Causal Inference meeting 1

The max-min hill-climbing Bayesian network structure learning algorithm

Tsamardinos - Brown - Aliferis

Sara Taheri

Outline

- **Introduction**
- Background
- The Max-Min Parents and Children (mmpc) algorithm
- Tests of conditional independence and measures of association
- What my research is about
- The Max-Min Hill Climbing algorithm
- Time complexity of the algorithms
- Empirical evaluation on few number of variables
- Empirical evaluation on thousands of variables
- Limitations of the algorithm

Introduction

- **Bayesian Networks** are graphical models that can efficiently represent and manipulate n-dimensional probability distributions (Pearl 1988). This representation has 2 components:

  - **A graphical structure**, or more precisely a DAG, G = (V, E), where the nodes in V={X1, X2,..., Xn} represent the random variables from the problem we are modeling, and the topology of the graph (the arcs in $E \subseteq V \times V$) encodes conditional (in)dependence relationships among the variables.

  - **A set of numerical parameters** (Θ), usually conditional probability distributions drawn from the graph structure: For each variable $Xi \in V$ we have a conditional probability distribution P(Xi |**pa(Xi)**), where pa(Xi) represents any combination of the values of the variables in **Pa(Xi)**, and **Pa(Xi)** is the parent set of Xi in G. The joint probability distribution over V:

$$P(X1, X2,\ldots, Xn) = \prod_{i=1}^{n} P(Xi \mid Pa(Xi))$$

Introduction

- Learning a Bayesian network from observational data is an important problem especially in bioinformatics to find regulatory pathways.
- Learning a Bayesian network is being used for inferring possible causal relations.
- Learning Bayesian network from data is an NP-Hard problem.(Chickering, 1996; Chickering,Meek & Hecherman,2004)
- MMHC is a structure learning algorithm that can learn the structure of network over thousands of variables.
- It first learns the structure (undirected) of the network with an algorithm called MMPC and then orients the edges with a greedy Bayesian-scoring hill climbing search.

Outline

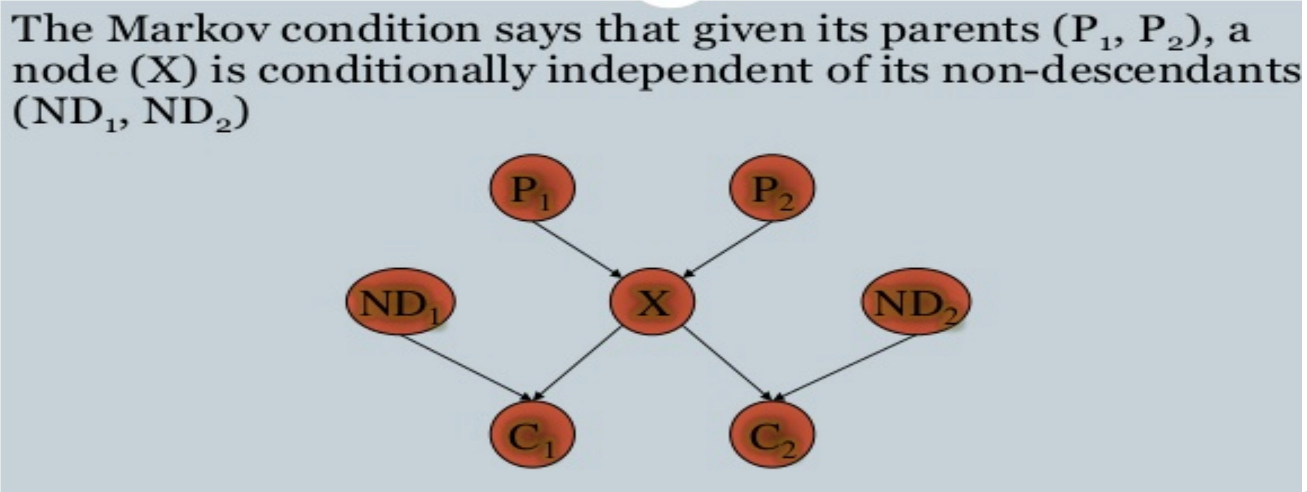
- Introduction
- **Background**
- The Max-Min Parents and Children (mmpc) algorithm
- Tests of conditional independence and measures of association
- What my research is about
- The Max-Min Hill Climbing algorithm
- Time complexity of the algorithms
- Empirical evaluation on few number of variables
- Empirical evaluation on thousands of variables
- Limitations of the algorithm

