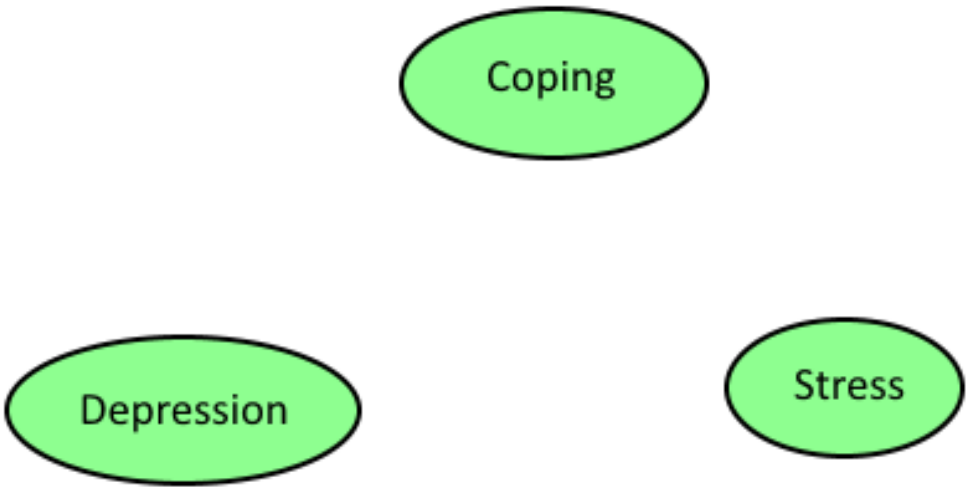
Causal Clustering for 2-Factor Measurement Models

