Incongruence Between the Statistical Theory and Substantive Application of Growth Mixture Models in Psychological Research

October 8, 2005

Meeting of the Society of Multivariate Experimental Psychology

Daniel J. Bauer

L.L. Thurstone Psychometric Laboratory

The University of North Carolina at Chapel Hill

Outline

Act I: Character Introduction

- **The latent curve model**
- **The growth mixture model**
- **The prototypical empirical application**

Act II: The Challenge

- **Methodological concerns with applications of growth mixture models**
- **Theoretical concerns with applications of growth mixture models**

Act III: The Dénouement

- **Doing better science with and without growth mixture models.**

The Latent Curve Model

Unconditional Model

Matrix expression:

$$\mathbf{y}_i = \mathbf{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$$
$$= \mathbf{\Lambda}\boldsymbol{\alpha} + \mathbf{\Lambda}\boldsymbol{\zeta}_i + \boldsymbol{\varepsilon}_i$$

Marginal PDF:

$$f(\mathbf{y}_i) = \phi\left(\mathbf{\Lambda}\boldsymbol{\alpha},\ \mathbf{\Lambda}\mathbf{\Psi}\mathbf{\Lambda}' + \mathbf{\Theta}\right)$$

Conditional Model

Matrix expression:

$$\mathbf{y}_i = \mathbf{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$$
$$= \mathbf{\Lambda}(\boldsymbol{\alpha} + \mathbf{\Gamma}\mathbf{x}_i + \boldsymbol{\zeta}_i) + \boldsymbol{\varepsilon}_i$$
$$= \mathbf{\Lambda}\boldsymbol{\alpha} + \mathbf{\Lambda}\mathbf{\Gamma}\mathbf{x}_i + \mathbf{\Lambda}\boldsymbol{\zeta}_i + \boldsymbol{\varepsilon}_i$$

Marginal PDF:

$$f(\mathbf{y}_i \mid \mathbf{x}_i) = \phi\left(\mathbf{\Lambda}\boldsymbol{\alpha} + \mathbf{\Lambda}\mathbf{\Gamma}\mathbf{x}_i,\ \mathbf{\Lambda}\mathbf{\Psi}\mathbf{\Lambda}' + \mathbf{\Theta}\right)$$

Key Assumptions

- **iid random effects (exchangeability / single population)**
- **Normally distributed random effects and residuals**
  - **Implies marginal normality of repeated measures**
- **Properly specified mean and covariance structure**
- **Linear relationships between repeated measures and exogenous predictors**
- **Simple random sample**
- **Data missing at random**

The Growth Mixture Model

- Elaborates the LCM by allowing latent classes, relaxing assumption of single population

The Growth Mixture Model

Unconditional Model

Marginal PDF: $f(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_k \phi_k \left(\mathbf{\Lambda}_k \boldsymbol{\alpha}_k, \ \mathbf{\Lambda}_k \mathbf{\Psi}_k \mathbf{\Lambda}_k' + \mathbf{\Theta}_k \right)$

Conditional Model

Marginal PDF: $f(\mathbf{y}_i \mid \mathbf{x}_i) = \sum_{k=1}^{K} \pi_{ik}(\mathbf{x}_i) \phi_k \left(\mathbf{\Lambda}_k \boldsymbol{\alpha}_k + \mathbf{\Lambda}_k \mathbf{\Gamma}_k \mathbf{x}_i, \ \mathbf{\Lambda}_k \mathbf{\Psi}_k \mathbf{\Lambda}_k' + \mathbf{\Theta}_k \right)$

$$\pi_{ik}(\mathbf{x}_i) = \frac{\exp(\alpha_{c_k} + \gamma_{c_k}' \mathbf{x}_{ik})}{\sum_{k=1}^{K} \exp(\alpha_{c_k} + \gamma_{c_k}' \mathbf{x}_{ik})}$$

Variants on the Growth Mixture Model

Identical functional form:

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_k \phi_k \left(\mathbf{\Lambda} \boldsymbol{\alpha}_k, \ \mathbf{\Lambda} \mathbf{\Psi}_k \mathbf{\Lambda}' + \mathbf{\Theta}_k \right)$$

Homogeneous class covariance matrices:

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_k \phi_k \left(\mathbf{\Lambda} \boldsymbol{\alpha}_k, \ \mathbf{\Lambda} \mathbf{\Psi} \mathbf{\Lambda}' + \mathbf{\Theta} \right)$$

No random effects (latent trajectory class analysis):

$$f(\mathbf{y}_i) = \sum_{k=1}^{K} \pi_k \phi_k \left(\mathbf{\Lambda} \boldsymbol{\alpha}_k, \mathbf{\Theta} \right)$$

Caveats to Assumptions

- Observed variables can be binary, ordinal or counts, but random effects must be normal.
- Complex samples can be handled given clustering information and sampling weights
- Some nonlinear effects can indeed be modeled

