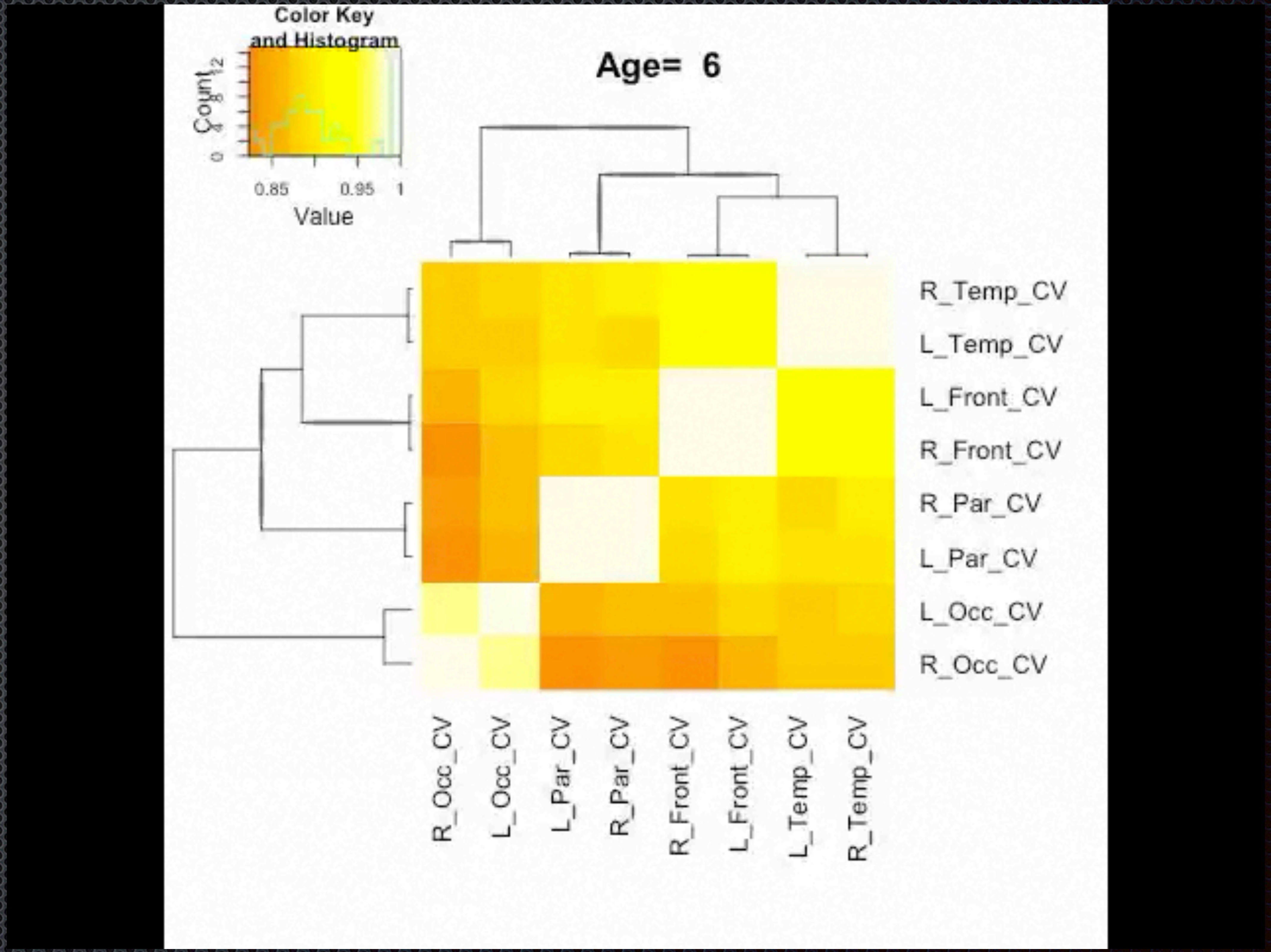
Multivariate Longitudinal Modeling with Genetically Informative Data

Michael C Neale, PhD
Virginia Institute for Psychiatric and Behavioral Genetics
Virginia Commonwealth University
Boulder Colorado NIMH Workshop June 2022

