Longitudinal latent variable modeling

Michael Hallquist
Penn State University

DEPENd
Developmental Personality Neuroscience Laboratory

Thinking about age (and time)

- The good news is that time marches forward in a principled fashion. We can quantify time elapsed, and time it is a ratio-distributed variable.

- The difficulty is resolving *how* brain and behavior change with development, and *what form* the age-related change takes.

- Much of the time, age is a proxy for unknown latent causes. We are not interested in age per se, but in the causal *processes* underlying development.

- Example: observing a peak in ventral striatum activity in adolescence invites the question, *why* does this occur?

Thinking about age

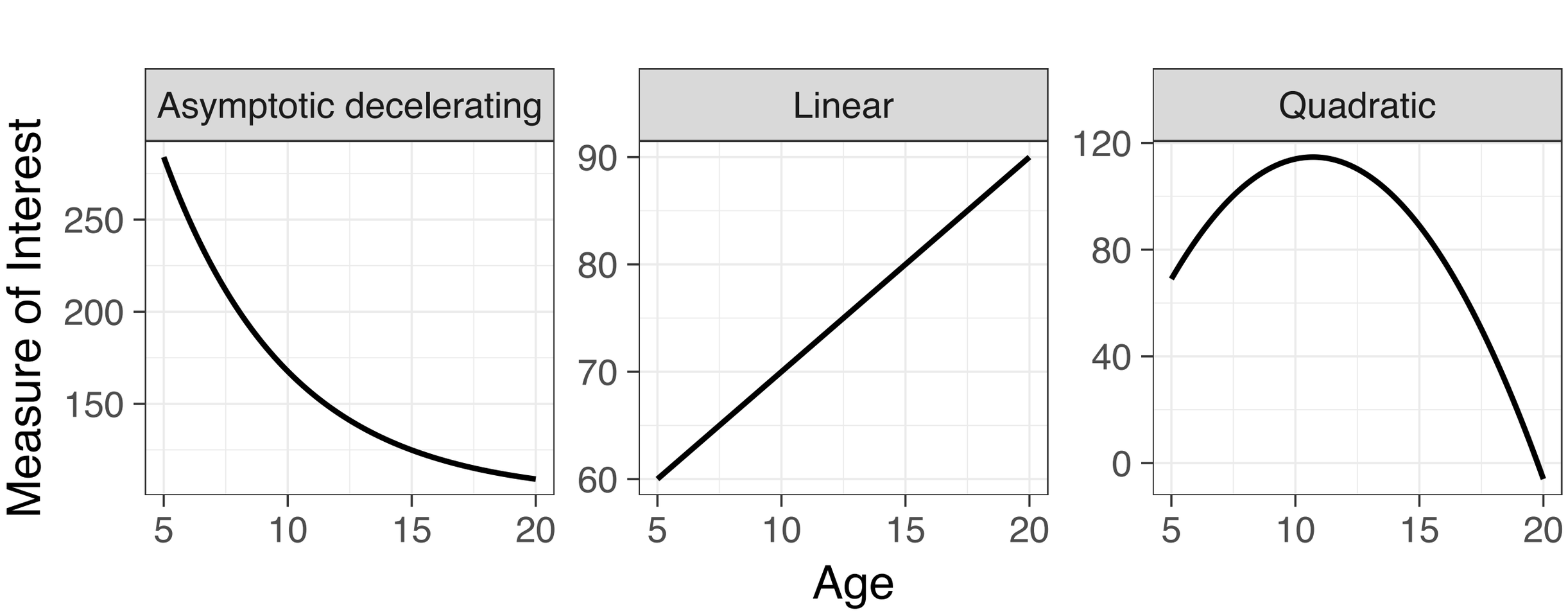
- Thus, make time to think about the *causal processes* that drive a developmental phenomenon. Don't shy away from developing an account, even if it is tentative and hypothetical.

- If your thinking is clear, the analytic strategy will follow. There are established statistical models of change for virtually any question.

How to model age?

- Ensure that your model of age captures your hypothesis about the *functional form* of change.

A few examples



Types of change

- Structural: do the variables reflect the same constructs over time? (aka measurement invariance)

- Rank-order: how similar are the relative rankings among individuals over time (aka differential)

- Normative: does the average level of a variable change? (aka mean-level)

- Individual: person-level variability around the sample mean trajectory parameters

- Ipsative: profile stability on multiple variables over time

Latent variable modeling

aka Structural Equation Modeling (SEM)

- An approach for testing and comparing alternative models of the latent structure underlying your data

- Built on the foundation of regression

- Multivariate: related equations

- Composed of *measurement* and *structural* models

- Distinguishes between *latent* and *observed* variables

- Seeks to represent the *mean* and *covariance* structure of your data.

Strengths of longitudinal SEM

- The form of change can be controlled using a flexible set of parameters ranging from simple (linear) to complex (nonparametric).

- Categorical (class) and continuous (factor) latent variables can be combined into a single model (*caveat emptor!*)

- Ability to constrain individual parameters to enforce specific hypotheses about the form of developmental change (e.g., AR1).

