Introduction to Latent Transition Analysis

Calum MacGillivray – University of Dundee (PhD Student)

Housekeeping

- We will take a short break every hour and an hour for lunch from 12:00-13:00
- In the morning we will be talking about the theory behind LTA and getting started with R
- In the afternoon we will get more hands on with the LTA code
- Questions: we will stop regularly for questions but if you have any as we go then feel free to jump in, raise your hand or put them in the chat.
- Afterwards: feel free to email me with any questions: c.y.macgillivray@dundee.ac.uk
- Resources for this workshop, including the notebook with the annotated code, data files and this presentation can be found on my website here: https://calummacgillivray.github.io/Training/

Introductions

Who am I?

Name: Calum MacGillivray

University: University of Dundee, School of Humanities, Social Science and Law

Discipline: Education (with a developmental psychology background)

PhD Project: Primary-secondary transitions experiences and associated educational outcomes

Now who are you?

What's your name? | Your discipline/background? | What are you working on?

Learning Objectives

1. To understand the concepts and assumptions behind LCA

2. To understand the data structure required to conduct LCA

3. To be confident in reading and interpreting LCA papers

4. To learn how to conduct, visualise and report LCA using R

A Quick Note on R notebooks

- We will be working from an R notebook
- R notebooks are great for sharing code as you can write in legible text around the chunks of code. This makes it easier to follow.
    - When aesthetic legibility is less of a concern I tend to work in an R script in Rstudio. The code works the same regardless.
- You can find the notebook we will be working with here:
- I have also provided a knitted workbook, meaning that the code has been ran and the outputs presented in a permanent html format:

Latent Transition Analysis

Useful resources

- Lund and Ritz (2025) – the flexible, modular, design we will be using in R
- Nylund-Gibson et al., (2023) – a great resource to learn more about LTA
- Sinha et al., (2022) – an excellent deep dive into LCA

- Some manuals for relevant R packages
  - Tidyverse – for data handling and much more
  - poLCA – for running LCA
  - mice – for multiple imputations
- And a great book for getting started with R
  - https://r4ds.hadley.nz/

What is latent transition analysis?

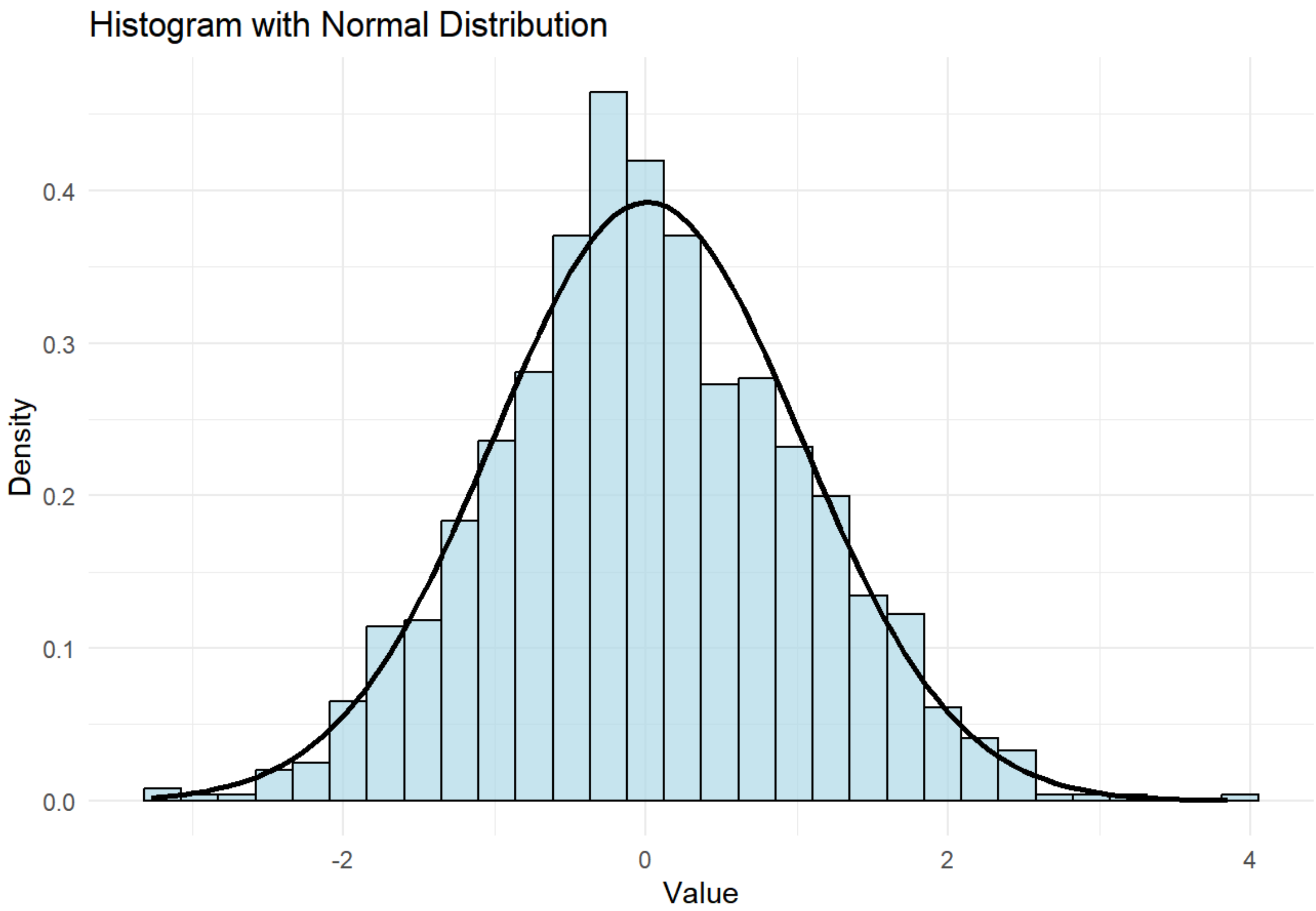
- Latent transition analysis (LTA)
  - A longitudinal extension of latent class analysis (LCA)
- Concerned with latent classes
  - Not directly observed
  - Based on indicators
- Mixture model approach
- Examines changes in latent class membership over time
  - How many latent classes and their characteristics
  - Probability of membership in a class given membership at previous time point
    - The same but for different covariates

