A SINful approach to Gaussian graphical model selection

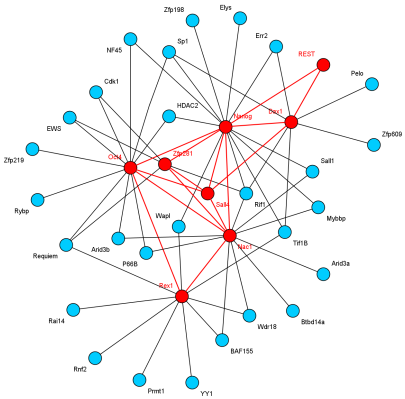
Mathias Drton and Michael Perlman

Adam Gustafson

Stat 572: Methodology Project
University of Washington

May 28, 2013

- In many data sets, characterizing the conditional independencies is of great interest.
- *Gene Regulatory Networks*: collection of DNA segments which interact with each other indirectly through RNA and protein expression; wish to learn network structure.



- Authors' methods: infer structure for Gaussian data.

Graphical Models

- A *graphical model* is a probabilistic model in which conditional independencies are encoded via graphical properties.
- Formally, we have
  - A Graph $G = (V, E)$ where V is set of nodes and $E \subset V \times V$ is a set of edges.
  - If $i, j \in V$, then $(i, j) \in E$ means that nodes i and j are connected (may or may not be directed edge).
  - One-to-one correspondence between nodes and a set of random variables.
  - A graphical model describes a family of distributions $p(x_V)$ over X_V.
  - A rule $r \in R$ is a predicate on a graph: $r(p, G) \in \{\text{true}, \text{false}\}$.
  - The set of distributions which a graphical model describes is

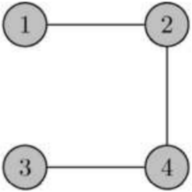
$$\mathcal{F}(G, R) = \{p : p \text{ is a distribution over } X_V \text{ and } r(p, G) = \text{true}, \forall r \in R\}$$

Undirected Graphical Models

- Encode conditional independence properties.
- Authors use *pairwise Markov property* rule:

$$\mathcal{F}(G, R^p) = \{p : X_u \perp\!\!\!\perp X_v | X_{V\setminus\{u,v\}} \quad \text{for all non-adjacent pairs } u, v \in V(G)\}$$

- Ex: $X_1 \perp\!\!\!\perp X_4|(X_2, X_3)$, $X_1 \perp\!\!\!\perp X_3|(X_2, X_4)$, $X_2 \perp\!\!\!\perp X_3|(X_1, X_4)$, etc.



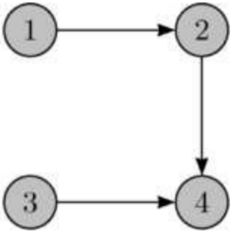
- *Gaussian Case*: $X_u \perp\!\!\!\perp X_v|X_{V\setminus\{u,v\}} \iff K_{uv} = 0$, where $K = \Sigma^{-1}$ is the *precision* or *concentration* matrix.

Directed Acyclic Graphical Models

- Conditional independence properties for *directed acyclic graphs* (DAGs) can be stated in terms of conditioning on parents, or directed factorization:

$$\mathcal{F}(G, R^{df}) = \{p : p(x) = \prod_{v \in V} p(x_v | x_{pa(v)})\}$$

- Ex: $p(x_{1:4}) = p(x_1)p(x_2|x_1)p(x_4|x_2,x_3)p(x_3)$



Directed Acyclic Graphical Models 2

- Partial order: renumber nodes such that $u \preceq v$ if $u = v$ or there is a directed path $u \to \cdots \to v$. Not all nodes comparable.
- Total ordering: Can show always possible to create *Well-numbered* ordering from partial order. Oreder satisfies $i \leq j$ whenever $i \preceq j$ – may need to renumber again. Not unique.
- Example from DAG on previous slide: swapping nodes 2 and 3 leads to well-numbering as well.

Directed Acyclic Graphical Models 3

- The *well-numbered pairwise directed Markov property* states:

$$\mathcal{F}(G, R^{wn}) = \{p : V \text{ is well-numbered }, X_u \perp\!\!\!\perp X_v | X_{\{1,\dots,v\}\setminus\{u,v\}} \iff \text{no directed path from } u \text{ to } v,\ u \leq v\}$$

- Equivalent to directed factorization, but allows identifiability of structure.
- Consider the following three Bayesian Networks:
  - $u \rightarrow v \rightarrow w$
  - $u \leftarrow v \leftarrow w$
  - $u \leftarrow v \rightarrow w$
- Can show all three are equivalent in terms of their joint distribution under directed factorization.

