A Study on Comparison of Bayesian Network Structure Learning Algorithm for Selecting Appropriate Model

유재성

Dept. of Statistics

November 21, 2014

Title

목표

변수에 대한 정보만 존재하고 그들의 구조를 잘 모를 때 베이지안 네트워크 구조 학습을 진행하게 된다.

다양한 분야에서의 각 이론적 배경에 맞는 "적합한" 모형 선택을 돕기 위해,

베이지안 네트워크의 구조 학습 방법을 선택하기 위한 여러가지 방법을 비교하고자 한다.

Outline

비교 기준

점수를 이용한 비교 BDe, Log Likelihood, AIC, BIC
주의! 점수가 "클수록" 좋다.(*1)

목표 네트워크와 학습된 네트워크와의 직접 비교(*2) 학습된 모형 자체가 실제 기대한 모형과 일치한지 자체에 주목한다.

- C (Correct Arcs) : 목표 네트워크 O, 학습 네트워크 O, 방향 일치
- M (Missing Arcs) : 목표 네트워크 O, 학습 네트워크에 X
- WO (Wrongly Oriented Arcs) : 목표 네트워크 O, 학습 네트워크 O, 방향 불일치
- WC (Wrolgly Connected Arcs) : 목표 네트워크 X, 학습 네트워크 O

(*1) Silvia Acid 외 5명,
"A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service",
Artificial Intelligence in Medicine 30 (2004) 215–232.

(*2) Reference : Fadhl, M. Al-Akwaa, Mohammed M. Ikhawlani, (2012),
"Comparison of the Bayesian Network Structure Learning Algorithms",
International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 3.

어떻게 비교할 것인가?

- 어떤 Algorithm을 적용했는가에 따라
- Topology의 유형에 따라
- Node 개수가 증가함에 따라
- Sample Size가 증가함에 따라

전제

- 각 Node에 대한 Cardinality는 2로 제한한다.
  (즉, 이 모든 Node가 Binary Data인 경우만 다룬다.)
- 이 연구에서 사용하는 구조 학습 알고리즘은 R의 bnlearn 패키지에서 제공하는 것으로 한정한다.
- 모든 실험은 100번 반복하였다. 그 결과들을 받아 종합하였다.
- Synthetic Data로 분석할 때, 확률 관계를 정의할 때, 확률값을 U(0,1) 사이의 값에서 임의로 주었다.

Outline

Asia DataSet

Description Small synthetic data set from Lauritzen and Spiegelhalter (1988) about lung diseases (tuberculosis, lung cancer or bronchitis) and visits to Asia.