What is latent class analysis?

- Latent class analysis
  - Based on categorical indicators
    - LTA with continuous indicators utilises latent profile analysis instead
  - Probability of an individual belonging to a latent class
  - Usually repeated for different numbers of latent classes
    - Then compared with fit statistics

- Finite mixture modelling:
  - Determining unobserved groups from the structure of the indicators
    - Does not assume one normal distribution, but rather multiple distributions

What we may be used to:



Finite Mixture Models:

How does latent class analysis work?

- Latent class analysis works by fitting the pre-specified number of classes to the data.
  - It does this based on probabilities of being in each class based on patterns of answers.
  - This involves multiple random starts to fit the data with a maximum likelihood algorithm.
- Once the classes are created individual's get a probability of being in each class based on their indicators, this must sum to one: e.g. prob class 1 = 0.8, prob class 2 = 0.2.
- Each indicator gets item response probabilities per class:

Indicator 1 (yes/no)	Pr (Yes)	Pr (No)
Class 1	0.8767	0.1233
Class 2	0.5224	0.4776
Class 3	0.0961	0.9039

Common Fit Statistics

- Akaike Information Criterion (AIC): $2k - 2\ln(L)$
  - k = no. parameters $(C - 1) + C \cdot I$
    - C = Classes; I = Indicators
    - However, more parameters when not binary
  - L = likelihood of model – maximum likelihood
  - ln = likelihood is logarithmic
  - (2*parameters) – (2*log likelihood)

  - Comparative: lower = better
  - Lower with goodness of fit, higher with increasing parameters
- Corrected AIC (AICc): $\frac{2k(k+1)}{n-k-1}$
  - Penalises for complexity with small sample sizes

Common Fit Statistics

- Bayesian Information Criterion (BIC): $k \ln(n) - 2\ln(L)$
  - n = sample size
  - (parameters*log sample size) - (2*log likelihood)

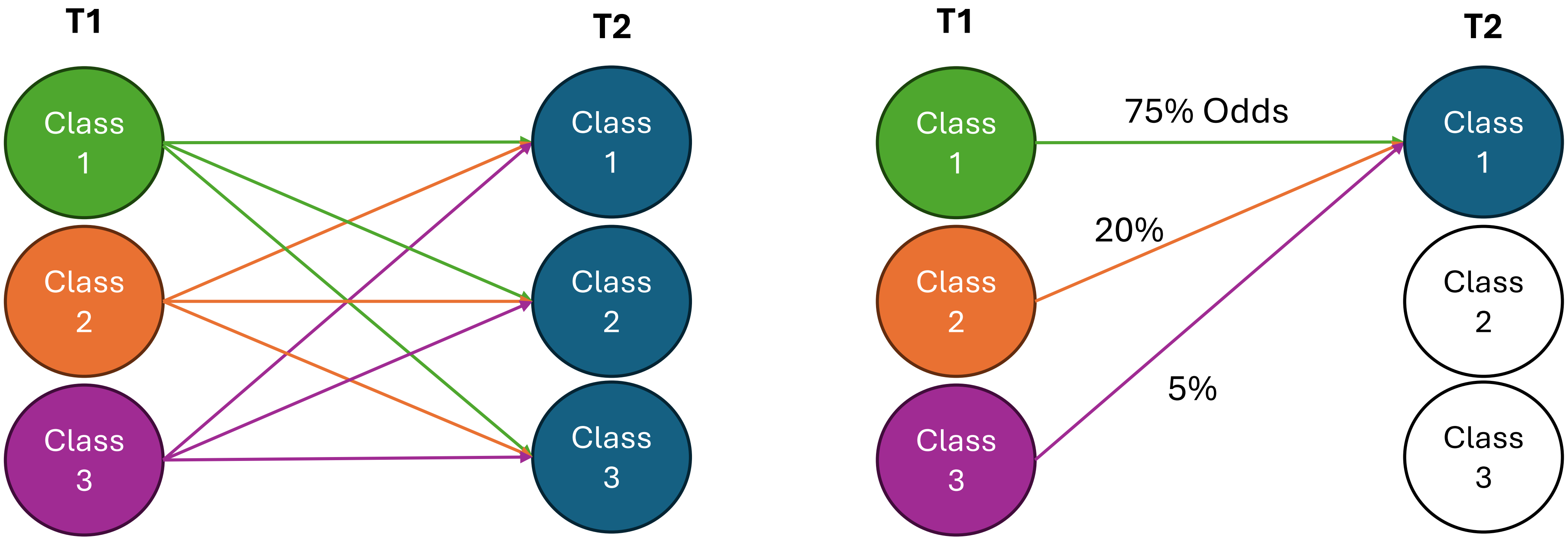
  - Also comparative
  - Lower with goodness of fit, higher with increasing parameters multiplied by log sample size
  - Penalises complexity more – especially as sample size increases

