

Estimating high-dimensional directed acyclic graphs with the PC-algorithm

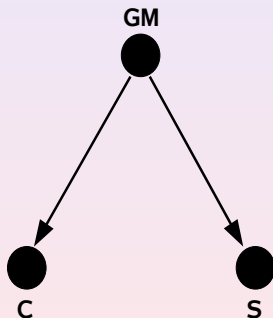
Markus Kalisch

Seminar für Statistik, ETH Zürich, Switzerland

Overview

- 1 DAG and its skeleton
- 2 PC-algorithm
- 3 Consistency
- 4 Simulation
- 5 Application

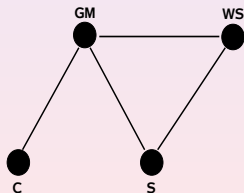
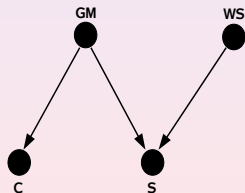
Directed Acyclic Graphs (DAGs)



- Nodes: Random Variables
- Edges: Some Dependence
- Recursive factorization:
 $f(GM, C, S) = f(GM)f(C|GM)f(S|GM)$
- We assume Multivariate Normal Distribution

Directed Global Markov Property

- DAG implies conditional independence relations
- $C \perp S | GM \iff C, S$ are separated by GM in $(G_{An(C \cup S | GM)})^m$



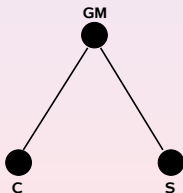
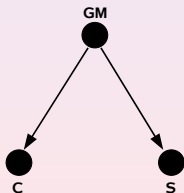
- Ancestral set
- Moralize
- Drop directions

Faithfulness

Conditional independence relations implied by DAG
=
Conditional independence relations of distribution

Skeleton of a DAG

- Ignore directions of arrows
- Edge between two nodes A and $B \iff A, B$ are dependent given every subset of remaining nodes



The PC-algorithm for finding a DAG

- **Finding the skeleton:**

Form complete graph G

$l = -1$

repeat

$l = l + 1$

 repeat

 select (new) ordered pair of adjacent nodes A, B in G

 select (new) neighborhood N of A with size l (if possible)

 if A, B are cond. indep. given N

 save N in \mathbf{N}

 delete edge A, B in G

 until all ordered pairs have been tested

 until all neighborhoods are of size smaller than l

- **Finding the DAG:** The skeleton can be directed using \mathbf{N} and four simple rules.

Sample Version of the PC-algorithm

- Real World: Cond. Indep. Relations $A \perp B|S$ are not known
- Instead: Test for partial correlation $\rho_{AB|S} = 0$ (due to Gaussian assumption)

Therefore:

Remove edge if test for $\rho_{AB|S} = 0$ cannot be rejected for some S on level α .

Consistency: Assumptions

n : Number of samples, p : Number of nodes

- Multivariate Normality, Faithfulness
- Nodes: $p_n = O(n^a)$ $0 \leq a < \infty$ (**high-dimensional**)
- Max number of neighbors is $O(n^{1-b})$ $0 < b \leq 1$ (**sparse**)
- Bounded partial correlations ($0 < d < \frac{b}{2}$):
 $\inf\{|\rho_{ij|\mathbf{k}}|; \rho_{ij|\mathbf{k}} \neq 0\} \geq c_n, c_n^{-1} = O(n^d)$ (**larger than $\frac{1}{\sqrt{n}}$**)
 $\sup\{|\rho_{ij|\mathbf{k}}|\} \leq M < 1$

Consistency: Main Result

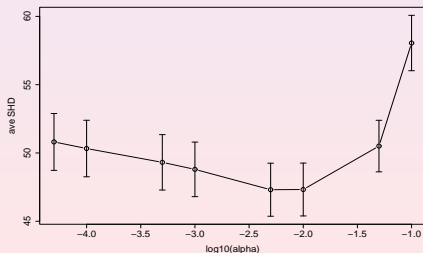
Under these assumptions:

There exists some $\alpha_n \rightarrow 0$ ($n \rightarrow \infty$) so that

$$P(\text{estimated DAG} = \text{true DAG}) = 1 - O(\exp(-Cn^{1-2d})) \rightarrow 1 \\ (n \rightarrow \infty) \quad \text{for } 0 < C < \infty$$

Choice of α

- Structural Hamming Distance (SHD) measures distance between estimated and true graph.
- Over a wide range of parameters the average SHD is minimized for significance levels between $\alpha = 0.005$ and $\alpha = 0.001$.
- In practice: Either choose default values for α or generate priority list of edges



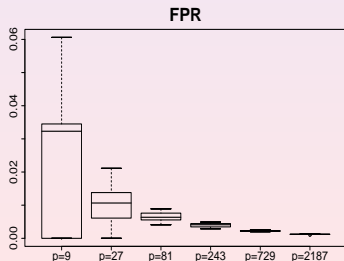
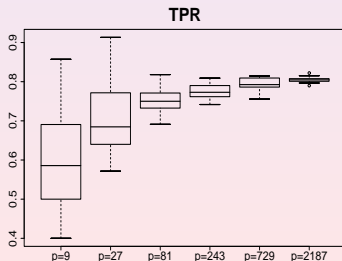
Performance

Computing Time: $p = 1000, n = 1000, E[N] = 8 \rightarrow t \sim 1h$

Estimation:

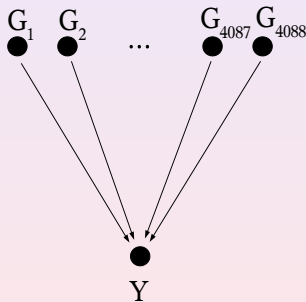
- Number of variables p increases exponentially
- Number of samples n increases linearly
- Expected size of neighborhood $E[N] = \sqrt{n}$ increases sublinearly

Then: TPR increases, FPR decreases



Application

Production of Riboflavin (Vitamin B_2) in *Bacillus Subtilis*

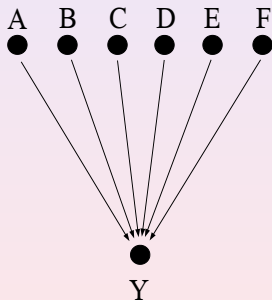


- **Goal:** Maximize output of Riboflavin Y by manipulating genes
- Data obtained by Affymetrix *B. subtilis* GeneChips from DSM Nutritional Products
- Number of Variables $p = 4088$, number of samples $n = 50$

Which genes have an influence on Y ?

Application 2

Result



- Small number of stable candidates extracted
- They are a subset of genes found with other techniques (Lasso, Elastic Net, . . .)
- Findings promising from a biological point of view
- Experimental testing in progress

Conclusion

- DAG, Skeleton, Dependence
- PC-algorithm finds true DAG/skeleton consistently (under some assumptions)
- PC-algorithm is fast for sparse graphs
- More information:
M. Kalisch and P. Bühlmann
Estimating High-Dimensional Directed Acyclic Graphs with the PC-algorithm
JMLR 8 (2007)
- R-package `pcalg` for the PC-algorithm (including robust version)