Applications of Growth Mixture Models

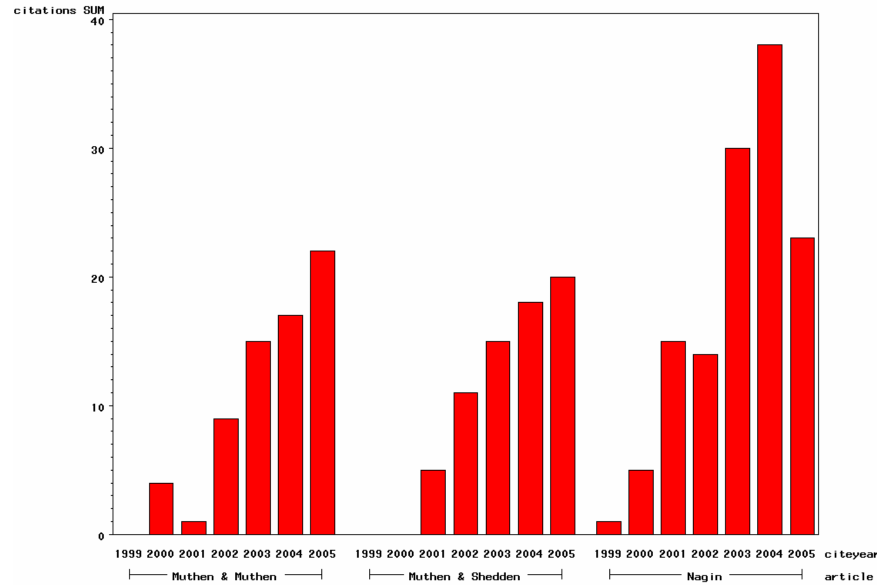
- **Number of applications of growth mixture models is accelerating.**
- **An imperfect index: citation counts for**

Muthén, B. O., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth Mixture Modeling with Latent Trajectory Classes. *Alcoholism: Clinical and Experimental Research, 24*, 882-891.

Muthén, B. & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*, 463-469.

Nagin, D. (1999) Analyzing developmental trajectories: A semi-parametric, group-based approach. *Psychological Methods, 4*, 139-157.

The Increasing Application of Growth Mixture Models



The Prototypical Application

An informal survey of 7 applications published in 2005 citing Muthen & Muthen (2001):

- **Content is aggressive / deviant behavior or substance use**
- **School-based saturation samples and convenience samples**
- **Median consent rate 75%, median attrition rate 25%**
- **Ad hoc measurement of outcome variable:**
  - **Ordinal item, sum or average of ordinal items, logged sum of counts**
- **Number of latent classes estimated by BIC: 2 to 6, mode of 4**
- **Latent classes directly interpreted as types and used to draw policy implications.**

Tenability of Model Assumptions

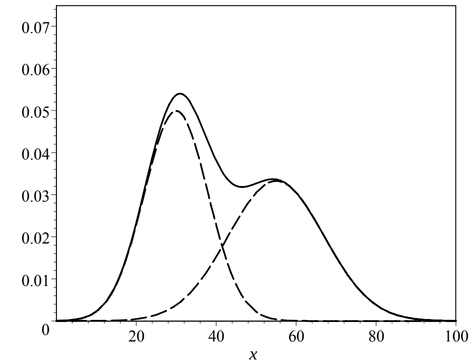
Assumptions of the Model:	**Typical Application:**
Conditional normality	**clear floor effects, typically poor measurement, distributions likely to be skewed in any case.**
Properly specified model	**rarely evaluated for 1-class model**
Linearity of relationships	**never evaluated**
Simple random sample	**never random, nesting within school**
Missing at random	**Non-response, non-random attrition.**

When the Assumptions are Wrong

Assumptions of the Model:	Consequence if Wrong:
Conditional normality	**???**
Properly specified model	**???**
Linearity of relationships	**???**
Simple random sample	**???**
Missing at random	**???**

Conditional Normality

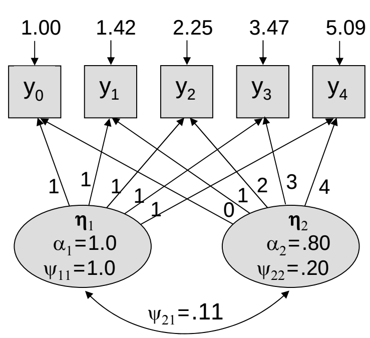
- **A mixture of normals is necessarily non-normal (except in degenerate cases)**