- Adjusted BIC (ABIC): $k \ln\left(\frac{n+2}{24}\right) - 2\ln(L)$
  - Similar but reduces the penalty when sample size is lower
  - Exact formula can vary

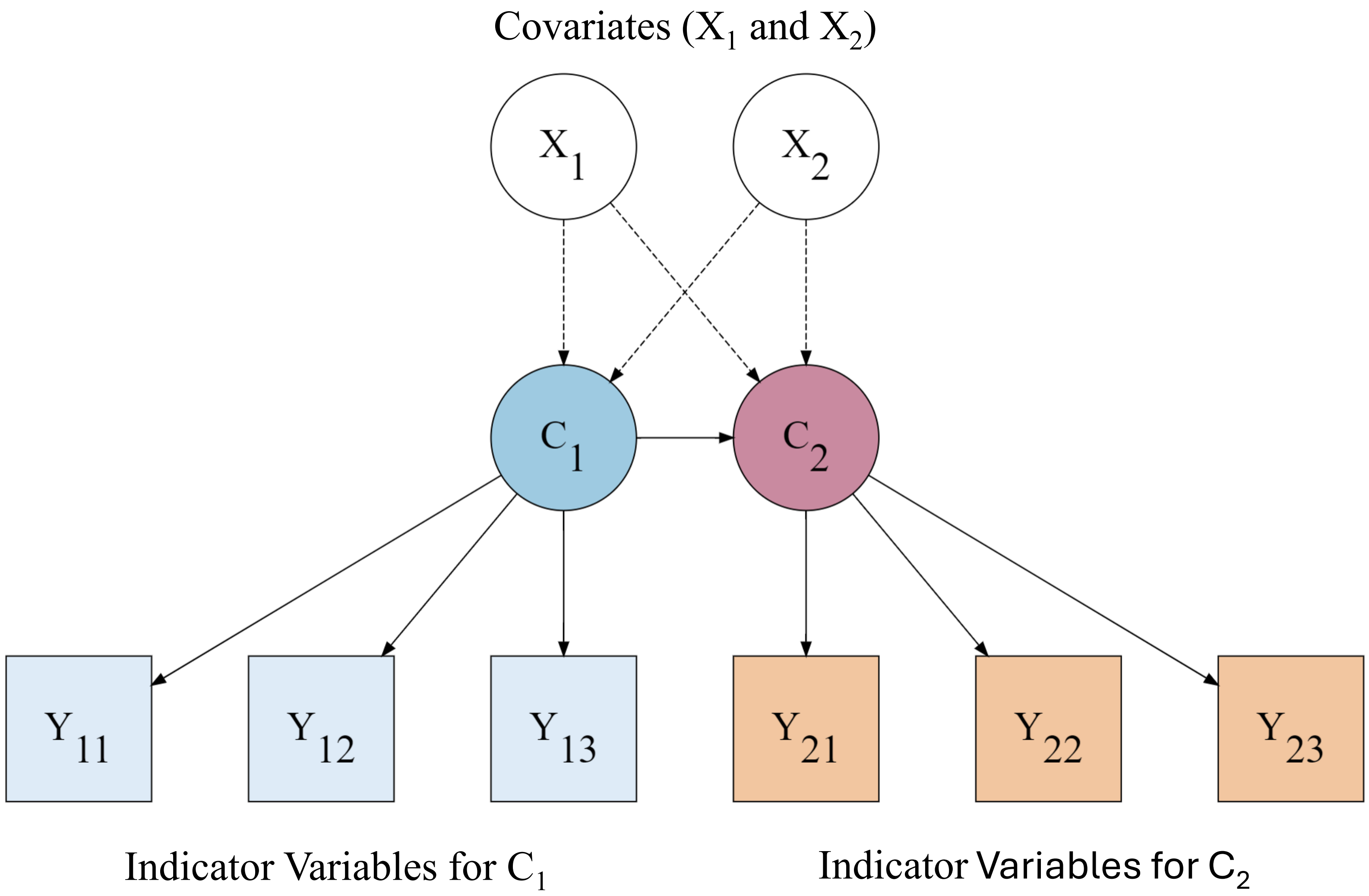
Logistic regression

- We will use logistic regression in this configuration of LTA
  - Based on categorical data – outcome and predictors
- Probability of belonging to a class given class at previous stage
  - We run a series looking at each transition of interest
  - For example – probability of being in class 3 given membership at previous time point
  - We transform to percentage odds and put them all together
- Later we add an additional predictor for each covariate to look at the effect of being in a subgroup compared to a reference subgroup