- Testing changes in multiple processes

- Fit statistics provide feedback about model quality

Disadvantages of SEM

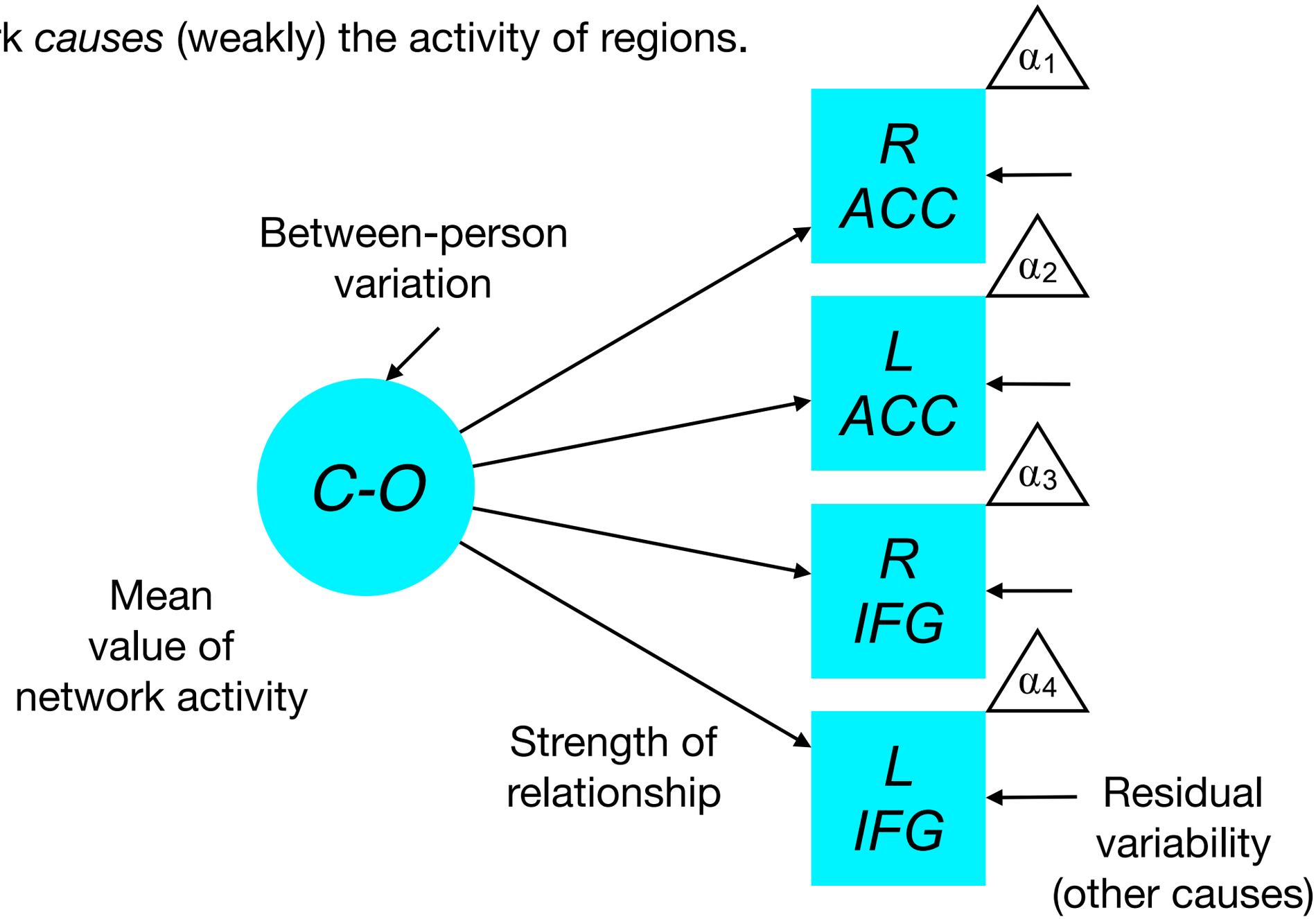
- A ‘large sample’ technique typically requiring 100+, if not 100s of participants.
  - More observations per person, better power (Muthén & Curran, 1997)
- Conventionally, latent curve modeling (LCM) assumes similar/identical spacing of observations.
  - Can overcome this. Mplus TSCORES method
- The flexibility of SEM is also its curse. You can test many models, but you need to understand what each represents, and sources of misfit and bias in your results.

Common Factor Model

Reflective latent model:

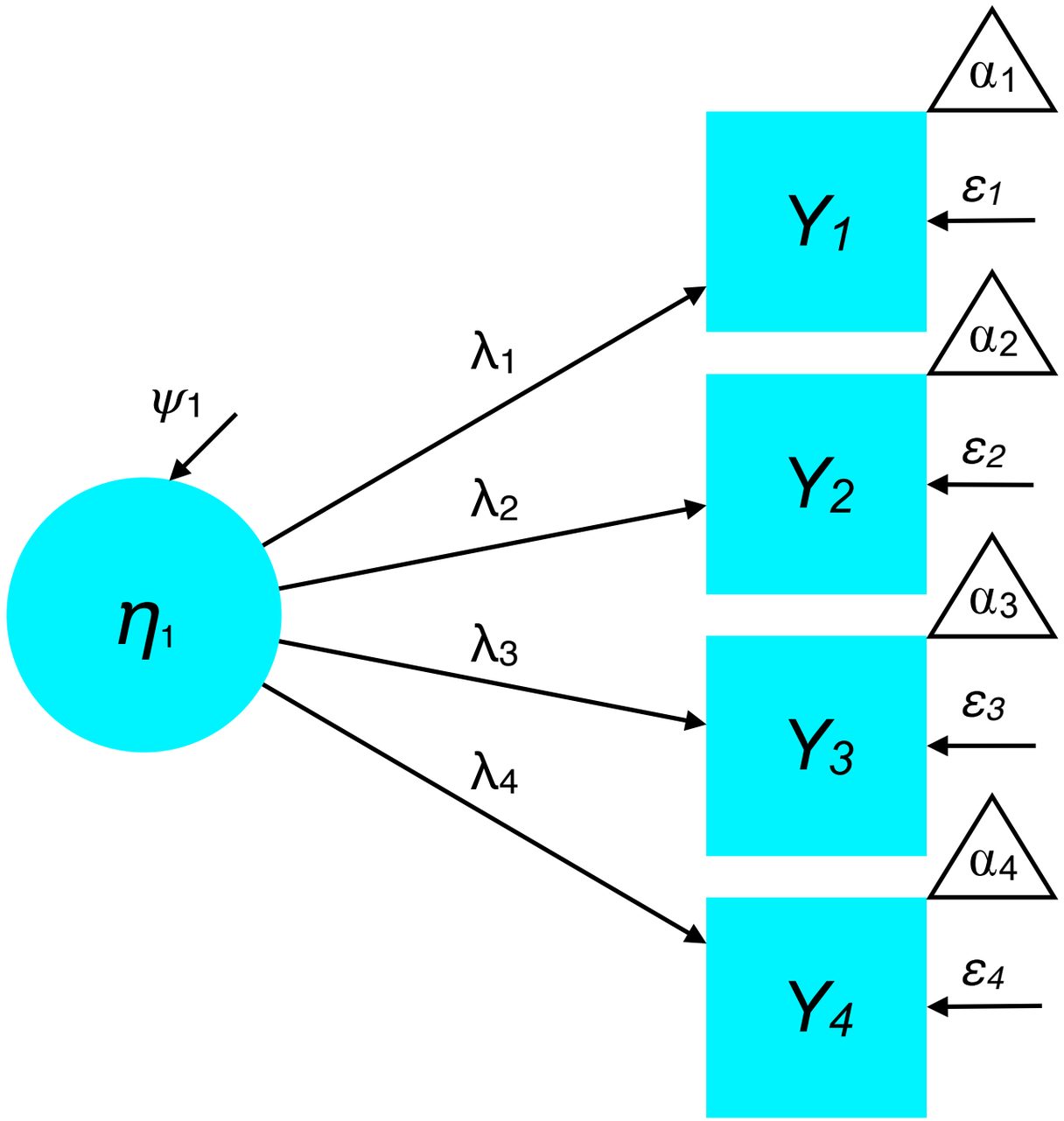
The observed data (regions) *reflect* the influence of the latent variable.

Here, the network *causes* (weakly) the activity of regions.