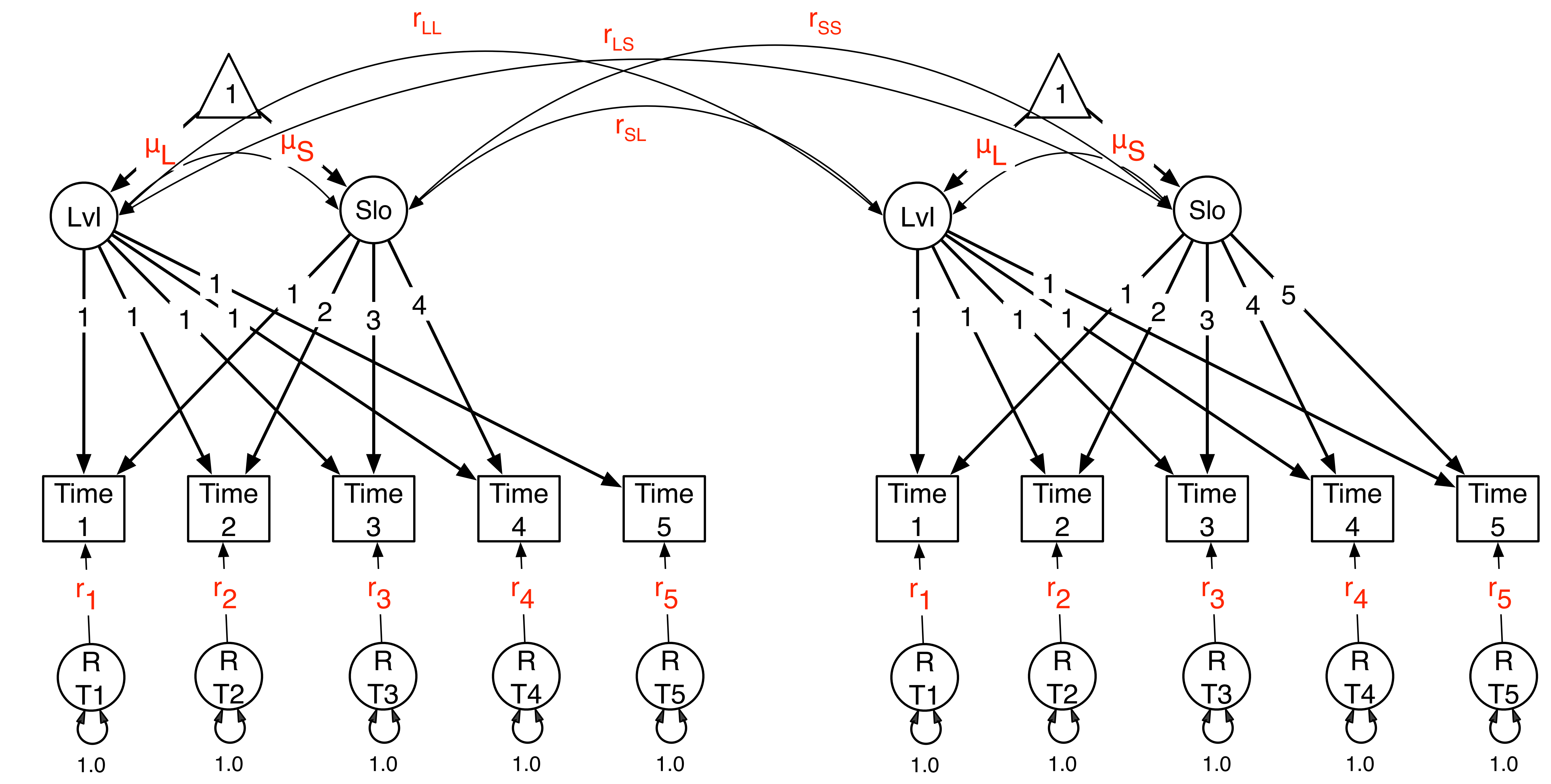
Overview

- Bivariate Latent Growth Curve
- Multivariate Markov Models
- Regime switching
- Long time series



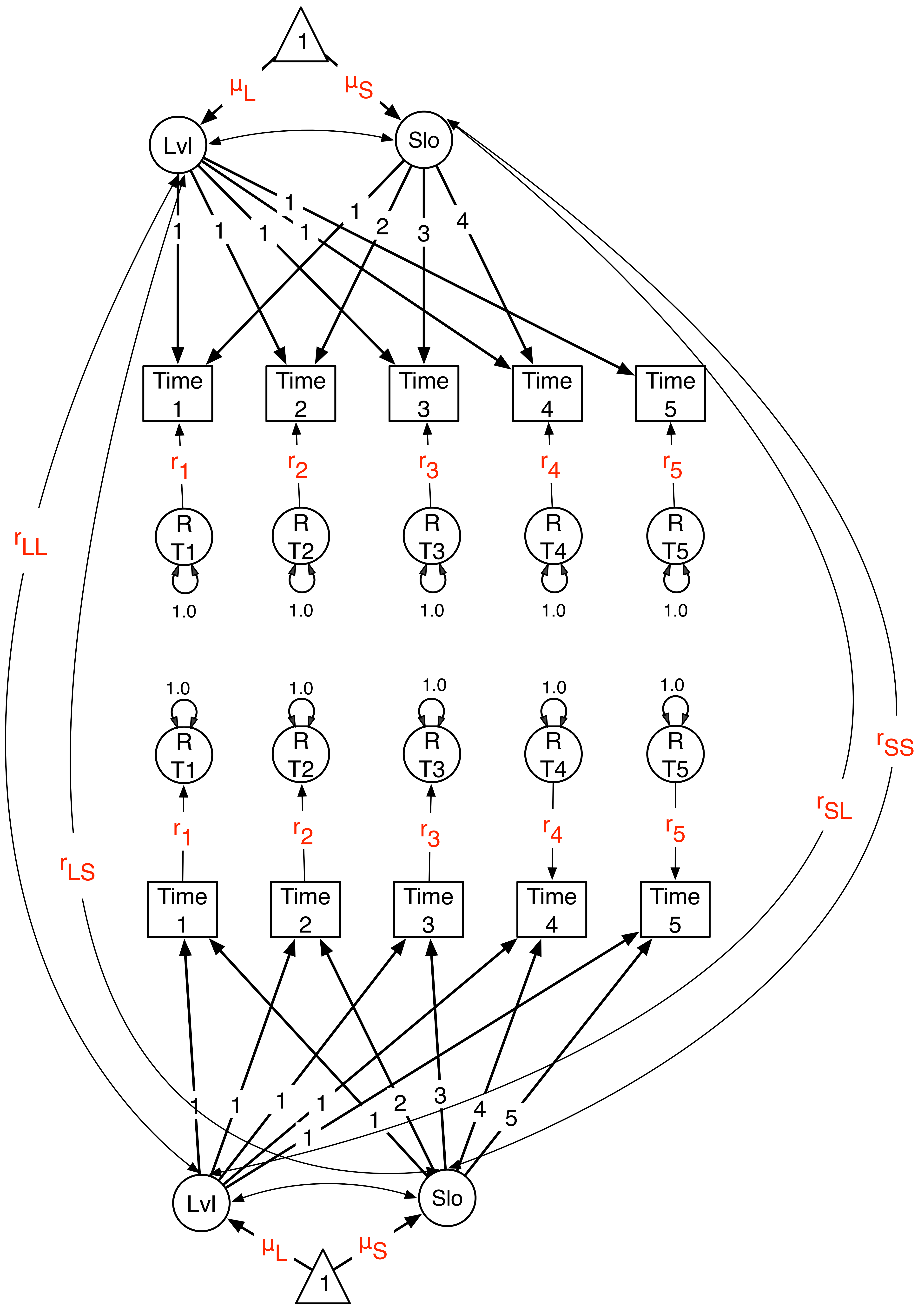
J Eric Schmitt, Jay N Giedd, Armin Raznahan, Michael C Neale, The Genetic Contributions to Maturational Coupling in the Human Cerebrum: A Longitudinal Pediatric Twin Imaging Study, Cerebral Cortex, Volume 28, Issue 9, September 2018, Pages 3184–3191, https://doi.org/10.1093/cercor/bhx190