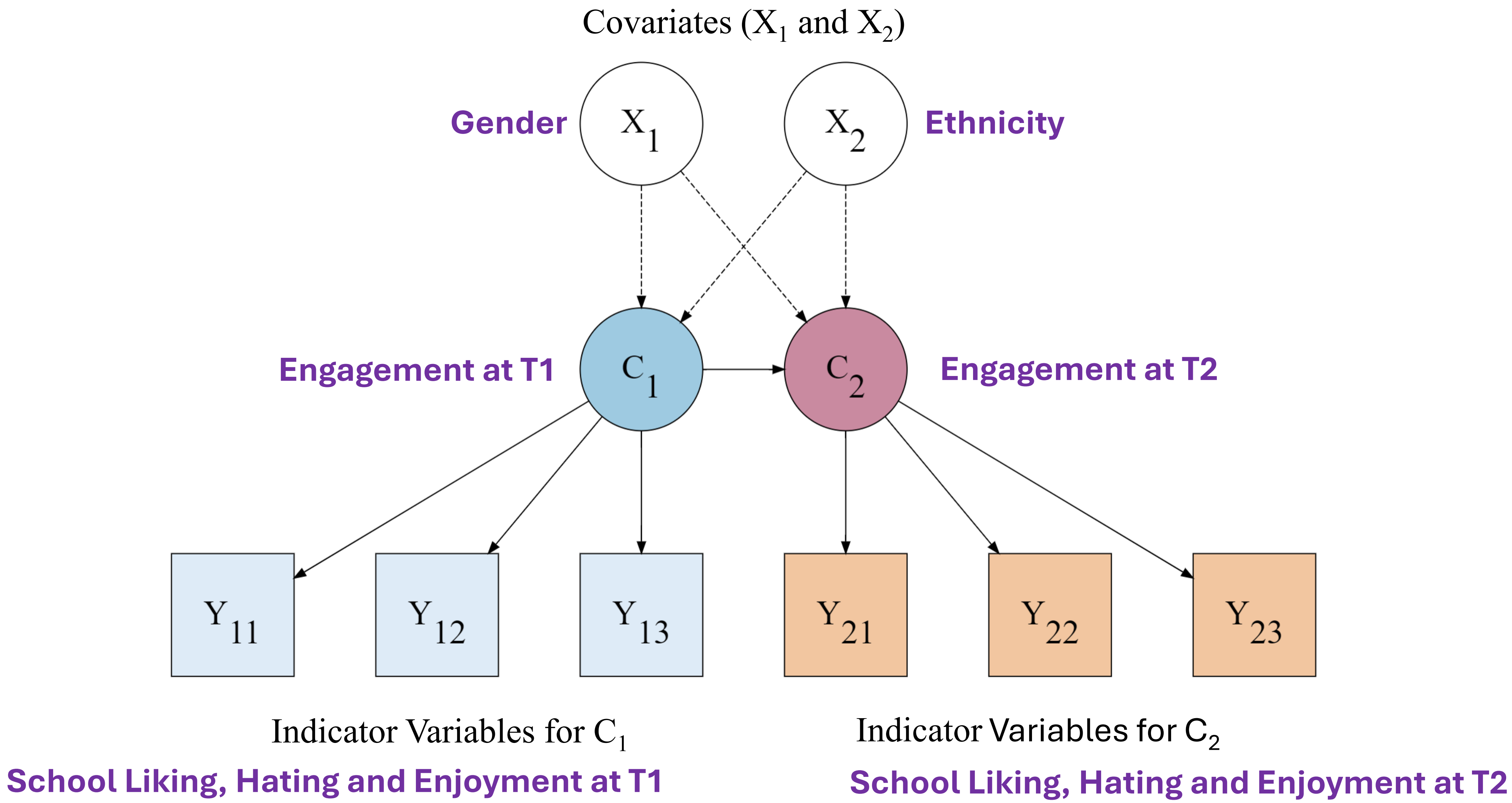
Logistic regression



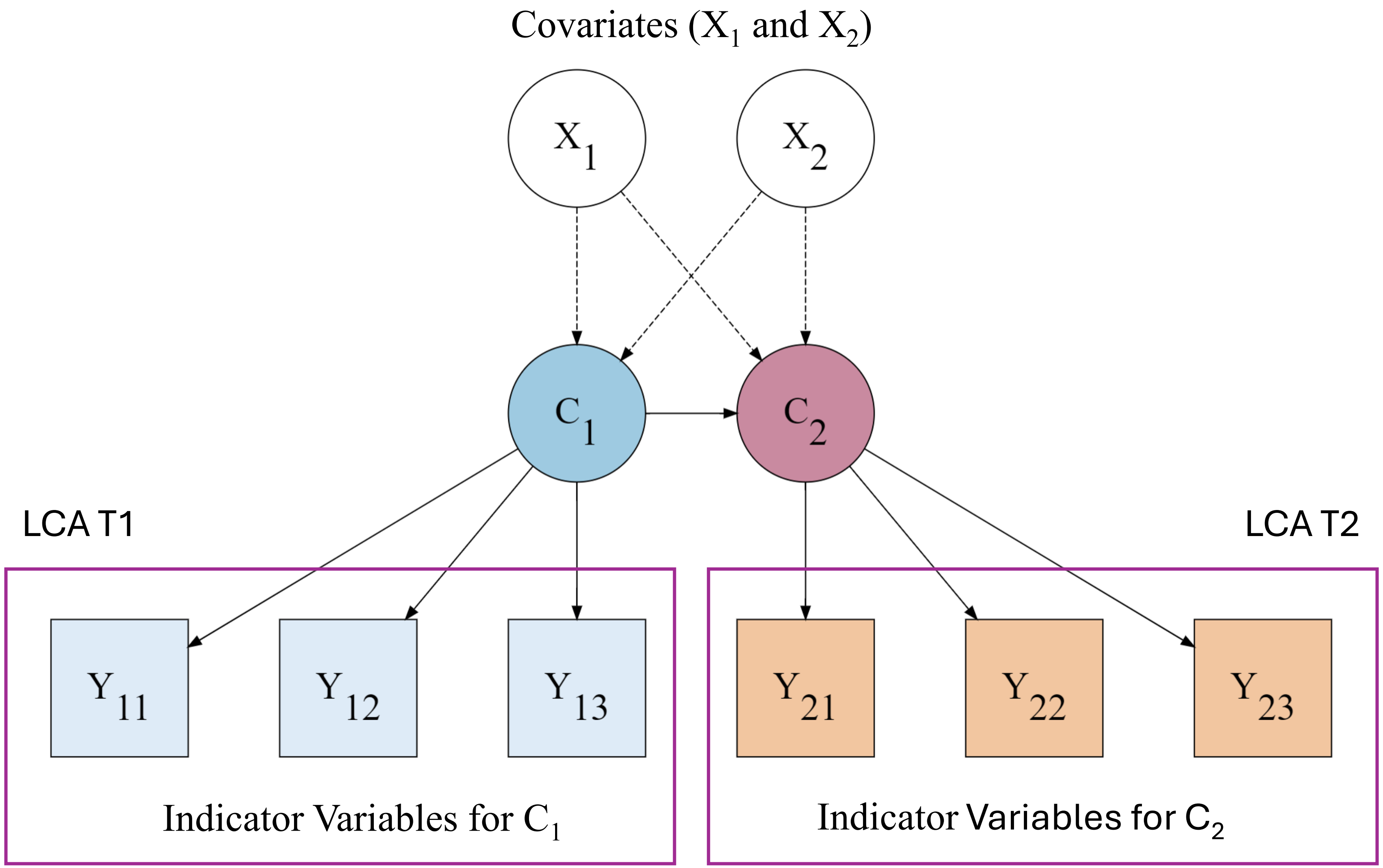
LTA example for 2 time points, with 2 covariates