Common Factor Model

$$y_1 = \alpha_1 + \lambda_1\eta_1 + \varepsilon_1$$
$$y_2 = \alpha_2 + \lambda_2\eta_1 + \varepsilon_2$$
$$y_3 = \alpha_3 + \lambda_3\eta_1 + \varepsilon_3$$
$$y_4 = \alpha_4 + \lambda_4\eta_1 + \varepsilon_4$$



A book worth reading…

If you want to learn longitudinal SEM

Latent Curve Models

A Structural Equation Perspective

KENNETH A. BOLLEN

University of North Carolina
Department of Sociology
Chapel Hill, North Carolina

PATRICK J. CURRAN

University of North Carolina
Department of Psychology
Chapel Hill, North Carolina

Ergodicity

- Most of our analyses focus on variability *between subjects*. For example, how much does activity in a region change, on average, with development?

- We are interested in making inferences about the *population* from which the data are *sampled*.

- A strong assumption of this approach is that there is homogeneity within the population.

- Indeed, the tenability of this assumption has motivated techniques for estimating subpopulation (e.g., mixture models).

Adapted from Peter Molenaar

Ergodicity

- Implicitly, we hope that between-person results generalize to individuals (otherwise, why do the science?!)
- But: is this plausible? (Molenaar, 2004). This is a question of *ergodicity*.
- Definition: A process is *ergodic* if results from analyses of inter-individual variation validly generalize to the level of intra-individual change in time, and vice versa.
- As pertains to modeling longitudinal change, does the *form* of functional change modeled at the between-person level generalize to the form of change at the person level?

Adapted from Peter Molenaar

Ergodicity

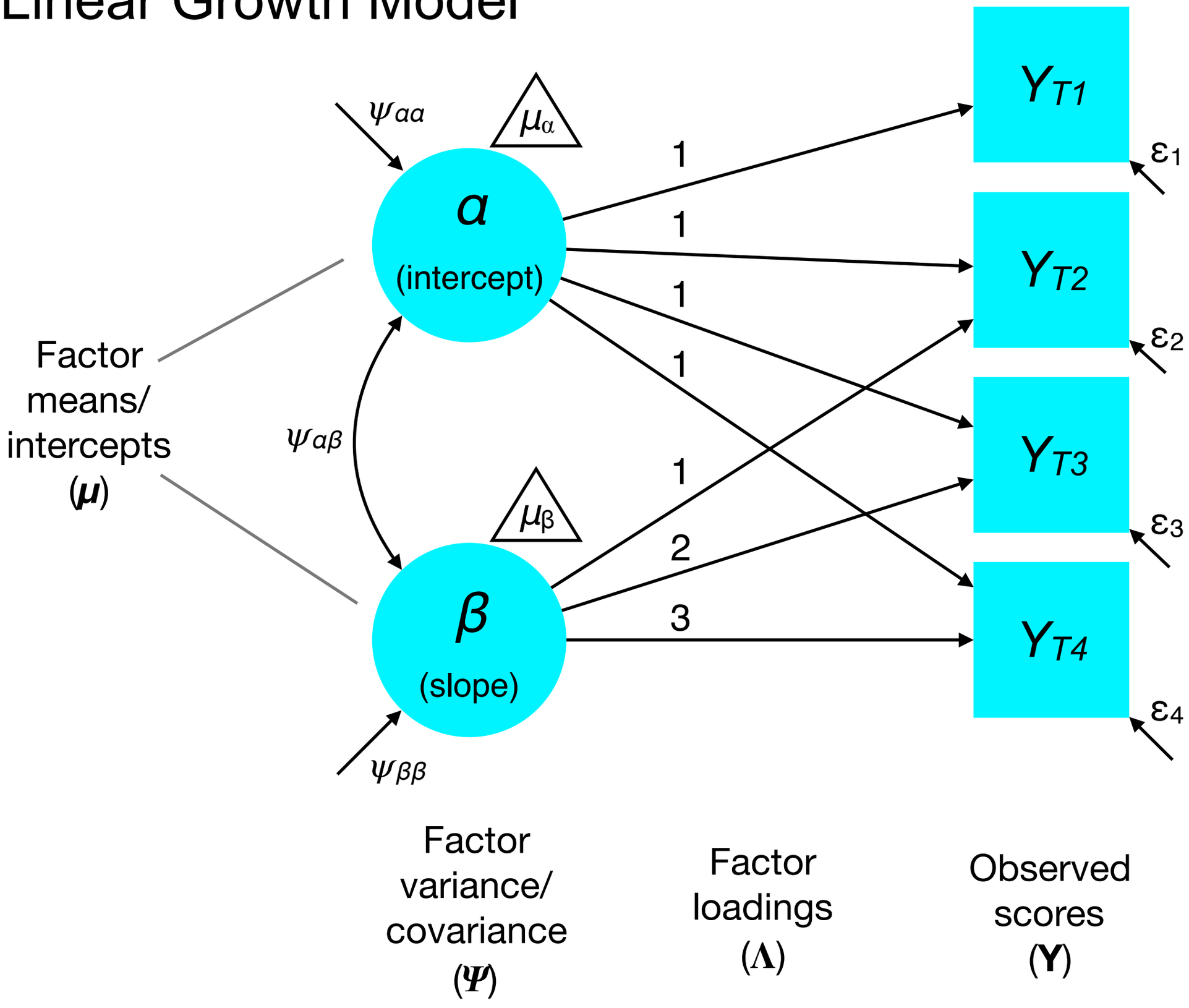
- One implication of ergodicity is that fitting a disparate function of age/time at the *group level* (i.e., fixed effects), but a simpler form for individual differences from the group (i.e., random effects), may constrain your conclusions.

- This is especially true for data where starting age (e.g., all 10-year-olds) and assessment schedules are similar.

