Fast, Accurate Causal Search Algorithms From the Center for Causal Discovery (CCD)

Gregory Cooper, University of Pittsburgh*; David Danks, Carnegie Mellon University; Joseph Ramsey, Carnegie Mellon University; Peter Spirtes, Carnegie Mellon University; Clark Glymour, Carnegie Mellon University*

Computational procedures for extracting from data causal relationships represented as directed graphs in a variety of scenarios (non-experimental data; no confounding; unknown confounding; sample selection bias; feedback) have been available for very low dimensional problems for two decades. In the last two years the CCD has, among other things, redesigned, parallelized, combined, and re-implemented these procedures so that they produce accurate results for problems with very high dimensions (up to 106 variables) and low sample sizes (103).

Sparse, Gaussian models without latent variables or cycles and with a million variables can now be recovered using the Fast Greedy Search (FGS) algorithm (a redesign and parallelization of the Greedy Equivalence Search method) to datasets with 1,000,000 variables with better than 98% precision in less than 18 hours using Pittsburgh Supercomputing Center resources. 50,000 variable problems can be solved with comparable precision on a 4-core laptop in less than 15 minutes. These procedures have been used on voxel level fMRI data, which are presumably very dense; to show that enforcing sparsity does not reduce precision. Another innovation, PC-Max provides similar improvements in speed and accuracy for PC, the oldest correct search algorithm for directed acyclic graphs.

The Fast Causal Inference (FCI) algorithm and its variants recover causal relations among measured variables when there may be latent confounders of the measured variables. Using FGS as a preprocessor, FCI has been speeded up and its causal-discovery accuracy much improved.

Current work includes developing and testing improvements of the Cyclic Causal Discovery algorithm, generalizations of FGS and PC-Max to non-Gaussian, non-linear systems, SAT solver search algorithms, under-sampled time series, mixed data types, and a Bayesian algorithm for finding expression pathways.