{A}X} \pi_{jt}^{\overline{B}X}$$

→ $\pi_{ijt}^{\overline{A}\overline{B}X}$: 잠재변수 X의 t범주에 속한 사례들이, 관찰변수 A의 i범주 & B의 j범주에 속할 조건부확률.

- 관찰변수 A, B간 상관성이 잠재변수 X에 의해 발생하는 것.
- LCA는 두 범주형 변수의 지역독립성을 성립하는 잠재계층을 찾는 것.

※ 지역독립성 가정이 충족하지 않으면 LCA model은 성립 안됨.

LCA 기초 이론

- **Parameter Estimates**

- 모형모수: 잠재계층확률($\pi_i^X : T$ 개), 조건확률($\pi_{it}^{\bar{A}X} : I \times T$ 개, $\pi_{jt}^{\bar{B}X} : J \times T$ 개)
→ 결국, 추정해야 할 모수는 $T + (T \times I) + (T \times J)$ 개
- 과파악(over-identification) 모형 필요. 즉 자유도가 있어야....
- 최대우도추정
→ π_{ij}^{AB} 의 추정치는 ① 관찰비율(p_{ij}), ② 모형의 예측 확률($\hat{\pi}_{ij}^{AB}$)

$\text{Min} | p_{ij} - \hat{\pi}_{ij}^{AB} |$ 을 충족하는 모형모수들을 추정.

- EM-알고리즘 활용하여 Iterative proportional fitting.
변화의 크기가 무의미할 때까지..
- 우도함수(ex. 2개 문항): n_{ij} 가 다항분포한다는 가정.

$$\frac{N!}{n_{11}!n_{12}!n_{21}!n_{22}} (\pi_{11}^{AB})^{n_{11}} (\pi_{12}^{AB})^{n_{12}} (\pi_{21}^{AB})^{n_{21}} (\pi_{22}^{AB})^{n_{22}}$$

※ 모수는 우도함수의 편미분값=0. (다만, 지역최대 vs. 전체최대: 초기값이 중요)

LCA 기초 이론

- **Model testing**

- 모형검증: 모형과 자료의 합치도 검증(p_{ij} 와 $\hat{\pi}_{ij}^{AB}$ 의 합치도). χ^2 검증.

- ① Pearson Chi-square

$$N \sum_{ij} \frac{(p_{ij} - \hat{\pi}_{ij}^{AB})^2}{\hat{\pi}_{ij}^{AB}}$$

→ No Significant !

- ② Likelihood ratio Chi-square

$$2N \sum_{ij} (p_{ij}) \ln \left(\frac{p_{ij}}{\hat{\pi}_{ij}^{AB}} \right)$$

- 검증에 관련된 두 가지 특성

- ① p_{ij} 와 $\hat{\pi}_{ij}^{AB}$ 의 차이가 클수록 χ^2 값이 커짐: 모형의 자료설명력 ↓.
 - ② N이 클수록 χ^2 값 커짐: 모형설명력과 무관.

LCA 기초 이론

- **Model selection**

- Akaike's Information Criterion(AIC)
 - 정보량과 복잡성을 동시에 반영. 모형의 간명성 강조.
 - 0에 가까울수록 좋은 모형. \therefore 절대값이 작은 것이 간명모형.
 - 단, 우도비 X^2 검증에서 유의하지 않은 것만을 비교.
- Bayesian Information Criterion(BIC)
 - 복잡성의 가중치를 조정한 기준. AIC와 동일 기준.
- Lo-Mendell-Rubin Adjusted Likelihood Ratio Test(LRT)
 - K개 계층 우도비와 K-1개 계층의 우도비를 비교
 - $2 \times (\text{우도비차이})$ 값을 검증. 유의하면, K개 계층 유의.
- 기타: Entropy 지수(1에 근접)

LCA 기초 이론

- **Posterior Probability & Case Assignment**

- 사후확률: 추정된 모수를 활용, 사례들이 각 잠재계층에 속할 확률 도출 (Bayes's theorem).

$$\hat{\pi}_{tij}^{\bar{X}AB} = \frac{\hat{\pi}_{ijt}^{\bar{A}\bar{B}X} \hat{\pi}_t^X}{\hat{\pi}_{ijt}^{\bar{A}\bar{B}X} \hat{\pi}_t^X + \hat{\pi}_{ijt'}^{\bar{A}\bar{B}X} \hat{\pi}_{t'}^X} \quad (t': \text{잠재계층})$$

→ 관찰변수 A의 i & B의 j에 반응한 사례가 잠재계층 t에 속할 확률.