Erich Kummerfeld, Joe Ramsey, Renjie Yang, Peter Spirtes, Richard Scheines

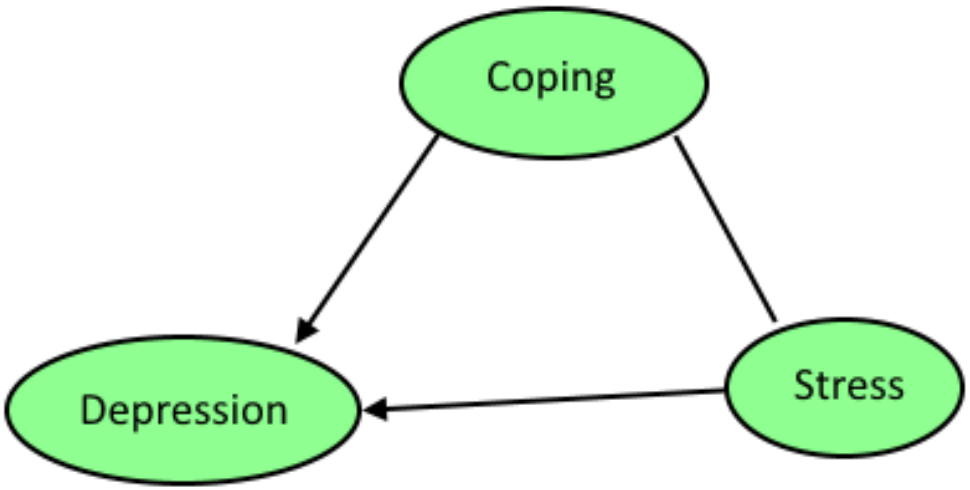
Carnegie Mellon University

September 16, 2014

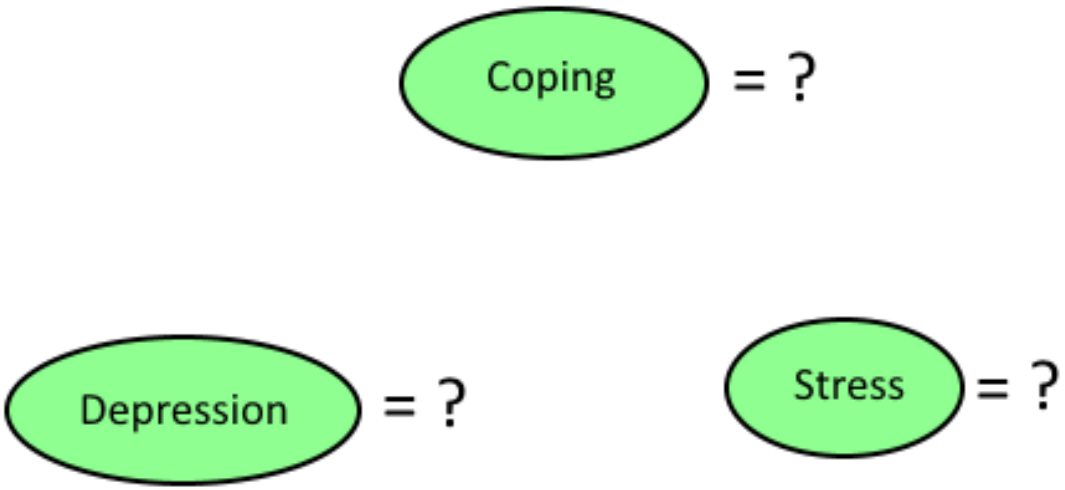
Psychometric Variables



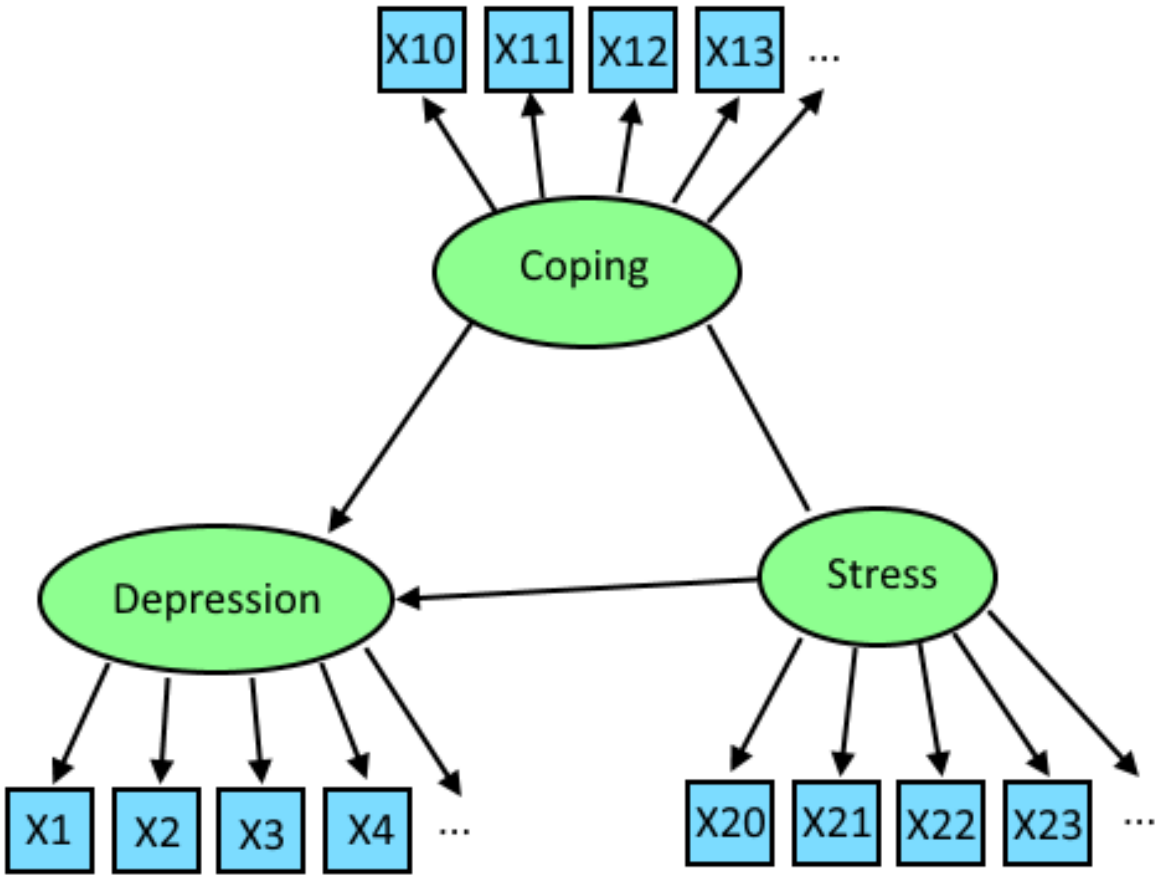
Structural Model



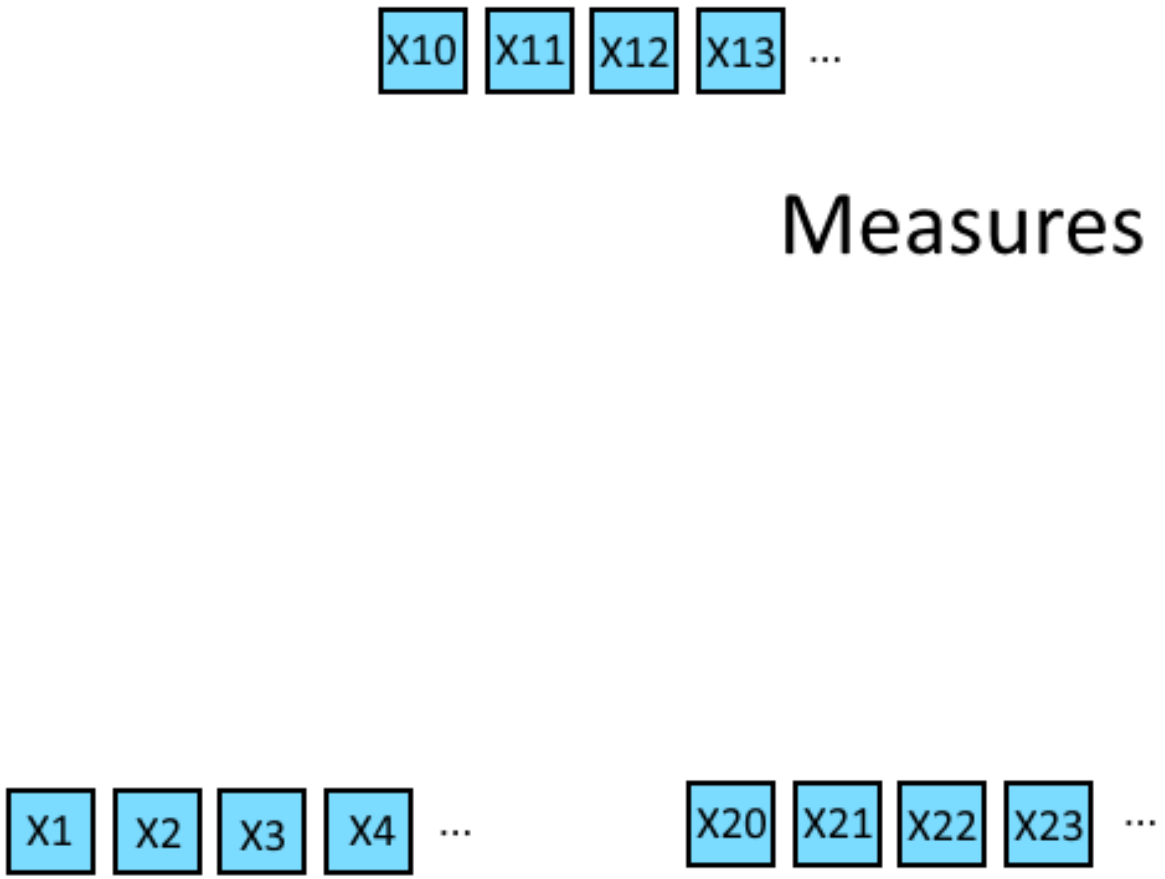
Important Variables Not Directly Measurable