Bivariate Latent Growth Curve: Unrelated Individuals



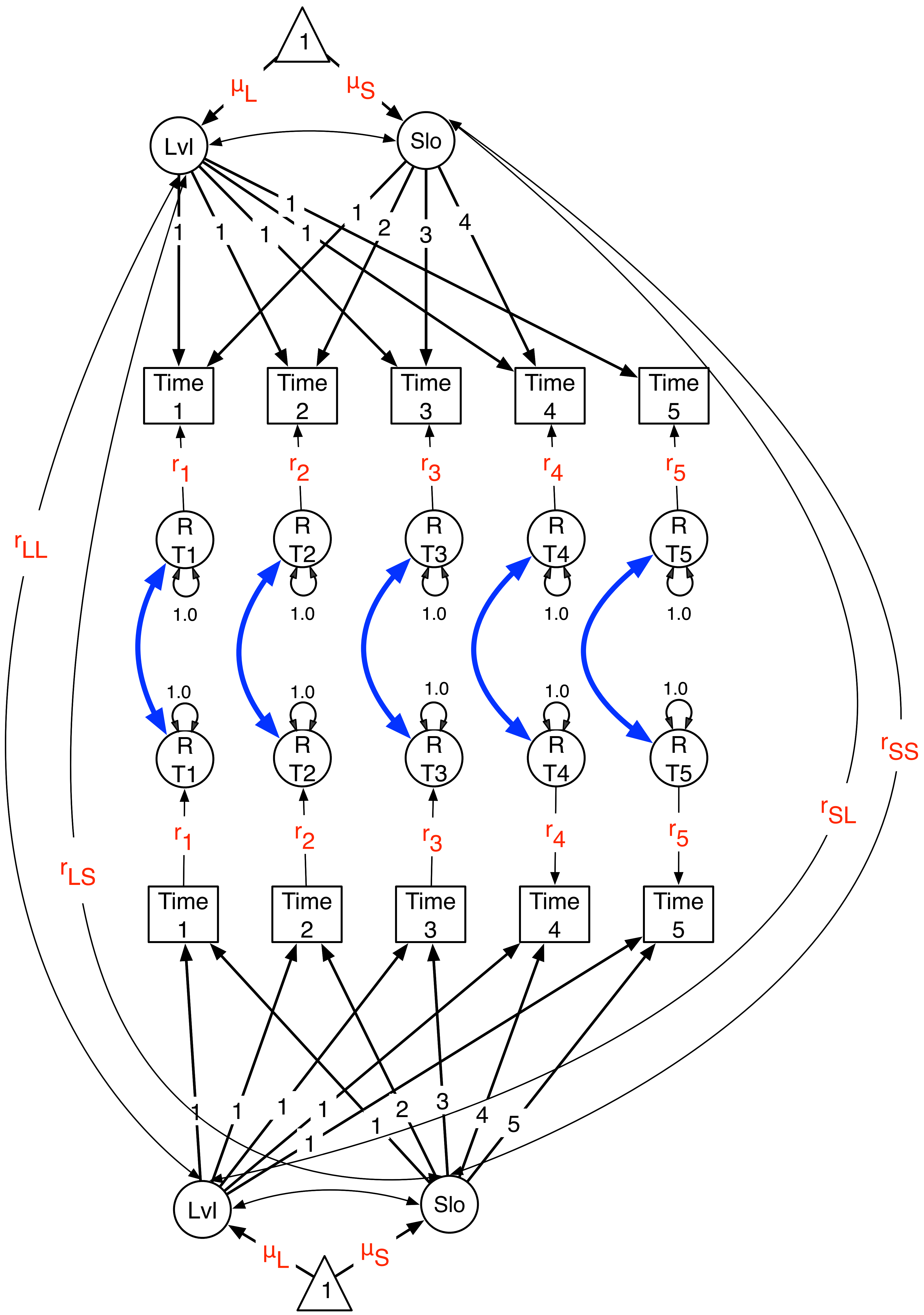
Bivariate Latent Growth Curve

- Covariance exclusively via factors
- Level-slope cross correlations
- Simple model, few parameters
- If it fits well it's a miracle