Background

- **Definition 1.** Two variables X and Y are conditionally independent given **Z** with respect to a probability distribution P, denoted as $Ind_p(X;Y|Z), if\ \forall x, y, \boldsymbol{z}$ where $P(\boldsymbol{Z} = \boldsymbol{z}) > \boldsymbol{0},$

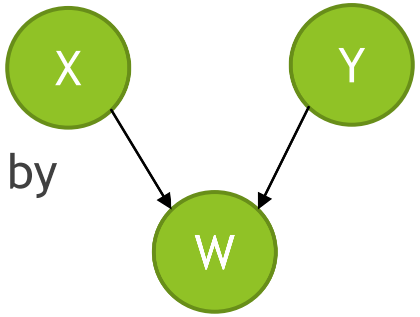
$$P(X = x, Y = y|\boldsymbol{Z} = \boldsymbol{z}) = P(X = x|\boldsymbol{Z} = \boldsymbol{z})P(Y = y|\boldsymbol{Z} = \boldsymbol{z})$$

- **Definition 2.** Let P be a discrete joint probability distribution of the random variables in some set *V* and *G* = <*V*,*E*> be a Directed Acyclic Graph (DAG). We call <G,P> a discrete Bayesian network if <*G*,*P*> satisfies the Markov Condition.



Background

- **Definition 3.** A node W of a path p is a **collider** if p contains two incoming edges into W.



- **Definition 4.** A path p from node X to node Y is blocked by a set of nodes **Z,** if there is a node W on p, which,

  1. W is not a collider and $W \in Z$, or
  2. W is a collider and neither W or its descendants are in **Z** (Pearl,1988)

- **Definition 5.** Two nodes X and Y are **d-separated** by **Z** in graph G if and only if every path from X to Y is blocked by **Z.** Two nodes are ***d-connected*** if they are not d-separated.

- A pair of nodes d-separated by a variable set in network <G,P> is also conditionally independent in P given the set.

Background

- **Definition 6.** If all and only the conditional independencies true in the distribution P are entailed by the Markov condition applied to G, we will say that P and G are ***faithful*** to each other.[1]
- **Definition 7.** A Bayesian network <G,P> satisfies the ***faithfulness condition*** if P embodies only independencies that can be represented in the DAG G. [1] We call such a Bayesian network a ***faithful network***.
- **Theorem 1.** In a faithful BN <G,P> (Pearl,1988),

$$Dsep_G(X;Y|\boldsymbol{Z}) \Leftrightarrow Ind_{\boldsymbol{p}}(X;Y|\boldsymbol{Z})$$

- **Theorem 2.** In a faithful BN <G,P> on variables *V* there is an edge between the pair of nodes X and Y in *V iff* $Dep_p(X;Y|\boldsymbol{Z})$, for all $\boldsymbol{Z} \subseteq V$. [1]
- [1] (Spirtes, Glymour & Scheines, 1993)

Assumptions

- Data set is complete.
- Network is faithful.
- The distribution of the data can be arbitrary.
- The data set is discrete.

Outline

- Introduction
- Background
- **The Max-Min Parents and Children (mmpc) algorithm**
- Tests of conditional independence and measures of association
- What my research is about
- The Max-Min Hill Climbing algorithm
- Time complexity of the algorithms
- Empirical evaluation on few number of variables
- Empirical evaluation on thousands of variables
- Limitations of the algorithm

Max-Min Parents and Children algorithm

- $\boldsymbol{PC_T^G}$: The set of parents and children of node T in a graph G.

- **Definition 9.** We define the minimum association of X and T relative to a feature subset **Z,** denoted as MinAssoc(X;T|**Z),** as

$$MinAssoc(X;T|\boldsymbol{Z}) = \min_{\boldsymbol{S}\subseteq\boldsymbol{Z}}[Assoc(X;T|\boldsymbol{S})]$$

  i.e., as the minimum association achieved between X and T over all subsets of **Z**.

- If $Assoc(X;T|\boldsymbol{S}) = 0$ for any S, then X and T are conditionally independent from each other and there is no edge between them.

