Mixture Modeling in Mplus

Gitta Lubke

University of Notre Dame
VU University Amsterdam

Mplus Workshop John's Hopkins 2012

Outline

Overview

- first part: theory and example cross sectional mixture models
  - Latent Class Analysis (LCA)
  - Factor Mixture Modeling (FMM)
  - example: FMM analysis of CBCL attention data
- second part: theory and example longitudinal mixture models
  - Growth Mixture Model (GMM)
  - Latent Transition Analysis Model (LTA)
  - (discrete time) Survival Mixture Analysis Model (SMA)
  - example: SMA of substance use initiation with covariates
- the theory of each model type is structured as follows:
  - conceptual basis
  - what does the model look like
  - a bit more statistical detail
  - how is it done in Mplus
  - what are the critical issues & potential pitfalls
  - interpretation of results

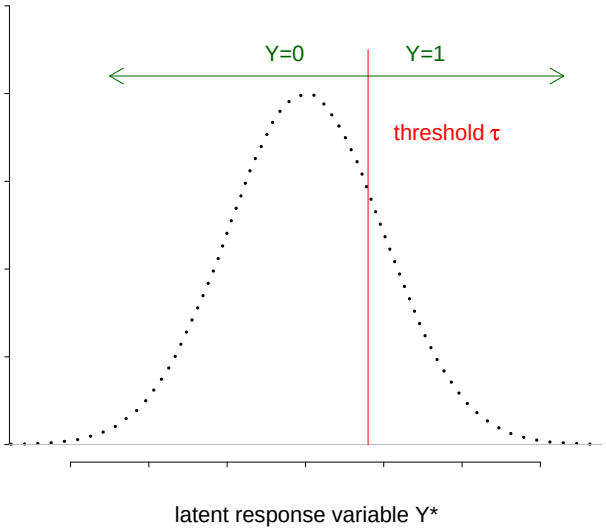
Before we get started ...(1)

- continuous and categorical observed and latent variables

- first the observed variables
- continuous observed variables $\mathbf{Y}$
  - as a default we assume $\mathbf{Y} \sim mvn(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- categorical observed variables $\mathbf{Y}$
  - as a default we assume an underlying unobserved $\mathbf{Y}^*$
  - $\mathbf{Y}^* \sim mvn(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - relation between $\mathbf{Y}$ and $\mathbf{Y}^*$ established through thresholds $\boldsymbol{\tau}$

Illustration

- here a single threshold is shown
- can be extended to $p - 1$ thresholds for p ordered response categories



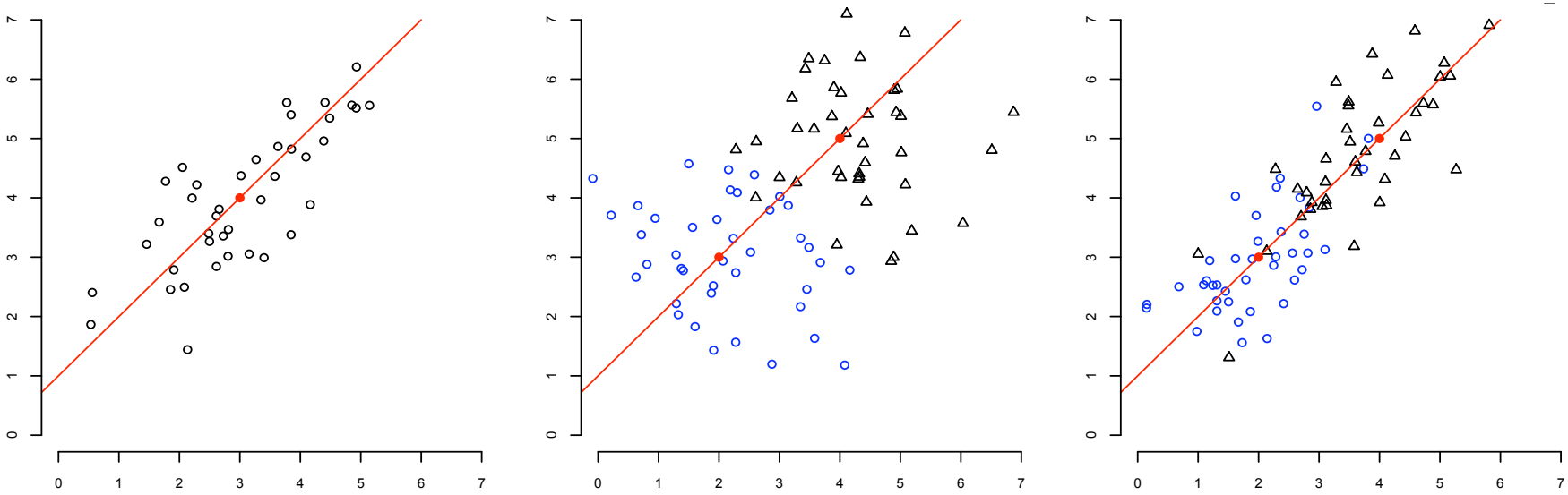
- Mplus can handle a variety of other types of observed variables

Before we get started ...(2)

- so we have multiple observed variables **Y**
  - general goal is to find out why the variables covary
  - different models impose different structures to 'explain' covariation
- EFA
  - continuous latent variables explain covariances between the **Y**
  - several anxiety items: the higher you score on an underlying anxiety dimension, the higher you score on the observed items
- LCA
  - categorical latent variable(s) explain covariances between the **Y**
  - mean differences between classes
- FMM (including GMM)
  - continuous and categorical latent variable(s) explain covariances between the **Y**
  - mean differences between classes and continuous factor(s) within class

Illustration

- covariation due to continuous latent factor
- covariation due to mean differences between latent classes
- covariation due to both factors and class mean differences



Latent Class Analysis: The concept

- historically, for continuous **Y** this has been called latent profile analysis
- multiple observed items (usually less than 10)
- assumption that covariation is due to mean differences between classes
- within class covariation is zero: local independence
- interest in finding qualitatively or quantitatively different classes
  - for instance high and low scoring classes, or class specific patterns (high on some items but low on others)
- exploratory analysis, similar to EFA

What does the modeling look like?

- fit several models with increasing numbers of classes
  - one class, two classes, three classes, É
- within class covariances between **Y** are fixed to zero
- within each class the means and residual variances of the **Y** are estimated
- or, for categorical **Y**, the thresholds are estimated
- models are compared with respect to BIC, or bootstrapped LRT
- model with lowest BIC = best fitting LCA model

A bit more statistical detail

- when fitting a factor model we assume data come from a single homogeneous population, and that $\mathbf{Y} \sim mvn$
  - accordingly, we use a single mvnormal
- when fitting a k-class model we assume there are k classes
  - so now we use a mixture distribution with k mvnormal components
- the joint distribution is a mixture (=weighted sum) of the component distributions
- $f(y) = \Sigma_{k=1}^{K} \pi_k f(y_k; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
  - note that each class has its own parameters
  - LI: Σ is diagonal
  - this constraint can be locally relaxed (example below)
- class weight π_k is the class proportion
- a posteriori it is possible to get the class probabilities for each subject (Bayes theorem, more comments on class assignment later)

How is it done in Mplus? (1)

- from the Mplus User's guide:

```
TITLE:     this is an example of a LCA with
           continuous latent class indicators using
           automatic starting values with random
           starts
DATA:      FILE IS ex7.9.dat;
VARIABLE:  NAMES ARE y1-y4;
           CLASSES = c (2);
ANALYSIS:  TYPE = MIXTURE;
OUTPUT:    TECH1 TECH8;
```