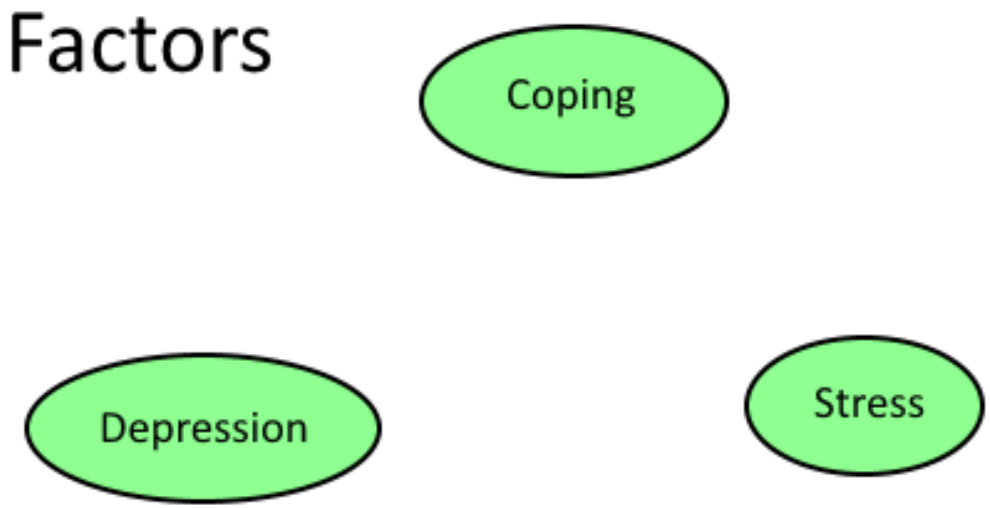
Hypothesis: Factor Model With Measures



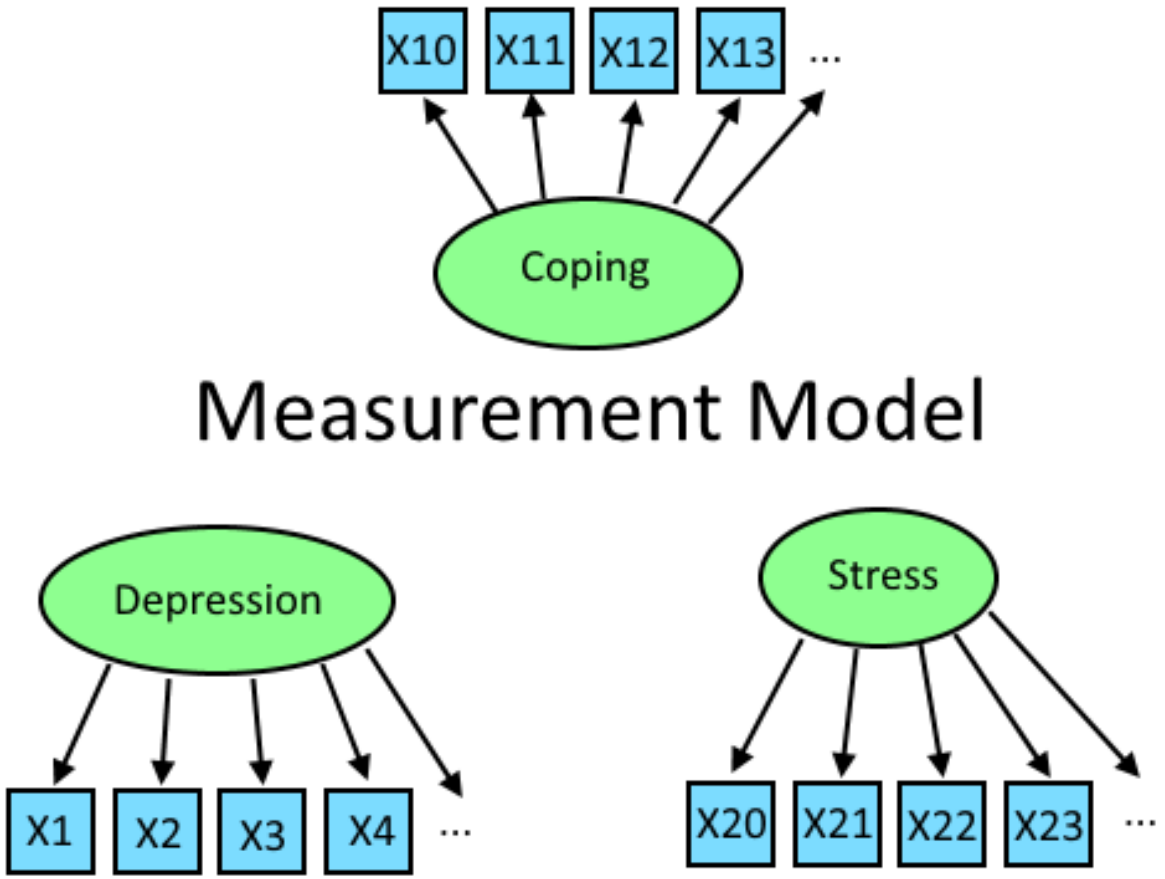
The Parts of a Factor Model



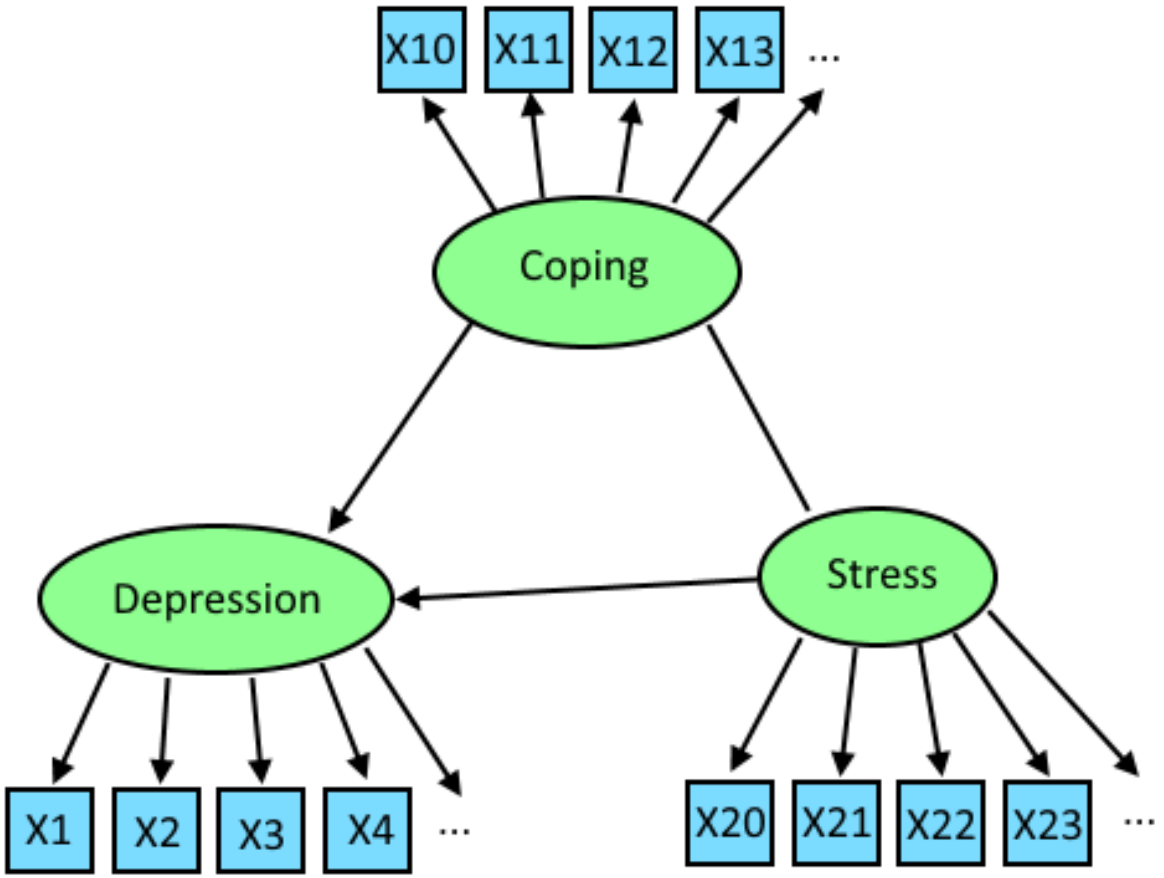
The Parts of a Factor Model



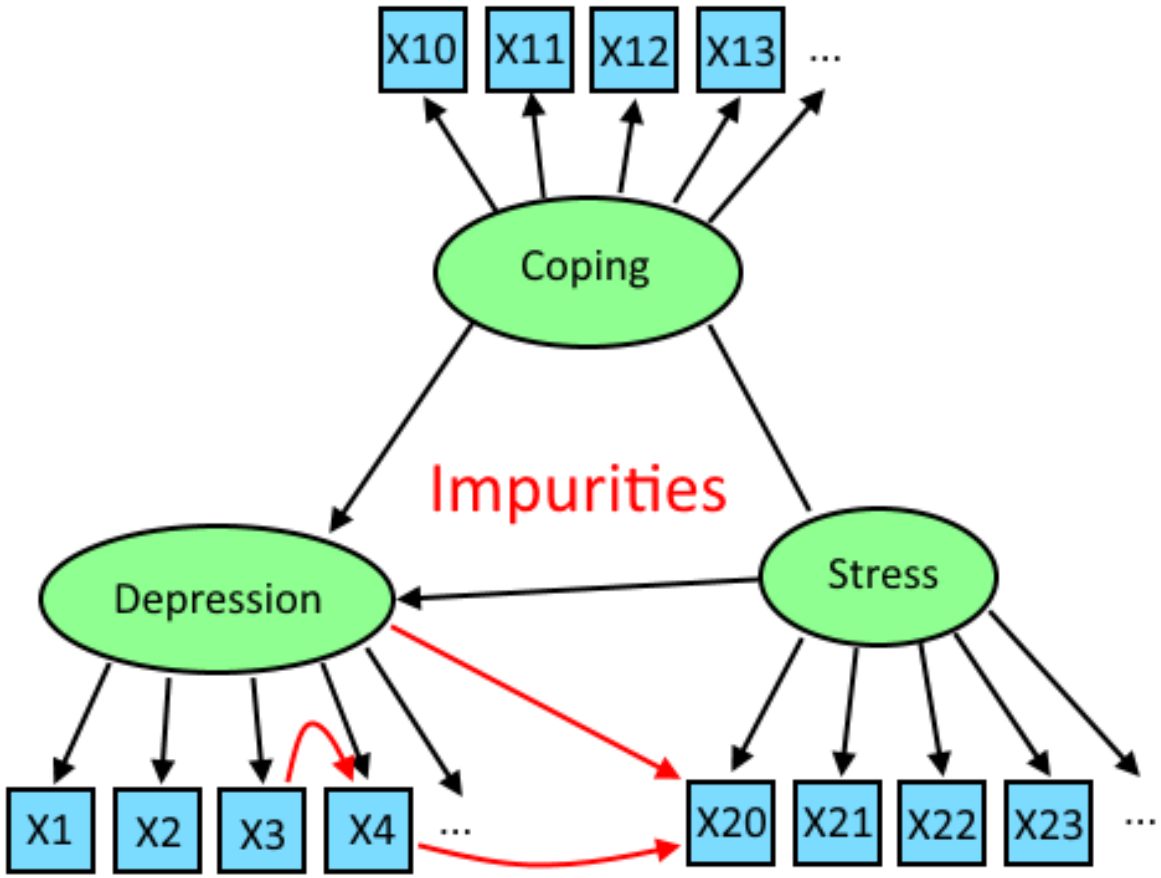
The Parts of a Factor Model



Hypothesized Factor Model



Hypothesized Factor Model



Measurement Model Inference

- **GOAL:** Find factor models that pass χ^2 tests

Measurement Model Inference

- **GOAL:** Find factor models that pass χ^2 tests
- Many hypothesized factor models fail χ^2 test
- Potential Strategies:

Measurement Model Inference

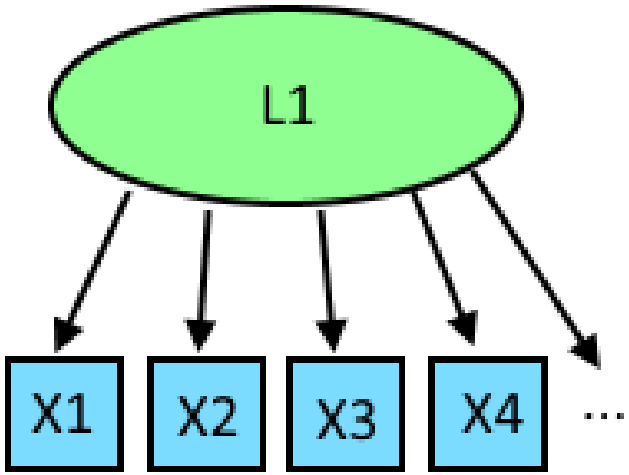
- **GOAL:** Find factor models that pass χ^2 tests
- Many hypothesized factor models fail χ^2 test
- Potential Strategies:
  1. Ignore impurities
  2. Find impurities without knowing structural model, # factors

Measurement Model Inference

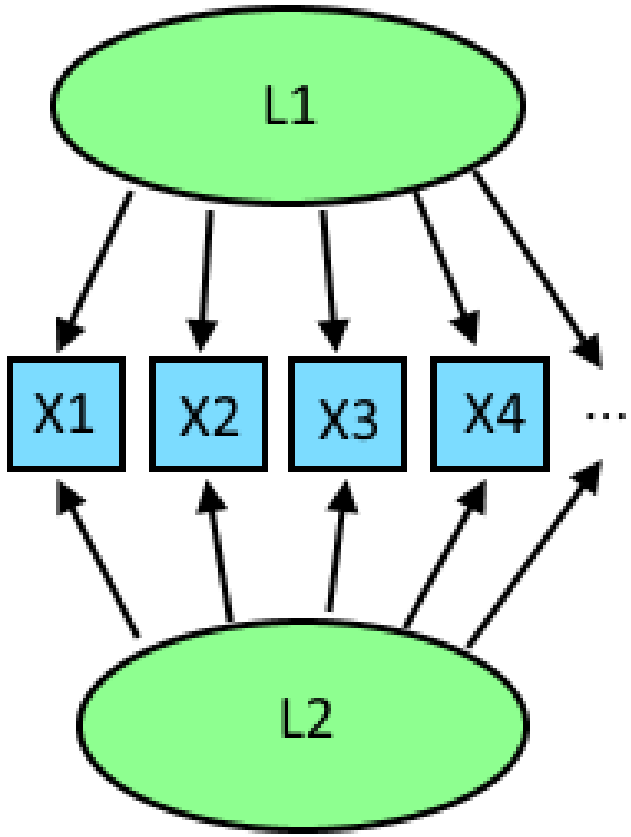
- **GOAL:** Find factor models that pass χ^2 tests
- Many hypothesized factor models fail χ^2 test
- Potential Strategies:
  1. Ignore impurities
  2. Find impurities without knowing structural model, # factors
- Our Strategy:
  1. Find *measurement model* with no impurities
  2. Use *measurement model* to learn structural model

Strategy previously employed by BPC algorithm (Silva et al., 2006)

Two Factor Measurement Models



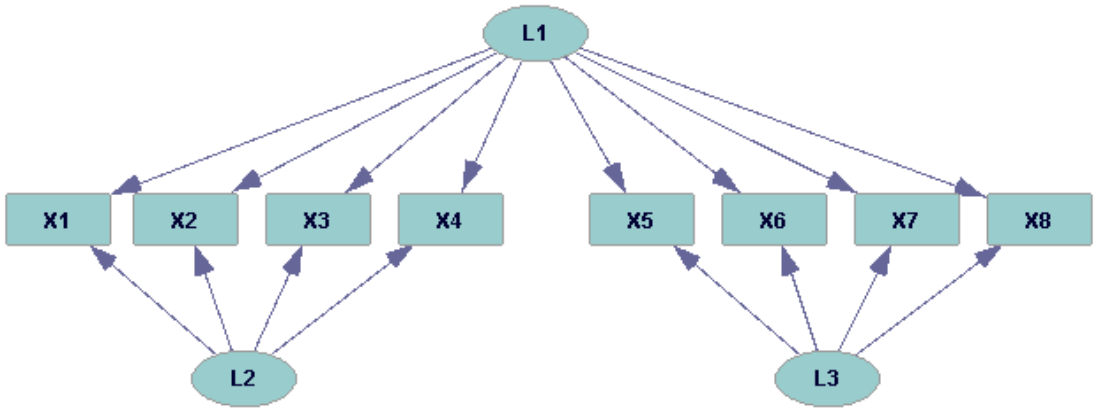
Two Factor Measurement Models



Factor analysis

Factor analysis can find bi-factor models

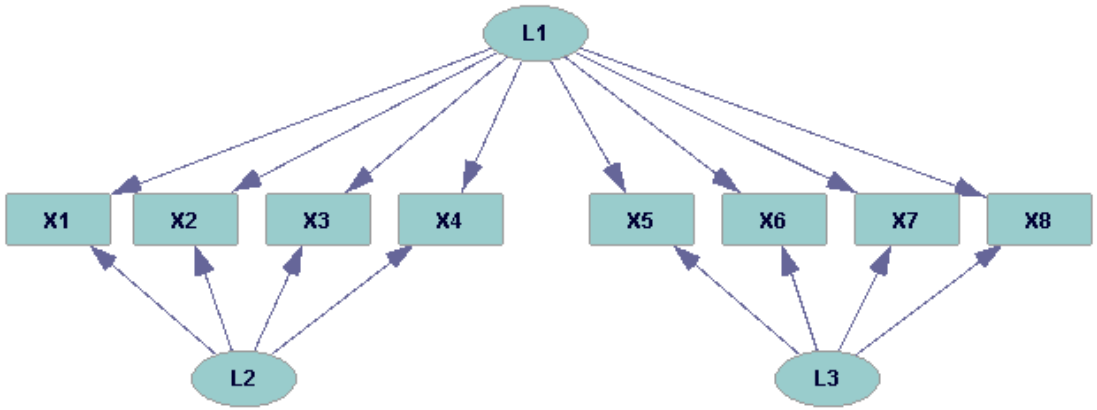
Doesn't pass χ^2 tests on data we've tried



Factor analysis

Factor analysis can find bi-factor models

Doesn't pass χ^2 tests on data we've tried

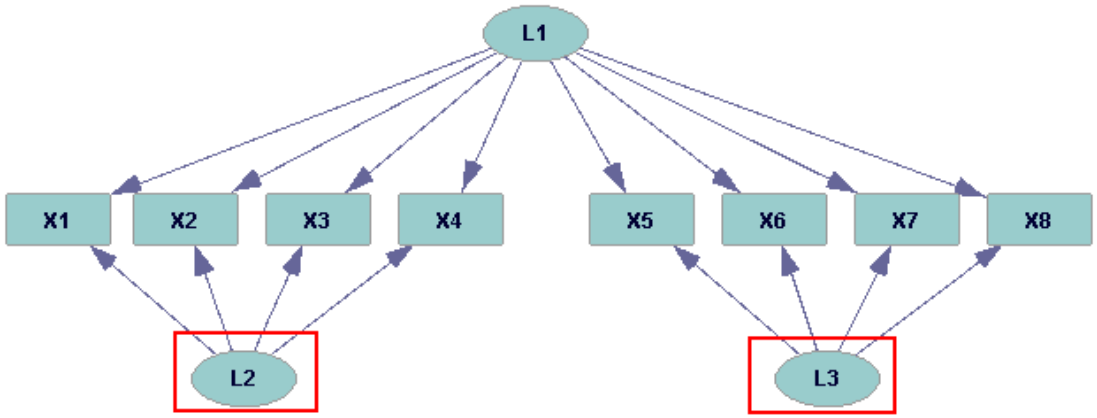


Various problems:

Factor analysis

Factor analysis can find bi-factor models

Doesn't pass χ^2 tests on data we've tried

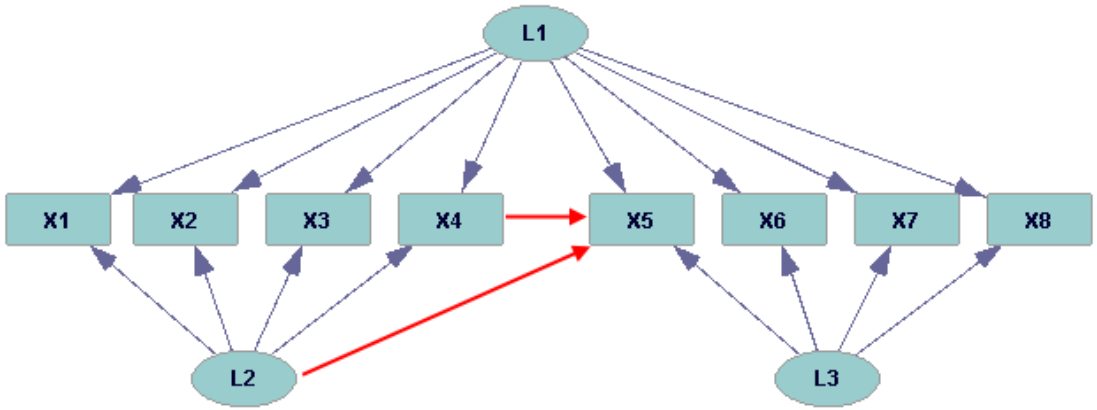


Various problems: # of factors?

Factor analysis

Factor analysis can find bi-factor models

Doesn't pass χ^2 tests on data we've tried

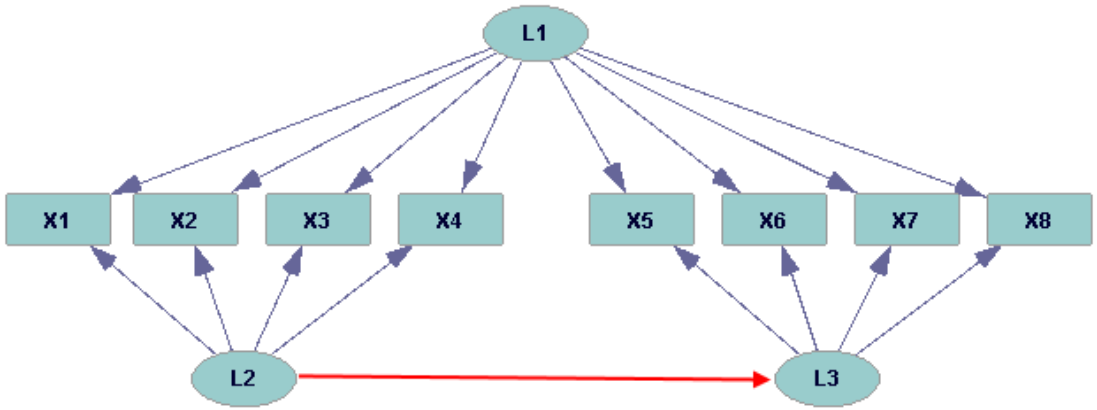


Various problems: # of factors? Impurities?

Factor analysis

Factor analysis can find bi-factor models

Doesn't pass χ^2 tests on data we've tried



Various problems: # of factors? Impurities? Structural edges?

Find Two-Factor Clusters

- **Purpose:** Find two factor measurement models within set of given measures.
- **Strategy:** Find subsets of measures satisfying higher-order algebraic constraints
- **Advantages:**
  - structural edges can be nonlinear
  - input measures may have impurities
  - number of factors not required as input

Trek Separation

t-separation (Sullivant et al., 2010) is a graphical generalization of *d*-separation for sets of variables

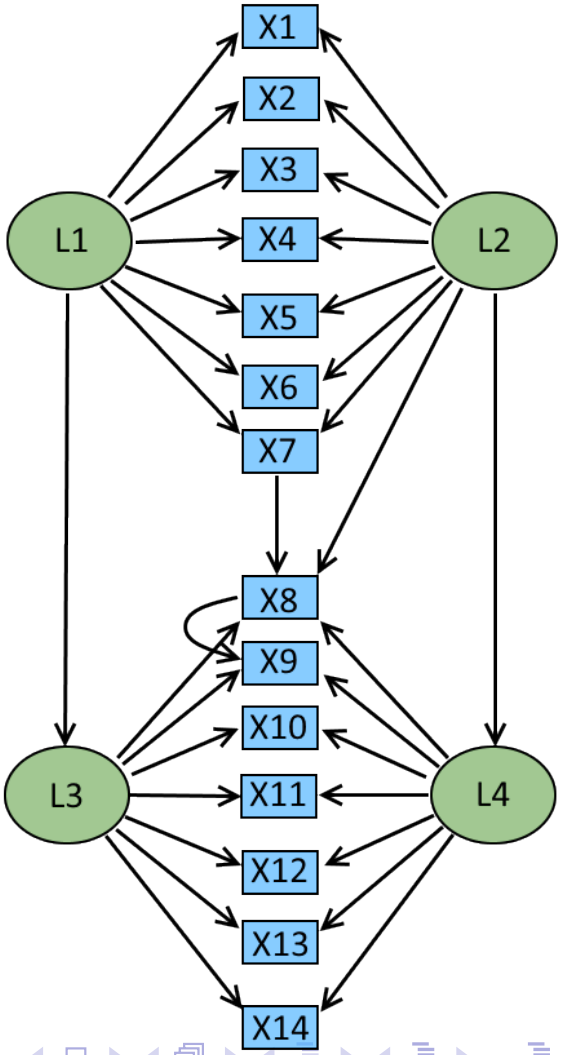
Trek Separation

t-separation (Sullivant et al., 2010) is a graphical generalization of *d*-separation for sets of variables

- Assumes linearity on some edges (Spirtes, 2013)
- sets of variables *t*-separated by an ordered pair of sets of variables
- *t*-sep set size is bounded iff vanishing determinants of submatrices of covariance matrix (Sullivant et al., 2010, Spirtes, 2013)
- Can use statistical tests for vanishing determinants of submatrices of cov matrix!