$\overline{MMPC}$ Algorithm

Algorithm 1 $\overline{MMPC}$ Algorithm

1: **procedure** $\overline{MMPC}$ $(T,\mathcal{D})$

Input: target variable T; data $\mathcal{D}$

Output: the parents and children of T in any Bayesian network faithfully representing the data distribution

%Phase I: Forward

2: $\mathbf{CPC} = \emptyset$

3: **repeat**

4: $\langle F, assocF \rangle = MaxMinHeuristic(T; \mathbf{CPC})$

5: **if** $assocF \neq 0$ **then**

6: $\mathbf{CPC} = \mathbf{CPC} \cup F$

7: **end if**

8: **until** $\mathbf{CPC}$ has not changed

$\overline{MMPC}$ Algorithm

%Phase II: Backward
for all $X \in \mathbf{CPC}$ **do**
    if $\exists \mathbf{S} \subseteq \mathbf{CPC}$, s.t. $Ind(X; T|\mathbf{S})$ **then**
        $\mathbf{CPC} = \mathbf{CPC} \setminus \{X\}$
    end if
end for

return CPC
end procedure

$\overline{MMPC}$ Algorithm

procedure MaxMinHeuristic(T,**CPC**)

Input: target variable T; subset of variables **CPC**

Output: the maximum over all variables of the minimum association with T relative to **CPC**, and the variable that achieves the maximum

$assocF = \max_{X \in \mathcal{V}} MinAssoc(X; T|\mathbf{CPC})$

$F = \arg\max_{X \in V} MinAssoc(X; T|\mathbf{CPC})$

return $\langle F, assocF \rangle$

end procedure

Reminder:

$$MinAssoc(X; T|\boldsymbol{Z}) = \min_{\boldsymbol{S} \subseteq \boldsymbol{Z}}[Assoc(X; T|\boldsymbol{S})]$$

MMPC algorithm

Algorithm 2 Algorithm *MMPC*

procedure MMPC(T,$\mathcal{D}$)
    CPC $= \overline{MMPC}(T, \mathcal{D})$
    for every variable $X \in$ **CPC do**
        if $T \notin \overline{MMPC}(X, \mathcal{D})$ **then**
            CPC $=$ **CPC** $\setminus X$
        end if
    end for

    return CPC
end procedure

Outline

- Introduction
- Background
- The Max-Min Parents and Children (mmpc) algorithm
- **Tests of conditional independence and measures of association**
- What my research is about
- The Max-Min Hill Climbing algorithm
- Time complexity of the algorithms
- Empirical evaluation on few number of variables
- Empirical evaluation on thousands of variables
- Limitations of the algorithm

Tests of conditional independence $Ind(X_i; X_j|\boldsymbol{X_k})$ for discrete variables

- G^2statistic **Null hypothesis** : Conditional independence holding
- The G^2statistic is defined as:

$$G^2 = 2 \sum_{a,b,\mathbf{c}} S_{ijk}^{ab\mathbf{c}} \ln \frac{S_{ijk}^{ab\mathbf{c}} \quad S_k^{\mathbf{c}}}{S_{ik}^{a\mathbf{c}} \quad S_{jk}^{b\mathbf{c}}}.$$

- $S_{ijk}^{ab\boldsymbol{c}}$: the number of times in the data where $X_i = a, X_j = b, and\ \boldsymbol{X_k} = \boldsymbol{c}.$
- The G^2statistic is asymptotically distributed as χ^2 with degrees of freedom:

$$df = (|D(X_i)| - 1)(|D(X_j)| - 1) \prod_{X_l \in \mathbf{X}_k} |D(X_l)|$$

D(X) is the domain (number of distinct values) of variable X.

- If p-value is less than a significance level α, the null hypothesis is rejected.
- P-value less than α, is considered to indicate zero association.

Tests of conditional independence for continuous variables

- Consider the test $Ind(X_i; X_j|\boldsymbol{X_k})$.
- Null hypothesis : Conditional independence holding.
- The student t test is define as:

$$t(X_i, X_j \middle| \boldsymbol{X_k}) = \rho_{X_i,X_j|S} \sqrt{\frac{p - |\boldsymbol{X_k}| - 2}{1 - (\rho_{X_i,X_j|S})^2}}$$

- $\boldsymbol{\rho_{X_i,X_j|X_k}}$: partial correlation or conditional correlation of X_i and X_j given $\boldsymbol{X_k}$.
- $|\boldsymbol{X_k}|$: total number of variables in $\boldsymbol{X_k}$.
- If there is a subset $\boldsymbol{X_k} \subseteq X \backslash \{X_i , X_j\}$ that $Ind(X_i, X_j \middle| \boldsymbol{X_k})$ is true, there is no edge between X_i and X_j.