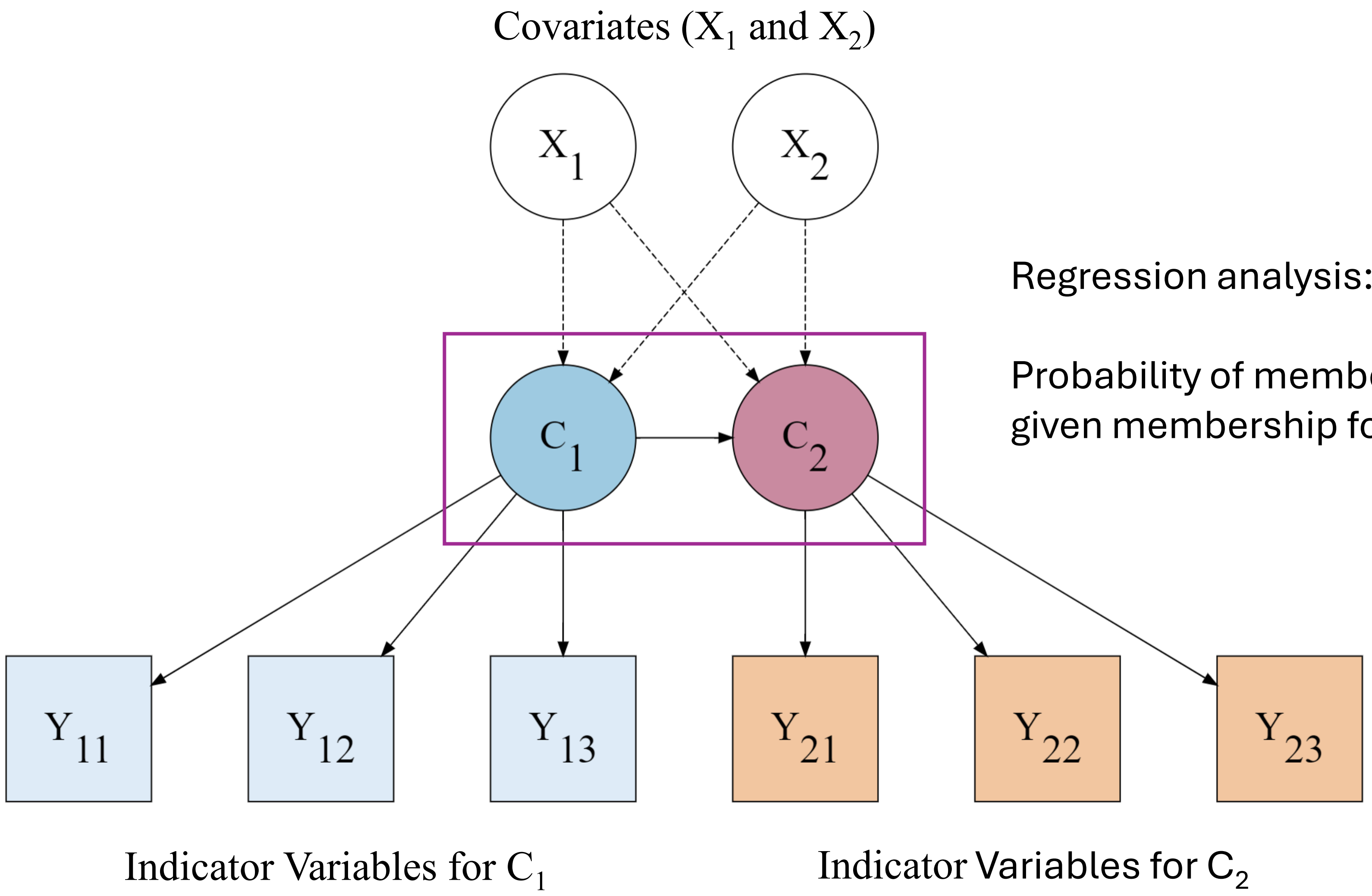
For example:


Covariates (X_1 and X_2)
Gender
X_1
X_2
Ethnicity
Engagement at T1
C_1
C_2
Engagement at T2
Y_{11}
Y_{12}
Y_{13}
Y_{21}
Y_{22}
Y_{23}
Indicator Variables for C_1