- Mplus has defaults based on common scenarios
  - you can check these using `TECH1`
- add your own within class specification
  - `Model %overall% ... %c#1% ... %c#2%`
  - similar to multi-group model specification
- output: save class probabilities
- make plots: class profiles
- boostrapped LRT in `TECH14`

How is it done in Mplus? (2)

- specifying your own within class model
- from the Mplus User's guide:

```
TITLE:     this is an example of a LCA with binary
           latent class indicators using user-
           specified starting values with random
           starts
DATA:      FILE IS ex7.5.dat;
VARIABLE:  NAMES ARE u1-u4;
           CLASSES = c (2);
           CATEGORICAL = u1-u4;
ANALYSIS:  TYPE = MIXTURE;
           STARTS = 100 10;
           STITERATIONS = 20;
MODEL:
           %OVERALL%
           %c#1%
           [u1$1*1 u2$1*1 u3$1*-1 u4$1*-1];
           %c#2%
           [u1$1*-1 u2$1*-1 u3$1*1 u4$1*1];
OUTPUT:    TECH1 TECH8;
```

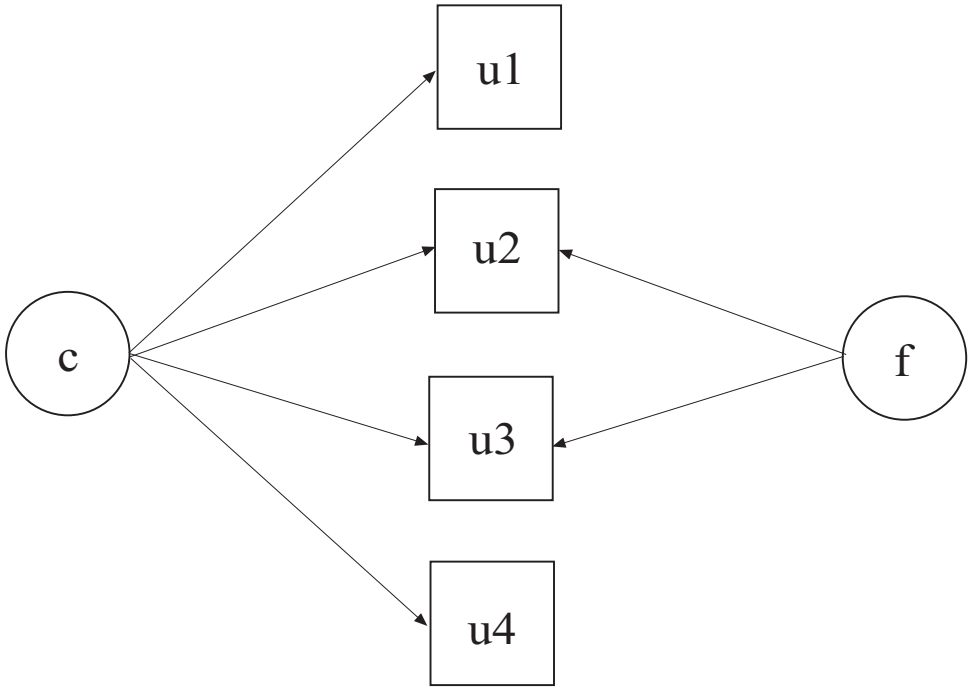
How is it done in Mplus? (3)

- allowing for local dependence
- from the Mplus User's guide:

```
TITLE:     this is an example of LCA with partial
           conditional independence
DATA:      FILE IS ex7.16.dat;
VARIABLE:  NAMES ARE u1-u4;
           CATEGORICAL = u1-u4;
           CLASSES = c(2);
ANALYSIS:  TYPE = MIXTURE;
           ALGORITHM = INTEGRATION;
MODEL:
           %OVERALL%
           f by u2-u3@0;
           f@1; [f@0];
           %c#1%
           [u1$1-u4$1*-1];
           f by u2@1 u3;
OUTPUT:    TECH1 TECH8;
```

How is it done in Mplus? (4)

- allowing for local dependence: path model
- from the Mplus User's guide:

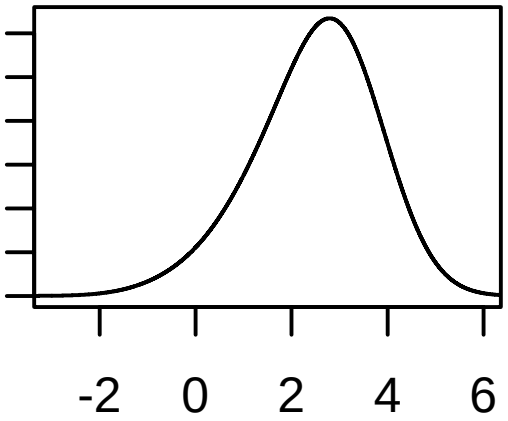


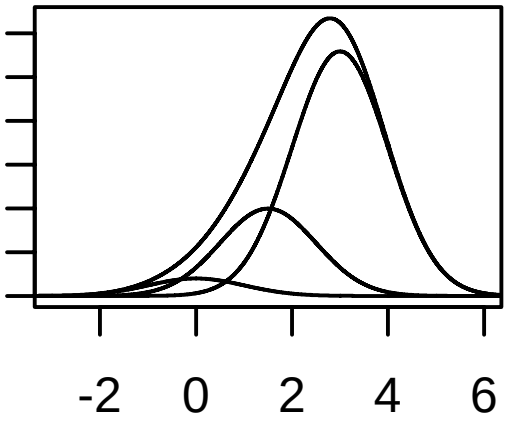
Critical issues & potential pitfalls (1)

- "likelihood function can have multiple local maxima"
- what does that mean?
- ideally when optimizing a function the function is convex, and only has one optimum
  - so it does not matter where you start the iterative procedure to find the maximum
- mixture distributions are known to have many local maxima
- need to use random starts
- no guarantee the solution is not local
- replication of maximum for multiple random starts is reassuring

Critical issues & potential pitfalls (2)

- mixtures of nov normal components can be used to approximate any distribution
- indirect interpretation of mixtures

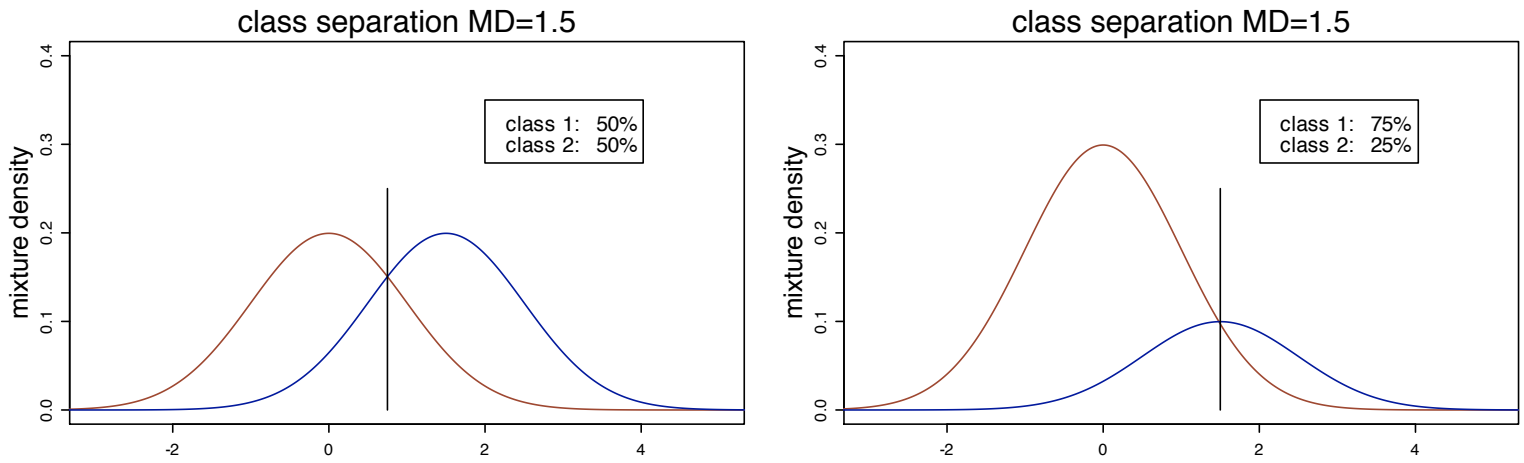




- note that here the variances of the normal components are equal
- if allowed to vary: less components needed for approximation