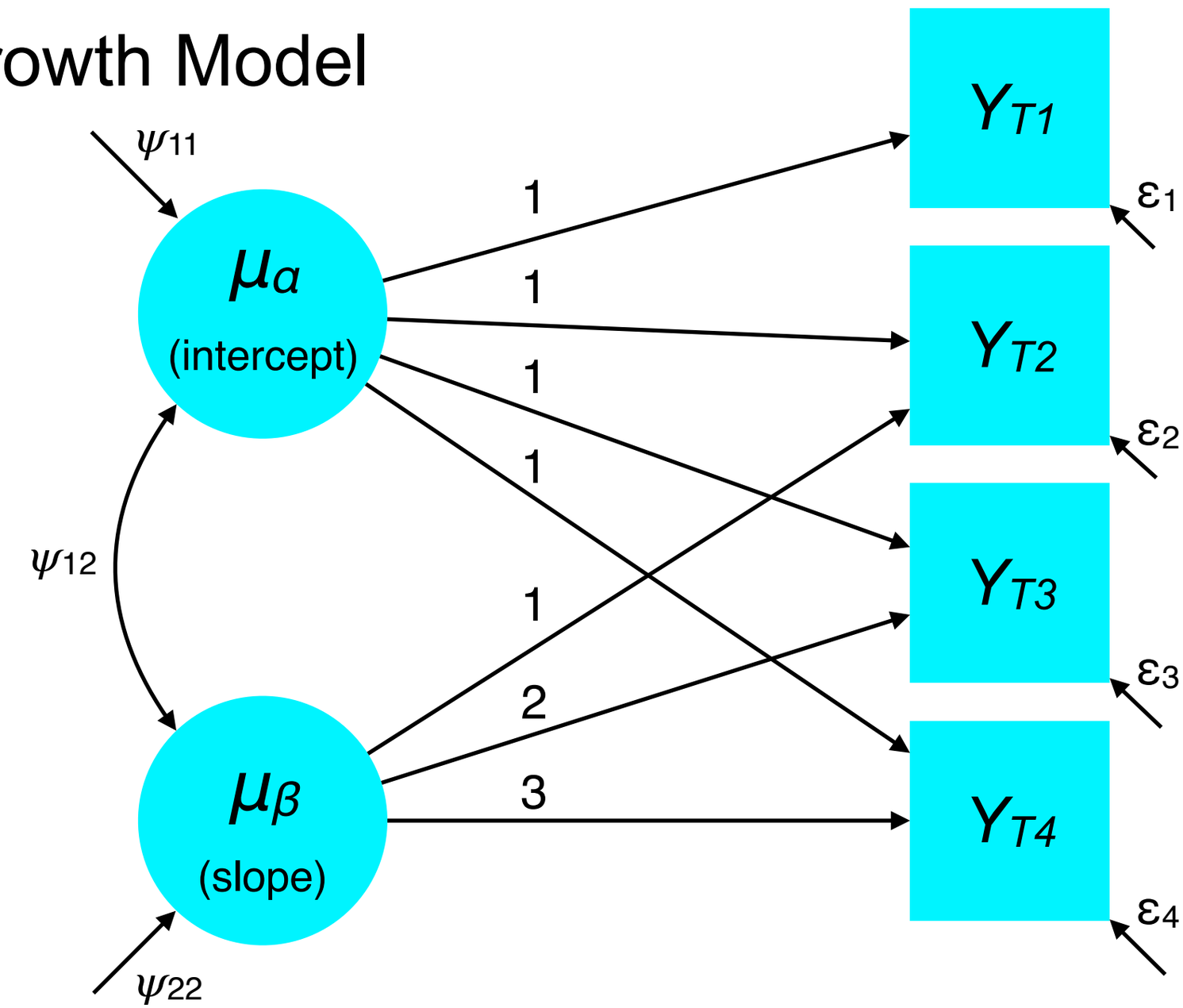
- But, if we sample individuals longitudinally, but at different starting ages (or assessment schedules), a complex model (e.g., cubic) at the group level may be built from simpler models (e.g., linear).

  - But we still assume ergodicity in this case…

Linear Growth Model



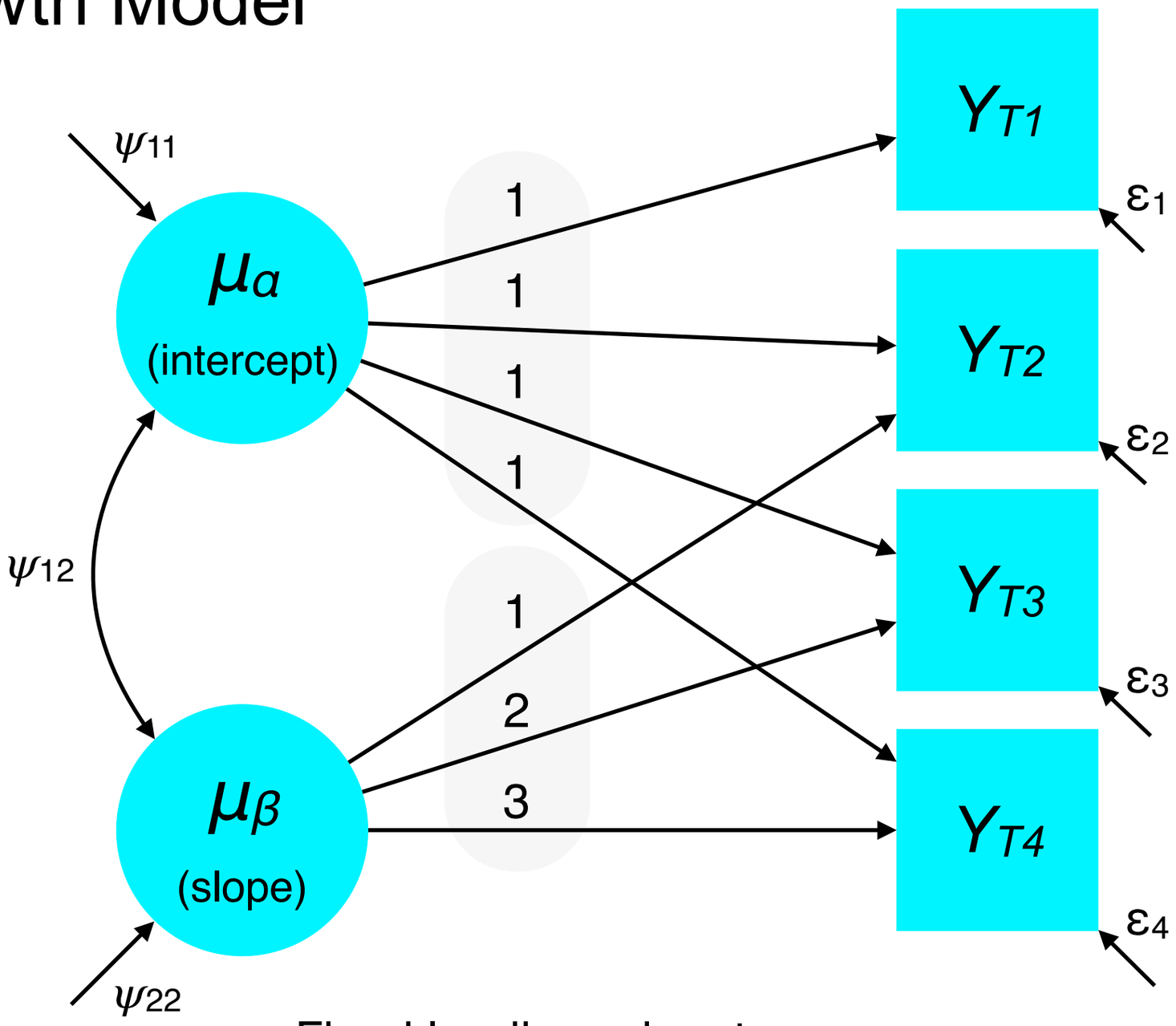
Linear Growth Model



$$y_{i,T1} = \alpha_i + \beta_i \lambda_{T1} + \varepsilon_{i,T1}$$
$$y_{i,T2} = \alpha_i + \beta_i \lambda_{T2} + \varepsilon_{i,T2}$$
$$y_{i,T3} = \alpha_i + \beta_i \lambda_{T3} + \varepsilon_{i,T3}$$
$$y_{i,T4} = \alpha_i + \beta_i \lambda_{T4} + \varepsilon_{i,T4}$$