- 사례할당: 사후확률이 가장 높은 잠재계층에 할당.
- 오류할당: 설정된 잠재계층 모형이 사례들을 잘 못 할당하는 것.

LCA 기초 이론

- **Posterior Probability & Case Assignment**

- 사후확률: 추정된 모수를 활용, 사례들이 각 잠재계층에 속할 확률 도출 (Bayes's theorem).

$$\hat{\pi}_{tij}^{\bar{X}AB} = \frac{\hat{\pi}_{ijt}^{\bar{A}\bar{B}X} \hat{\pi}_t^X}{\hat{\pi}_{ijt}^{\bar{A}\bar{B}X} \hat{\pi}_t^X + \hat{\pi}_{ijt'}^{\bar{A}\bar{B}X} \hat{\pi}_{t'}^X} \quad (t': \text{잠재계층})$$

→ 관찰변수 A의 i & B의 j에 반응한 사례가 잠재계층 t에 속할 확률.

- 사례할당: 사후확률이 가장 높은 잠재계층에 할당.
- 오류할당: 설정된 잠재계층 모형이 사례들을 잘 못 할당하는 것.

M-plus 실행

- **준비사항**

- ① data file

- text or ASCII file. 변수명 삭제.
 - Excel의 CSV(comma separated values) file 유용.