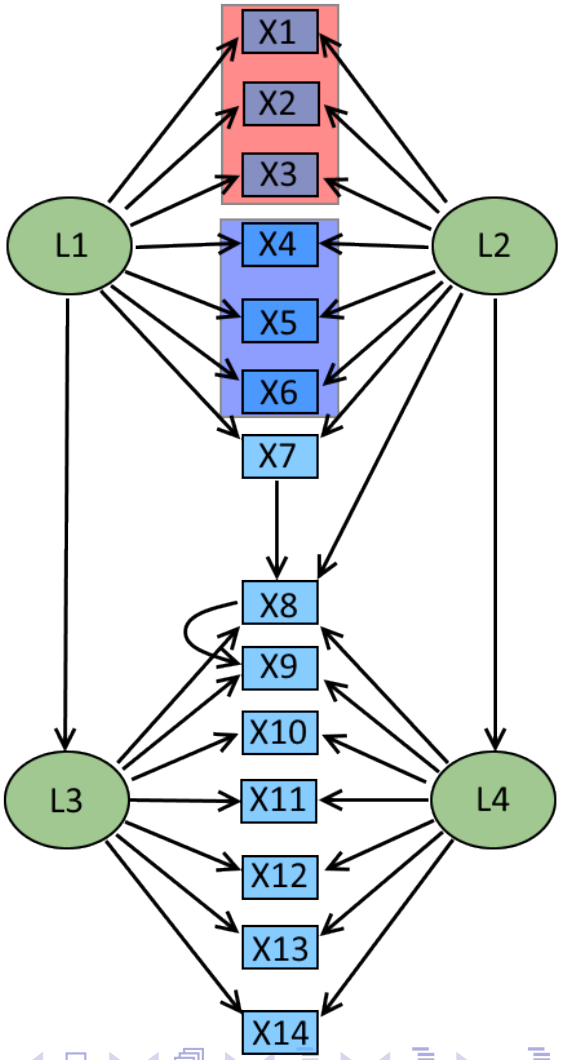
Trek Separation

What can *t*-separate



Trek Separation

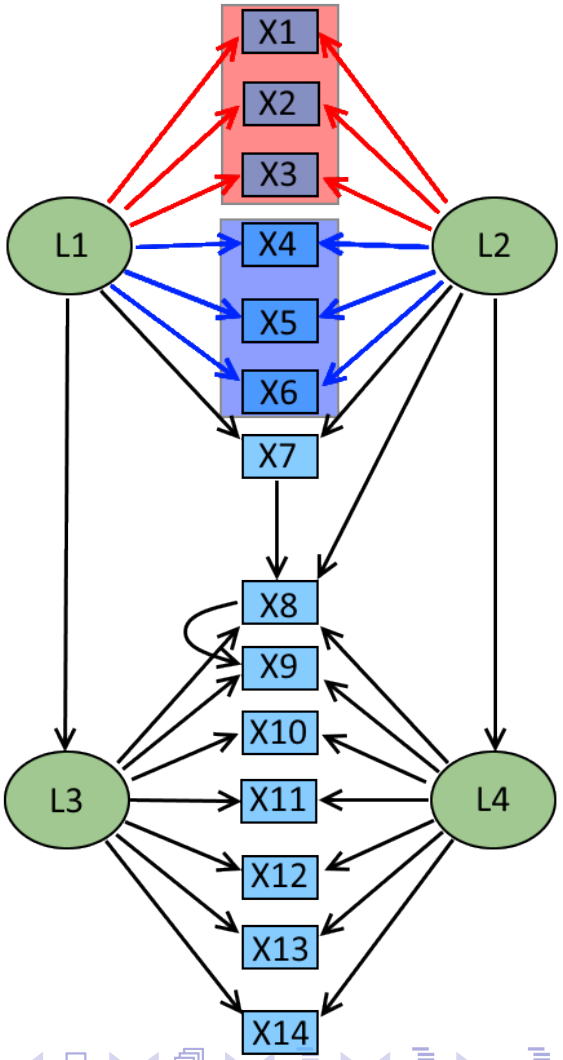
What can *t*-separate
{X1, X2, X3} from {X4, X5, X6}?



Trek Separation

What can *t*-separate
{X1, X2, X3} from {X4, X5, X6}?

All treks from red to blue must pass across a red edge and a blue edge.



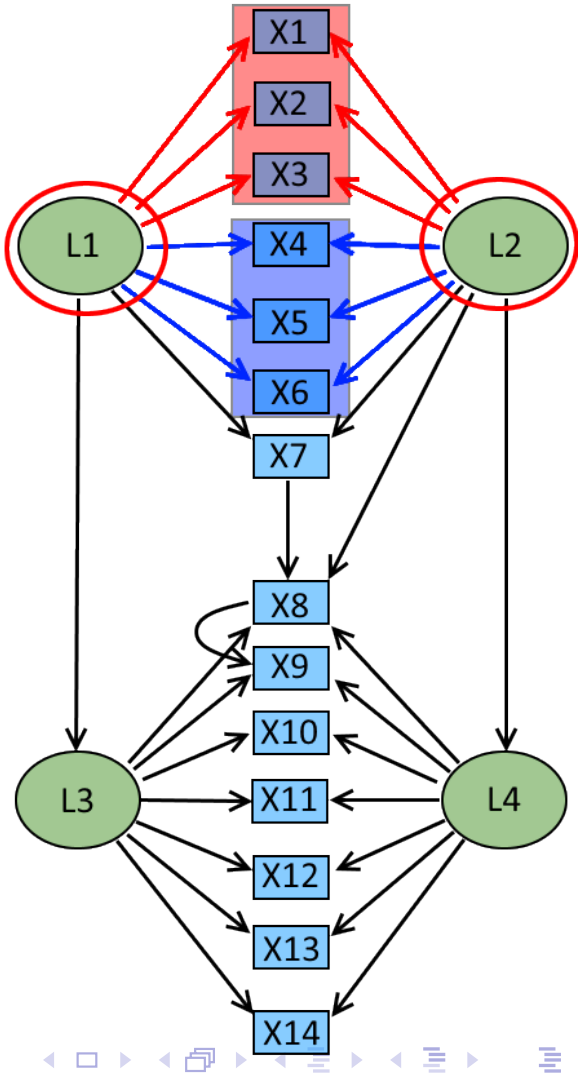
Trek Separation

What can *t*-separate
{X1, X2, X3} from {X4, X5, X6}?

All treks from red to blue must pass across a red edge and a blue edge.

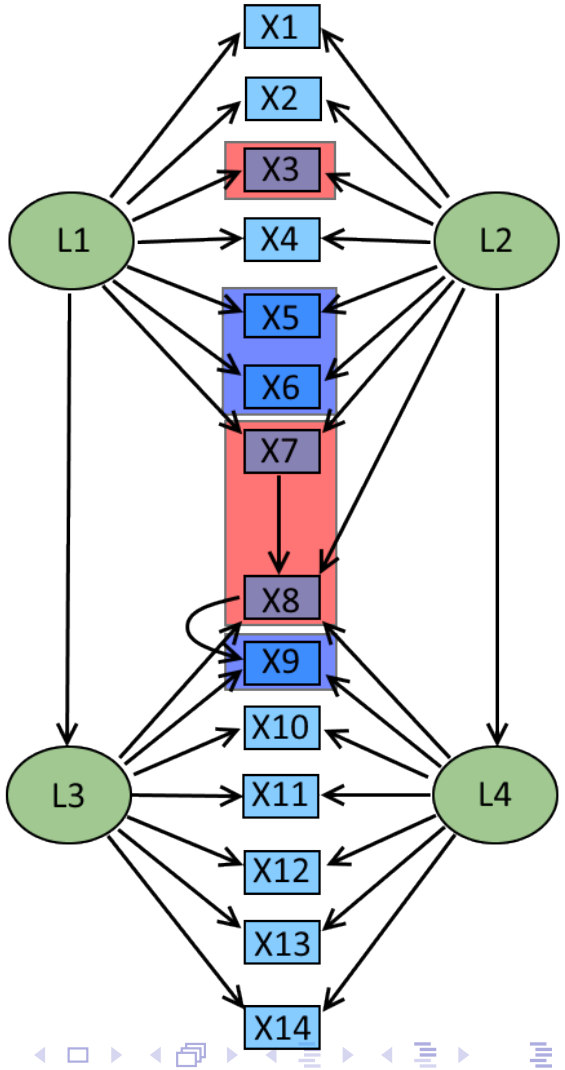
All red, blue edges cross L1 or L2.