Calculating the partial correlation

- Assume $X = (X_1, X_2, \dots, X_p) \sim N(\mu, \Sigma)$
- $S_1 = \{X_i, X_j\} \subseteq X, S_2 \subseteq X \backslash \{X_i, X_j\}$
- - $\Sigma = \begin{bmatrix} \Sigma_{S_1 S_1} & \Sigma_{S_1 S_2} \\ \Sigma_{S_1 S_2} & \Sigma_{S_2 S_2} \end{bmatrix}$

  - Calculate the inverse.

  - Convert it to a correlation matrix.

  - The (i,j) element in that matrix is the conditional correlation of Xi and Xj given
  the rest of variables.

Outline

- Introduction
- Background
- The Max-Min Parents and Children (mmpc) algorithm
- Tests of conditional independence and measures of association
- **What my research is about**
- The Max-Min Hill Climbing algorithm
- Time complexity of the algorithms
- Empirical evaluation on few number of variables
- Empirical evaluation on thousands of variables
- Limitations of the algorithm

Noisy data affects the covariance matrix

- Let's assume that $X' = \left({X_1}', {X_2}', \dots, {X_p}'\right) \sim N(\mu, \Sigma')$ are observed variables.
- $X_i' = X_i + \varepsilon_{X_i}$, where $\varepsilon_{X_i} \sim N(0, \tau_{X_i})$, technical noise, $i \in \{1,2,\dots,p\}$

  τ_{X_i} : Noise level of X_i

- Var(X_i') = Var(X_i) + ${\tau_{X_i}}^2$
- $$\Sigma' = \Sigma + \underbrace{\begin{pmatrix} \tau_{X_1} & 0 & 0 & 0 & 0 & 0 \\ 0 & \tau_{X_2} & 0 & 0 & 0 & 0 \\ 0 & 0 & . & 0 & 0 & 0 \\ 0 & 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & 0 & \tau_{X_p} \end{pmatrix}}_{\Gamma}$$

Cancel the effect of noise

$$\Sigma' = \Sigma + \underbrace{\begin{pmatrix} \tau_{X_1} & 0 & 0 & 0 & 0 & 0 \\ 0 & \tau_{X_2} & 0 & 0 & 0 & 0 \\ 0 & 0 & . & 0 & 0 & 0 \\ 0 & 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & 0 & \tau_{X_p} \end{pmatrix}}_{\Gamma}$$

τ_{X_i} : Noise level of X_i

If we can estimate the noise level of each variable, we can estimate the true covariance matrix :

$$\hat{\Sigma} = \widehat{\Sigma'} - \hat{\Gamma}$$

Outline

- Introduction
- Background
- The Max-Min Parents and Children (mmpc) algorithm
- Tests of conditional independence and measures of association
- What my research is about
- **The Max-Min Hill Climbing algorithm**
- Time complexity of the algorithms
- Empirical evaluation on few number of variables
- Empirical evaluation on thousands of variables
- Limitations of the algorithm

The Max-Min Hill-Climbing algorithm

Algorithm 3 *MMHC* Algorithm

procedure MMHC($\mathcal{D}$)
    Input: data $\mathcal{D}$
    Output: a DAG on the variables in $\mathcal{D}$
    % Restrict
    for every variable $X \in \mathcal{V}$ **do**
        $\mathbf{PC}_X = \text{MMPC}(X, \mathcal{D})$
    end for
    % Search
    Starting from an empty graph perform Greedy Hill-Climbing with operators *add-edge*, *delete-edge*, *reverse-edge*. Only try operator *add-edge* $Y \rightarrow X$ if $Y \in \mathbf{PC}_X$.
    Return the highest scoring DAG found
end procedure

Outline

- Introduction
- Background
- The Max-Min Parents and Children (mmpc) algorithm
- Tests of conditional independence and measures of association
- What my research is about
- The Max-Min Hill Climbing algorithm
- **Time complexity of the algorithms**
- Empirical evaluation on few number of variables
- Empirical evaluation on thousands of variables
- Limitations of the algorithm