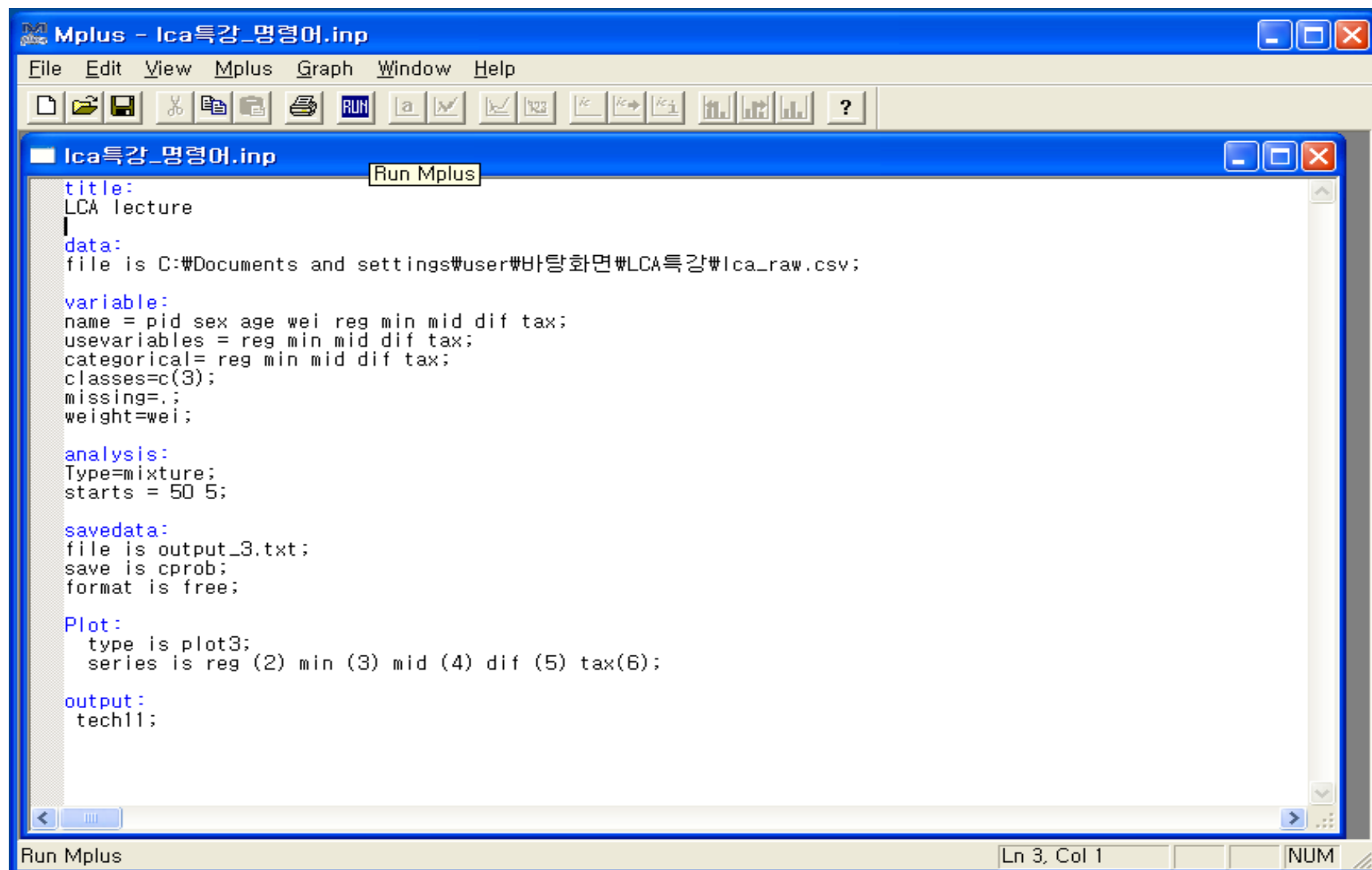
- ② syntax file

- title, data, variable(name, usevariables, classes, categorical...) analysis, model, output, plot & savedata로 구성.

- **분석종류**

- 일요인 모형: 한 변수의 계층 수를 고려하는 분석
 - covariate 모형: 잠재변수와 다른 변수와의 관계성 분석.