Number of nodes 8

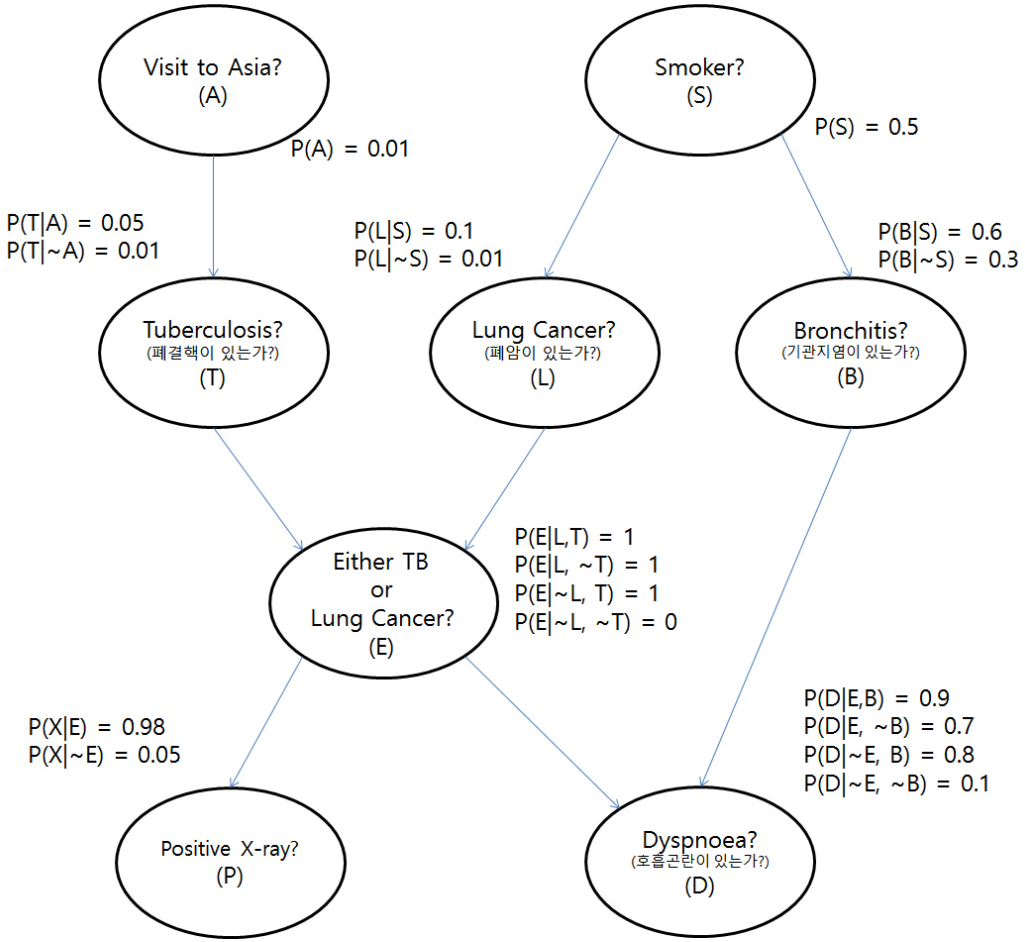
Number of arcs 8

Number of parameters 18

Source Lauritzen S, Spiegelhalter D (1988).
"Local Computation with Probabilities on Graphical Structures and their Application to Expert Systems (with discussion)".
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 50(2), 157-224.

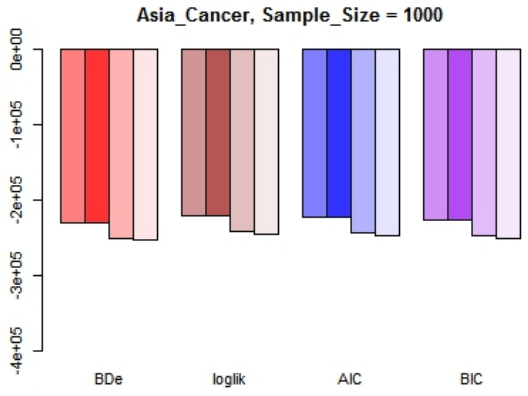
Asia DataSet by Lauritzen and Spiegelhalter
Insurance Evaluation Network DataSet
ALARM Monitoring System DataSet
The HailFinder Weather Forecast System DataSet

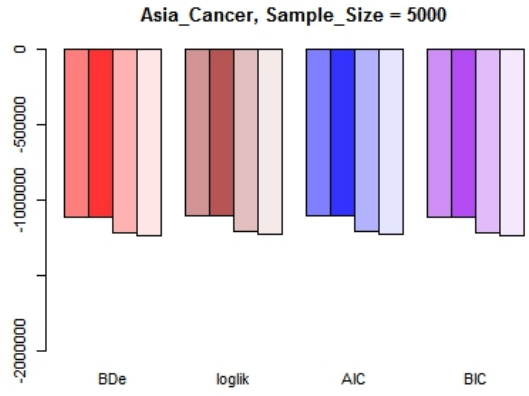
Asia DataSet

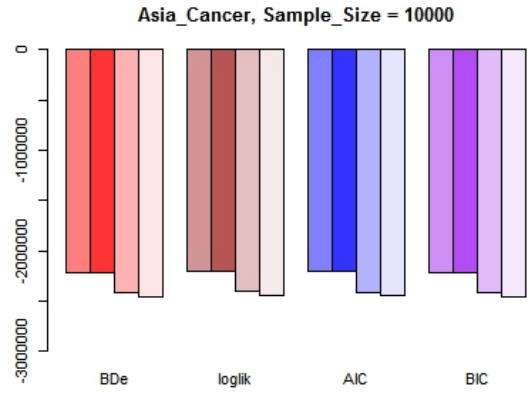


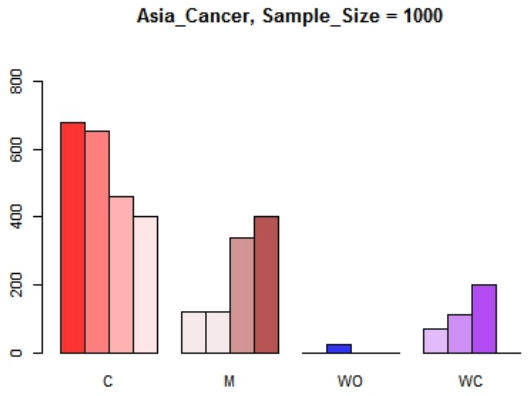
Asia DataSet by Lauritzen and Spiegelhalter
Insurance Evaluation Network DataSet
ALARM Monitoring System DataSet
The HailFinder Weather Forecast System DataSet