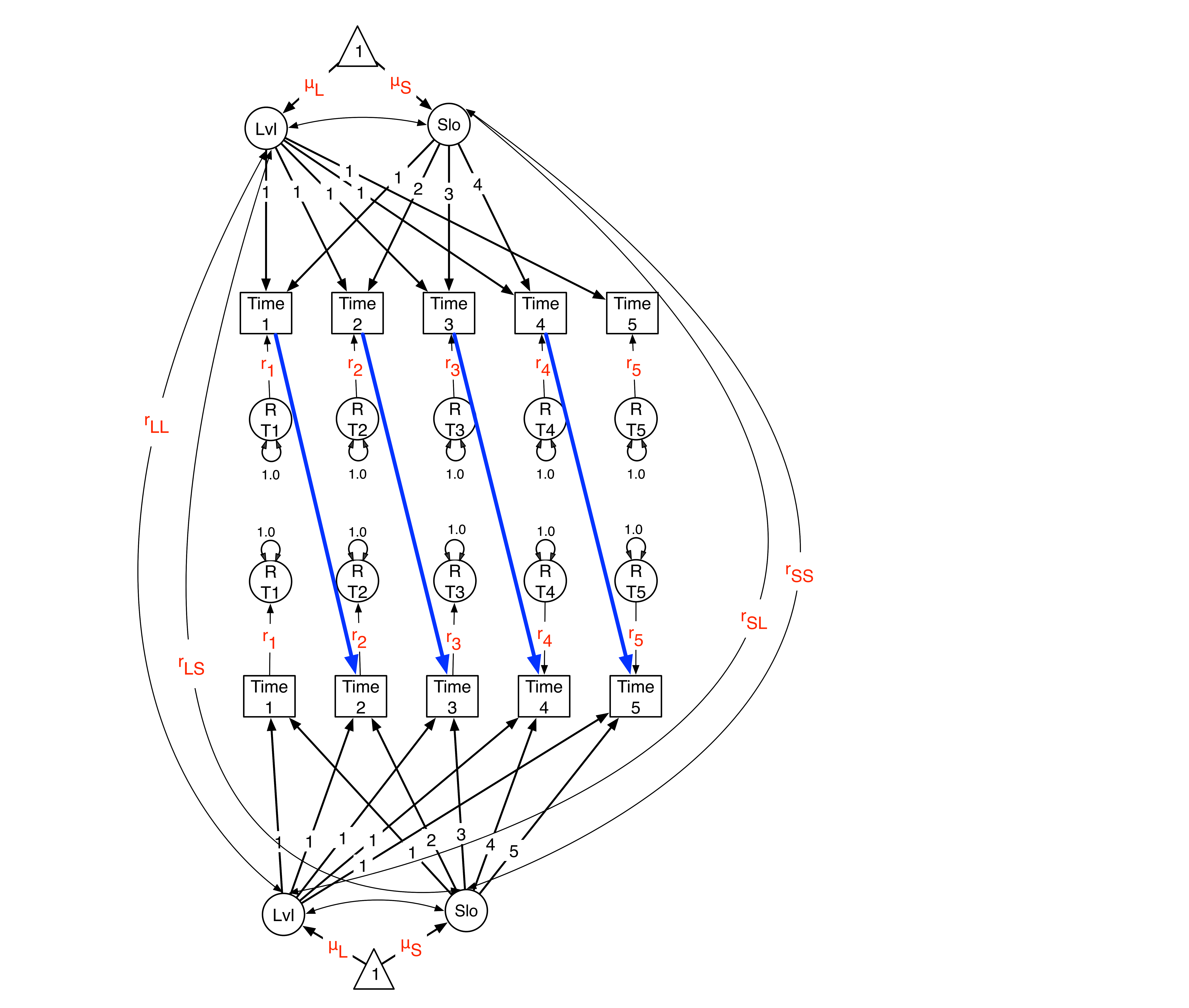
Bivariate Latent Growth Curve

- Covariance via factors and residuals
- Level-slope cross correlations
- Simple model, few parameters
- If it fits well it's still a miracle