Critical issues & potential pitfalls (3)

- class assignment accuracy
- mixture models not so good as subject classifier
- much better to estimate structure



Critical issues & potential pitfalls (4)

- number of classes depends on within class parameterization
- if within class variances are allowed to differ, vs. constrained to be the same can affect number of classes
- so when a 4 class model fits better than a 3 class model what does this mean?
- interpretation always cautious
- more on this when discussing FMMs
- LCA a very useful first step: fast and easy
- but it is a very constrained model: local independence

Interpretation of results

- k class model = best fitting model
- mean differences between classes are usually the most interesting
- differences in residual variances can also be of interest
- classes do not necessarily correspond to meaningful subpopulations: indirect use of mixture distributions
- interpretation: k classes describe the observed sample data best
- also: interpretation depends a bit on sample size
  - small sample = sampling fluctuation can give weird classes
  - large samples have more power to detect small classes
- may need to check whether LCA is not too restrictive by comparing to more lenient FMM
- FMMs might fit better, often have less classes because some of the covariances between items is explained by common within class covariation

Factor Mixture Modeling: The concept

- in LCA observed variables are assumed to be independent within class
- what if not all covariation is due to class differences?
- covariation within class can be due to continuous factors
  - for instance gradual severity differences within class
- interest is in finding classes and factors within class
- total covariation between observed variables is decomposed into
  - what is due to class mean differences
  - continuous factors within class

What does the modeling look like?

- fit several models with increasing numbers of classes
- if the number of factors is not known, also fit models with increasing numbers of factors
  - exploratory factor mixture model
- FMM's are usually exploratory with respect to the number of classes
- within each class a factor model is estimated
- models are compared with respect to BIC, or bootstrapped LRT

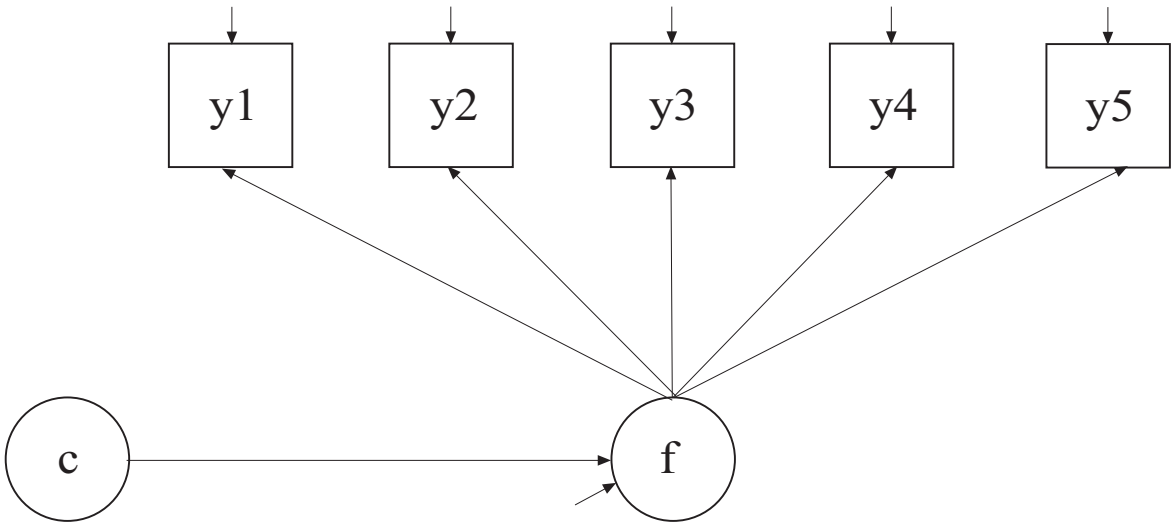
A bit more statistical detail

- within class parameterization affects number of classes
  - $f(y) = \Sigma_{k=1}^{K} \pi_k f(y_k; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- more constraints on $\boldsymbol{\Sigma}_k$ often means more classes
  - measurement invariance
- $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ can be structured in many different ways
- structure can vary across classes
  - great flexibility
  - example: GMM
- number of classes can be extended as well
- useful for multi-group factor mixture modeling
  - using `knownclass` option
  - example: GMM
- useful also for longitudinal data where class membership can change over time
  - sequential process models
  - latent transition models (see later)

How is it done in Mplus? (1)

- from the Mplus User's guide:

```
TITLE:     this is an example of CFA mixture modeling
DATA:      FILE IS ex7.17.dat;
VARIABLE:  NAMES ARE y1-y5;
           CLASSES = c(2);
ANALYSIS:  TYPE = MIXTURE;
MODEL:     %OVERALL%
           f BY y1-y5;
           %c#1%
           [f*1];
OUTPUT:    TECH1 TECH8;
```



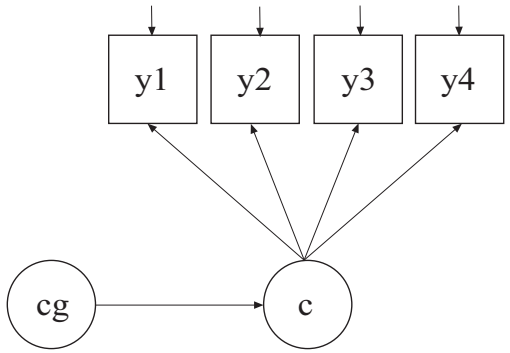
How is it done in Mplus? (2)

- many defaults
  - observed variables means and residual variances
  - factor loadings
  - factor variances
  - factor means
- check `TECH1` and add your own within class specification

How is it done in Mplus? (3a)

- multiple group FMM using option `knownclass`
- from the Mplus User's guide:

```
TITLE:     this is an example of mixture modeling
           with known classes (multiple group
           analysis)
DATA:      FILE IS ex7.21.dat;
VARIABLE:  NAMES = g y1-y4;
           CLASSES = cg (2) c (2);
           KNOWNCLASS = cg (g = 0 g = 1);
ANALYSIS:  TYPE = MIXTURE;
MODEL:
           %OVERALL%
           c ON cg;
MODEL c:
           %c#1%
           [y1-y4];
           %c#2%
           [y1-y4];
MODEL cg:
           %cg#1%
           y1-y4;
           %cg#2%
           y1-y4;
OUTPUT:    TECH1 TECH8;
```



How is it done in Mplus? (3b)

- using option `knownclass` is handy to investigate MI for observed and latent groups
- for categorical observed variables: thresholds and loadings
- for continuous observed variables: intercepts, loadings, residual variances
- note: BIC might incorrectly favor more constrained MI model
- especially for categorical items with more than 2 response categories

How is it done in Mplus? (4)

- extension to Structural Equation Mixture Models
- from the Mplus User's guide:

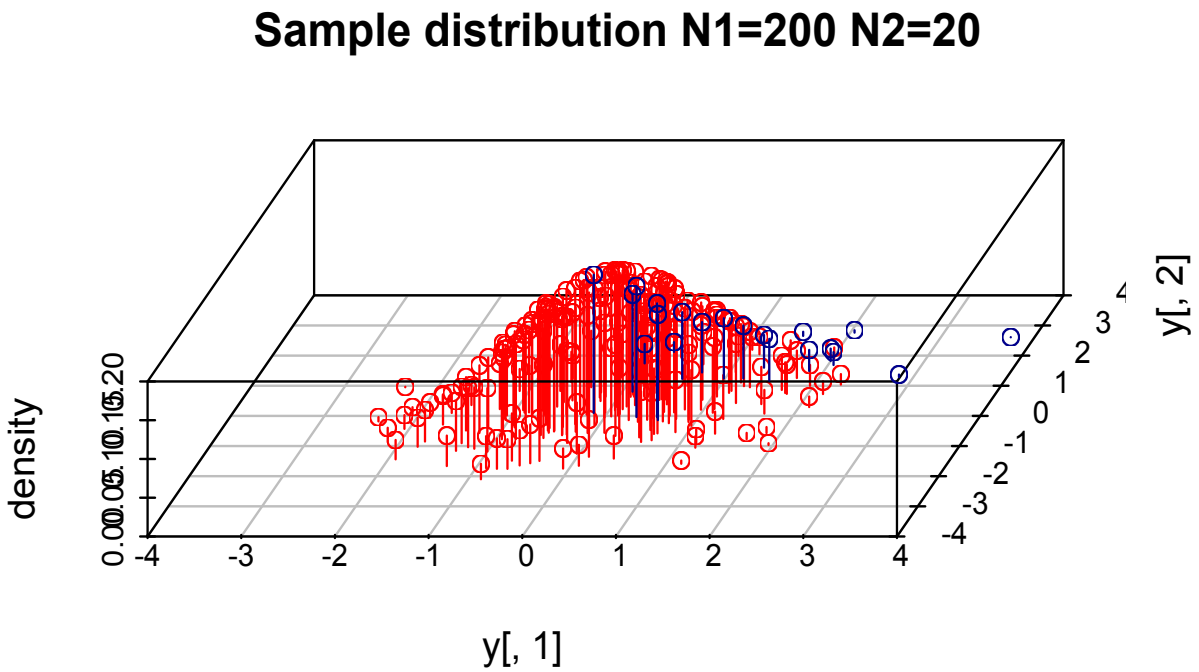
```
TITLE:     this is an example of structural equation
           mixture modeling
DATA:      FILE IS ex7.20.dat;
VARIABLE:  NAMES ARE y1-y6;
           CLASSES = c (2);
ANALYSIS:  TYPE = MIXTURE;
MODEL:
           %OVERALL%
           f1 BY y1-y3;
           f2 BY y4-y6;
           f2 ON f1;
           %c#1%
           [f1*1 f2];
           f2 ON f1;
OUTPUT:    TECH1 TECH8;
```

Critical issues & potential pitfalls (1)

- recall: number of classes depends on within class parameterization
- very lenient models can take a lot of computation time and may not converge
- start with more constrained models: LCA
  - number of classes required in LCA are an upper bound
- start relaxing constraints
  - factors within class
  - factor variance differences across classes
  - loading differences
  - item mean or threshold differences

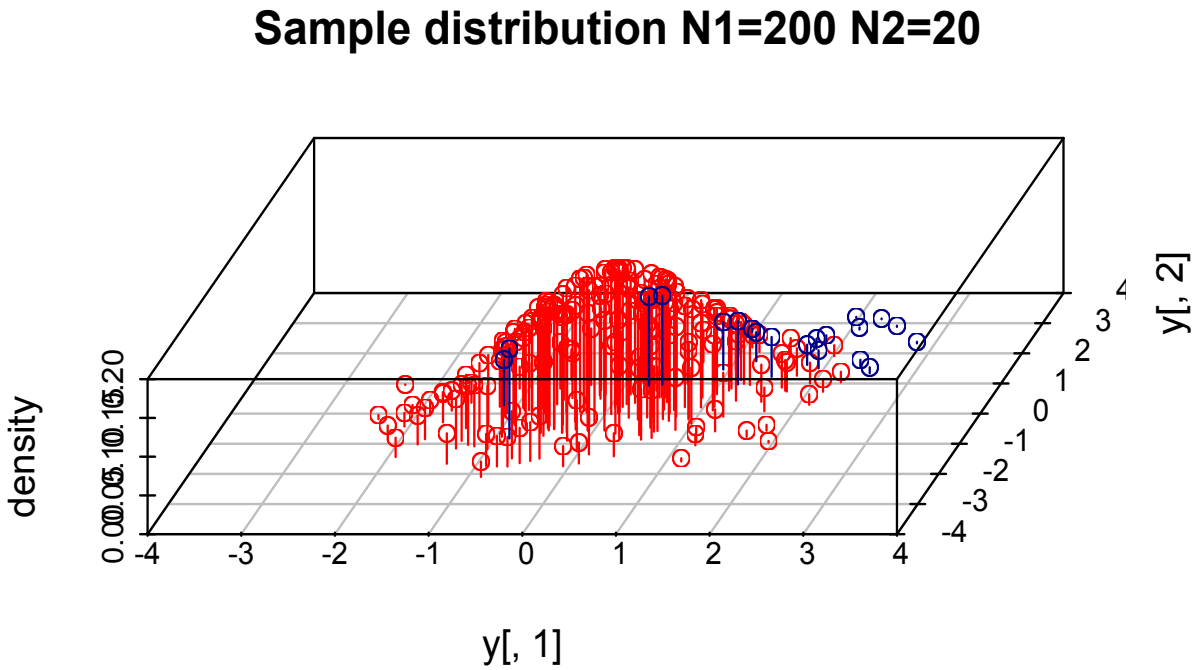
Critical issues & potential pitfalls (2a)

- sample size and sampling fluctuation
  - especially sample size in the smaller class



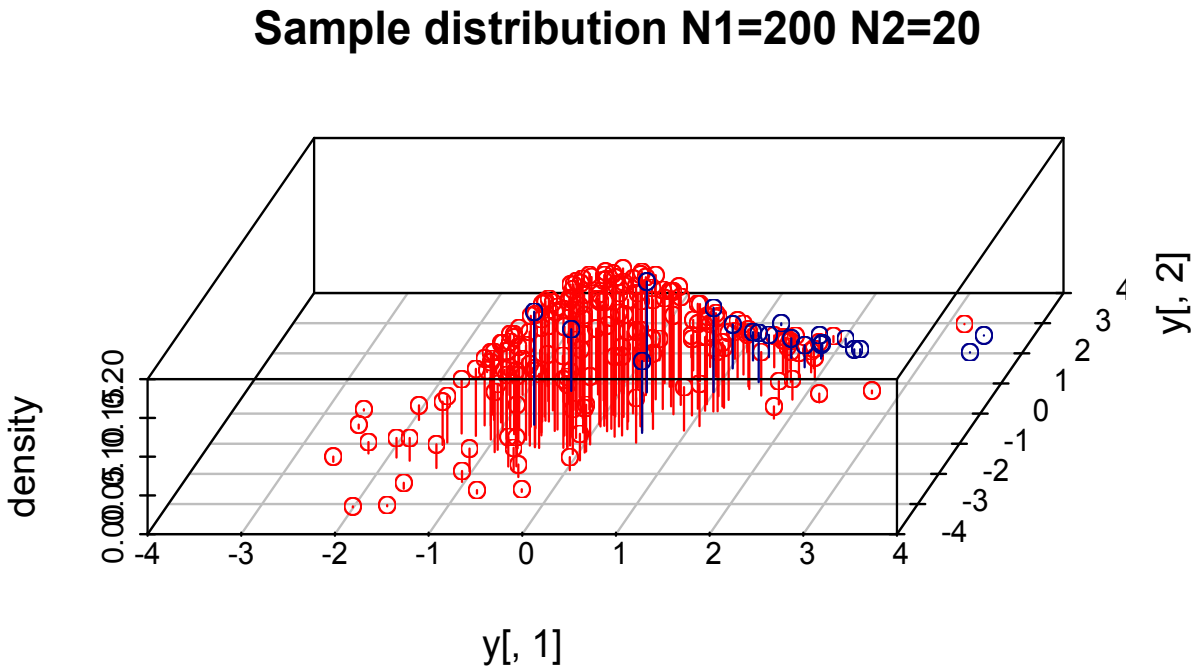
Critical issues & potential pitfalls (2b)