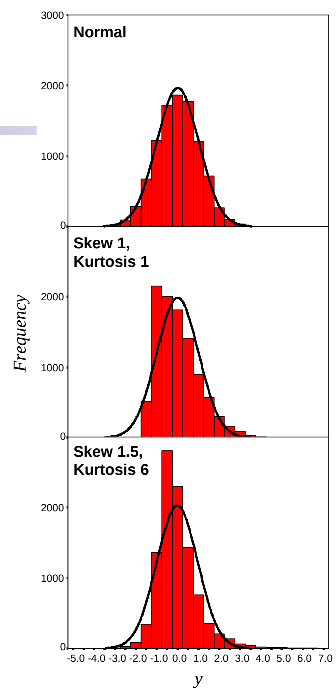


- **A non-normal distribution does not necessarily arise from a mixture.**

Conditional Normality

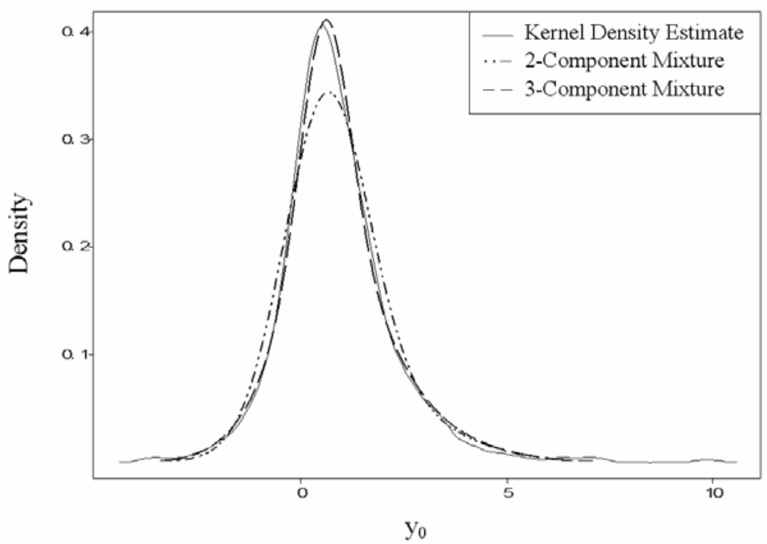
- **500 Datafiles Generated From a Single Population Latent Curve Model (with *N*=200 or *N*=600 each)**
- **Three Distributional Conditions**





Conditional Normality

- **When marginal distributions were normal, minimum BIC occurred with 1 class 100% of the time**
- **When marginal distributions were nonnormal, minimum BIC occured with 2 classes 100% of the time (spurious classes)**
- **Spurious latent classes served to approximate the nonnormal repeated measures via a normal mixture.**



When the Assumptions are Wrong

Assumptions of the Model:	Consequence if Wrong:
Conditional normality	**Spurious latent classes** (Bauer & Curran, 2003)
Properly specified model	**???**
Linearity of relationships	**???**
Simple random sample	**???**
Missing at random	**???**

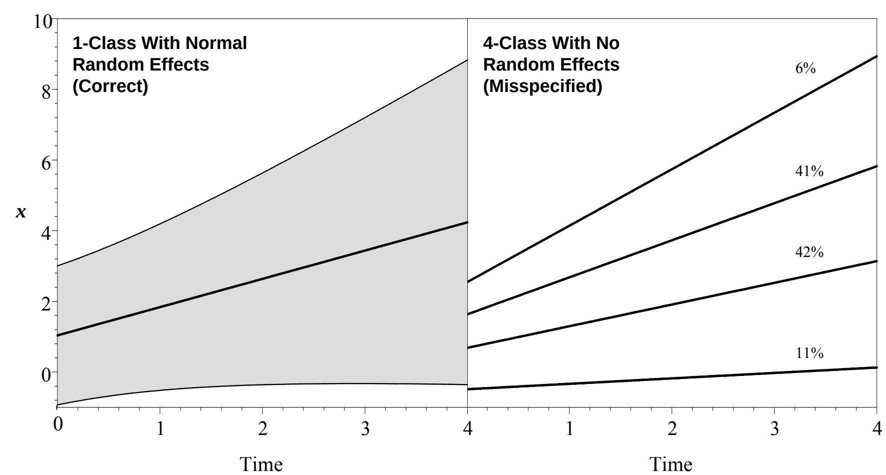
Properly Specified Covariance Structure

- **For an unconditional GMM, the implied aggregate covariance matrix for the repeated measures is:**

$$\boldsymbol{\Sigma}(\boldsymbol{\pi},\boldsymbol{\theta}_1\cdots\boldsymbol{\theta}_K)=\underbrace{\sum_{k=1}^{K}\sum_{j=k+1}^{K}\pi_k\pi_l\left(\boldsymbol{\Lambda}_k\boldsymbol{\alpha}_k-\boldsymbol{\Lambda}_j\boldsymbol{\alpha}_j\right)\left(\boldsymbol{\Lambda}_k\boldsymbol{\alpha}_k-\boldsymbol{\Lambda}_j\boldsymbol{\alpha}_j\right)'}_{\text{between class covariance}}+\underbrace{\sum_{k=1}^{K}\pi_k\left(\boldsymbol{\Lambda}_k\boldsymbol{\Psi}_k\boldsymbol{\Lambda}_k'+\boldsymbol{\Theta}_k\right)}_{\text{within class covariance}}$$

- **Given this, the estimation of spurious latent classes can "compensate" for improper specification of the within-class structure.**