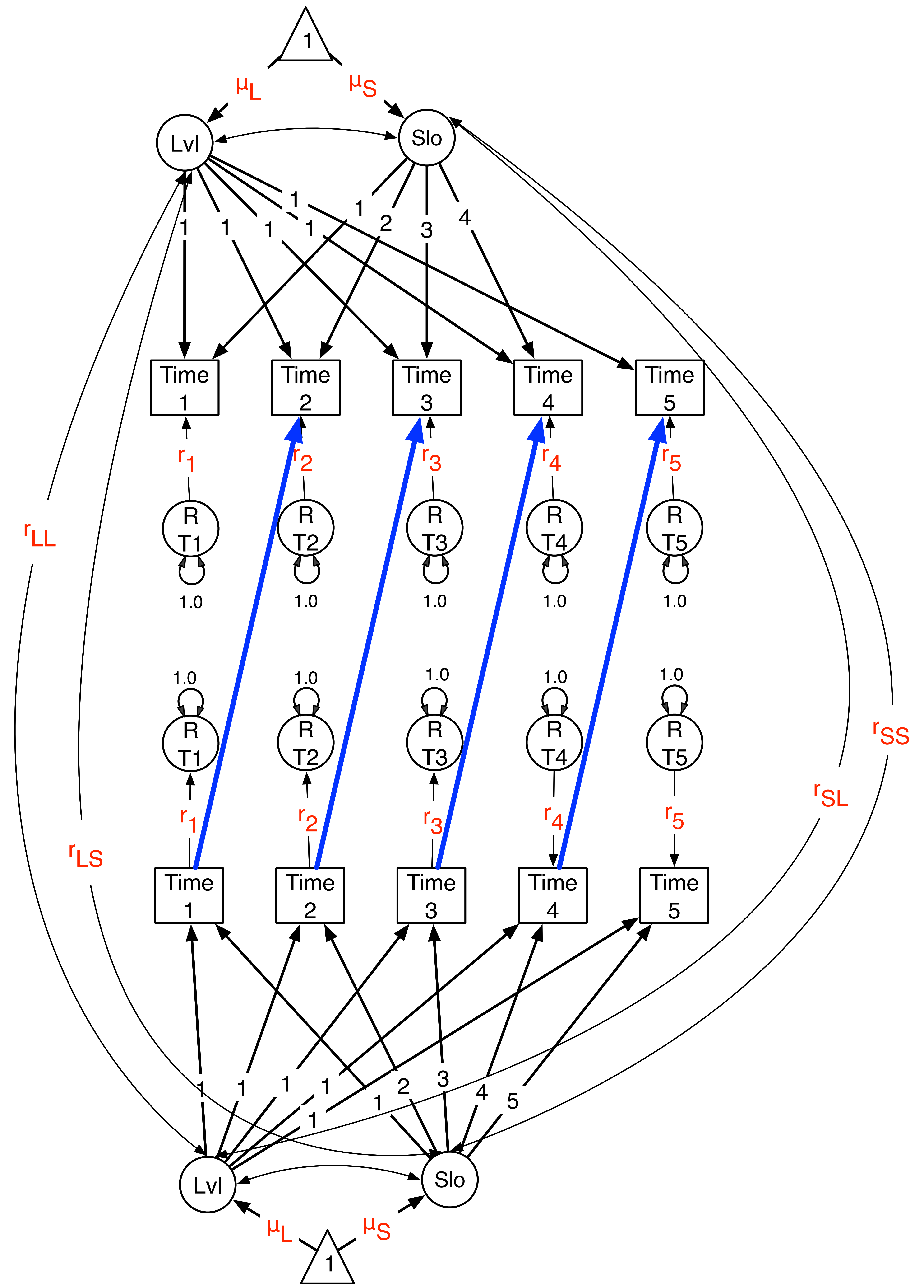
Bivariate Latent Growth Curve

- Covariance exclusively via factors
- Level-slope cross correlations
- Simple model, few parameters
- If it fits well it's a miracle