- another draw from the same mixture distribution



Critical issues & potential pitfalls (2c)

- and another draw from the same mixture distribution

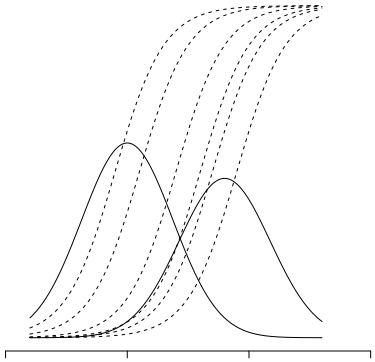


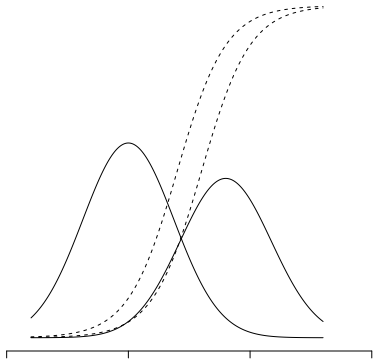
Critical issues & potential pitfalls (2d)

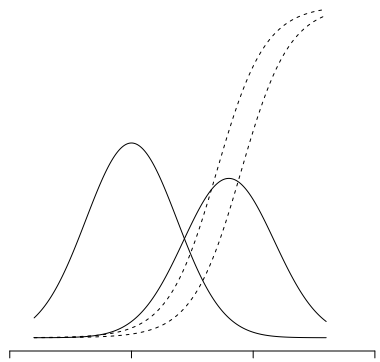
- so, what are the implications?
  - there are limits on the parametrization
  - replication in an independent data set is necessary before drawing conclusions
  - without replication interpretation of results is tentative

Critical issues & potential pitfalls (3)

- measuring all hypothesized latent classes







Interpretation of results

- compare models featuring several different within class structures with increasing numbers of classes
- decision based on BIC, bootstrapped LRT, and theory
  - not necessary that a single model is the best fitting model!
- contextualize interpretation:
  - model complexity: consider number of estimated parameters (power, BIC)
  - sample size
  - class separation
  - quality measurement model
- without replication interpretation very cautious

Overview mixture models for longitudinal data

- Latent Class Growth Analysis Model (LCGA)
- Growth Mixture Model (GMM)
- Latent Transition Analysis Model (LTA)
- (discrete time) Survival Mixture Analysis Model (SMA)
- example: SMA of substance use initiation with covariates

Growth Mixture Modeling: The concept

- for each subject data are collected at several time points
- interest in development over time
- mixture models are an extension of the linear of quadratic growth model
- find classes that differ with respect to their average trajectories
- classic GMM has a single latent categorical variable
  - each subject stays in one class

What does the modeling look like?

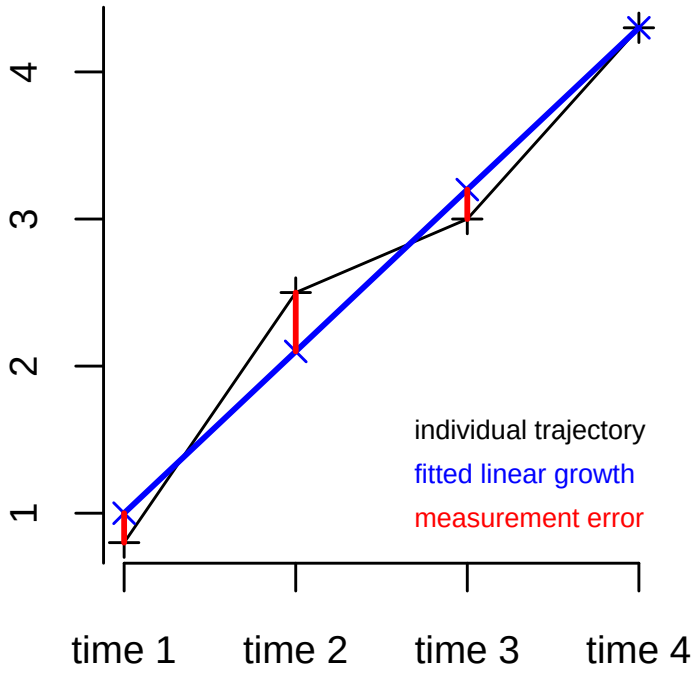
- there are different growth mixture models
  - differences with respect to how individual trajectories are modeled
- latent class growth analysis (LCGA)
  - estimate average growth trajectory in each class
  - regard individual variation around trajectory as random error
  - structure similar to LCA
- growth mixture model (GMM)
  - estimate average growth trajectory in each class
  - estimate random effects: intercept and slopes
  - structure similar to FMM

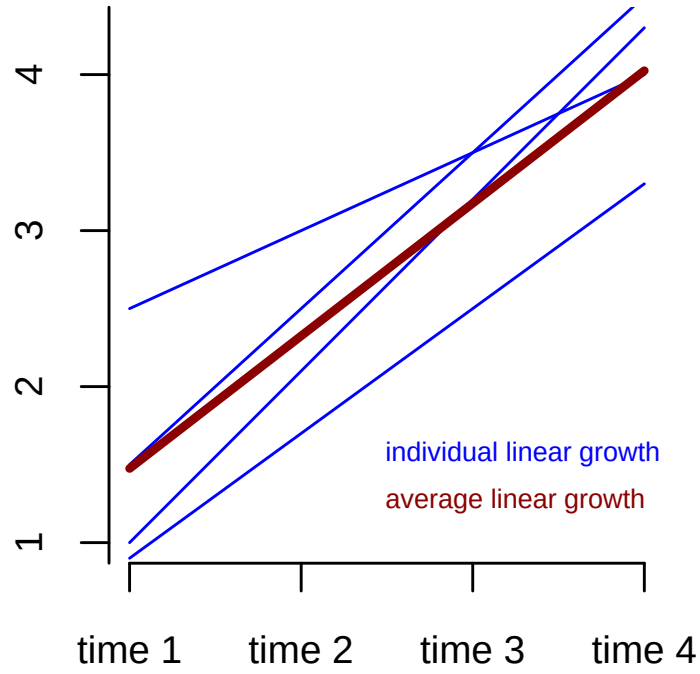
A bit more statistical detail

- both the LCGA and the GMM can be accommodated using the more general FMM framework
  - the within class factors are used to model intercept and slope of the growth curve
  - loading matrix used to model time axis (as in latent growth model)
- $Y = \lambda_i i + \lambda_s s + \varepsilon$
- $E(Y) = \lambda_i E(i) + \lambda_s E(s)$
- $Var(Y) = \lambda_i^2 Var(i) + \lambda_s^2 Var(s) + \lambda_i \lambda_s Cov(i, s) + Var(\varepsilon)$
- fixing the variances of intercept and slope factors to zero in each class gives the LCGA
  - variance of Y equals error variance
  - just as in LCA
- estimating the variances of i and s, and covariance between i and s gives the GMM
  - just as in an FMM