Gaussian Graphical Models

- Gaussian data: can be represented as any type of graphical model.
- 0's in the precision matrix $K = \Sigma^{-1}$ correspond to missing links in the undirected graphical model.
- Cholesky decomposition: $K = U'DU$, where U is upper triangular with ones on the diagonal, and D is diagonal.
- Directed factorization: Let $B = (I - U)$. $B_{ij} = 0$ for $j > i$ implies no edge from node j to node i. Can be well-numbered.
- Previous methods:
  - Backwards selection starting with full graph, controlling error at each stage.
  - Example: PC algorithm.
- *Problem*: Overall error rate for false edge inclusion not controlled.

Authors' Method: Asymptotics

- *Gaussian Data*: n samples, p-dimensional likelihood:

  $$L(\mu, K) \propto (\det K)^{n/2} \exp\{-\mathrm{tr}(KX'X/2) + \mu' KX' 1_p - n\mu' K\mu/2\}$$

- *Exponential Family*: canonical paramter $\theta = (K, K\mu)$, canonical statistic $T(X) = (-X'X/2, X'1_p)$.

  $$L(\mu, K) \propto \exp\{\langle T(X), \theta\rangle - [n\theta' K^{-1}\theta/2 - (n/2)\log\det K]\},$$

  where $\langle (S_1, s_1), (S_2, s_2)\rangle = \mathrm{tr}(S_1 S_2) + s_1' s_2$.

- *Maximum Likelihood*: For $n > p$, we have

  $$\widehat{\mu} = X' 1_p / n \qquad \widehat{\Sigma} = \frac{1}{n}(X - \bar{x}' \otimes 1_n)'(X - \bar{x}' \otimes 1_n)$$

- *Delta Method*: MLE asymptotically normal implies that $\widehat{K} = \widehat{\Sigma}^{-1}$ also asymptotically normal. Depends on unknown population covariance.

Authors' Method: Model Selection

- MLE is functionally invariant: MLE for the partial correlations given by sample partial correlations, since a function of sample covariance.
- Sample partial correlations $r_{ij \cdot C(i,j)}$ asymptotically normal.
- Fisher's $\mathcal{Z}$-transform: improves asymptotics; free of population partial correlations.

$$z_{ij \cdot C(i,j)} = \frac{1}{2} \log \left(\frac{1 + r_{ij \cdot C(i,j)}}{1 - r_{ij \cdot C(i,j)}} \right)$$

- *Model Selection*: $p(p-1)/2$ testing problems:

$$H_{ij} : \rho_{ij \cdot C(i,j)} = 0 \quad \text{vs.} \quad H_{ij} : \rho_{ij \cdot C(i,j)} \neq 0$$

  Note: $\mathcal{Z}(\rho_{ij}) = 0 \iff \rho_{ij} = 0$.
- *Undirected Graphical Model*: $C(i,j) = V \setminus \{i,j\}$.
- *Directed Acyclic Graphical Model*: $C(i,j) = \{1, \ldots, j\} \setminus \{i,j\}$.

Authors' Method: Model Selection Continued

- Simultaneous testing: authors derive p-values to control overall rate of false edge inclusion at level α.
- Unadjusted p-values:

$$\pi_{ij} = 2\left(1 - \Phi\left(\sqrt{n_{C(i,j)} - 3} \cdot \left|z_{ij \cdot C(i,j)}\right|\right)\right).$$

- Bonferroni correction:

$$\pi_{ij}^{\text{Bonf}} = \min\left[\binom{p}{2}\pi_{ij},\, 1\right], \quad 1 \le i < j \le p$$

- Bonferroni too conservative: improve with Holm's step-down procedure to find adjusted p-values.
- Estimation procedure for estimating presence of edge e_{ij} given adjusted p-values π_{ij}^*:

$$\widehat{e}_{ij}(\alpha) = \begin{cases} 0, & \pi_{ij}^* \ge \alpha \\ 1, & \pi_{ij}^* < \alpha. \end{cases}$$

Authors' Method: Guarantees

- Distribution *faithful* to graph if all of its conditional independencies encoded by the graph. In general, distributions encode more conditional independencies than the graph.
- Undirected (pairwise) Gaussian case: minimal number of edges.
- Directed Acyclic Graph case: well-numbering guarantees faithfulness.
- $1-\alpha$ 'consistency' result:

$$\liminf_{n\to\infty} \Pr(\widehat{G}(\alpha) = G_{\text{faithful}}) \geq 1-\alpha$$

- False edge inclusion result (not necessarily tight):

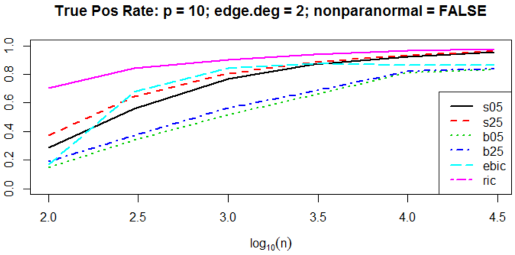
$$\limsup_{n\to\infty} \Pr(\widehat{G}(\alpha) \not\subseteq G)) \leq \alpha.$$