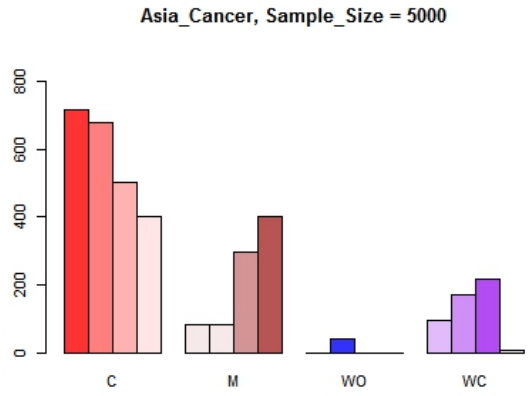
선행연구와 결과 비교 : Asia DataSet

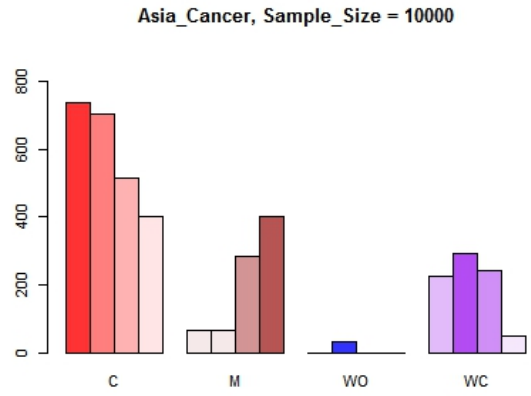












Insurance DataSet

Description Insurance is a network for evaluating car insurance risks.

Number of nodes 27

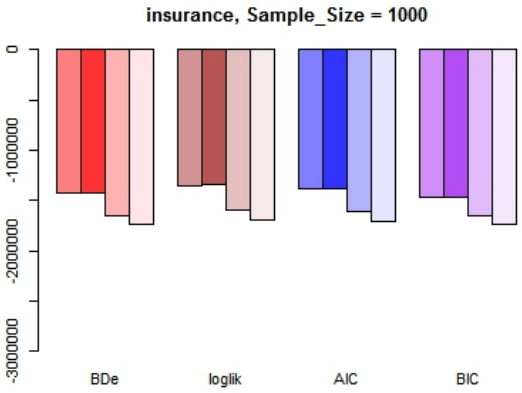
Number of arcs 52

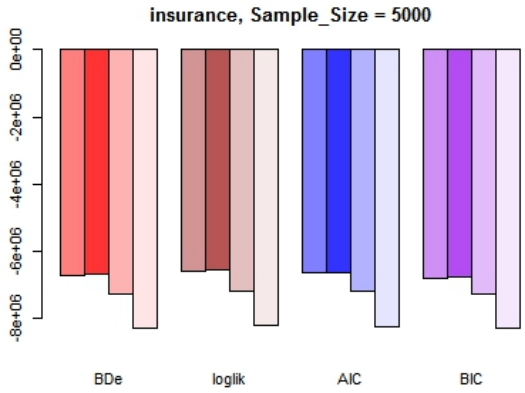
Number of parameters 984

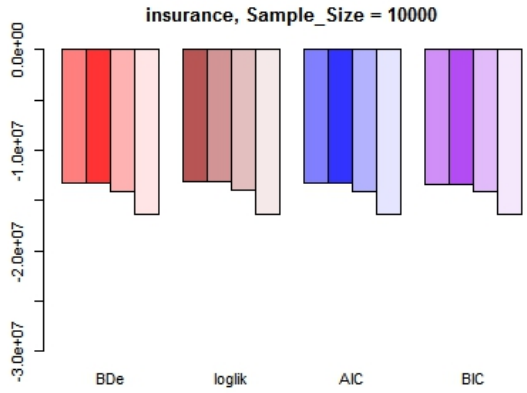
Source Binder J, Koller D, Russell S, Kanazawa K (1997).
"Adaptive Probabilistic Networks with Hidden Variables".
Machine Learning, 29(2-3), 213-244.

