Causal Clustering for 2-Factor Measurement Models

Erich Kummerfeld, Joseph Ramsey, Renjie Yang, Peter Spirtes, & Richard Scheines

Dept. of Philosophy, Carnegie Mellon University

Problem

- Social scientists interested in variables they cannot directly measure
- Factor models used to relate unobserved variables of interest to measurable indicators
- Existing inference algorithms' output fails tests

Our Strategy

1. Find *pure measurement model* with weak assumptions about the factor model
2. Use *pure measurement model* to learn about the factors (future work)

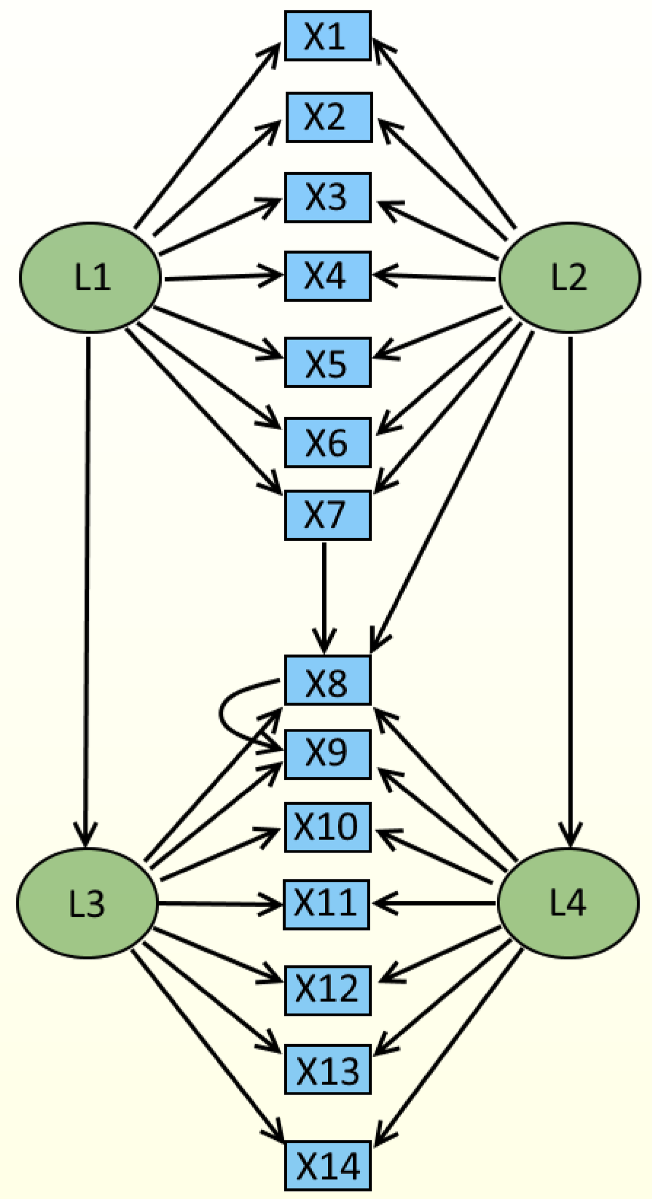
Algorithm: FTFC

FTFC runs three modules in sequence: FindPureClusters, GrowClusters, and SelectClusters.

FindPureClusters: brute force search to find all subsets of **V** of size 5 such that any sextad containing all 5 vanishes.