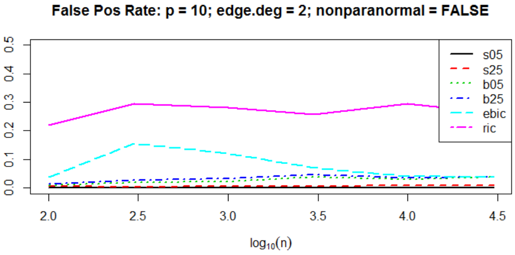
- Asymptotic result: for finite n, may have low power, may not control rate of false edge inclusion, or both.

Experiment: Undirected case

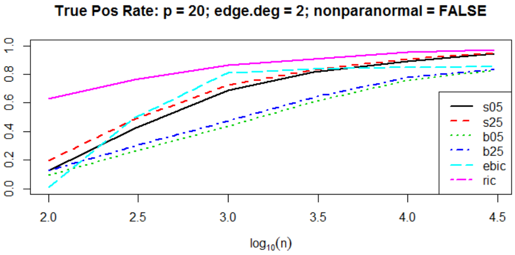
- Simulate random undirected graph with various expected probabilities and various p
- Sample from multivariate normal or nonparanormal (random sign transformation) with various n
- Record the true positive rate and false positive rate
- Compare with graphical lasso and three parameter selection criteria:
    - λ_α in Banerjee et. al. (2008) such that estimated graph satisfies:
      $$\limsup_{n\to\infty} \Pr(\widehat{G}(\alpha) \nsubseteq G)) \leq \alpha.$$
    - "Extended" Bayesian Information Criterion from `HUGE` package.
    - Rotation Information Criterion from `HUGE` package.