Illustration

- individual and average trajectories

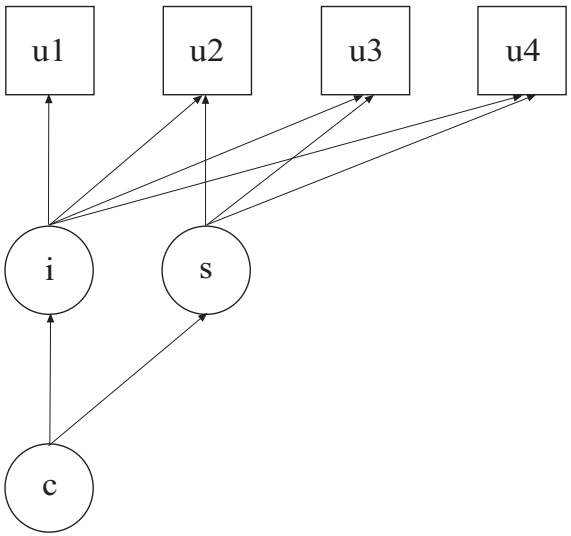




How is LCGA done in Mplus?

- note that without `Algorithm=integration` LCGA is the default
- since *i* and *s* have zero variance, their covariance equals zero

```
TITLE:      this is an example of a LCGA for a binary
            outcome
DATA:       FILE IS ex8.9.dat;
VARIABLE:   NAMES ARE u1-u4;
            CLASSES = c (2);
            CATEGORICAL = u1-u4;
ANALYSIS:   TYPE = MIXTURE;
MODEL:
            %OVERALL%
            i s | u1@0 u2@1 u3@2 u4@3;
OUTPUT:     TECH1 TECH8;
```



How is GMM done in Mplus? (1)

- here variances of *i* and *s* and their covariance are estimated
- it is possible to regress *i* and/or *s* on a covariates **X**

```
TITLE:     this is an example of a GMM for a
           categorical outcome using automatic
           starting values and random starts
DATA:      FILE IS ex8.4.dat;
VARIABLE:  NAMES ARE u1-u4 x;
           CLASSES = c (2);
           CATEGORICAL = u1-u4;
ANALYSIS:  TYPE = MIXTURE;
           ALGORITHM = INTEGRATION;
MODEL:
           %OVERALL%
           i s | u1@0 u2@1 u3@2 u4@3;
           i s ON x;
           c ON x;
OUTPUT:    TECH1 TECH8;
```

How is GMM done in Mplus? (2)

- there are numerous options to treat different types of outcome variables appropriately
  - counts, zero-inflated poisson
  - especially the latter is often a good choice for empirical count data
- accommodate non-linear growth with a quadratic factor
- incorporation of covariates, and class-predicted outcomes
- multiple groups using `knownclass` option.

Critical issues & potential pitfalls

- always start with a spaghetti plot
- proceed with the more constrained LCGA model, then relax constraints
  - first only estimate variance of i, then add variance s and factor covariance
  - sometimes slope variance is close to zero, causing difficulties estimating factor covariance
- in general, the critical issues discussed for LCA and FMMs carry over
  - random starts
  - over extraction of classes / indirect interpretation of mixtures
  - sampling fluctuation and small classes
  - class assignment
- GMMs are a bit more stable than the more general FMMs since Λ is usually fixed

Critical issues: A note on confirmatory GMMs

- exploratory and confirmatory mixture models
- generally the number of classes, and the level of leniency within class is not known a priori
  - exploratory mixture analysis
- in Developmental Psychology (and other areas) there are theories implying different trajectory classes
- these theories can in principle be translated into confirmatory GMMs
- a couple of considerations
  - use constraints on parameters corresponding to theory
  - if possible use covariates and distal outcomes to further define the theory-implied classes
  - compare the confirmatory model to more unconstrained GMMs to check whether this is a reasonable model for the data
- if the theory implied model is complex, start with fitting parts

Interpretation of results

- interest is usually in the average trajectories in each class
  - check the estimates of the mean and variance of the slope factors and their standard errors
  - should be reported in a paper
- acknowledge non-converged models when interpreting model comparisons
- if a series of increasing lenient models is fitted, the cause of non-convergence is easier to narrow down
- as in FMMs, contextualize results
  - sampling fluctuation
  - sample size of the smaller classes
- if possible replicate findings in an independent sample to validate classes
  - honeymoon for GMMs is over: more critical papers concerning interpretation of classes are coming up
- using distal outcomes is another nice option to validate classes

Latent transition models: The concept

- two or more latent class variables
- the focus here is on models permitting subjects to change classes over time
- examples are
  - classic latent transition
  - sequential processes
  - mover stayer mixture model
- interesting applications

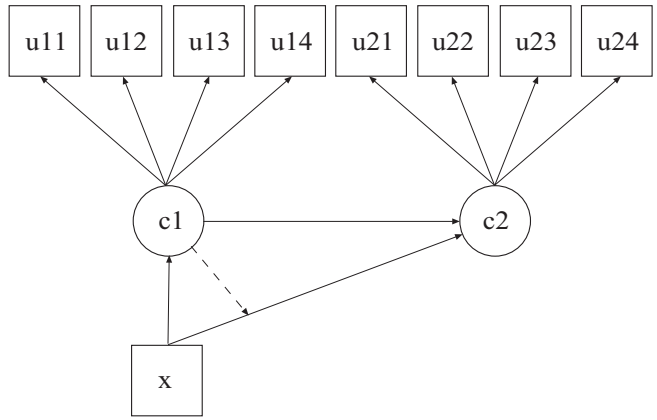
What does the modeling look like?

- it looks quite different for the classic latent transition, sequential process models, and mover stayer mixture models!
- classic LTA simply connects two or more LCA models
- sequential processes connects two or more growth mixture models
- mover stayer explicitly models the transition matrix using an underlying third latent class variable

How is a classic latent transition model done in Mplus?