Properly Specified Covariance Structure

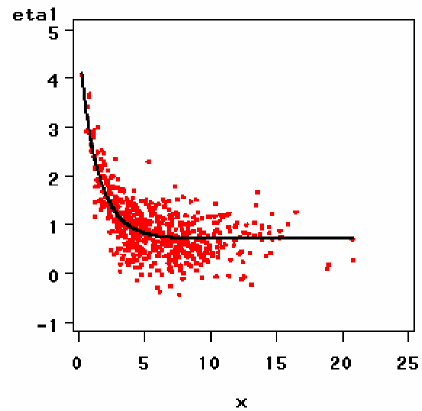


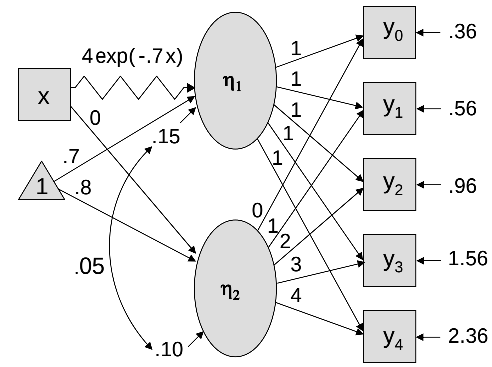
When the Assumptions are Wrong

Assumptions of the Model:	Consequence if Wrong:
Conditional normality	**Spurious latent classes** (Bauer & Curran, 2003)
Properly specified model	**Spurious latent classes** (Bauer & Curran, 2004)
Linearity of relationships	**???**
Simple random sample	**???**
Missing at random	**???**

Linearity

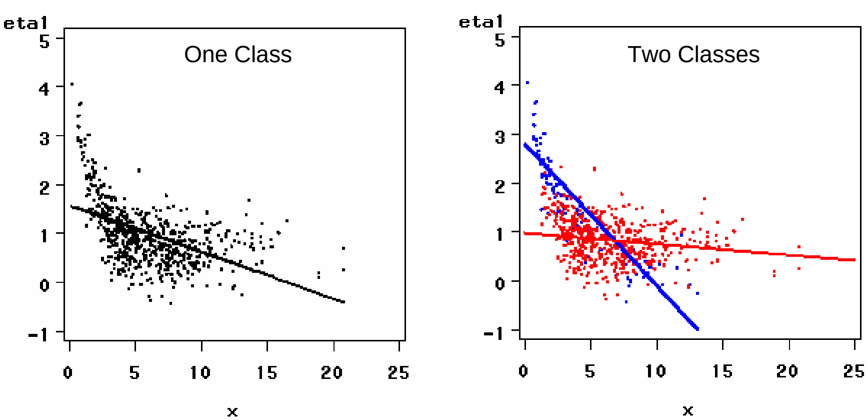
- **500 Datafiles Generated From a Single Population Latent Curve Model (*N*=600 each)**
- **Exogenous variable nonlinearly predicts the intercept**





Linearity

- **Minimum BIC favored 2 classes in 75% of replications**



When the Assumptions are Wrong

Assumptions of the Model:	Consequence if Wrong:
Conditional normality	**Spurious latent classes** (Bauer & Curran, 2003)
Properly specified model	**Spurious latent classes** (Bauer & Curran, 2004)
Linearity of relationships	**Spurious latent classes** (Bauer & Curran, 2004)
Simple random sample	**???**
Missing at random	**???**

Simple Random Sample

- **Most empirical applications include some nesting not taken account of either through fixed grouping variables or random effects.**
- **For a small number of groups, group differences in change over time may emerge as latent classes (i.e., omitted known grouping variable compensated for by a latent grouping variable).**
- **For a larger number of groups, latent classes may discretely approximate a continuous distribution of random effects.**

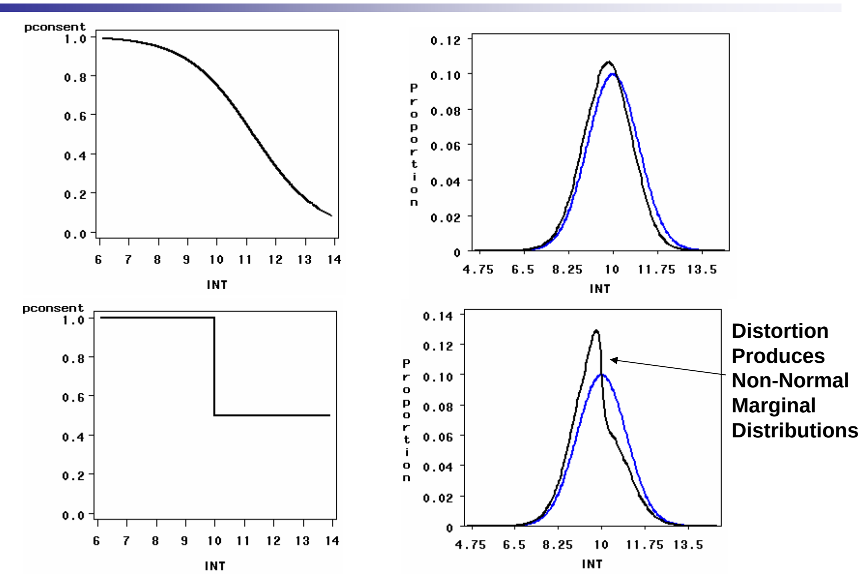
When the Assumptions are Wrong

Assumptions of the Model:	Consequence if Wrong:
Conditional normality	**Spurious latent classes** (Bauer & Curran, 2003)
Properly specified model	**Spurious latent classes** (Bauer & Curran, 2004)
Linearity of relationships	**Spurious latent classes** (Bauer & Curran, 2004)
Simple random sample	**Spurious latent classes** (Wedel, Hofstede, Steenkamp, 1998)
Missing at random	**???**