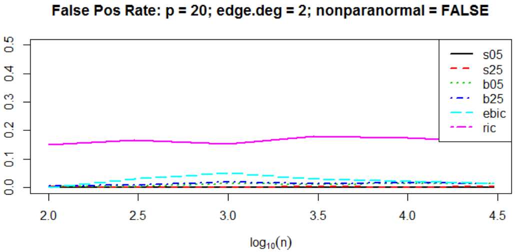
Experiment: Undirected case



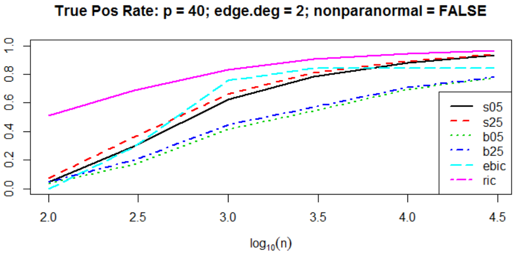


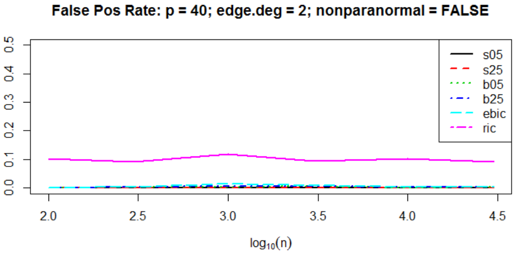
Experiment: Undirected case



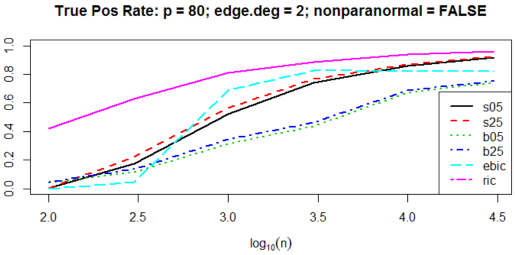


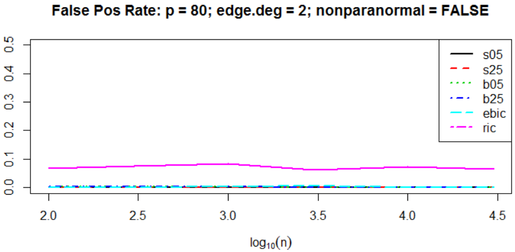
Experiment: Undirected case



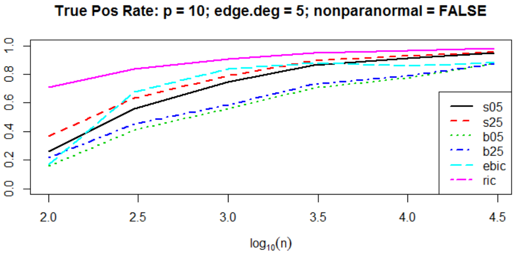


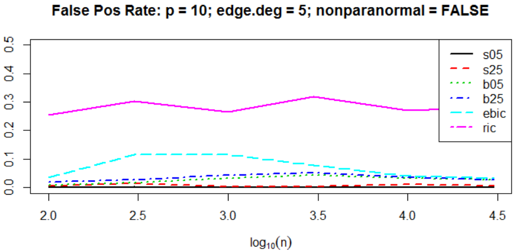
Experiment: Undirected case



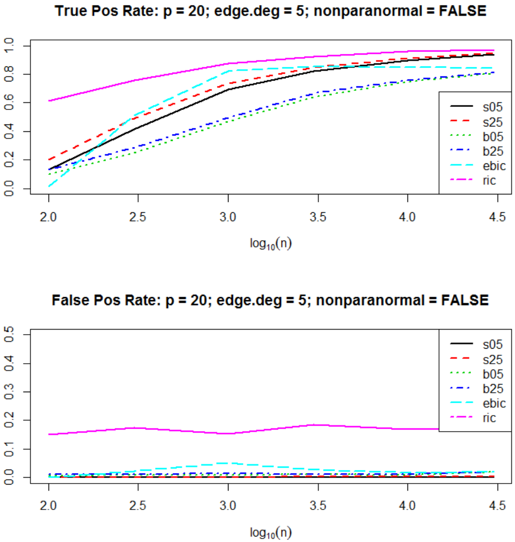


Experiment: Undirected case

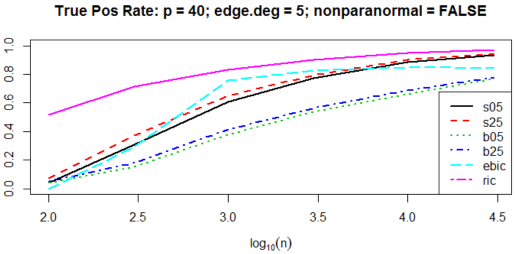


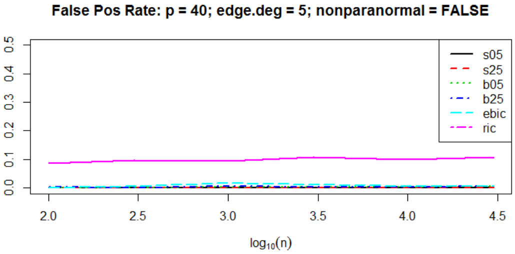


Experiment: Undirected case

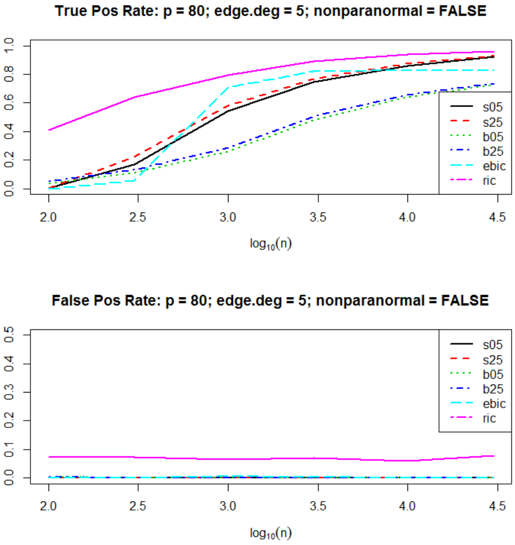


Experiment: Undirected case

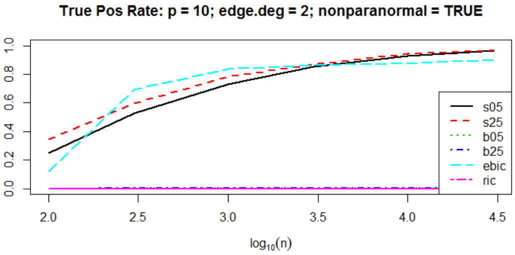


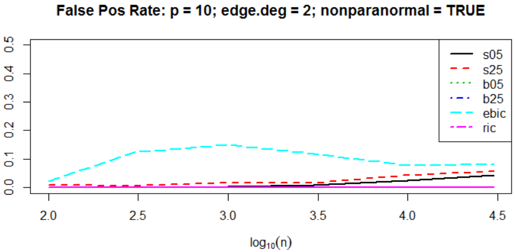