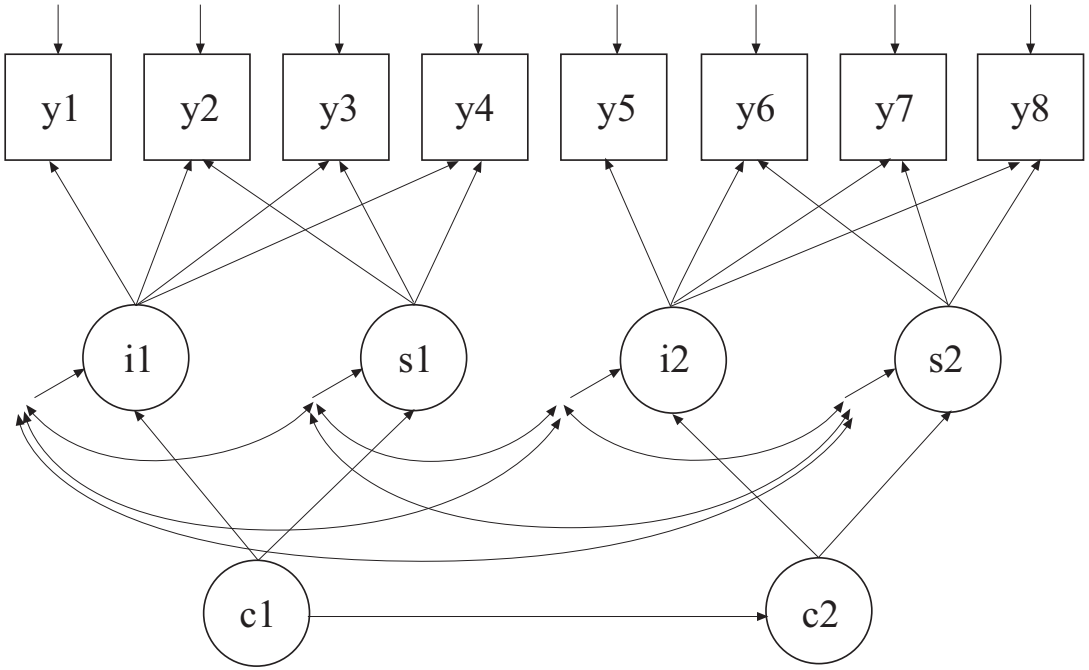
- LTA consists of two (or more) connected LCA models

```
TITLE:      this is an example of a LTA with a
            covariate and an interaction
DATA:       FILE IS ex8.13.dat;
VARIABLE:   NAMES ARE u11-u14 u21-u24 x;
            CATEGORICAL = u11-u14 u21-u24;
            CLASSES = c1 (2) c2 (2);
ANALYSIS:   TYPE = MIXTURE;
MODEL:
            %OVERALL%
            c2 ON c1 x;
            c1 ON x;
MODEL c1:
            %c1#1%
            [u11$1-u14$1*1]  (1-4);
            c2 ON x;
            %c1#2%
            [u11$1-u14$1*-1] (5-8);
MODEL c2:
            %c2#1%
            [u21$1-u24$1*1]  (1-4);
            %c2#2%
            [u21$1-u24$1*-1] (5-8);
OUTPUT:     TECH1 TECH8;
```



How are sequential processes modeled in Mplus? (1)

- two growth mixture models, with correlations between classes and factors
- correlations can be replaced by regressions



How are sequential processes modeled in Mplus? (2)

- the two growth mixture models don't have to have the same number of classes
- this would be a potential follow-up model for the CBCL attention data
- or: different instrument is used for older subjects

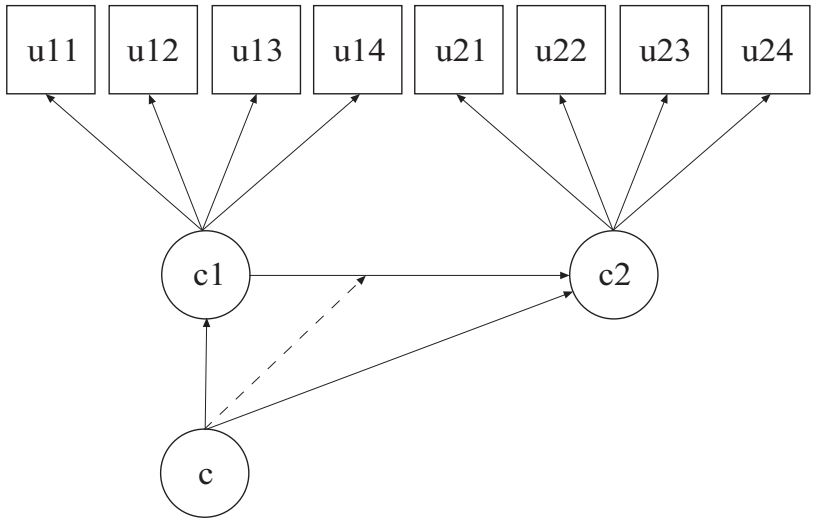
```
TITLE:      this is an example of a sequential
            process GMM for continuous outcomes with
            two categorical latent variables
DATA:       FILE IS ex8.7.dat;
VARIABLE:   NAMES ARE y1-y8;
            CLASSES = c1 (3) c2 (2);
ANALYSIS:   TYPE = MIXTURE;
MODEL:
            %OVERALL%
            i1 s1 | y1@0 y2@1 y3@2 y4@3;
            i2 s2 | y5@0 y6@1 y7@2 y8@3;
            c2 ON c1;
MODEL c1:
            %c1#1%
            [i1 s1];

            %c1#2%
            [i1*1 s1];

            %c1#3%
            [i1*2 s1];
MODEL c2:
            %c2#1%
            [i2 s2];

            %c2#2%
            [i2*-1 s2];
OUTPUT:    TECH1 TECH8;
```

How is the Mover Stayer model implemented in Mplus? (1)

- the previous two models are only connecting models we discussed before
- mover stayer script looks a bit more complicated
- again two LCA models as in LTA
- connection between the two class models done differently
  - not just a regression between the class variables
- instead, a 'higher-order' underlying class variable is used to predefine classes for subjects who stay in the same kind of class over time, and for subjects who move
  - this is accomplished by fixing transition probabilities to one or zero
- think of classes for 'never-users', 'always-users', 'quitters' and 'late-starters'
- data decide which subject joins which class

How is the Mover Stayer model implemented in Mplus? (2)

- first the path diagram



How is the Mover Stayer model implemented in Mplus? (3)

- class1 of c are movers, class 2 of c are stayers
- fixing class probabilities to one and zero is done by fixing the intercept of the corresponding regression of $c1$ and $c2$ on c to some high or low number, respectively

```
TITLE:     this is an example of latent transition
           mixture analysis (mover-stayer model)
DATA:      FILE IS ex8.14.dat;
VARIABLE:  NAMES ARE u11-u14 u21-u24;
           CATEGORICAL = u11-u14 u21-u24;
           CLASSES = c (2) c1 (2) c2 (2);
ANALYSIS:  TYPE = MIXTURE;
MODEL:
           %OVERALL%
           c1 ON c;
           [c1#1@10];
           c2 ON c;
           [c2#1@-10];
MODEL c:
           %c#1%
           c2 ON c1;
           %c#2%
           c2 ON c1@20;
```

```
MODEL c.c1:
           %c#1.c1#1%
           [u11$1-u14$1] (1-4);
           %c#1.c1#2%
           [u11$1-u14$1] (5-8);
           %c#2.c1#1%
           [u11$1-u14$1@15];
           %c#2.c1#2%
           [u11$1-u14$1@-15];
MODEL c.c2:
           %c#1.c2#1%
           [u21$1-u24$1] (1-4);
           %c#1.c2#2%
           [u21$1-u24$1] (5-8);
           %c#2.c2#1%
           [u21$1-u24$1@15];
           %c#2.c2#2%
           [u21$1-u24$1@-15];
OUTPUT:    TECH1 TECH8;
```

Critical issues & potential pitfalls

- check observed data
  - do the data seem to support change over time
  - do the data support classes
- the main potential pitfall is to start with the final complex model, and to end up with non-convergence
- build models stepwise
  - check each part separately
  - numbers of classes
  - standard errors of within class parameters
- when used appropriately these models have a wide range of very interesting applications

Interpretation of results

- main focus is obviously on transition
  - present transition matrices
- possibly also on included covariates that predict transition
- same caution as in other mixture models when it comes to interpreting classes of classic LTA and sequential process models
- classes in mover stayer are predefined by constraints

Survival Models: The concept

- the main idea is to predict the time until an event occurs using some suitable covariate
- examples of events
  - starting to date, smoke, work, ...
  - first arrest, or relapse
  - getting married, divorced
- event is the transition from one state to a different state
- the main difference with previous transition models is the outcome variable
- the outcome is rate of occurrence of an event conditional on the event not having occurred before