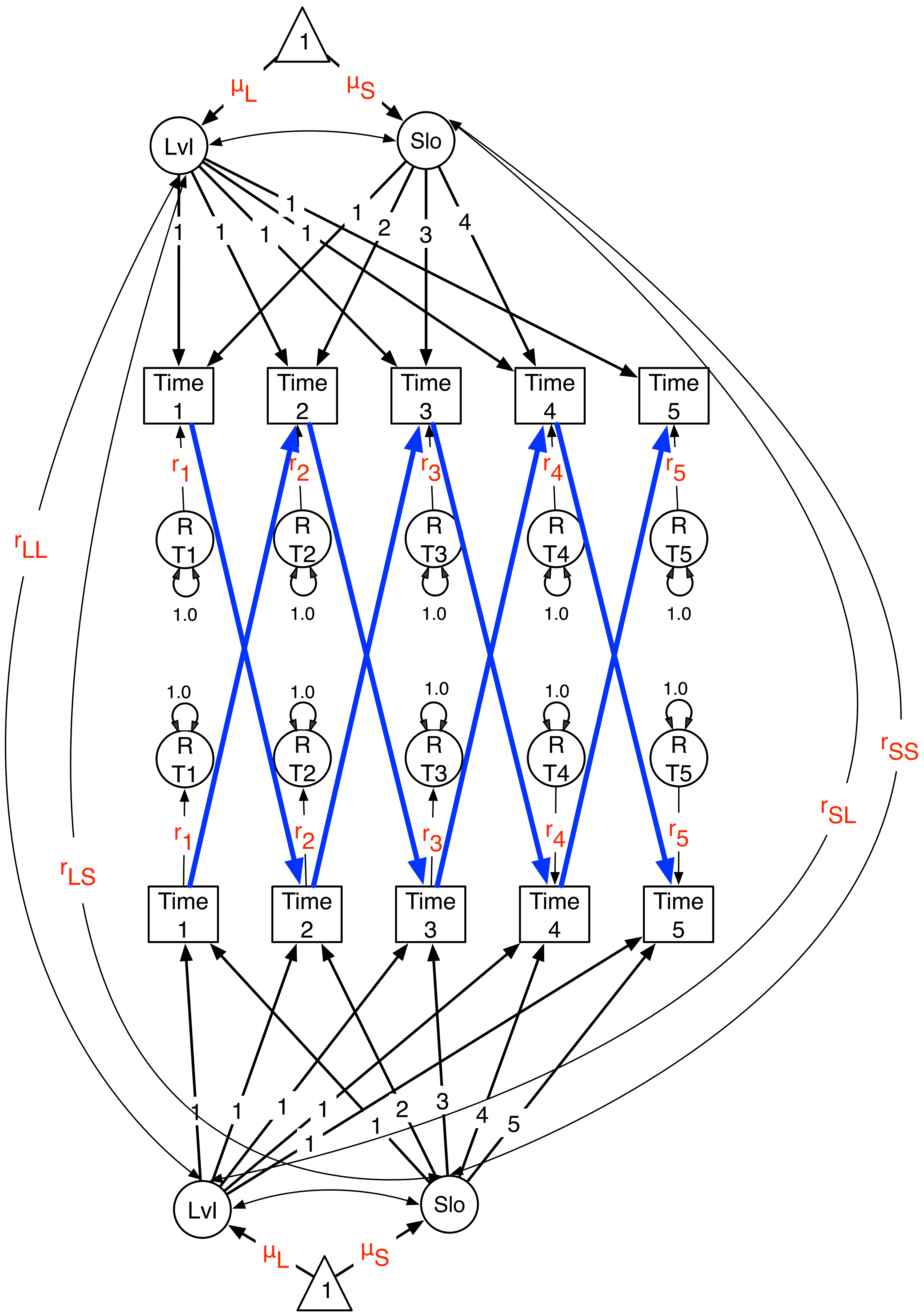
Bivariate Latent Growth Curve

- Covariance exclusively via factors
- Level-slope cross correlations
- Simple model, few parameters
- If it fits well it's a miracle



Bivariate Latent Growth Curve

- Covariance via factors and cross-lagged paths
- Level-slope cross correlations
- May be specified with few or many parameters
- If it fits well, interesting!



Testing Models for Identification

- If a model is identified, it implies that:
  - One and only one set of parameter values will yield the maximum likelihood solution
  - There are at least as many observed statistics as there are parameters
  - Optimization is likely to be successful if given reasonable starting values

Bollen, K. A., & Bauldry, S. (2010). Model Identification And Computer Algebra. Sociological methods & research, 39(2), 127–156.

A little calculus

- We want to know whether changing one parameter does exactly the same thing to the model-expected statistics as changing one or more of the other parameters
- How do we measure change? Gradients as a function of parameter change
- Organize: columns=parameters, rows=expected statistics
- This matrix of gradients (partial derivatives of each expected statistic, with respect to each parameter) is called the *Jacobian* after Carl Jacobi 1804-51