Experiment: Undirected case

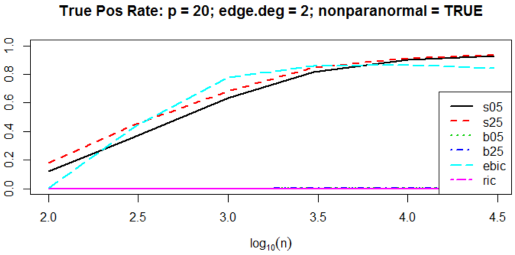


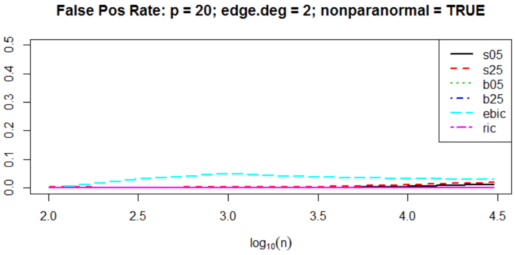
Experiment: Undirected case



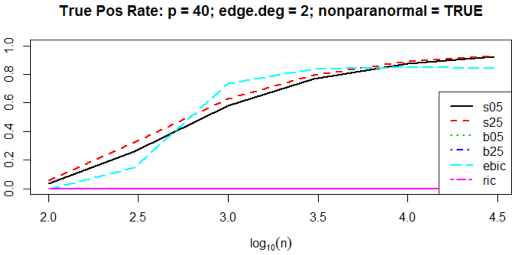


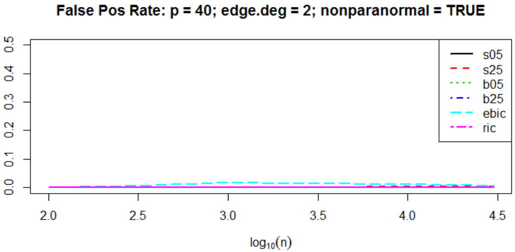
Experiment: Undirected case



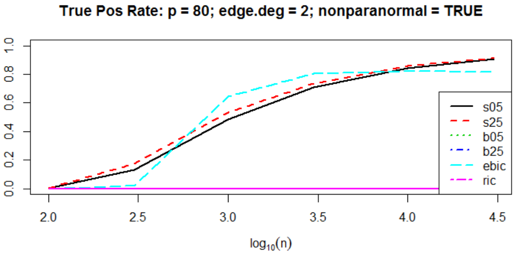


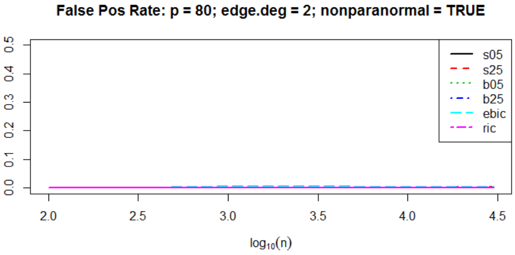
Experiment: Undirected case



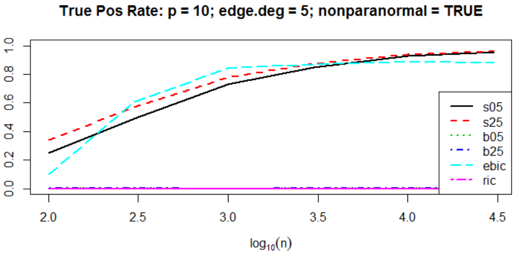


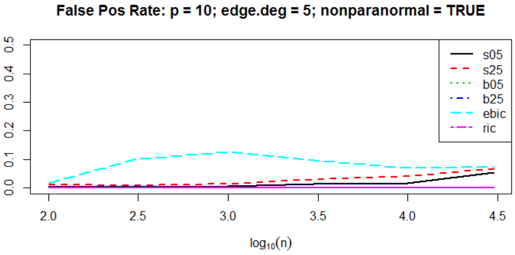
Experiment: Undirected case



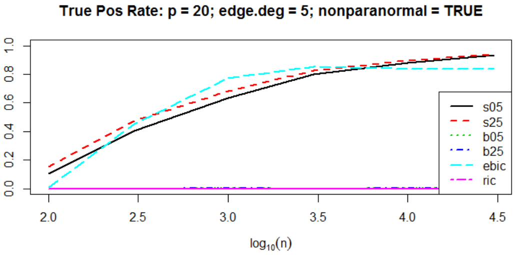


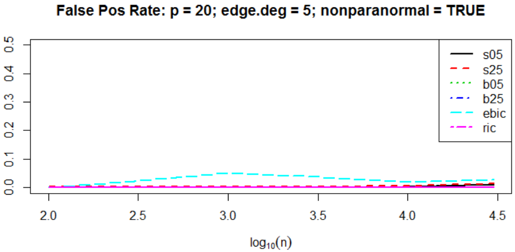
Experiment: Undirected case



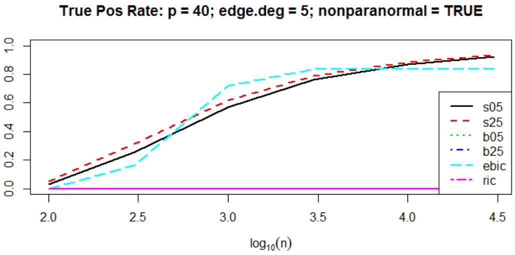


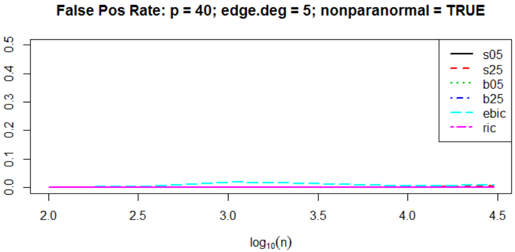
Experiment: Undirected case



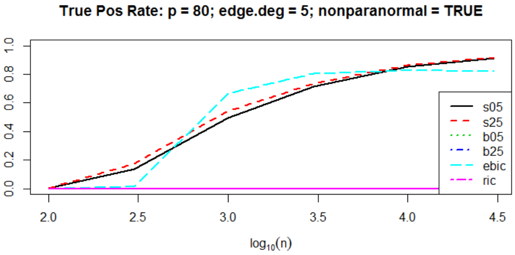


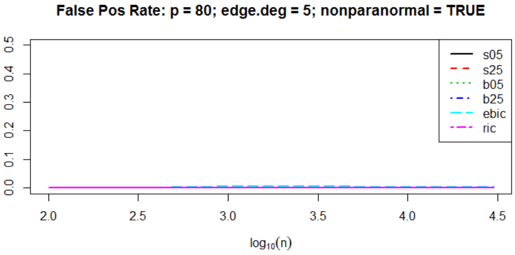