Time complexity of the algorithm

- In first phase : in worst case, it calculates the association of every variable with target variable (T) conditioned on all subsets of CPC. $O(|V|.2^{|CPC|})$
- In second phase : it calculates the independence of any variable in the CPC with the target T conditioned on all subsets of the rest of variables in the CPC. $O(|CPC|.2^{|CPC|-1})$

$$\left.\begin{array}{r} O(|V|.2^{|CPC|}) \\ O(|CPC|.2^{|CPC|-1}) \end{array}\right\} \quad O(|V|.2^{|CPC|})$$

Outline

- Introduction
- Background
- The Max-Min Parents and Children (mmpc) algorithm
- Tests of conditional independence and measures of association
- What my research is about
- The Max-Min Hill Climbing algorithm
- Time complexity of the algorithms
- **Empirical evaluation on few number of variables**
- Empirical evaluation on thousands of variables
- Limitations of the algorithm

Evaluation Study

Table 1 Bayesian networks used in the evaluation study

Network	Num. vars	Num. edges	Max In/Out-degree	Min/Max $\|PCset\|$	Domain range
Child	20	25	2 / 7	1 / 8	2–6
Child3	60	79	3 / 7	1 / 8	2–6
Child5	100	126	2 / 7	1 / 8	2–6
Child10	200	257	2 / 7	1 / 8	2–6
Insurance	27	52	3 / 7	1 / 9	2–5
Insurance3	81	163	4 / 7	1 / 9	2–5
Insurance5	135	281	5 / 8	1 / 10	2–5
Insurance10	270	556	5 / 8	1 / 11	2–5
Alarm	37	46	4 / 5	1 / 6	2–4
Alarm3	111	149	4 / 5	1 / 6	2–4
Alarm5	185	265	4 / 6	1 / 8	2–4
Alarm10	370	570	4 / 7	1 / 9	2–4
Hailfinder	56	66	4 / 16	1 / 17	2–11
Hailfinder3	168	283	5 / 18	1 / 19	2–11
Hailfinder5	280	458	5 / 18	1 / 19	2–11
Hailfinder10	560	1017	5 / 20	1 / 21	2–11
Mildew	35	46	3 / 3	1 / 5	3–100
Barley	48	84	4 / 5	1 / 8	2–67
Munin	189	282	3 / 15	1 / 15	1–21
Pigs	441	592	2 / 39	1 / 41	3–3
Link	724	1125	3 / 14	0 / 17	2–4
Gene	801	972	4 / 10	0 / 11	3–5

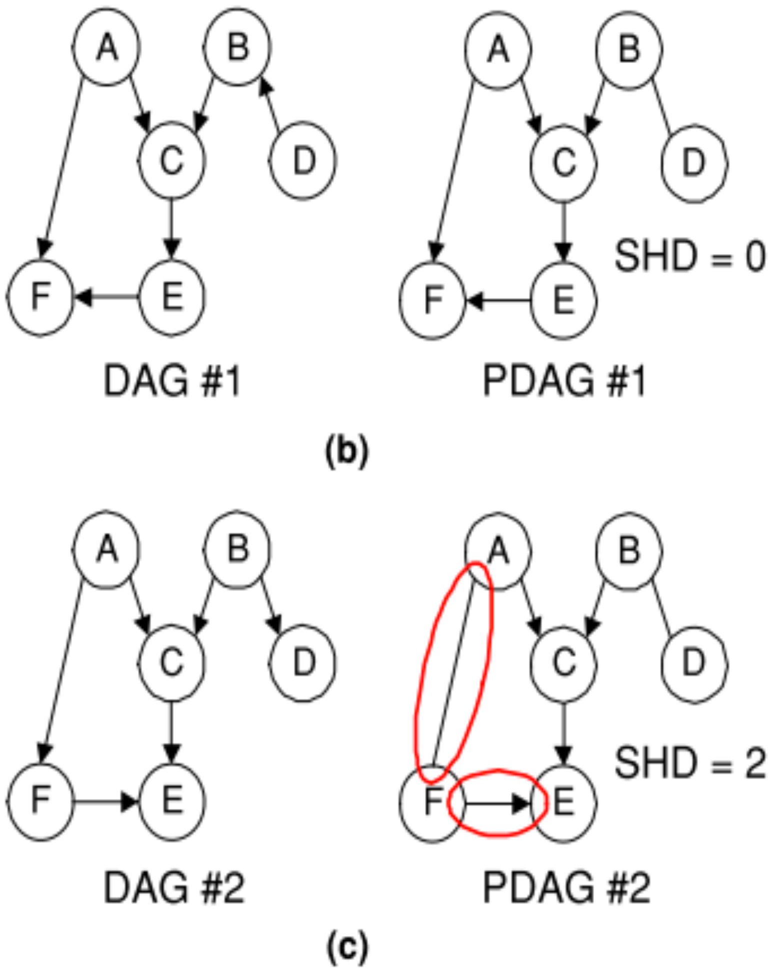
- 4290 Networks
- Using a year’s single-CPU time
- Networks are from medicine, agriculture, weather forecasting, biology,…