$$\alpha_i = \mu_\alpha + \zeta_{\alpha_i}$$
$$\beta_i = \mu_\beta + \zeta_{\beta_i}$$
$$Var(\zeta_{\boldsymbol{\alpha}}) = \psi_{11}$$
$$Var(\zeta_{\boldsymbol{\beta}}) = \psi_{22}$$

Linear Growth Model



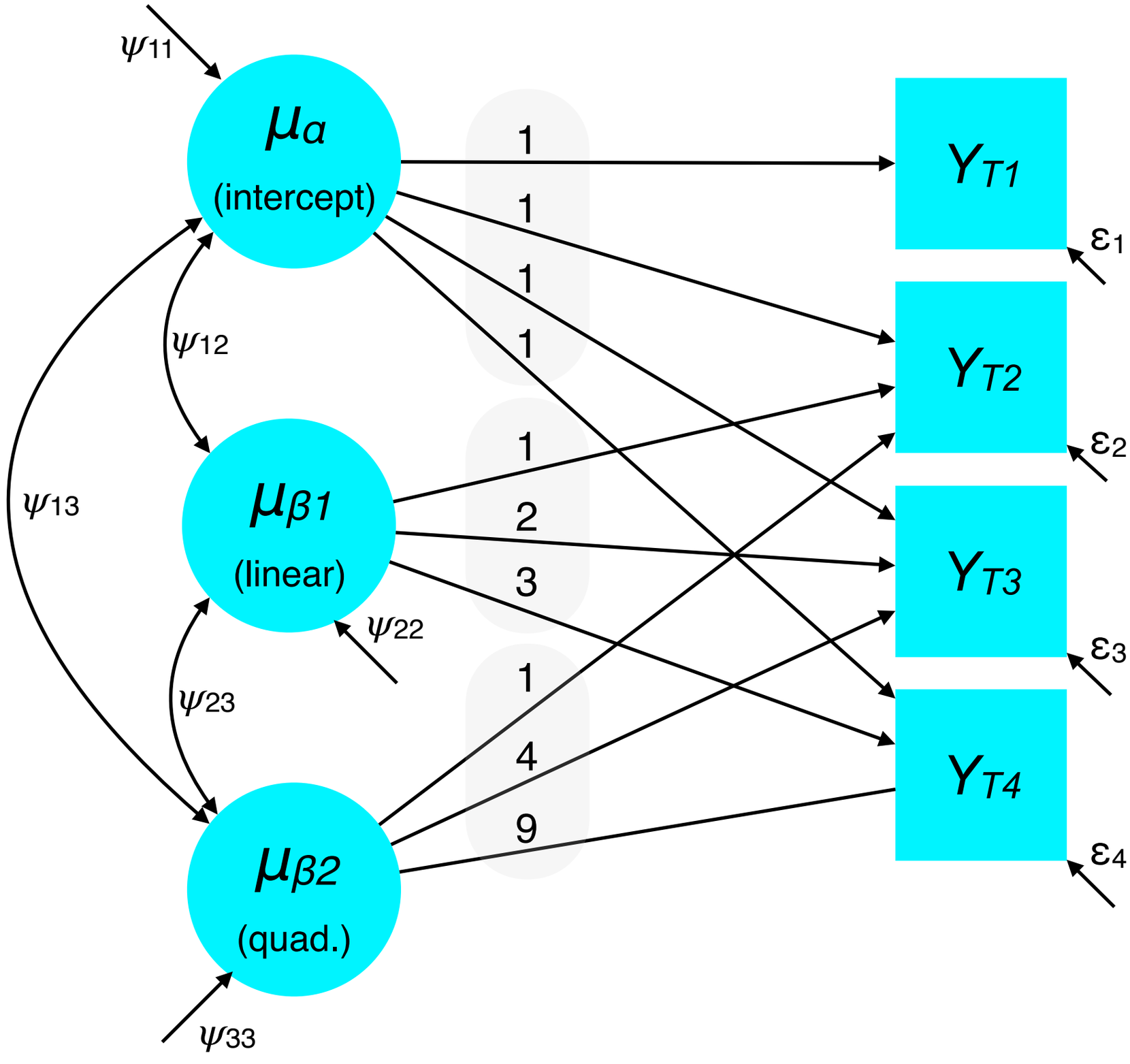
Fixed loadings denote:

a) functional form of change

and

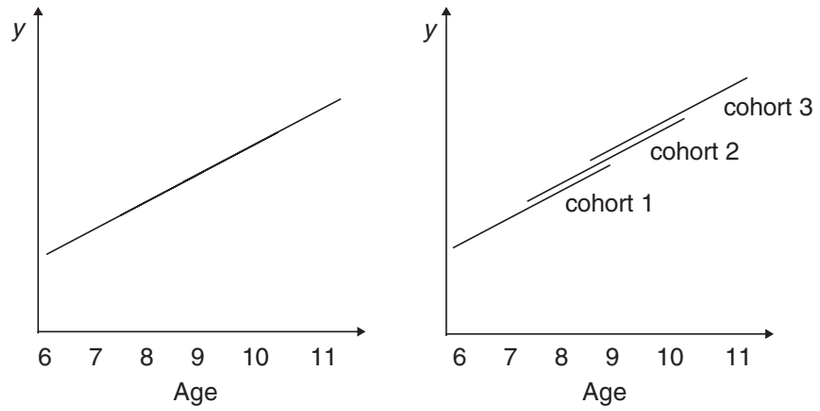
b) the spacing between observations

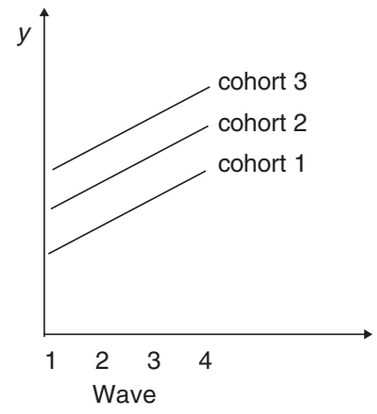
Quadratic Growth Model



Organizing data by age or wave

- Conventional latent curve models usually organize the data by wave (time 1, time 2, time 3).
- This is generally a good idea, but breaks down if a) assessment schedules are substantially different, or b) individuals at each wave are substantially different due to an *unmodeled* time-related process.