Experiment: Undirected case





Experiment: Undirected case

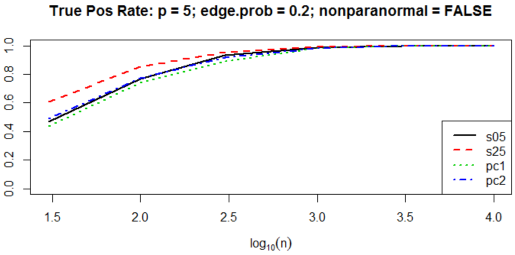


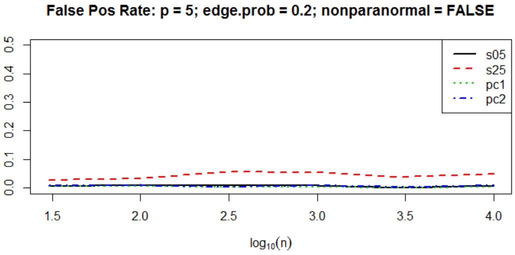


Experiment: Directed case

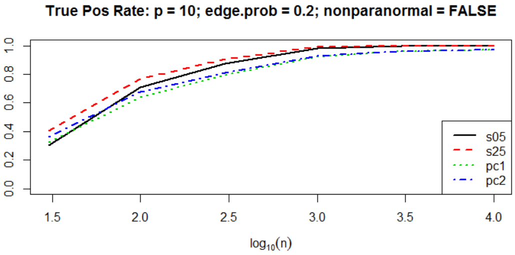
- Simulate a random DAG as follows:
  - The p nodes are arbitrarily well-ordered. Let number of nodes with higher order be K.
  - For each node, number of neighboring nodes is $\mathrm{Bin}(K, prob.)$.
- Simulate random graph with various expected probabilities and various p
- Sample from multivariate normal or nonparanormal (random sign transformation) with various n
- Record the true positive rate and false positive rate of the skeleton (graph with edges dropped).
- Compare with the PC algorithm with parameters (1) 0.005 and (2) 0.01