A bit more statistical detail (1)

- first some terminology
  - event
  - state
  - duration of nonoccurrence
  - hazard rate
- hazard rate is the rate of occurrence of an event during the risk period at a given time t
- If at time t there are I subjects at risk of making a transition to state J then

$$\texttt{rate}(t) = \frac{\texttt{\# subjects in state } J \texttt{ at time } t}{\sum_{i=1}^{I} \texttt{risk period}_i \texttt{ at time } t}$$

- example: the hazard rate is the ratio of divorced subjects at time t and the total number of marriage years counted at time t

A bit more statistical detail (2)

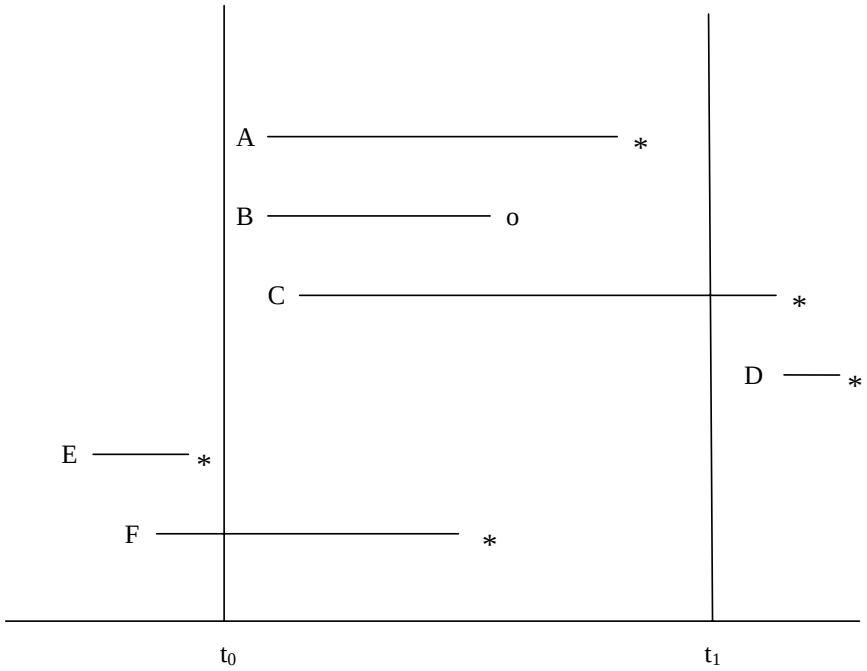
- for discrete time, the duration of non-occurrence T is a discrete random variable, observed at discrete time points
  - let $l = 1, \ldots, L$ indicate the number of time points
- if a state is observed at time point t_l, but not yet at time point t_{l-1}, one can deduct the event took place before t_l but not before t_{l-1}
  - therefore, the duration of non-occurrence is $T = t_{l-1}$
- the probability that a state is observed at time t_l is the probability that the event took place in the interval $(t_{l-1}, t_l]$
  - excludes t_{l-1} and includes t_l
- denote the probability as $f(t_l) = p(T = t_l)$
- the survival function gives the probability that the event has not occurred before or at t_{l-1}:
  - $S(t_{l-1}) = p(T > t_{l-1}) = \sum_{k=l}^{L} f(t_k)$
- the conditional probability that a state is observed at $T = t_l$ given that the event has not occurred prior to t_{l-1} is
  - $\lambda(t_l) = p(T = t_l \mid T \geq t_l) = \frac{f(t_l)}{S(t_{l-1})}$

Why not just use logistic regression (yes/no event)? (1)

- loss of information
  - whether or not event takes place before or after the observe time point
- no time dependent covariates
  - logit model can not include `time` $\times$ `covariate` effects
  - for example, the **effect of** educational level may change over time
  - highly educated people might have a higher probability of promotion in the first 3 years, whereas lower educated people might have a higher chance after three years
- no time varying covariates
  - time-varying covariates are a variables that for individual i changes value over time
  - for example, the number of children can change over a period of time, which can have an influence of promotions of women. Another example would be an increasing educational level resulting in better chances of promotion

Why not just use logistic regression (yes/no event)? (2)

- no possibility to deal with censoring
  - censoring is a form of partially missing data that occurs when subjects are not observed during the full time interval that is chosen for the logit model
  - for instance, subjects dropping out before time period, or subjects without event before data collection ends



What does the modeling look like: The data

- one column in the data file per time interval
- whether or not an event has taken place observed at t time points indicated by 0 and 1
- after the event has taken place data are 'missing'

```
1       11316754        0       1       999     999     999
2       12571478        0       0       1       999     999
3       14614574        0       0       0       0       0
4       14714942        0       0       0       0       0
```

What does the modeling look like: Different models

- for discrete time survival mixtures the hazard is modeled conditional on covariates
- the hazard can be modeled in different ways
- unconstrained
  - hazard can be different in each time interval,
  - covariates can have interval-specific regression weights
- proportional
  - hazard can increase and/or decrease over time
  - covariate effects follow the proportionality constraint
  - only one regression coefficient is estimated for each covariate
- constant
  - hazard is constant over time
  - only one regression coefficient is estimated for each covariate

What does the modeling look like: Mixture models

- 2-class discrete time survival model
- one class reflects the 'long-term survivors'
  - event is never observed
- the other class models the hazard for subjects with observed events
- based on covariate responses subjects without observed event can end up in the second class
  - this is the additional value of using a latent class variable

How is it done in Mplus? (1)

- 2-class discrete time survival model with proportional hazard

```
Title:  proportional hazard
Data: File is surv1109.txt;
Variable: Names are ... v8T2-v8T8;
Missing are all (991 993 995 999);
Categorical are v8T2-v8T8;
Classes = c(2);
Analysis:
      Type = Mixture;
Model:  %overall%
    f by v8T2-v8T8@1;
    f on bsex snp1-snp10;
    [f@0];
    c#1 on bsex snp1-snp10;
        %c#1%
    [f@0];
    [v8T2$1-v8T8$1@15];
    f on bsex@0 snp1-snp10@0;
        %c#2%
    [f@0];
    [v8T2$1-v8T8$1];
    f on bsex snp1-snp10;
Output: stand  Tech1 Tech8;
```

How is it done in Mplus? (2)

- unconstrained hazard model estimates intercepts of event variables in both classes
- unconstrained hazard model estimates regression of event variables on covariates in both classes
- constant hazard model constrains intercepts of event variables to be the same across classes
- estimates factor mean difference instead
- constant hazard model constrains regression of event variables on covariates to be the same across classes

Critical issues & potential pitfalls

- check the event data and covariates
- unconstrained hazard model is highly parameterized
  - regression coefficients for covariates at all time points
  - data might not contain sufficient information
  - non-convergence or unstable results (standard errors)
- still: comparison of different models provides more information than fitting one type of model
- survival (mixtures) should be considered more often due to advantages compared to logit models

Interpretation of results

- comparison using BIC as usual, compare to expectation based on theory
- interest in class sizes and patterns of covariate effects
- advantage of mixture is that subjects with unobserved event can still be in risk class
  - compare proportion of subjects with observe event to class size
- contextualize results with respect to precision of event observations
  - how dense were time points observed
  - observed events vs. retrospective self-reports
- replication if possible

Second last slide

- Thank you for listening!

Acknowledgements

- collaborators CBCL and ADD Health studies
- Netherlands Twin Registry team and participants
- ADD Health team and participants
- my NIH funding: DA018673 by NIDA