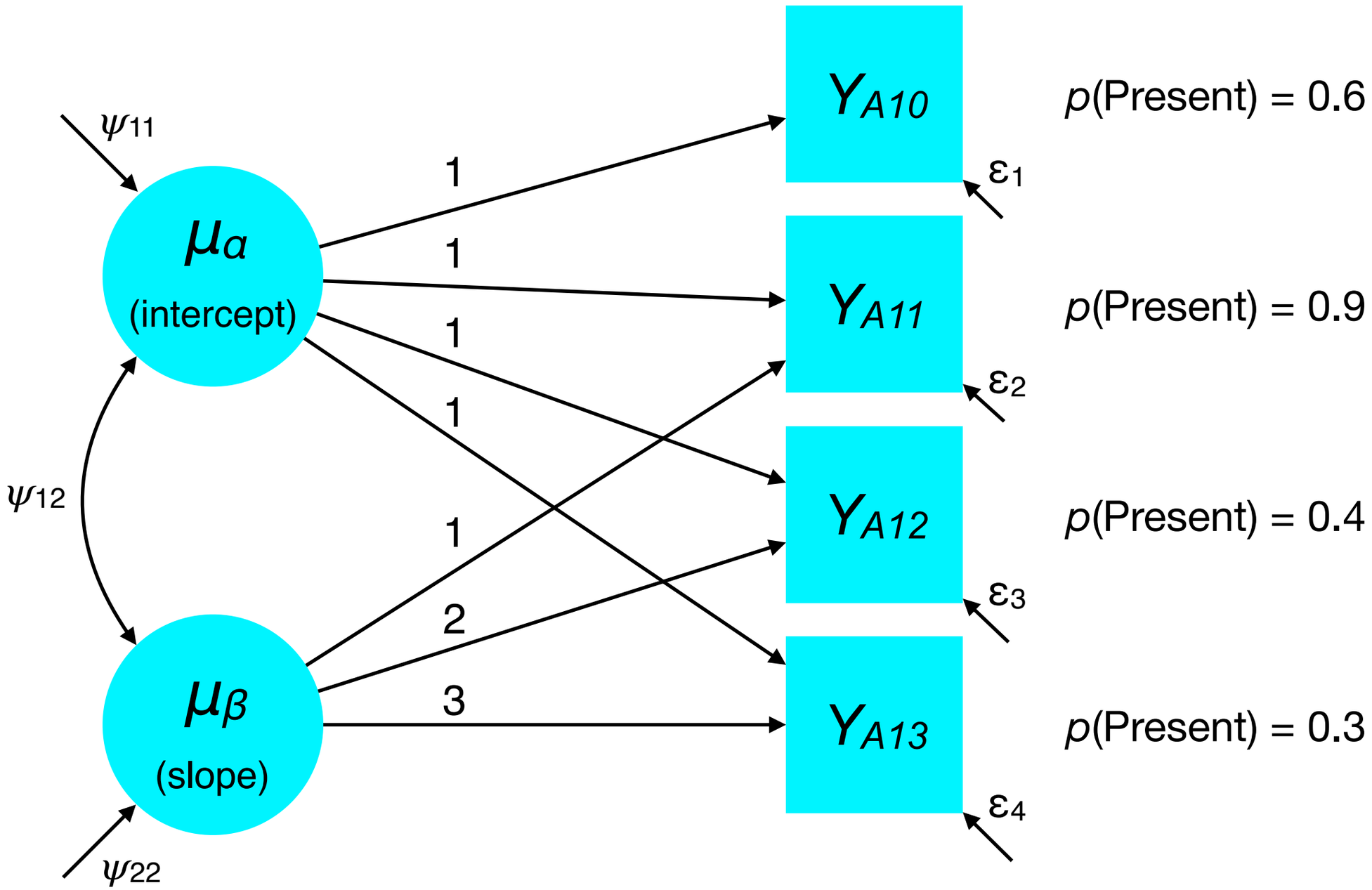


Bollen and Curran 2006, Fig. 3.2

Organizing data by age or wave

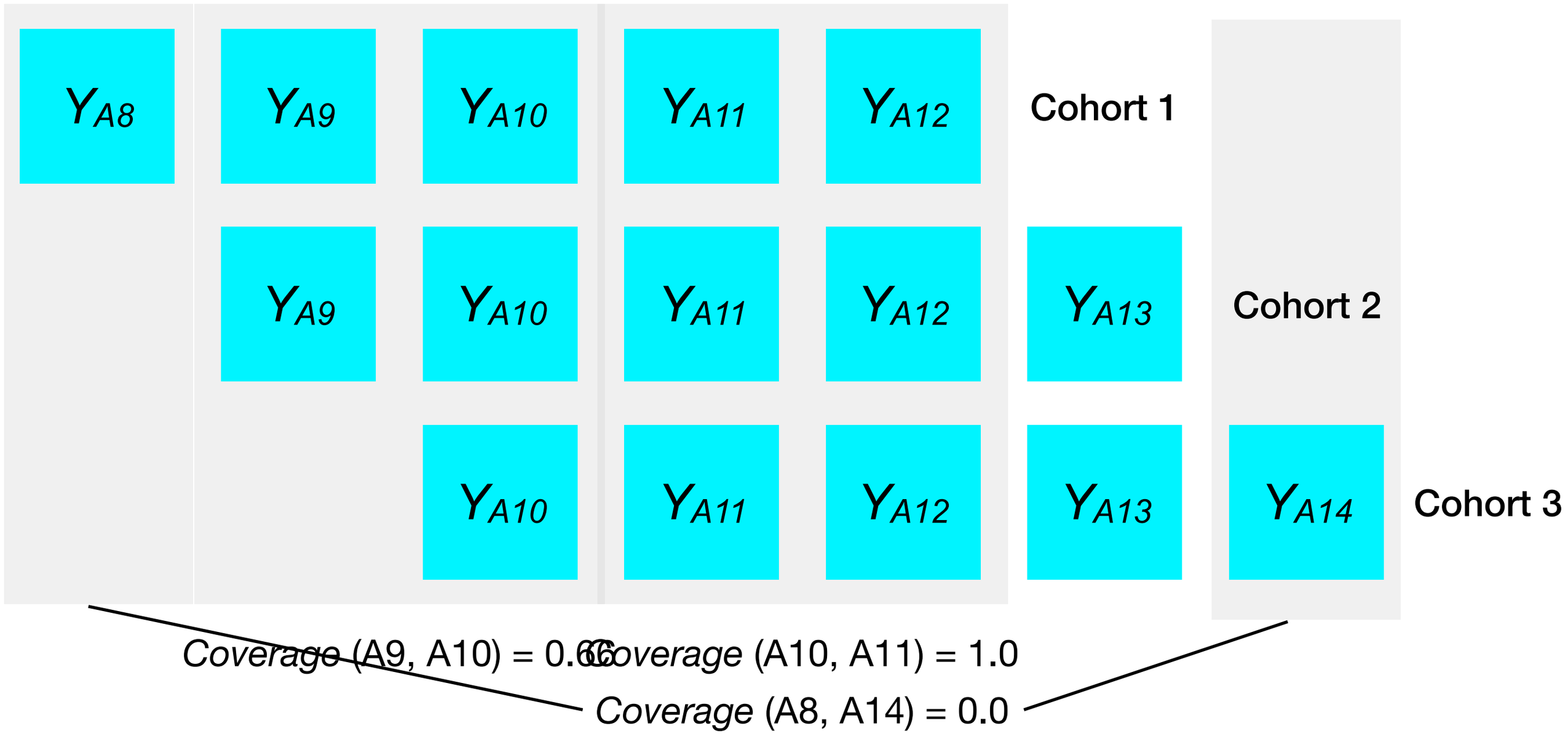
- tl;dr: if ages are quite different at each wave (e.g., accelerated design), one should probably organize the data by age.
    - This leads to substantial missingness at each age (e.g., if only some people were sampled at age 10). But this is not a problem in principle if the data are *missing by design.*
    - Full-information maximum likelihood estimation (FIML) will handle this case well, though it can strain convergence in estimation.