Algorithms that were used in the study

- Sparse Candidate (SC)
- PC
- Three Phase Dependency Analysis (TPDA)
- Optimal Reinsertion (OR)
- Greedy Hill Climbing Search (GS)
- Greedy Equivalent Search (GES)
- Max-Min Hill-Climbing (MMHC)

Measures of performance

- 1) Structural **hamming distance**.
- 2) **KL-divergence** : is a measure of how one probability distribution is different from a second probability distribution.
- 3) **Time results**



Average normalized structural hamming distance results

Table 5 Average normalized structural hamming distance results

Algorithm	Sample size (SS) 500	1000	5000	Average over SS
MMHC	1.00 (22)	1.00 (22)	1.00 (22)	1.00
OR1 $k = 5$	1.30 (19)	1.45 (18)	1.70 (17)	1.48
OR1 $k = 10$	1.29 (19)	1.37 (18)	1.78 (16)	1.48
OR1 $k = 20$	1.31 (19)	1.45 (18)	1.86 (16)	1.54
OR2 $k = 5$	1.18 (19)	1.33 (18)	1.66 (16)	1.39
OR2 $k = 10$	1.19 (18)	1.34 (18)	1.65 (16)	1.39
OR2 $k = 20$	1.22 (18)	1.34 (18)	1.71 (16)	1.42
SC $k = 5$	1.13 (21)	1.28 (22)	1.57 (18)	1.33
SC $k = 10$	1.18 (13)	1.28 (13)	1.35 (13)	1.27
GS	1.62 (20)	2.08 (20)	1.86 (20)	1.85
PC	8.85 (18)	10.07 (18)	2.82 (20)	7.25
TPDA	9.63 (21)	10.22 (21)	1.76 (22)	7.21
GES	1.18 (7)	0.94 (6)	1.19 (6)	1.10

Normalized Structural Hamming Distance (*SHD*) is the *SHD* of each algorithm for a particular sample size and network divided by *MMHC*'s *SHD* on the same sample size and network. The term in parentheses is the number of networks the algorithm was averaged across. Average normalized *SHD* values greater than one correspond to an algorithm with more structural errors than *MMHC*.

Outline

- Introduction
- Background
- The Max-Min Parents and Children (mmpc) algorithm
- What my research is about
- Tests of conditional independence and measures of association
- The Max-Min Hill Climbing algorithm
- Time complexity of the algorithms
- Empirical evaluation on few number of variables
- **Empirical evaluation on thousands of variables**
- **Limitations of the algorithm**

Scaling to thousands of variables

- 5000 variables
- 6845 edges
- Sample size 5000
- Running time : 13 days
- Reconstructed network : 1340 extra edges (Specificity 99.9%)
  1076 missing edges (sensitivity 84 %)
  1468 wrongly oriented
- Sensitivity : # of correctly identified edges over total # of edges
  Specificity : # of correctly identified non-edges over total # of non-edges.

Limitations of the algorithm

- The sample sizes analyzed is limited : 500, 1000, 5000 and a smaller set of tests at 20000
  It would be interesting to make adjustments to the algorithm to work for few (100) samples.
- The algorithm requires a network to be faithful, or close to faithful.
- Possible extension to algorithm would be to incorporate statistical tests targeting specific distributions, employing parametric assumptions, or incorporate background knowledge.