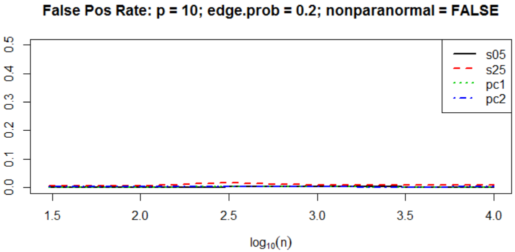
Experiment: Directed case



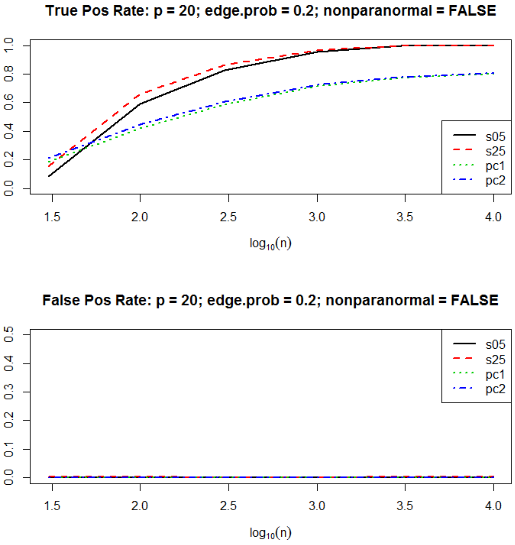


Experiment: Directed case

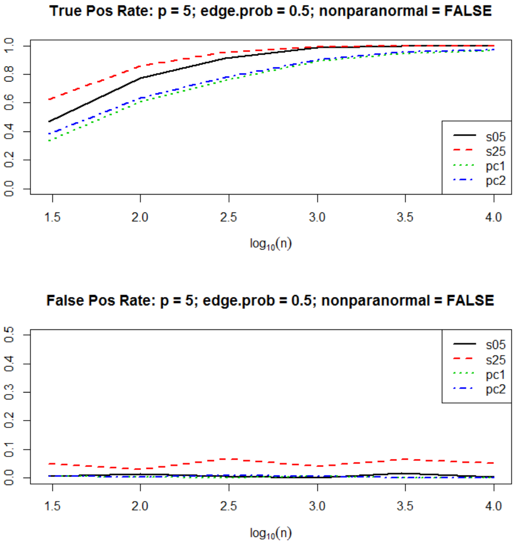




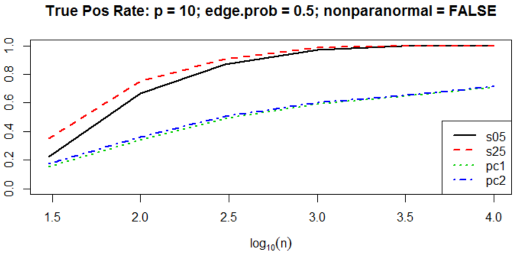
Experiment: Directed case

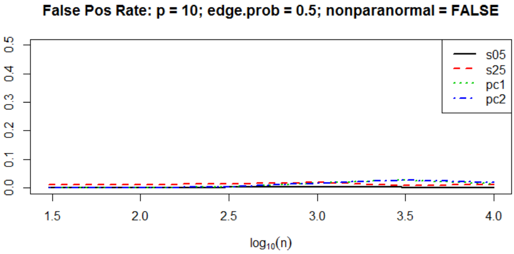


Experiment: Directed case

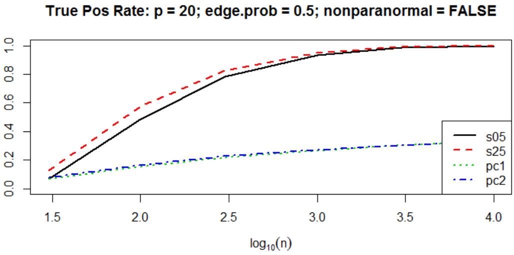


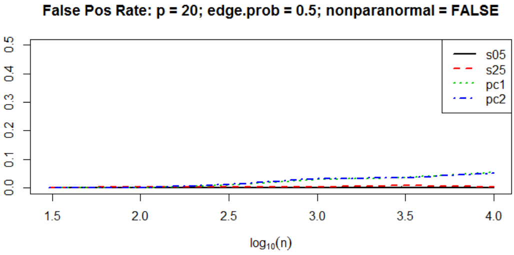
Experiment: Directed case



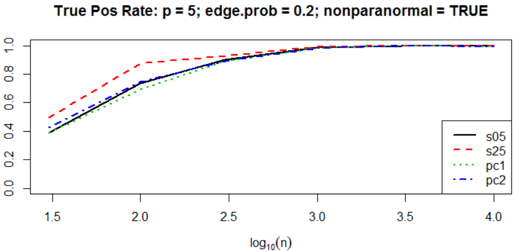


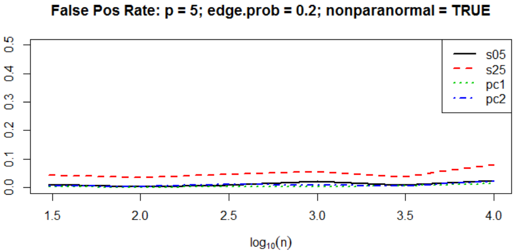
Experiment: Directed case



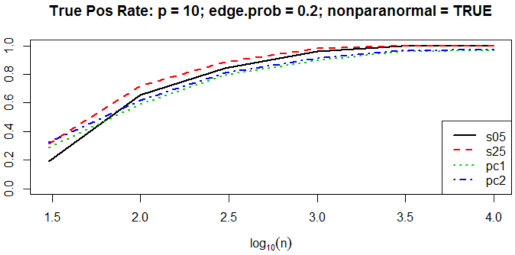


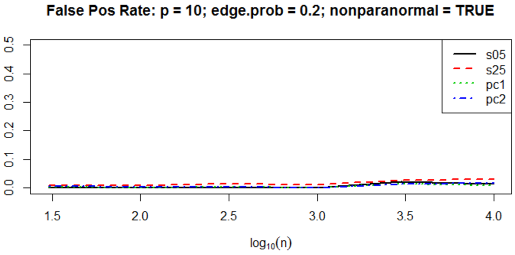
Experiment: Directed case



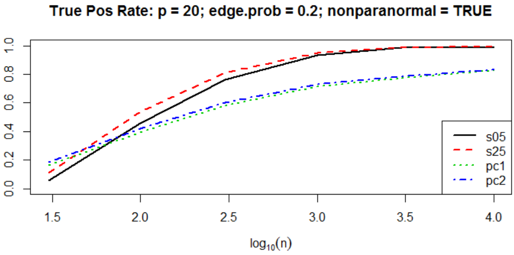


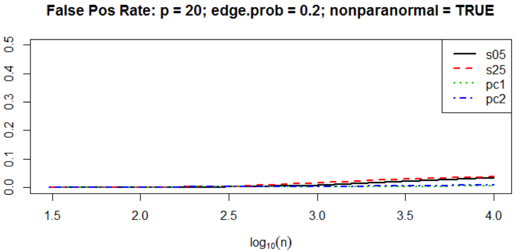