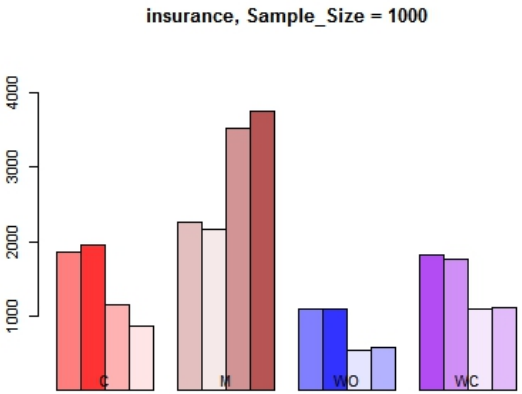
Insurance DataSet

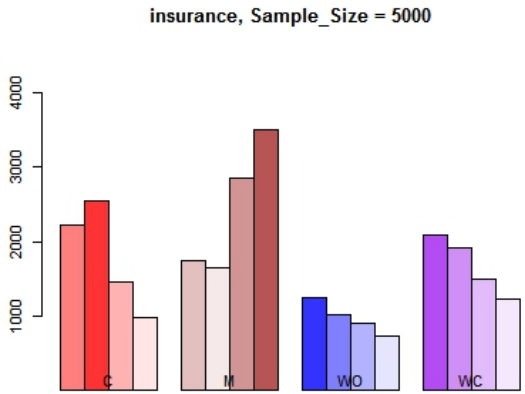
Insurance DataSet

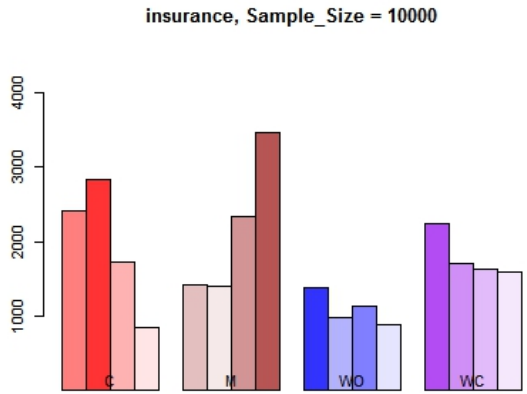












Alarm DataSet

Description The ALARM ("A Logical Alarm Reduction Mechanism") is a Bayesian network designed to provide an alarm message system for patient monitoring.