Linear Growth Model organized by age



Accelerated Longitudinal Cohort Design

aka Cohort-Sequential Design



(Assuming Equal n per cohort)

Why is organizing by age okay?

- Missing by design (e.g., sequential-cohort or ad hoc accelerated design) is MCAR by definition.
  - Thus, FIML is an acceptable estimation method

- This is based on the fact that the residual covariance between indicators is *not* a feature of the model — i.e., we assume conditional independence.

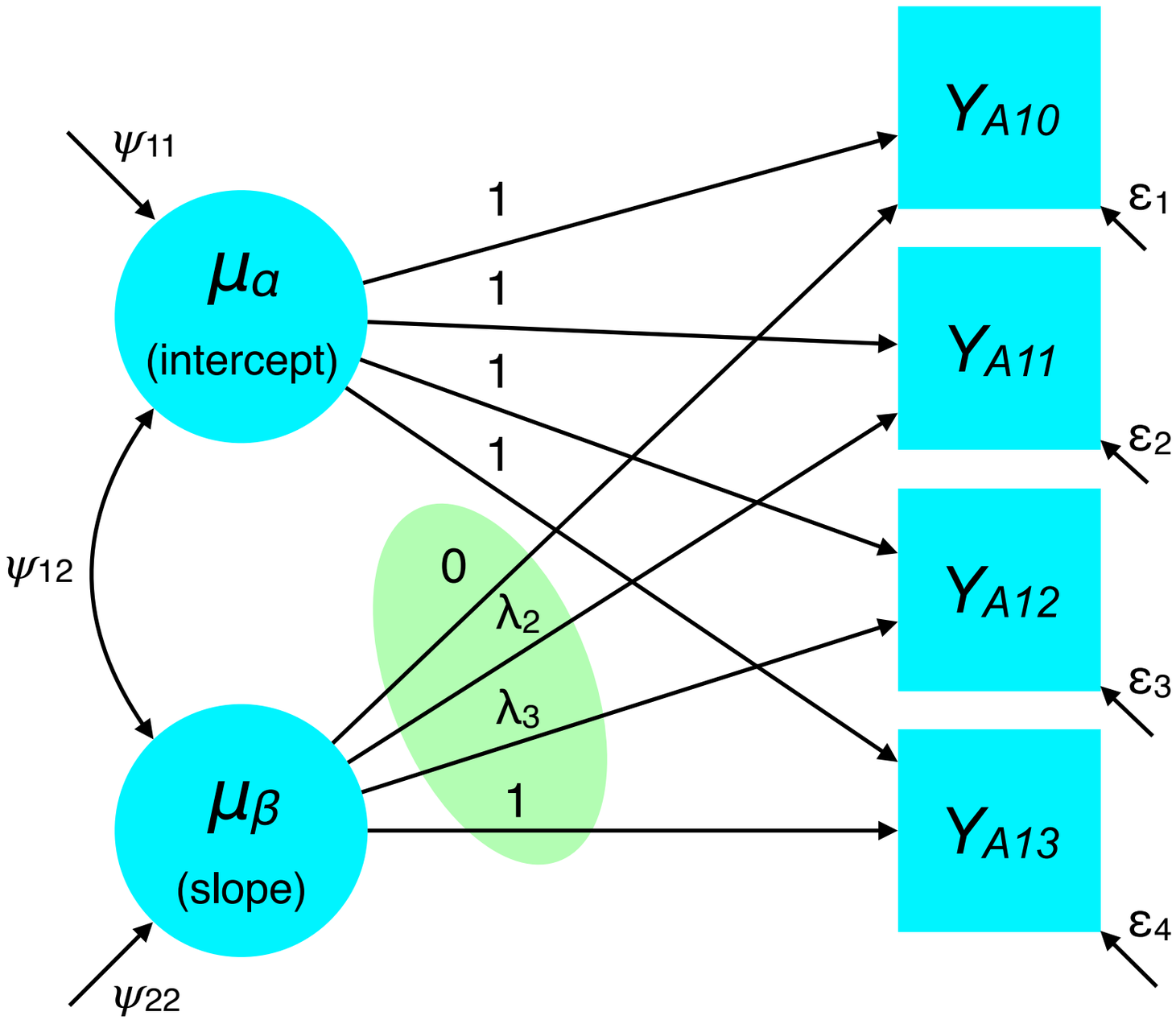
- If we thought there were some meaningful residual covariation (e.g., age 8 with age 14), we should worry about covariance coverage because the structural equations for **y** now depend on the theta $\boldsymbol{\Theta}$ matrix (residual covariance structure).

$$\boldsymbol{\theta} = \begin{bmatrix} \varepsilon_1 & 0 & 0 & 0 \\ 0 & \varepsilon_2 & 0 & 0 \\ 0 & 0 & \varepsilon_3 & 0 \\ 0 & 0 & 0 & \varepsilon_4 \end{bmatrix}$$

Time scores represent alternative models of change

- The time scores in LCMs specify a hypothesis about the form change
  - Example: if we measure once per year, then linear time scores are 0, 1, 2, 3, … t. And the slope is now in change in *y* per year.
    - But if we measure at 0, 8, and 24 months (on average), linear time scores are 0, 8, 24 and the slope is change in *y* per month.
- Although polynomial (linear, quadratic, cubic, etc.) change models may characterize some datasets, there are many parametric and semi-parametric alternatives (I will discuss just one).