School Liking, Hating and Enjoyment at T1
Indicator Variables for C_2
School Liking, Hating and Enjoyment at T2

Step 1: Latent Class analysis



Step 2: Transition Probabilities



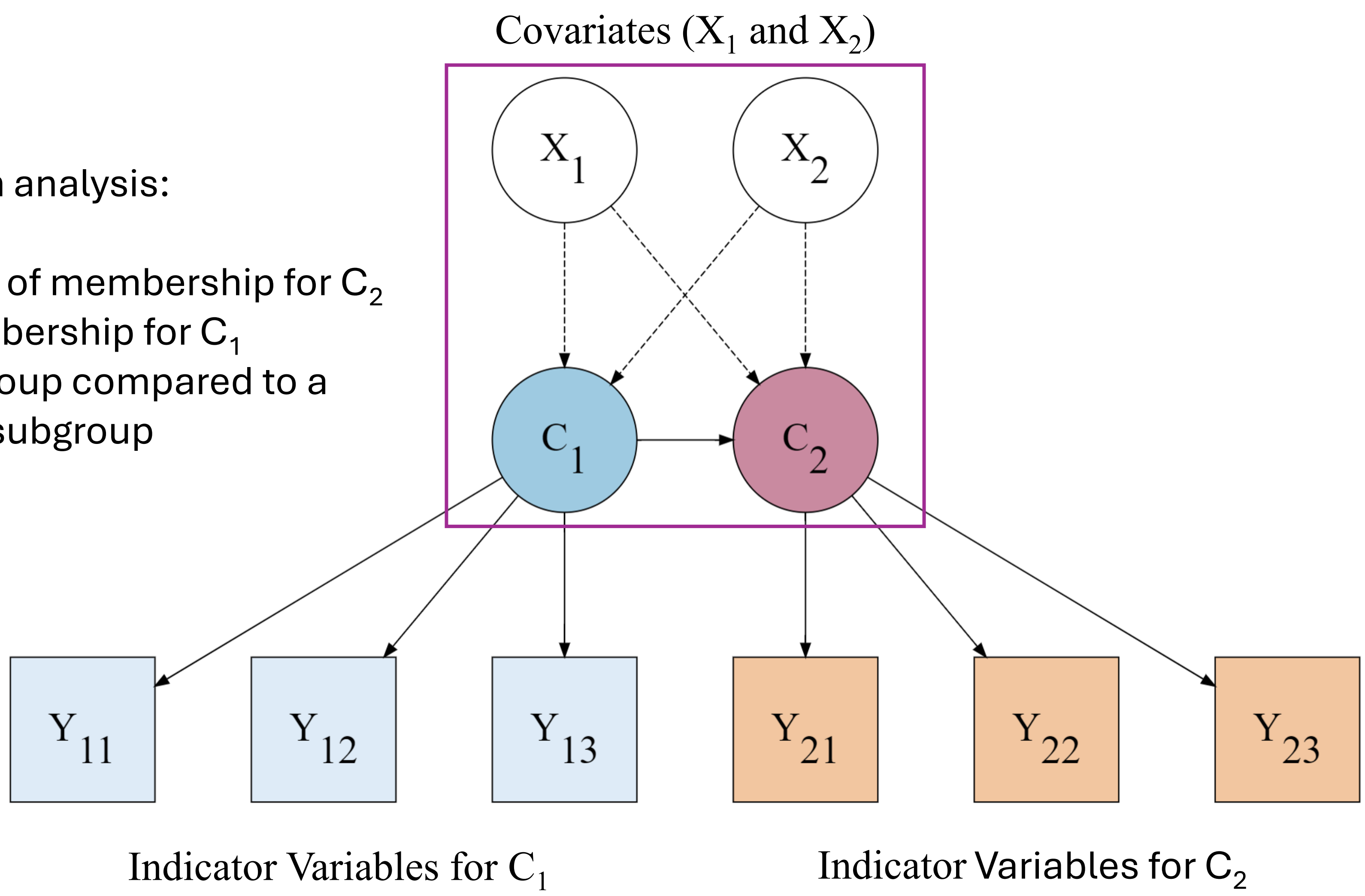
Regression analysis:

Probability of membership for C_2 given membership for C_1

Step 3: Covariates

Regression analysis:

Probability of membership for C_2 given membership for C_1 for a subgroup compared to a reference subgroup



Assumptions

- Latent classes should be mutually exclusive
  - No one can be a member of more than one
- Local independence
  - Indicators should be distinct
- Independence of observations
  - No repeated measures within each time point – only between
- Missingness
  - Missing at random (can approach missingness in a few ways)
- Measurement invariance
  - Classes should represent the same unobserved variable at each time point
  - Usually we look for a stable structure (e.g. same number of classes)
    - But disputed very slightly for this approach!

Quote from Lund and Ritz (2025)

"Why is measurement invariance needed in the first place? If latent classes changes somewhat over time then it may still be interesting to describe transitions between classes; the notion of a transition does not require measurement invariance.

Of course it makes life easier in terms of interpretation, but it's not always that life is so simple."

(Making this argument yourself to a reviewer is another thing entirely however!)

Why conduct LTA?

- LTA is individual based – useful for considering how states change over time
- Based on LCA which allows for examining unobserved variables
- Works well with large social surveys
- Can uncover hidden patterns obscured by population averages

But:

- Can struggle with lower sample sizes
- Must be able to draw together appropriate indicators
- Some subgroups can become very small

Examples

- Transitions experiences over time:
  - No direct variable
    - Constructed from indicators
      - E.g. engagement, peer relationships, wellbeing etc.
    - Could be: negative experiences, positive experiences with peer relationship problems, overall positive experiences

- Engagement over time
  - E.g. high engagement, middle engagement, low engagement

- Mental health profile
  - E.g. different symptom patterns

Additional Factors

- Entropy
  - Separation of latent classes:
    - Higher = more separation
    - But also high for overfit models
- Class size
  - Very small classes (e.g. *n* <10%) may be caused by outliers
    - Worth investigating

Conducting LTA in R

- Some programs will run LTA in one go, including Latent Gold. But it can also be done in M-plus and SAS. There are also alternative ways of doing this in R.
  - R is freely available
  - No need to relearn every time you move institution
  - Other R approaches are quite inflexible in my experience
    - Especially with covariates etc.
- 3-step approach – Lund and Ritz (2025)
  - Very flexible
    - Can incorporate covariates, weights, multiple imputations