Procedure

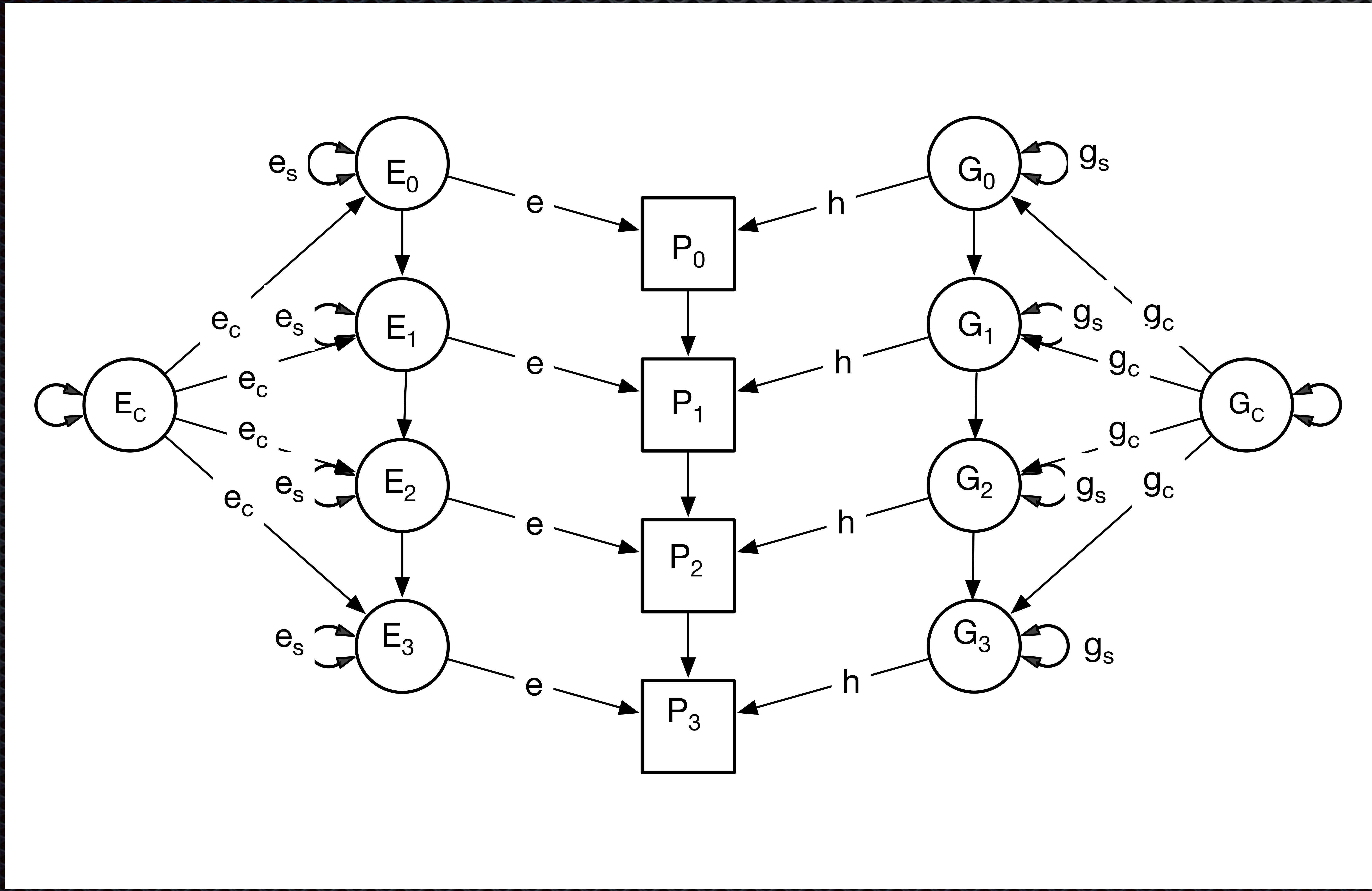
Not to be confused: Jacobean

1. Draw model in Onyx
2. Use “To Script” function to generate OpenMx script
3. Make some fake data for the model
4. Consider eliminating means and provide covariance matrix input, or use both means and covariances (e.g. LGC)
5. Fit model to data
6. Test local identification at the solution via Jacobian
    - mxCheckIdentifiction(fittedModel)

mxCheckIdentification()

- Numerically tests models for local identification
- Do not use for models with definition variables (likely fix is designed but not implemented - possibly late 2022)
- Works reliably, estimates Jacobian numerically
- Examines column rank to test for linear dependencies among the columns. Matrix::rankMatrix

Multivariate AR1 Model with Random Latent Intercepts



Markov Process
(time to time)

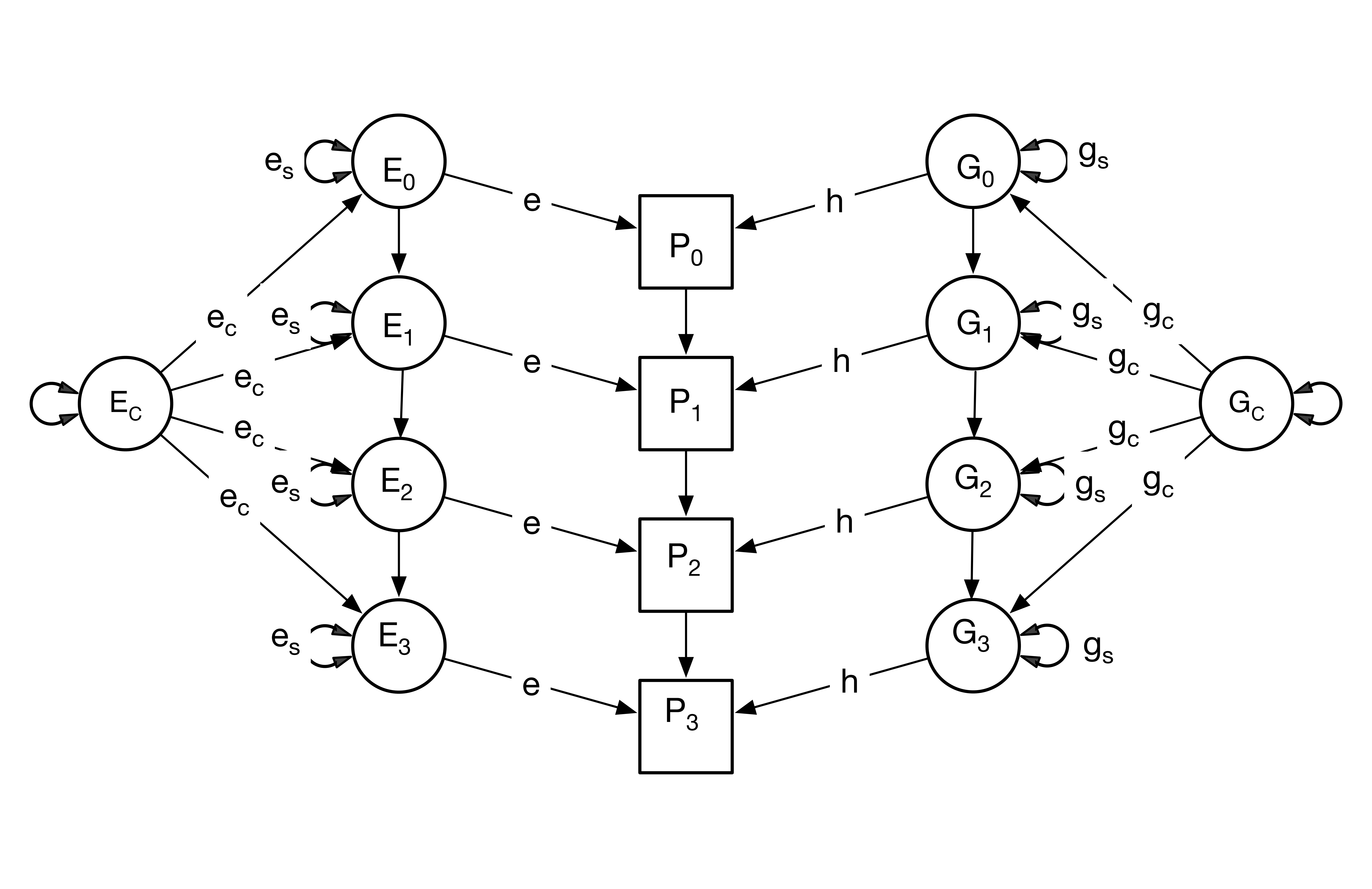
Factors Gc & Ec
(constant)