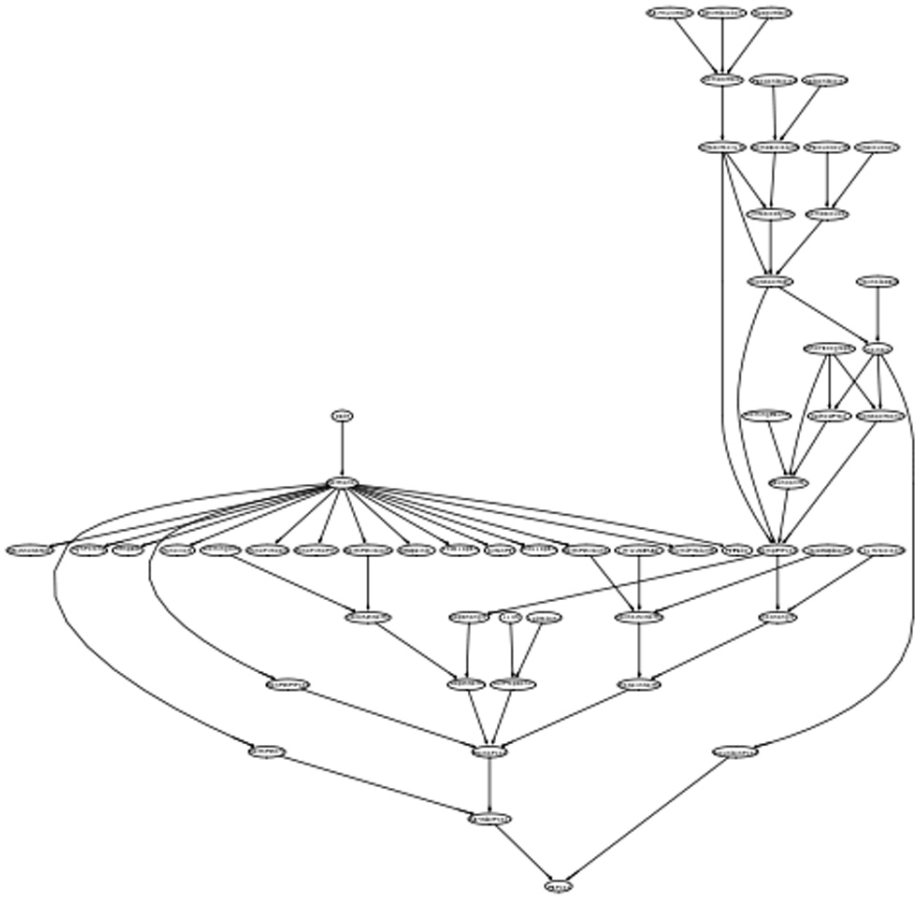
Number of nodes 37

Number of arcs 46

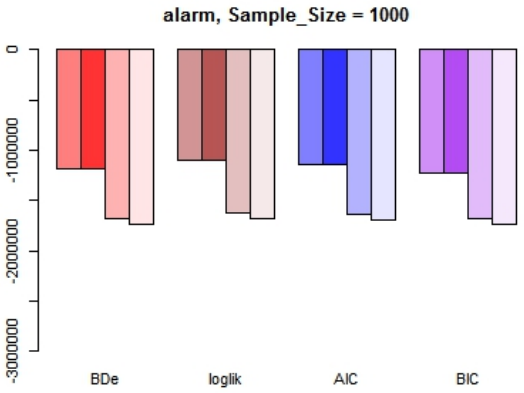
Number of parameters 509

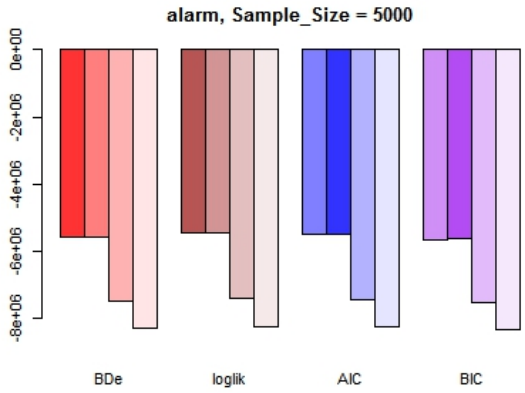
Source Beinlich I, Suermondt HJ, Chavez RM, Cooper GF (1989).
"The ALARM Monitoring System: A Case Study with Two Probabilistic Inference Techniques for Belief Networks."
In "Proceedings of the 2nd European Conference on Artificial Intelligence in Medicine", pp. 247-256. Springer-Verlag.