1.Latent Class Analysis → 2. Transition Probabilities → 3. Covariates

Interpreting LTA outputs

Latent class analysis

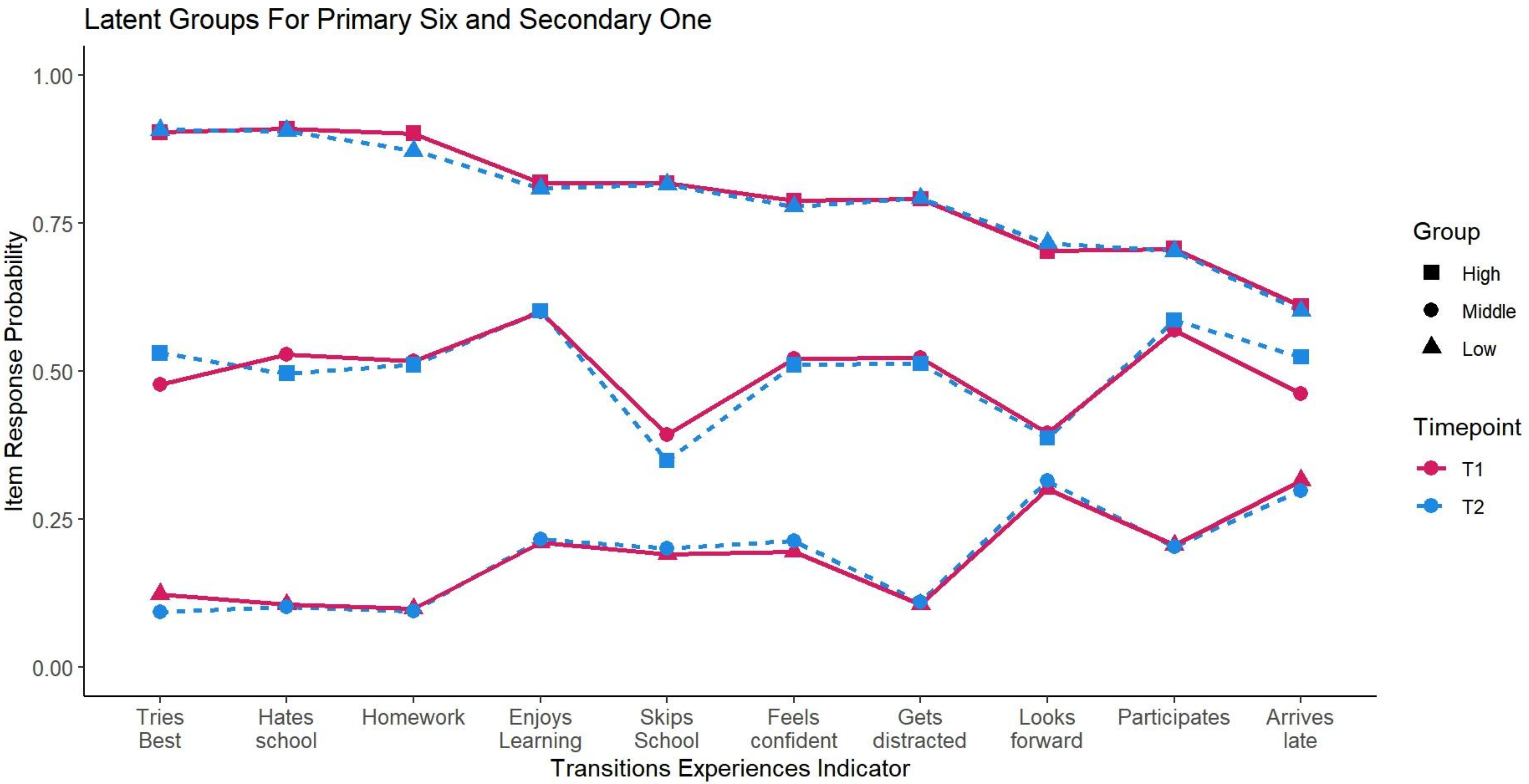
Model Fit Statistics

Classes	N	Parameters	Log-likelihood	AIC	AICc	BIC	aBIC
2	4,518	21	-26,742	53,527	53,527	53,661	53,595
3	4,518	32	-26,603	53,270	53,270	53,475	53,373
4	4,518	43	-26,589	53,265	53,266	53,541	53,404
5	4,518	54	-26,577	53,262	53,263	53,608	53,437

Elbow plots


T2 Fit Statistics
53600
53500
53400
53300
2
3
4
5
Classes
AIC
BIC

Latent class analysis



Latent Class Analysis

Class Sizes

Class	N	Percentage (%)
Low	2,353	52
Middle	804	17
High	1,361	30

Transition Probabilities

T1 Class	T2 Low	T2 Middle	T2 High
Low	75.8% (74.1–77.5)	26.5% (23.7–29.6)	14.6% (12.9–16.5)
Middle	13.1% (11.8–14.5)	38.3% (35.1–41.5)	18.5% (16.6–20.6)
High	11.1% (9.9–12.4)	35.2% (32.1–38.5)	66.9% (64.5–69.3)

Covariates

Transition	Odds Ratio (OR)	Lower 95% Interval	Upper 95% Interval	P-value
Gender (reference = male)				
Low to Middle	0.685	0.540	0.869	**0.002**
Low to High	0.886	0.688	1.140	0.346
Middle to High	0.819	0.523	1.283	0.384
Middle to Low	0.909	0.602	1.372	0.650
High to Low	0.645	0.436	0.955	**0.029**
High to Middle	1.074	0.702	1.642	0.742

Any Questions?

Useful resources

- Lund and Ritz (2025) – the flexible, modular, design we will be using in R
- Nylund-Gibson et al., (2023) – a great resource to learn more about LTA
- Sinha et al., (2022) – an excellent deep dive into LCA

- Some manuals for relevant R packages
  - Tidyverse – for data handling and much more
  - poLCA – for running LCA
  - mice – for multiple imputations
- And a great book for getting started with R
  - https://r4ds.hadley.nz/