Missing at Random

- **Most studies have lower than optimal consent rates and some attrition.**
- **Those not participating or dropping out may reside in particular regions of the population distribution (e.g., the "worst" cases in the upper tail).**
- **The observed distributions will then be distorted.**
- **GMMs fit to these observed distribution may not recover true latent class structure.**

Missing at Random: Non-Response



When the Assumptions are Wrong

Assumptions of the Model:	Consequence if Wrong:
Conditional normality	**Spurious latent classes** (Bauer & Curran, 2003)
Properly specified model	**Spurious latent classes** (Bauer & Curran, 2004)
Linearity of relationships	**Spurious latent classes** (Bauer & Curran, 2004)
Simple random sample	**Spurious latent classes** (Wedel, Hofstede, Steenkamp, 1998)
Missing at random	**Possibly spurious latent classes**

The Methodological Challenge

- **Aside from true population subgroups, latent classes can represent:**
  - **Non-normality**
  - **Misspecification of within-class model**
  - **Nonlinear effects**
  - **Complex sample**
  - **Non-response / Non-random attrition**
- ***Typically more than one of these will be present*.**

Improving Methods

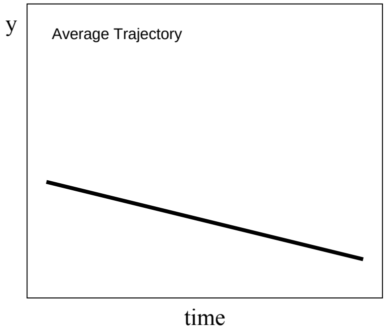
- **Most of these issues cannot be fixed by changing the statistical model.**
- **What is needed:**
  - **Better measurement (interval level)**
  - **Diagnostics for checking conditional normality, detecting misspecification of the covariance structure, and visualizing potentially nonlinear relationships.**
  - **More rigorous sampling procedures**
  - ***Sound methodology***
  - ***More careful interpretation***

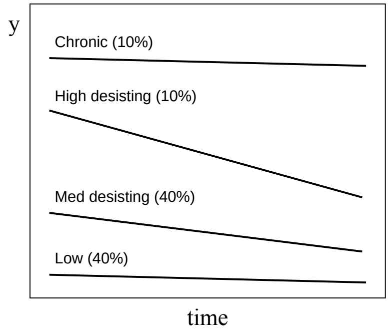
Theoretical Concerns

- **Even if methodology can be improved, theoretical reasons for skepticism remain.**

The “Selling” of Growth Mixture Models

- **Applied researchers are convinced that LCMs give them one trajectory while growth mixture models give them multiple trajectories.**



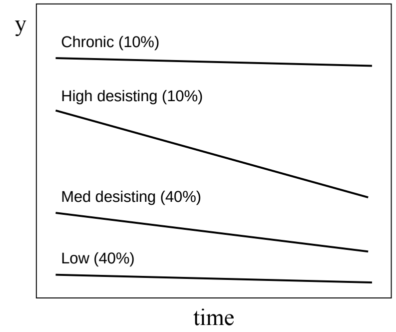


Modeling Heterogeneity with the GMM

A typical study identifies and then predicts the latent classes.

- **Worst case: assignment by modal probability then prediction.**
- **Best case: Prediction done in the model itself.**

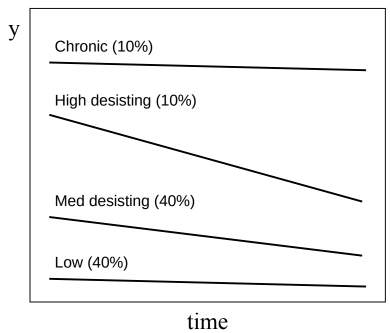
$$\pi_{ik} = \frac{\exp(\alpha_{c_k} + \gamma'_{c_k}\mathbf{x}_{ik})}{\sum_{k=1}^{K}\exp(\alpha_{c_k} + \gamma'_{c_k}\mathbf{x}_{ik})}$$



Modeling Heterogeneity with the GMM

- **If we predict only which of the four trajectories an individual belongs to, we limit ourselves to this taxonomy of four.**

- **But do we really believe these four trajectories represent a definitive taxonomy of heterogeneity in change over time?**



Returning to the Latent Curve Model

- **The conditional LCM can capture more heterogeneity in patterns of change.**

<u>Conditional Model</u>

Level 1: $y_{ti} = \eta_{1i} + \eta_{2i}\lambda_{ti} + \varepsilon_{ti}$

Level 2: $\eta_{1i} = \alpha_1 + \gamma_{11}x_{1i} + \gamma_{12}x_{2i} + \gamma_{13}x_{1i}x_{2i} + \zeta_{1i}$