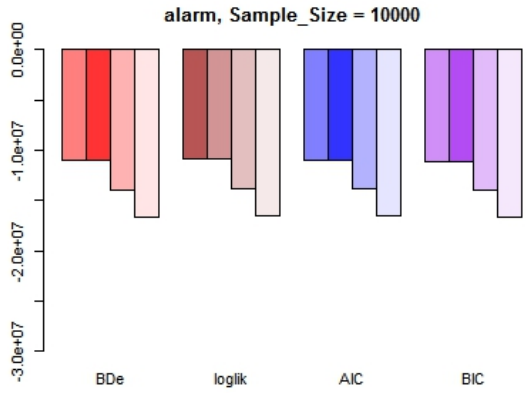
Alarm DataSet

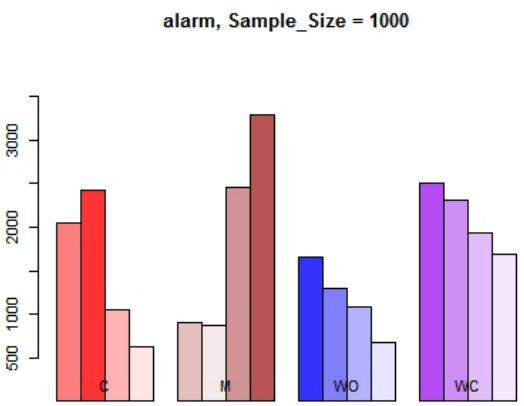


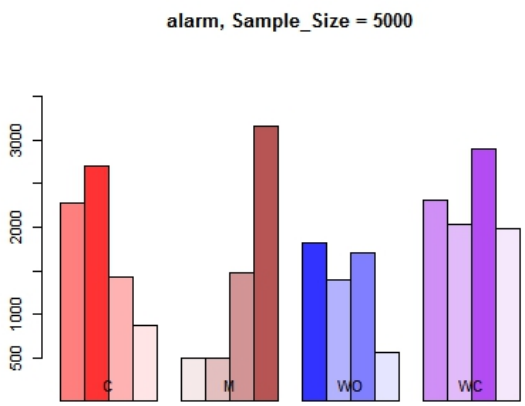
Alarm DataSet

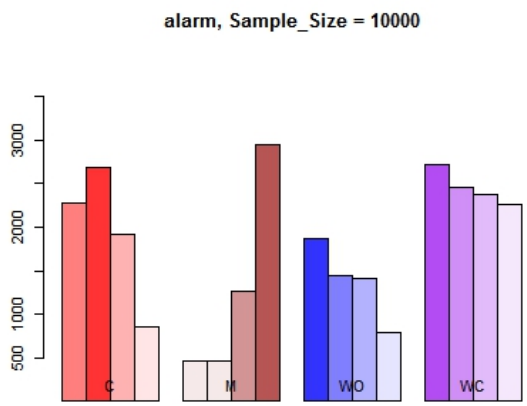












HailFinder DataSet

Description Hailfinder is a Bayesian network designed to forecast severe summer hail in northeastern Colorado.