Time-specific Factors
$G_0 \ldots G_m$ etc. (transient)

Eaves et al 1986 A theory of developmental change in quantitative phenotypes applied to cognitive development. Behavior Genetics 16(1):143-62

Multivariate Path Analysis

One Approach



Each node represents a vector of variables

Each path represents a matrix of paths

OpenMx mxPath() allows vectors of 'from' and 'to' variables, and connect options

Vogler GP. Multivariate path analysis of familial resemblance. Genet Epidemiol. 1985;2(1):35-53. doi: 10.1002/gepi.1370020105. PMID: 4054591.

Ergodicity

- Within-individual factor analysis
  - Rows of occasions, Columns of variables
- Between-individuals factor analysis
  - Rows of individuals, Columns of variables
- These two perspectives almost never agree about factor structure
- Agree if and only if the system is Ergodic
  - i.e., really boring, essentially no development

Very Long Time Series

Discrete time

A state-space model in discrete time can be stated generally as,

$$\mathbf{x}_t = \mathbf{g}[t, \mathbf{x}_{t-1}, \mathbf{u}_{t-1}, \mathbf{q}_{t-1}, \boldsymbol{\theta}_x] \tag{2}$$

$$\mathbf{y}_t = \mathbf{h}[t, \mathbf{x}_t, \mathbf{u}_t, \mathbf{r}_t, \boldsymbol{\theta}_y] \tag{3}$$

Linear state-space models, i.e, those in which observations (y) are proportionally related to latent states (x), and states are proportionally related to previous states (x_{t-1}), can be given in matrix form. Each matrix is subscripted by individual or family i to indicate that we will fit many such models in a twin study:

$$\mathbf{x}_{i,t} = \mathbf{A}_i \mathbf{x}_{i,t-1} + \mathbf{B}\mathbf{u}_{i,t} + \mathbf{q}_{i,t}, \quad \mathbf{q}_{i,t} \sim \mathcal{MVN}(0, \mathbf{Q}_i), \tag{4}$$

$$\mathbf{y}_{i,t} = \mathbf{C}\mathbf{x}_{i,t} + \mathbf{D}\mathbf{u}_{i,t} + \mathbf{r}_{i,t}, \quad \mathbf{r}_{i,t} \sim \mathcal{N}(0, \mathbf{R}), \tag{5}$$

$$t \geq 0, \quad i \in \{1..N\}. \tag{6}$$

The $\mathbf{A}$ matrix contains the auto and cross-regressive coefficients of the latent state, $\mathbf{x}_t$. $\mathbf{B}$ contains the regression coefficients on potentially time-varying covariates, or external forcing functions $\mathbf{u}_t$. $\mathbf{q}_t$ is a noise variable representing random disturbances to the state, most often taken to be normally distributed. The $\mathbf{C}$ matrix contains the factor loadings, or coefficients that relate the indicators to the latent state.

Continuous time

Continuous time state-space models use SDEs to specify change in the latent state $\dot{x}(t)$ as a function of the level $x(t)$. Other components remain virtually the same as discrete time models, except that $t \in \mathbb{R}$, so $x(t)$ refers to the solution of the differential equation at time t.

$$\dot{\mathbf{x}}(t) = \mathbf{g}[t, \mathbf{x}(t), \mathbf{u}(t), \mathbf{q}(t), \boldsymbol{\theta}_x] \tag{9}$$

$$\mathbf{y}(t) = \mathbf{h}[t, \mathbf{x}(t), \mathbf{u}(t), \mathbf{r}(t), \boldsymbol{\theta}_y] \tag{10}$$

For linear continuous time models, the state equation is given as a matrix SDE:

$$\dot{\mathbf{x}}_i(t) = \mathbf{A}_i \mathbf{x}_i(t) + \mathbf{B}\mathbf{u}(t) + \mathbf{q}_i(t). \tag{11}$$

A convenient way to fit differential equation models is to discretize matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{Q}$ according to time interval Δ., i.e., solve the SDE for each new successive measurement interval. The discretized matrices can then be used much like the discrete time model:

$$\mathbf{x}_i(t) = \mathbf{A^d}_i \mathbf{x}_i(t-\Delta) + \mathbf{B^d}\mathbf{u}(t) + \mathbf{q}_i(t). \tag{12}$$

Measurement models

McKee KL. Hierarchical Biometrical Genetic Analysis of Longitudinal Dynamics. Behav Genet. 2021 Nov;51(6):654-664. doi: 10.1007/s10519-021-10060-0. Epub 2021 May 12. PMID: 33978896.

Conclusion

- Multivariate longitudinal data offer great possibilities
- Relatively unexplored field with few fully developed models for multivariate genetically informative data
- Latent growth curve bivariate most well-established
- Markov/Simplex/AR1 models need work
- Promising new developments for analyses of very long time series (t=50+)