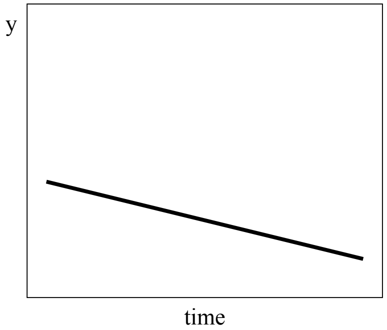
$\eta_{2i} = \alpha_2 + \gamma_{21}x_{1i} + \gamma_{22}x_{2i} + \gamma_{23}x_{1i}x_{2i} + \zeta_{2i}$

- **If these two predictors are dichotomous, we obtain four trajectories, if continuous, we obtain an infinite number of trajectories**

Unconditional Model

$$y_{ti} = \eta_{1i} + \eta_{2i}\lambda_{ti} + \varepsilon_{ti}$$

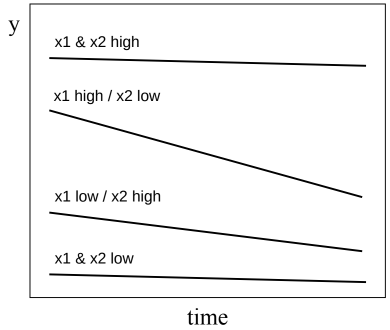
$$\eta_{1i} = \alpha_1 + \zeta_{1i}$$
$$\eta_{2i} = \alpha_2 + \zeta_{2i}$$



Conditional Model

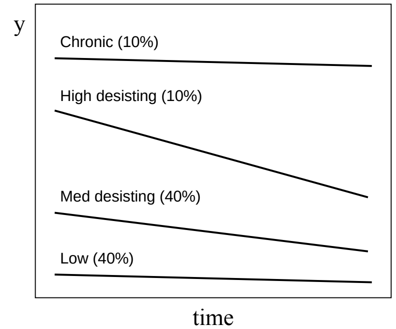
$$y_{ti} = \eta_{1i} + \eta_{2i}\lambda_{ti} + \varepsilon_{ti}$$

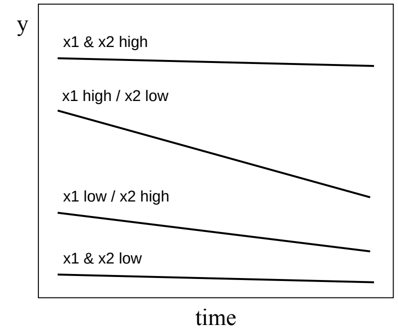
$$\eta_{1i} = \alpha_1 + \gamma_{11}x_{1i} + \gamma_{12}x_{2i} + \gamma_{13}x_{1i}x_{2i} + \zeta_{1i}$$
$$\eta_{2i} = \alpha_2 + \gamma_{21}x_{1i} + \gamma_{22}x_{2i} + \gamma_{23}x_{1i}x_{2i} + \zeta_{2i}$$



Comparison

- The GMM gives us four discrete trajectory types to predict.
- The conditional LCM gives us a potentially infinite family of model-implied trajectories, of which four are plotted.





Another option

- One other possibility in the GMM is to predict both class membership and random variability within classes.
- However this partitions the effects of the predictors into within- and between-class portions, making interpretation difficult.
- In practice this is rarely done.
- When done, the results are usually interpreted poorly.

Summary of Challenges

- The chief methodological challenge to the application of GMMs is that one does not know what the latent classes represent.
- The chief theoretical challenge is that it isn't clear that a finite set of trajectory classes well ever sufficiently capture heterogeneity in change in the population.

One must then ask, under what circumstances are growth mixture models scientifically useful?

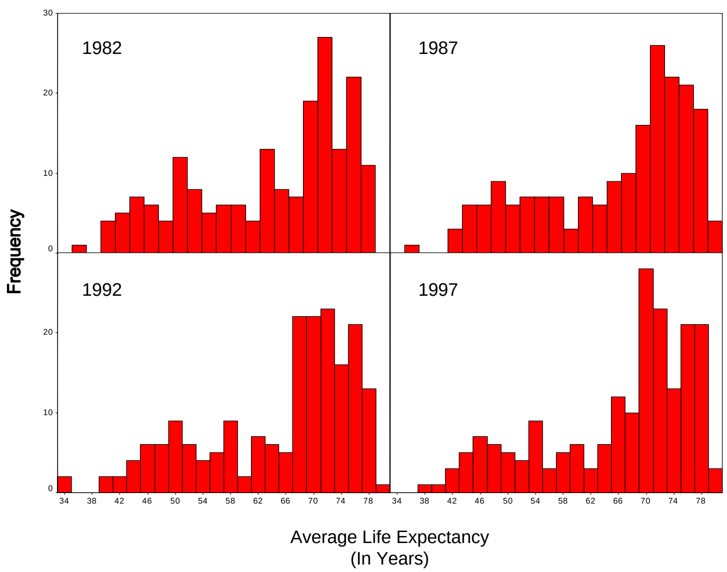
What a Scientifically Useful Application Might Look Like

- Let's consider the application of these models to a radically different kind of data that does not share the limitations of many psychological data sets.
  - World Bank data on average life expectancy within countries.
  - Assessed in 1982, 1987, 1992, and 1997.
  - Initial hypothesis: two trajectory classes will emerge representing developed and developing nations, respectively.