Latent basis curve model (McArdle)



First time score = 0
Last time score = 1

Other times are *free* parameters. The values of these parameters represent *proportion* of change that has occurred.

By constraining the first and last slopes on the [0 – 1] interval, we examine how much change (relative to total) has occurred at each age/wave.

The slope estimates (β_i) now represent an estimate of how much a person changed over the entire study.

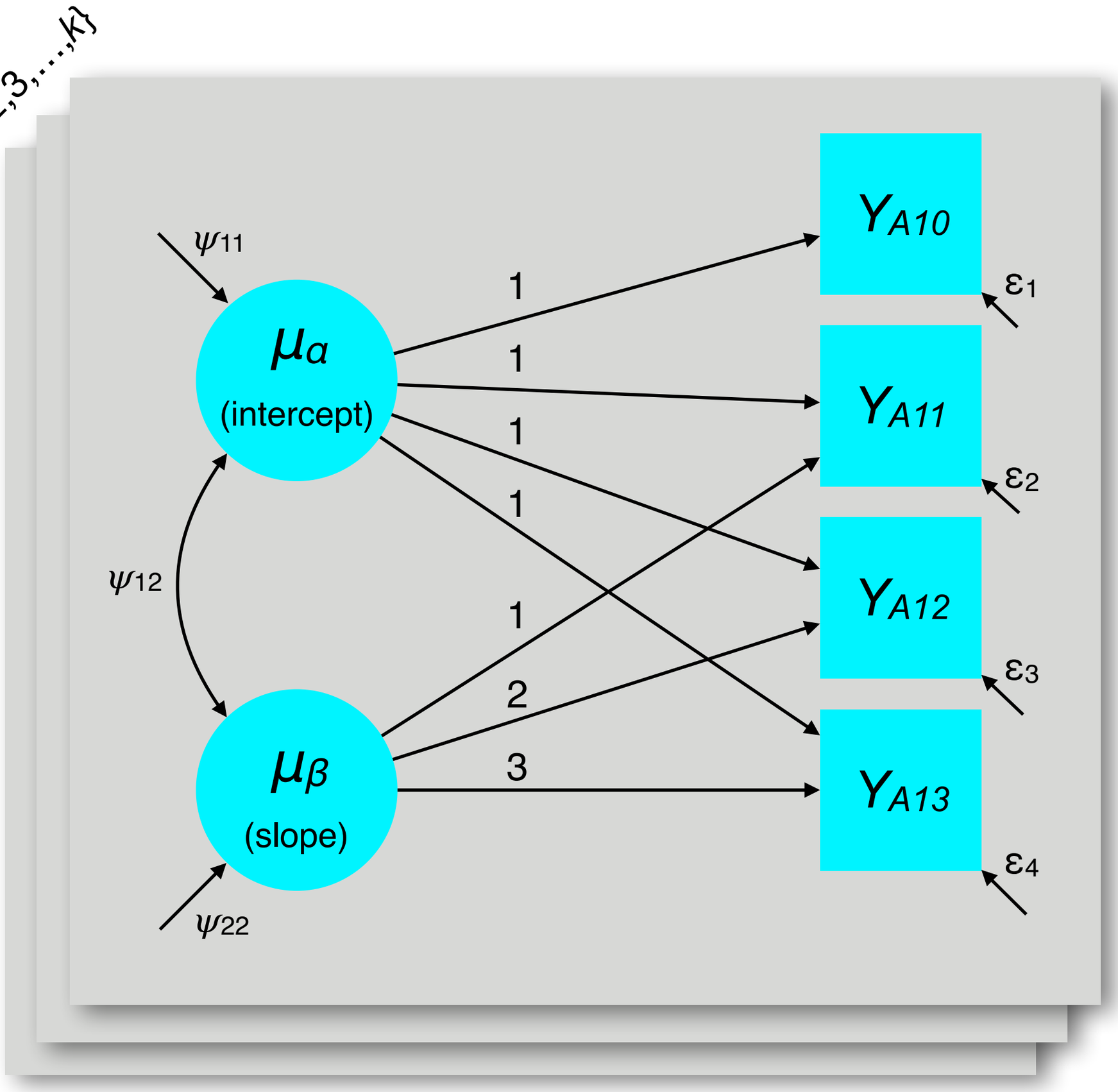
Growth mixture modeling

In growth mixtures, essentially any of the growth parameters can vary by *latent subgroup*.

That is, there could be subpopulations with meaningfully different growth trajectories.

Membership in each class can be predicted by covariates.

For example, being in the 'increasing' latent trajectory



Growth mixture modeling

- What parameters usually vary by subgroup?
- Typically, GMMs focus on subgroup differences in growth *means* ($\boldsymbol{\mu}$) and variances/covariances ($\boldsymbol{\Psi}$)
- Selection among GMMs usually based on model selection criterion (BIC, AIC, AIC_C), as well as likelihood ratio methods (esp. bootstrapped likelihood ratio test).
- Useful references
  - Bauer 2007, *Multivariate Behavioral Research*
  - Hallquist & Wright, 2014, *J Personality Assessment*
  - Kreuter & Muthén, 2008, *J Quant Criminol*
  - Masyn et al., 2010, *Social Development*
  - Wright & Hallquist, 2014, *J Personality Assessment*

GMM versus latent class growth

- Latent class growth analysis (LCGA; Nagin) is simply a GMM in which the variances of the growth factors (ψ_{11}, ψ_{22}, etc.) are fixed to zero.

- That is, the model explicitly disallows between-person variability of growth parameters in each subgroup.

- LCGA puts all the eggs in the trajectory basket, assuming homogeneity within trajectory.

- This likely overextracts classes in general. It may be useful for data description, but is dubious (in my opinion) for making substantive conclusions about latent groups.

General latent variable modeling

- If you want to test a standard latent curve model (including latent basis model), I would suggest the *lavaan* package (Rosseel, 2017).

- If you want to move beyond standard SEM latent curve models, especially incorporating latent classes (i.e., categorical latent variables), I would recommend *Mplus* (Muthén & Muthén, 2017), which is very powerful, but proprietary.

- And if you want to run a host of complex SEMs (not estimable in *R*), check out *MplusAutomation* (Hallquist & Wiley, 2017), which allows for *R* to run *Mplus* syntax and read *Mplus* outputs.