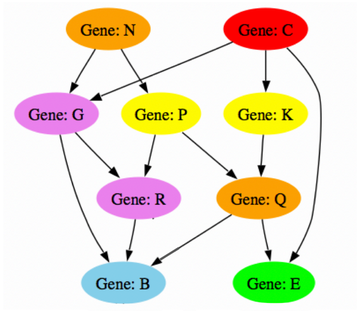
Causal Structure Learning in High Dimensions

Arjun Sondhi; Ali Shojaie
Department of Biostatistics, University of Washington

UNIVERSITY of WASHINGTON

Motivation

- Directed acyclic graphs (DAGs) are commonly used to represent causal networks
- Nodes/vertices represent random variables and directed edges represent causal effects



- Given observed data on a large set of variables, we are interested in learning a DAG skeleton (directions removed)

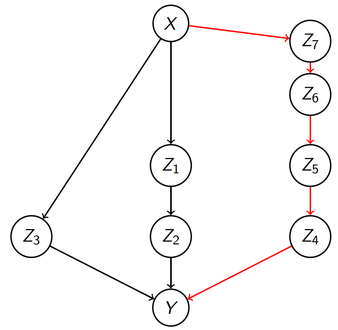
- If we can find a set of nodes Z such that $X \perp\!\!\!\perp Y \mid Z$, then there is no edge between X and Y
- The **PC-Algorithm** (Spirtes et al 1993; Kalisch and Buhlmann, 2007) starts with a complete graph, and searches for separating sets in local neighbourhoods to delete edges
- This becomes computationally infeasible for large graphs

Application: Gene Regulatory Networks

- Applied Reduced-PC and PC-Algorithm to gene expression data from prostate cancer subjects (parameters tuned to maximize BIC)
- Network estimated by Reduced-PC contains more highly-connected hub nodes than that estimated by PC
- A larger proportion of Reduced-PC hub nodes are also identified as hubs in the BioGRID database, and considered to be clinically associated with prostate cancer
- Over the subnetwork of non-hub nodes, Reduced-PC and PC give very similar estimates (F1-score = 0.86)

New Method: Reduced-PC (RPC)

- **Local separation:** for large (random) graphs, there exists only s short paths between any two nodes
  - In the graph below, only only 2 short paths between X & Y



- To find a set Z such that $X \perp\!\!\!\perp Y \mid Z$, we generally need to block all paths between X and Y
- However, correlation between two variables decays over long paths; therefore, we only need to block short paths
- Our algorithm (**Reduced-PC** or **RPC**) searches for separating sets Z, only up to a maximum size s
  - For many random graph families, s is very small – for example, $s = 2$ for Erdos-Renyi and power-law graphs
  - With less restrictive assumptions than PC, RPC consistently estimates the structure of sparse high-dimensional DAGs

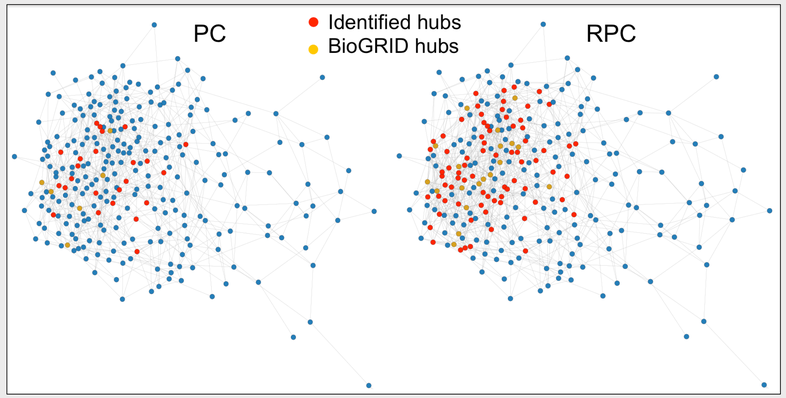


FIGURE 1: Estimated gene regulatory networks in prostate cancer

Simulation Studies

- Simulated data from Erdos-Renyi and power-law graph models; compared RPC and PC algorithms using partial ROC curves
- Both low-dimensional and high-dimensional settings

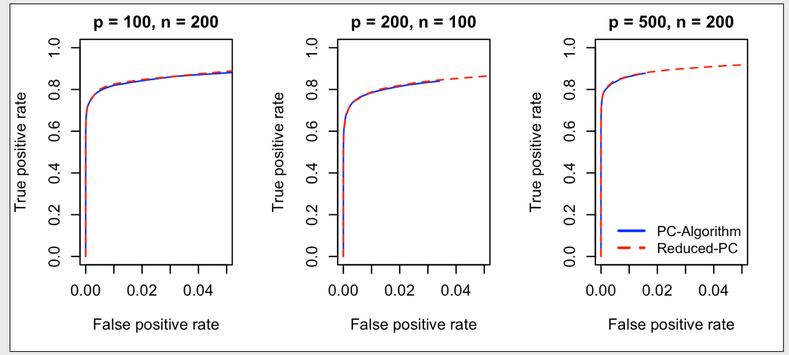


FIGURE 2: Erdos-Renyi graph simulations

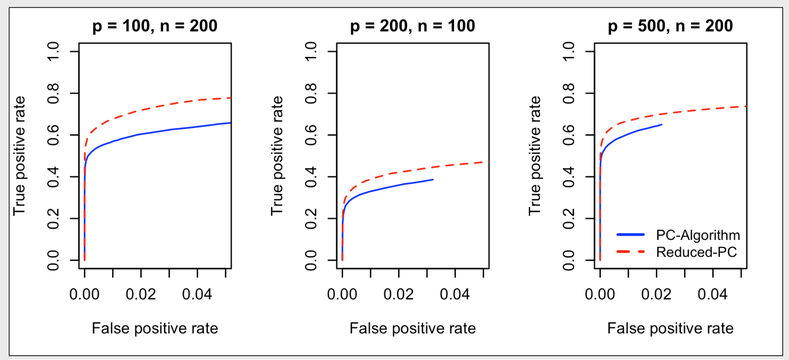


FIGURE 3: Power law graph simulations

Conclusions

- Reduced-PC generally outperforms the PC-Algorithm with lower computational and sample complexity for applicable graph families
- Theoretically consistent under a less restrictive set of assumptions on the underlying probability model
- Future work: nonlinear structural equation models (SEM) and extensions to cyclic graphs

Questions/comments? Please email ashojaie@uw.edu