Tenability of Model Assumptions

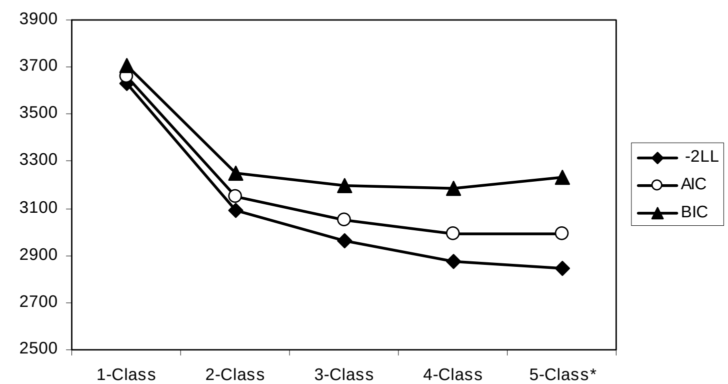
Assumptions of the Model:	Life Expectancy Application:
Conditional normality	Dependent variable is measured on a ratio scale, no apparent floor or ceiling effects
Properly specified model	Will begin with an unrestricted finite mixture
Linearity of relationships	Can use model diagnostics if this appears risky
Simple random sample	The biggest challenge: sample units likely spatially correlated, sample IS population
Missing at random	No, but too little to matter: 198 of 212 countries provided data (93% "consent" rate), 5% or less missing each year.

A Look at the Data

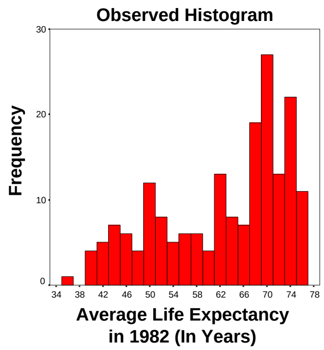


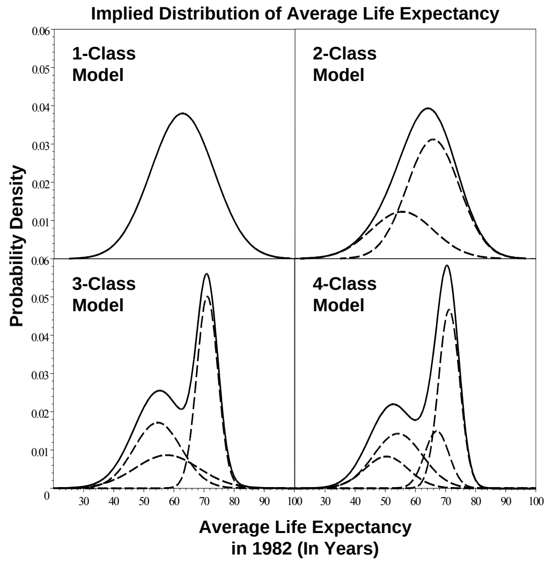
How Many Latent Classes?

- Comparative fit of unrestricted multivariate normal mixture models with 1 to 5 classes:



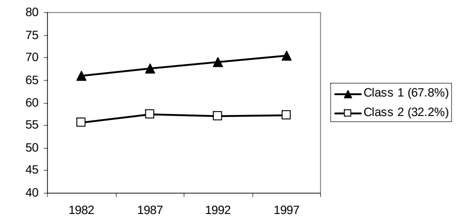
How well do the models reproduce the data?



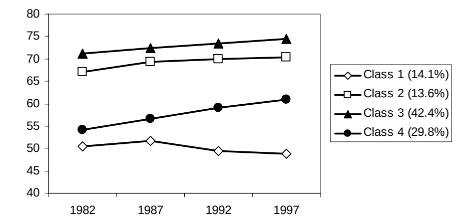


Choosing among models

- We expected to find two classes and indeed saw the greatest increase in model fit with the addition of the second class.



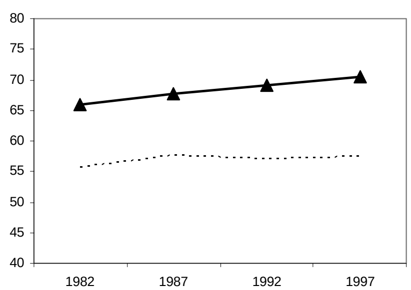
- The minimum BIC was obtained with 4 classes and this model subdivides high and low trajectories in a compelling way



Who's clustering where?

2 Class Model:

- **Class 1: France, Sweden, USA, Lebanon, Nepal, Philippines…**



Who's clustering where?