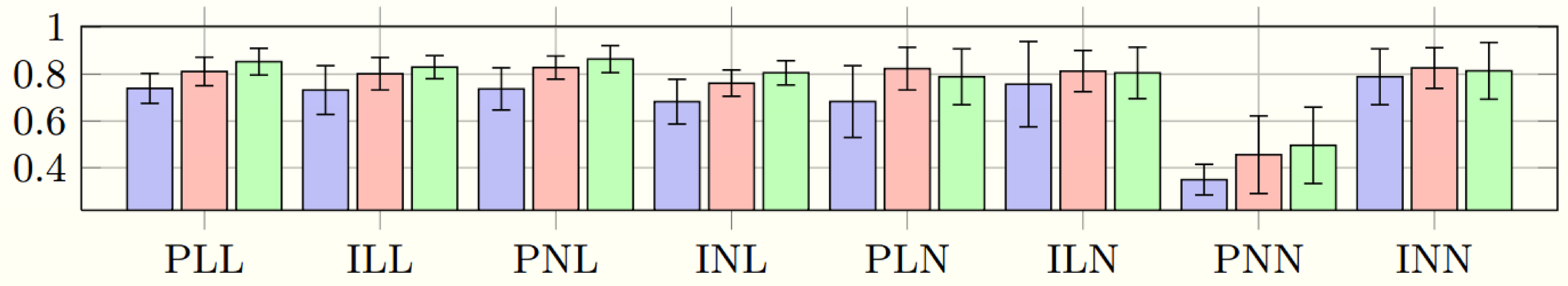
GrowClusters: merge overlapping pure clusters into larger clusters, if the larger clusters are still mostly pure

SelectClusters: choose a maximal set of disjoint clusters from Clusterlist

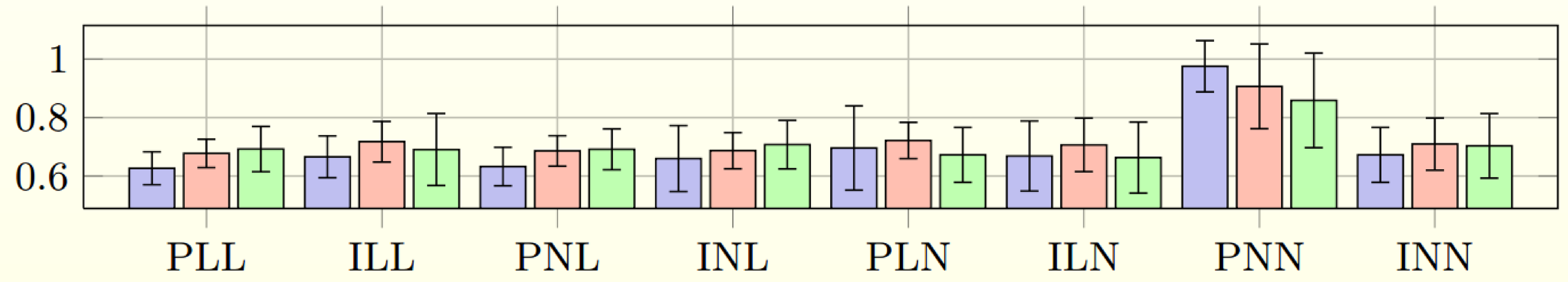


Simulation Results:

Precision

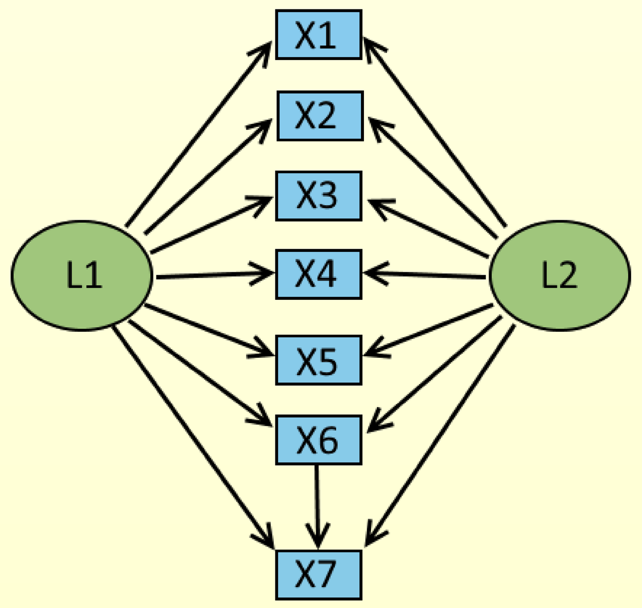


Sensitivity



We value precise clusters over sensitive measures

Trek-separation

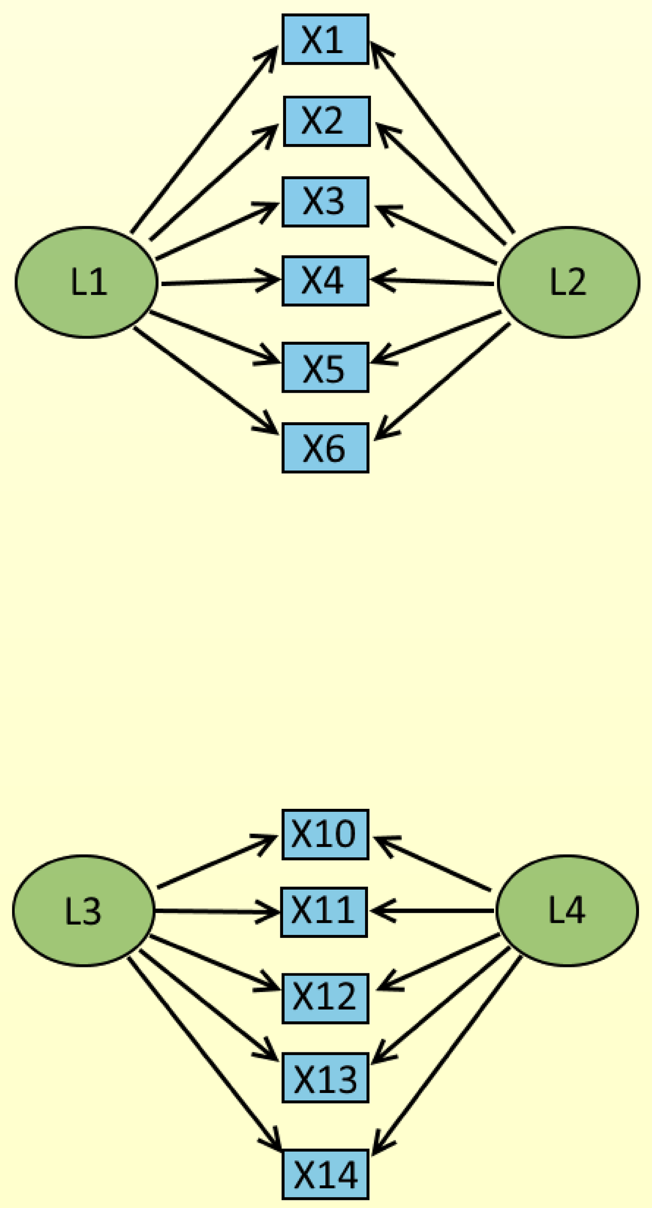


- *t*-separation: graphical generalization of *d*-sep
- *t*-sep set: *ordered pair* of sets of variables
- {{L1,L2},{}} *t*-seps {X1}, {X2}
- {{X6},{}} *t*-seps {X6}, {X7}
- No combination of L1 and L2 can *t*-sep {X6}, {X7}

Size of t-separating set for A and B is bounded above by rank of C(A,B). Rank of C(A,B) can be bounded if det(C(A,B))=0. When |A|=|B|=3, det(C(A,B)) is called a sextad.

A *vanishing sextad* equals 0. $C_{14}C_{25}C_{36} - C_{14}C_{26}C_{35} + C_{24}C_{35}C_{16} - C_{24}C_{15}C_{36} + C_{34}C_{15}C_{26} - C_{34}C_{25}C_{16} = 0$

Find Two-Factor Clusters



Find: {X1, X2, X3, X4, X5} is pure, {X2, X3, X4, X5, X6} is pure, {X1, X2, X3, X4, X10} is not pure. No set containing any of X7, X8, and X9 can be pure.

We are using statistical tests on finite data; GrowClusters increases robustness against noise and violations of faithfulness that induce anomalous, pure clusters in the sample population

We select largest clusters first, removing all other intersecting clusters, repeat.

Left: the output that FTFC converges to on this graph. There are no impurities present, but X7 removed unnecessarily.

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

FTFC finds models with good fit on data available in R

Summary

Advantages of FTFC:

- Does not assume linear factor-factor edges
- Permits impurities in data generating model
- Provably correct under fairly general conditions

Limitations of FTFC:

- Current proof of correctness assumes linear measures
- Computational limits prevent use when >50 measures
- Weird non-measurement models are not distinguished
- Still need to infer structural model
- FTFC removes more measures than is optimal