{{L1, L2},{}} *t*-seps red from blue, is of size 2+0=2. Determinant of covariance submatrix {X1, X2, X3} by {X4, X5, X6} is entailed to vanish



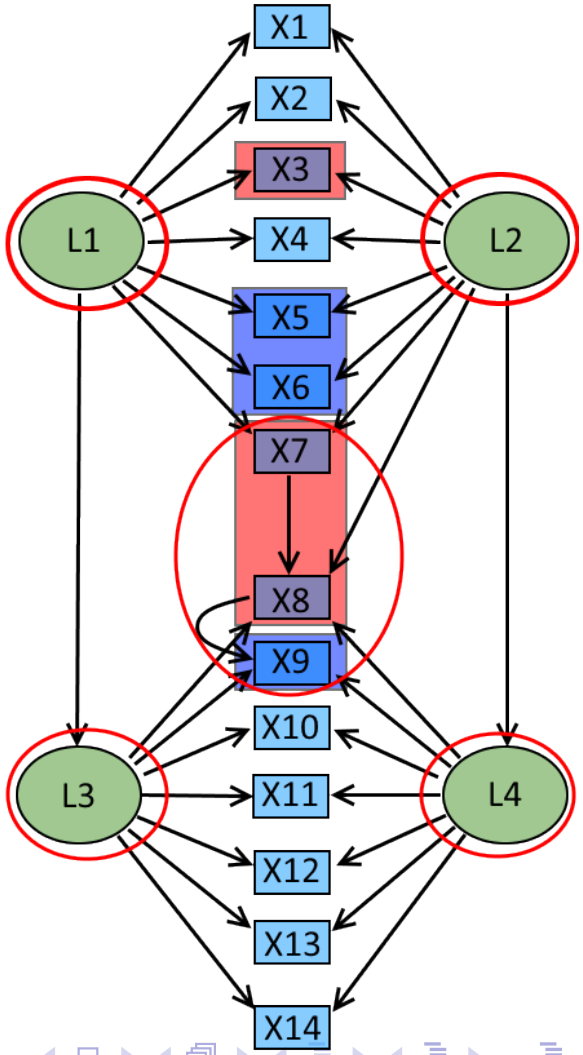
Trek Separation

- Multiple cross-cluster measures
- Measures with edges to other measures
-
-



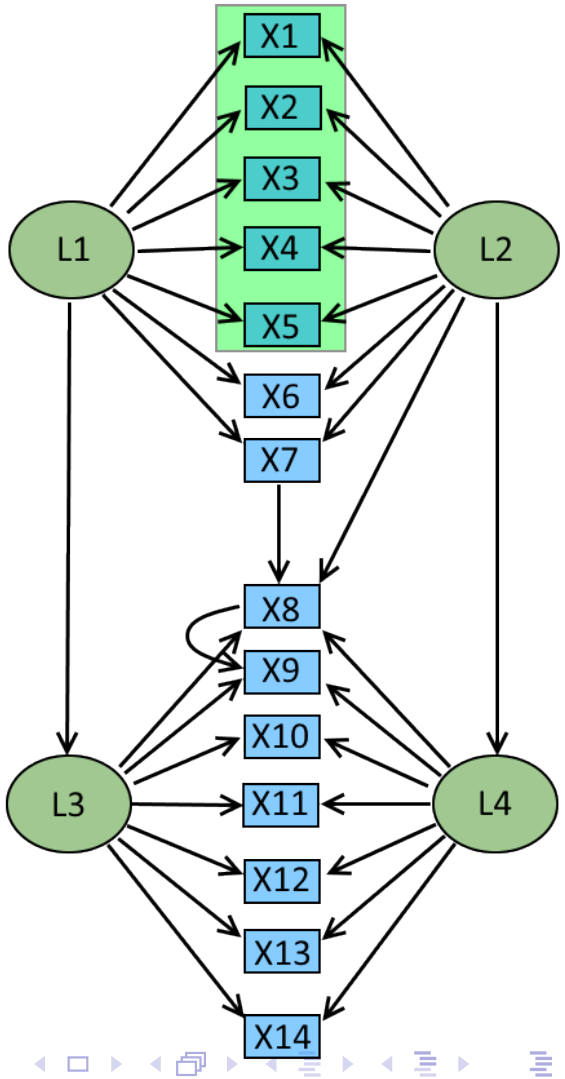
Trek Separation

- Multiple cross-cluster measures
- Measures with edges to other measures
- Requires larger *t*-sep sets
- Sextad not entailed to vanish



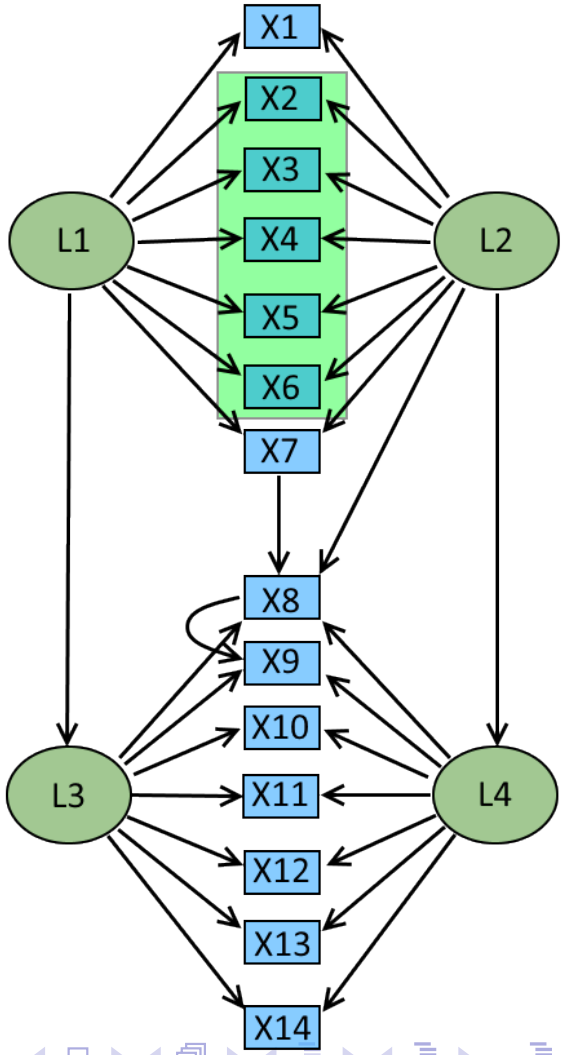
Algorithm intuition

- 5-pure set: If any sextad includes those 5, it's entailed to vanish
- Find 5-pure set
-
-
-
-



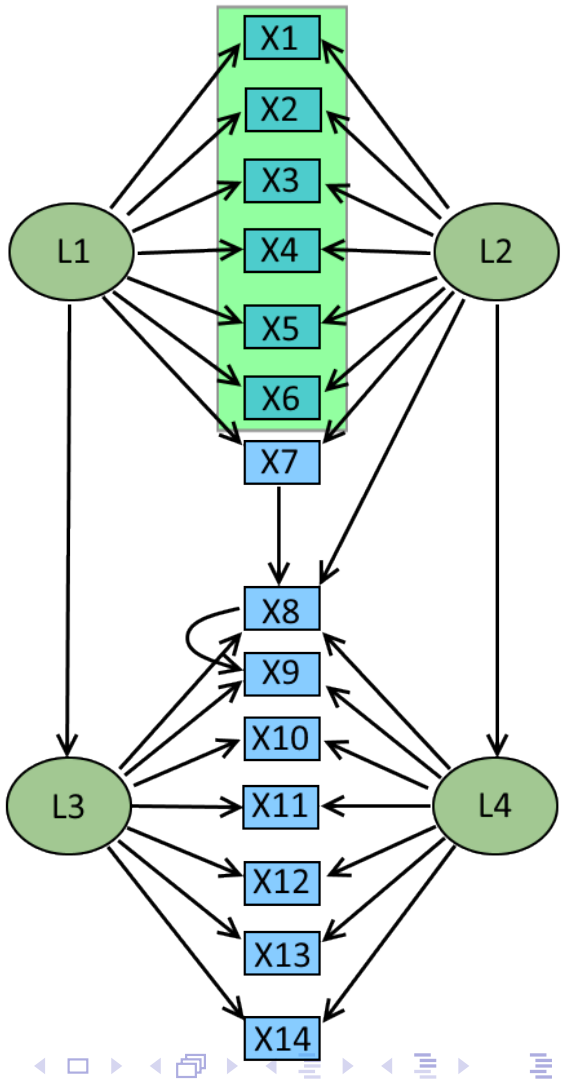
Algorithm intuition

- 5-pure set: If any sextad includes those 5, it's entailed to vanish
- Find 5-pure set
- Find another 5-pure set
-
-
-