2 Class Model:

- Class 1: France, Sweden, USA, Lebanon, Nepal, Philippines…
- **Class 2: Iraq, Kenya, Estonia, Ireland, Haiti, Ethiopia…**

Who's clustering where?

2 Class Model:

- Class 1: France, Sweden, USA, Lebanon, Nepal, Philippines…
- Class 2: Iraq, Kenya, Estonia, Ireland, Haiti, Ethiopia…

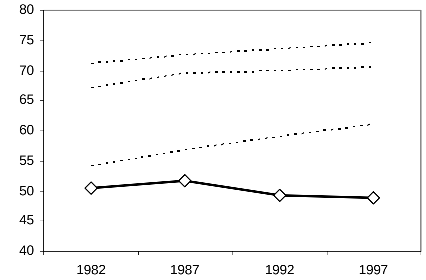
This classification is absurd

Ordinarily, however, we would have no way of knowing this

Who's clustering where?

4 Class Model:

- **Class 1: Botswana, Burundi, Congo, Zambia, Kenya, Zimbabwe, Lesotho, Liberia…**



Who's clustering where?

4 Class Model:

- Class 1: Botswana, Burundi, Congo, Zambia, Kenya, Zimbabwe, Lesotho, Liberia…
- **Class 2: Armenia, Belarus, Estonia, Lithuania, Latvia, Romania, Saudi Arabia, Portugal..**

Who's clustering where?

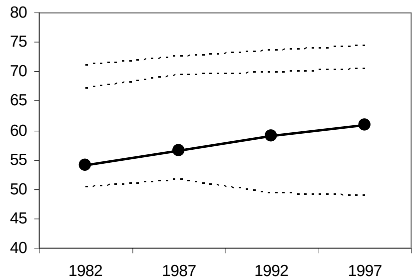
4 Class Model:

- Class 1: Botswana, Burundi, Congo, Zambia, Kenya, Zimbabwe, Lesotho, Liberia…
- Class 2: Armenia, Belarus, Estonia, Lithuania, Latvia, Romania, Saudi Arabia, Portugal…
- **Class 3: France, Sweden, Iceland, Japan, Singapore, United States…**

Who's clustering where?

4 Class Model:

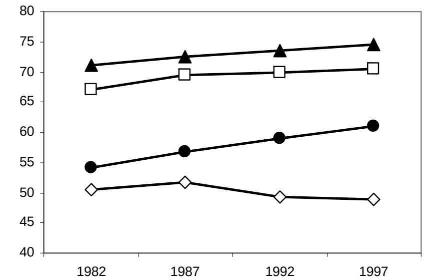
- Class 1: Botswana, Burundi, Congo, Zambia, Kenya, Zimbabwe, Lesotho, Liberia…
- Class 2: Armenia, Belarus, Estonia, Lithuania, Latvia, Romania, Saudi Arabia, Portugal…
- Class 3: France, Sweden, Iceland, Japan, Singapore, United States…
- **Class 4: Antigua, Panama, Argentina, Taiwan, Fiji, Paraguay, Kuwait, Czech Republic…**



Who's clustering where?

4 Class Model:

- Class 1: Botswana, Burundi, Congo, Zambia, Kenya, Zimbabwe, Lesotho, Liberia…
- Class 2: Armenia, Belarus, Estonia, Lithuania, Latvia, Romania, Saudi Arabia, Portugal…
- Class 3: France, Sweden, Iceland, Japan, Singapore, United States…
- Class 4: Antigua, Panama, Argentina, Taiwan, Fiji, Paraguay, Kuwait, Czech Republic…



This classification is more sensible

How was this exercise scientifically useful?

- **The exploratory mixture analysis showed that my naïve initial hypothesis was too simplistic.**
- **The class of countries experiencing decreases in life expectancy after 1987 shows a qualitative difference from the other class trajectories.**
- **Confidence in these results is bolstered by knowing what the latent classes are *not* representing (e.g., assumption violations).**
- **Nevertheless, an expert might say these results are still overly simplistic (e.g., 4 "types" of countries…)**

What about the old way?

- **With better theory, we can pursue confirmatory analyses.**
- **For this data, we might specify a conditional LCM:**
  - **Use status as <u>developed nations</u>, <u>transitional economies</u>, and <u>sub-Saharan</u> as time-invariant predictors.**
  - **Include <u>GDP</u>, <u>conflict</u>, and <u>HIV prevalence</u> as a time-varying predictors, possibly with lagged effects.**
- **Although a less complex statistical model, the LCM would likely capture and explain more heterogeneity in patterns of change than the GMM.**

Conclusions

- The statistical theory behind GMMs is incongruent with most applications.
  - Methodological problems range from measurement to sampling to model specification
  - At a theoretical level, GMMs lack verisimilitude.
- When methodological problems are minimized, GMMs can reveal unanticipated heterogeneity in patterns of change.
- Hypothesized heterogeneity in patterns of change is likely better evaluated with conditional LCMs.
- Indirect applications of mixtures also hold much promise.