Experiment: Directed case



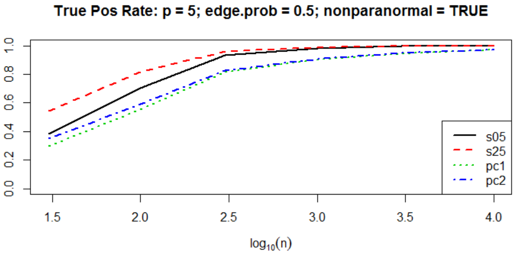


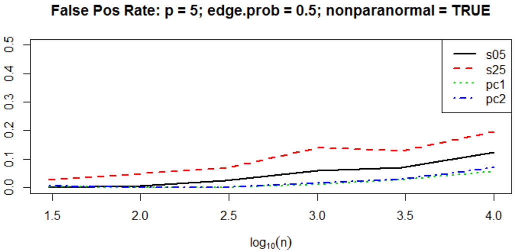
Experiment: Directed case



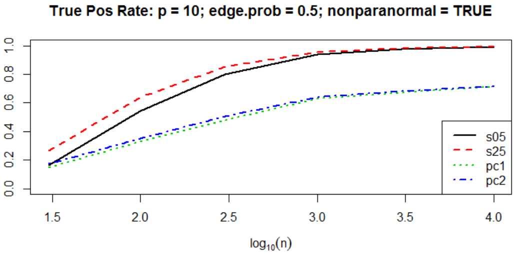


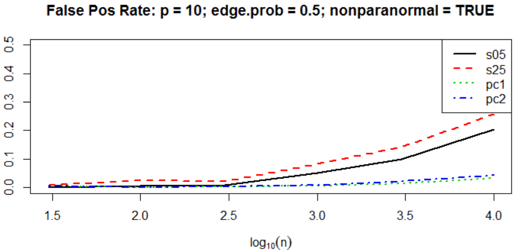
Experiment: Directed case



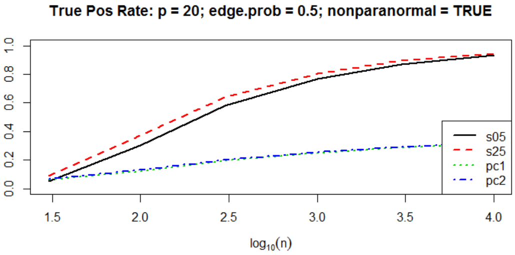


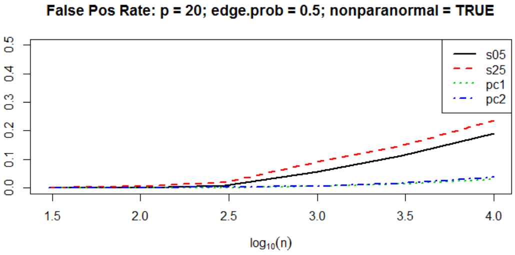
Experiment: Directed case





Experiment: Directed case





Comments and Conclusions

- Method is simple in terms of derivation and implementation.
- Method performs as well or nearly as well as other methods I tried.
- Method is robust to non-Guassian data with same conditional independence structure.
- Method does not extend to $n < p$ case.
- Method requires a priori knowledge of the well-numbering for directed graphs.