M-plus 실행: 일요인 모형

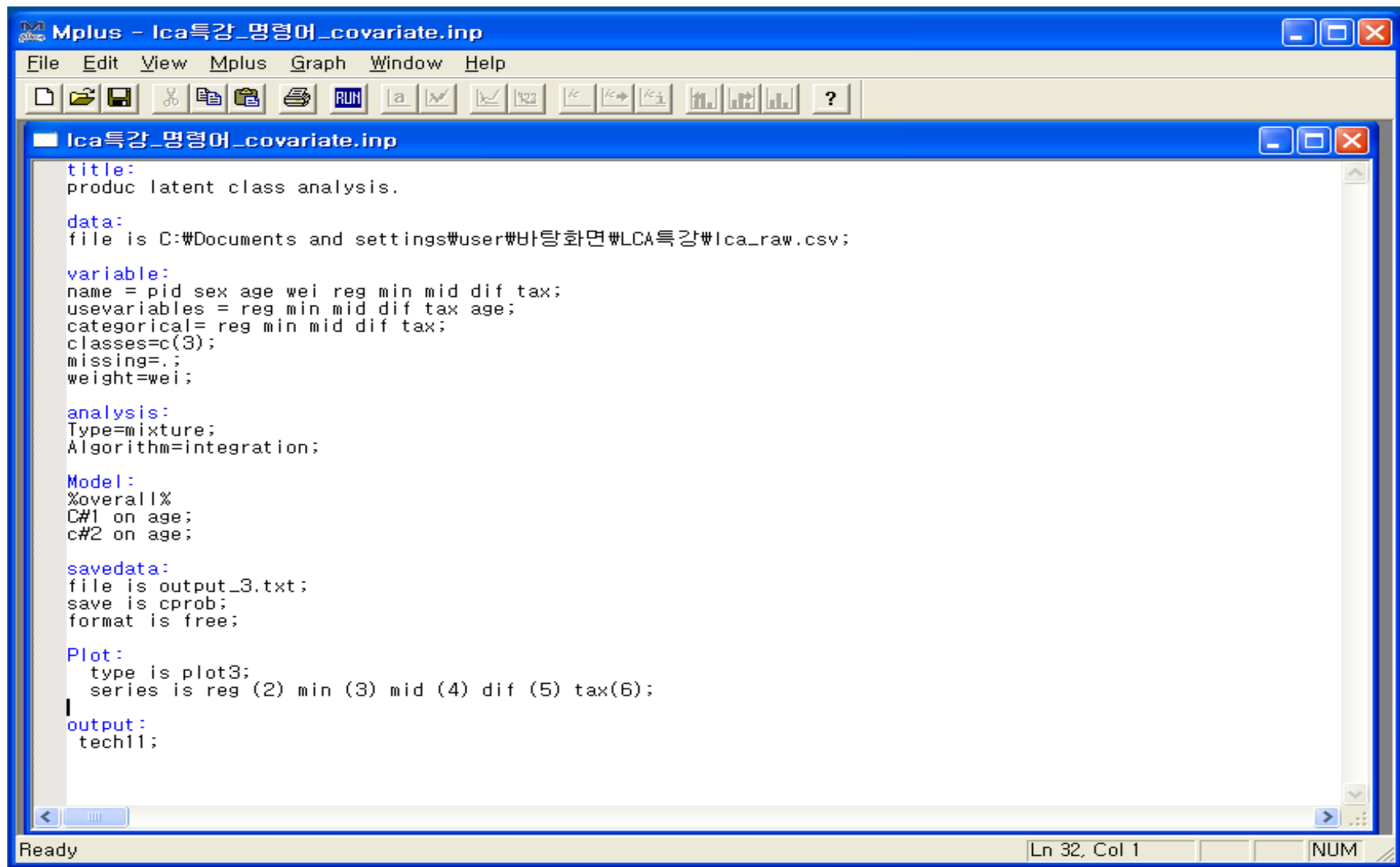


The screenshot displays the Mplus software window titled "Mplus - lca특강_명령어.inp". The window includes a menu bar (File, Edit, View, Mplus, Graph, Window, Help) and a toolbar with icons for file operations and analysis. The main text area contains the following command file content:

```
title:
LCA lecture
data:
file is C:\Documents and settings\user\바탕화면\LCA특강\lca_raw.csv;
variable:
name = pid sex age wei reg min mid dif tax;
usevariables = reg min mid dif tax;
categorical= reg min mid dif tax;
classes=c(3);
missing=.;
weight=wei;
analysis:
Type=mixture;
starts = 50 5;
savedata:
file is output_3.txt;
save is cprob;
format is free;
Plot:
type is plot3;
series is reg (2) min (3) mid (4) dif (5) tax(6);
output:
tech11;
```

The status bar at the bottom indicates "Run Mplus" and "Ln 3, Col 1".

M-plus 실행: covariate 모형



The screenshot displays the Mplus software window titled "Mplus - Ica특강_명령어_covariate.inp". The window includes a menu bar (File, Edit, View, Mplus, Graph, Window, Help) and a toolbar with icons for file operations and analysis. The main text area contains the following model specification code:

```
title:
produc latent class analysis.

data:
file is C:\Documents and settings\user\바탕화면\LCA특강\Ica_raw.csv;

variable:
name = pid sex age wei reg min mid dif tax;
usevariables = reg min mid dif tax age;
categorical= reg min mid dif tax;
classes=c(3);
missing=.;
weight=wei;

analysis:
Type=mixture;
Algorithm=integration;

Model:
%overall%
C#1 on age;
c#2 on age;

savedata:
file is output_3.txt;
save is cprob;
format is free;

Plot:
type is plot3;
series is reg (2) min (3) mid (4) dif (5) tax(6);

output:
tech11;
```

The status bar at the bottom indicates "Ready" and "Ln 32, Col 1".

M-plus Output 1

- **Estimate result**

- ① Model fit

- Pearson Chi-square, Likelihood ratio Chi-square의 검증결과
 - AIC, BIC, Lo-Mendell-Rubin LRT, 이론적 해석 가능성 등을 고려하여 잠재계층의 수 결정.
 - 반복적 작업이 불가피.

- ② Conditional probability

- 각 측정항목의 잠재계층에 대한 조건부 확률 확인
 - 조건부 확률값이 1.0이면 계층분류가 안된다는 의미.
 - 조건부 확률에 기초해 계층의 특성파악. 계층명 부여.

- ③ 기타

- 계층별 분포 확인. 그래프 등. ※ 자동저장 안됨. 직접 저장!

M-plus Output 2

- **Output data**

- 명령문 작성시 저장명령 해야...
- text file로 제공.
 - raw data, 계층별 조건부확률, 할당계층에 대한 정보 포함.
- 자료들을 활용해 다양한 분석 가능. ex) 기술통계, 다항로지트 등.

- **Graph**

- 각 잠재계층에 대한 조건부확률을 도표로 제시가능.
 - graph/view graph/estimates probability.
 - 필요에 따라 편집가능.