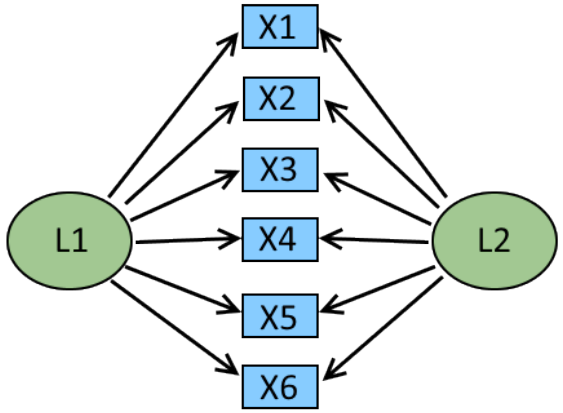
Algorithm intuition

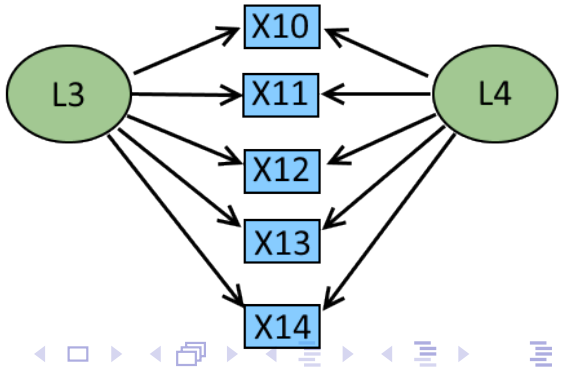
- 5-pure set: If any sextad includes those 5, it's entailed to vanish
- Find 5-pure set
- Find another 5-pure set
- If they overlap, merge them
-
-



Algorithm intuition

- 5-pure set: If any sextad includes those 5, it's entailed to vanish
- Find 5-pure set
- Find another 5-pure set
- If they overlap, merge them
- Variables without 5-pure sets are removed
- Measurement model as output





Algorithm correctness

Summarized theorem of correctness:

Theorem

If [S is an appropriate SEM], then [FTFC] outputs a [set of clusters of measures in S] in which any two variables in the same output cluster have the same pair of latent parents [and no impurities, except one technical detail].

Algorithm correctness

Summarized theorem of correctness:

Theorem

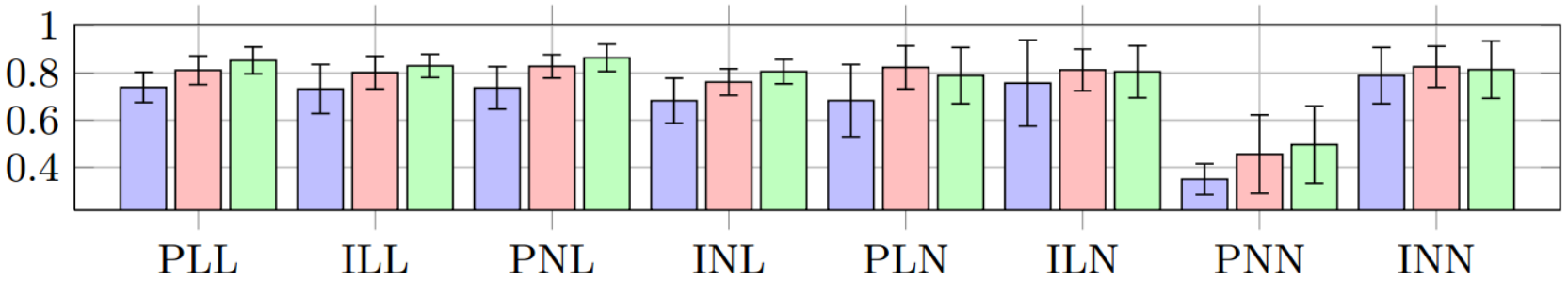
If [S is an appropriate SEM], then [FTFC] outputs a [set of clusters of measures in S] in which any two variables in the same output cluster have the same pair of latent parents [and no impurities, except one technical detail].

Limitations:

- Each measure must be a linear function of its parents
- SEM S must have at least six measures
- Clusters must contain at least 5 measures
- Algorithm is incorrect for specific strange SEM's

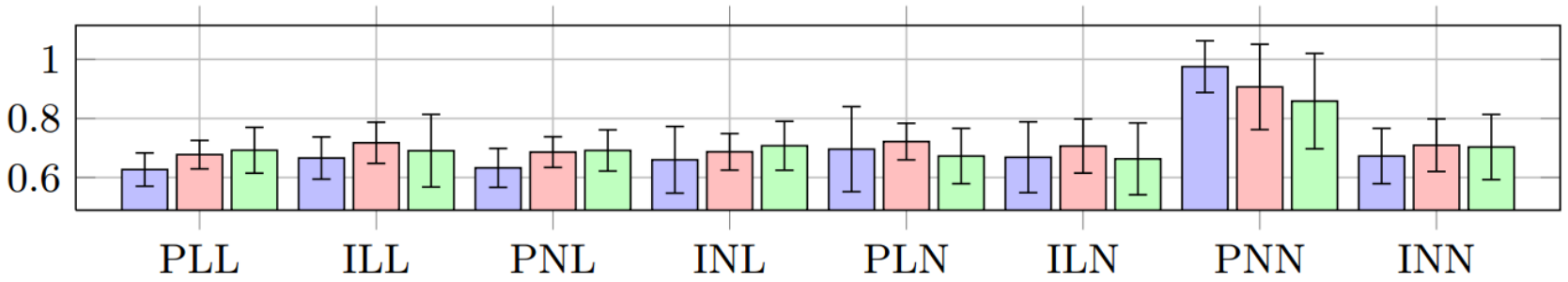
We believe full information tests distinguish problematic SEMs

Simulation Study



- Average precision of output clusters on simulated data
- Proportion of discovered measurement models that are submodels of the data generating model
- Precision is our most important success criteria

Simulation Study



- Average sensitivity (recall) of pure output clusters
- Proportion of measures in pure output cluster to total measures in the containing pure measurement model of data generating model

Real data

Data Set	p	n	$indicators$	$clusters$	$p-value$
Thurstone	9	213	6	1	0.96
Thurstone.33	9	417	5	1	0.52
Holzinger	14	355	7	1	0.23
Holzinger.9	9	145	6	1	0.82
Bechtholdt.1	17	212	8	1	0.59
Reise	16	1000	13	2	0.32

Thanks!

Thanks!

Algorithm correctness

Full version:

Theorem

If a SEM S is a 2-factor model that has a 2-pure measurement sub-model in which each indicator X is a linear function of $L_1(X)$ and $L_2(X)$, S has at least six indicators, and at least 5 indicators in each cluster, then the population FTFC algorithm outputs a clustering in which any two variables in the same output cluster have the same pair of latent parents. In addition, each output cluster contains no more than two impure indicators X_1 and X_2, one of which is on a trek whose source is a common cause of $L_1(X_1)$ and X_1, and the other of which is on a trek whose source is a common cause of $L_2(X_1)$ and X_2.