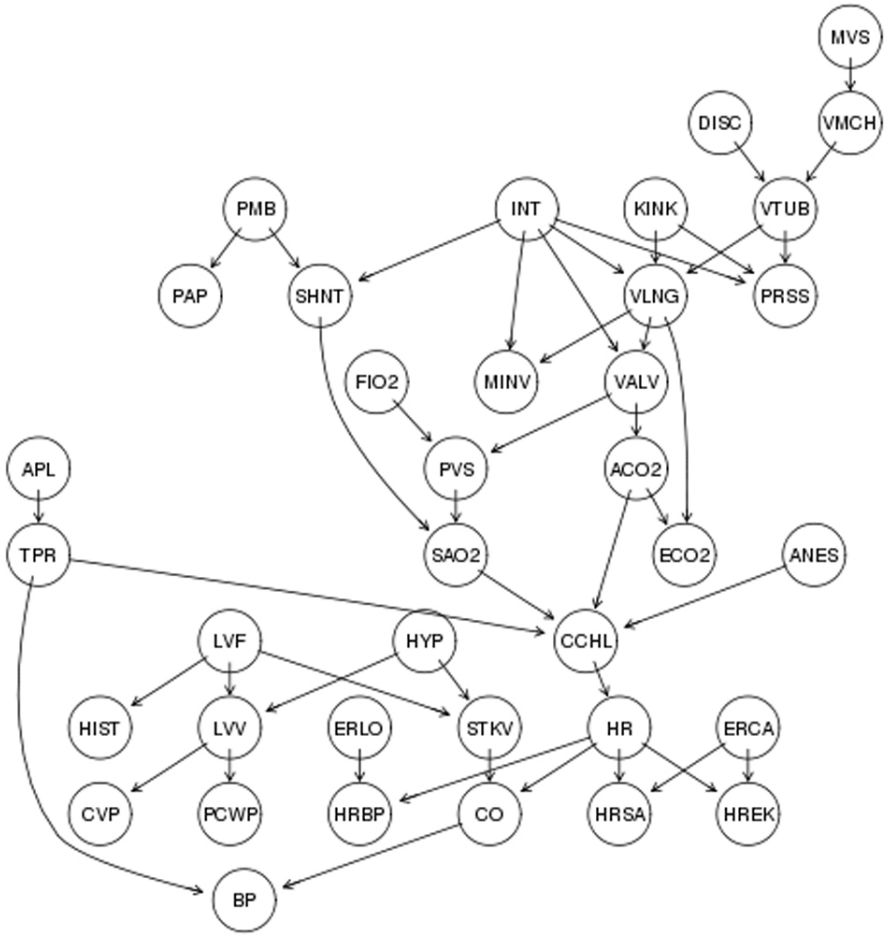
Number of nodes 56

Number of arcs 66

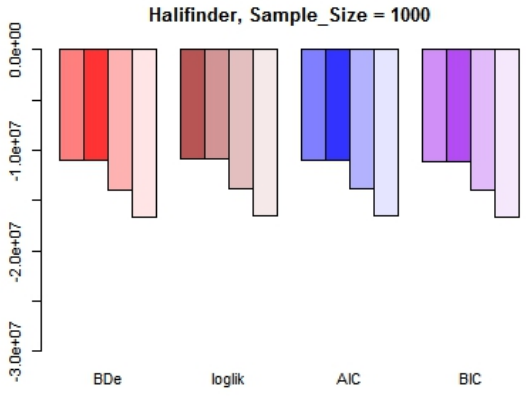
Number of parameters 2656

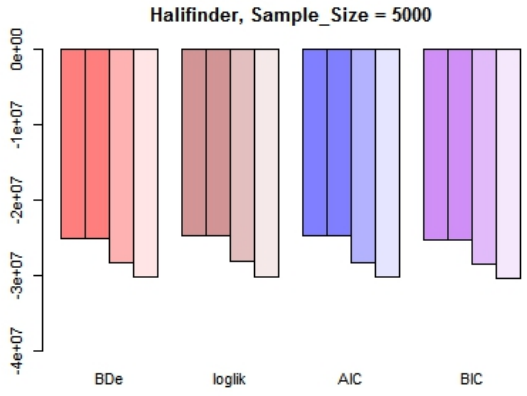
Source Abramson B, Brown J, Edwards W, Murphy A, Winkler RL (1996). "Hailfinder: A Bayesian system for forecasting severe weather". International Journal of Forecasting, 12(1), 57-71.

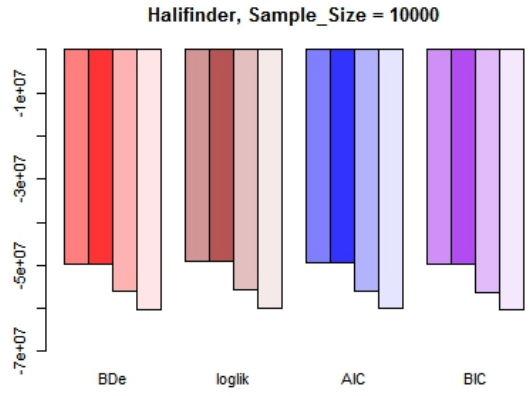
HailFinder DataSet

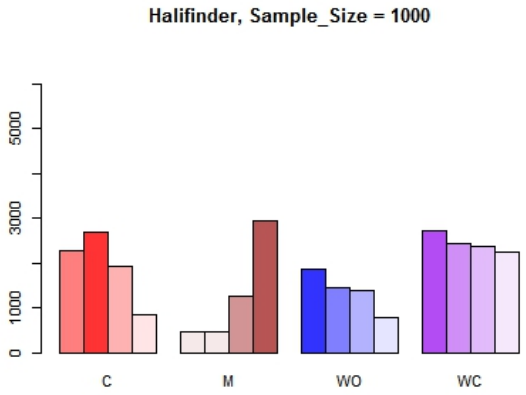


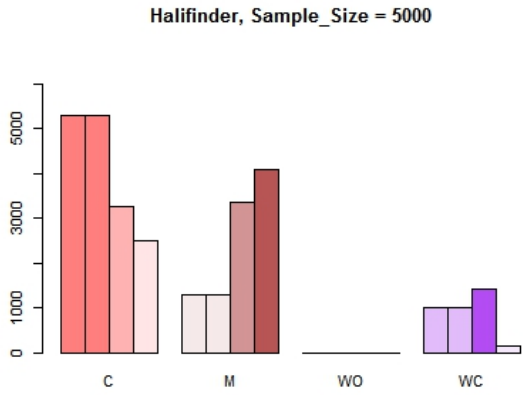
HailFinder DataSet

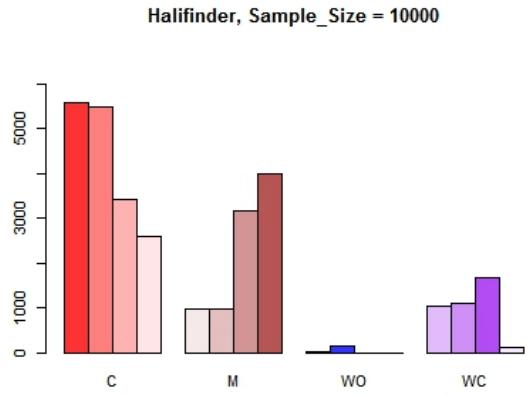












Asia DataSet by Lauritzen and Spiegelhalter
Insurance Evaluation Network DataSet
ALARM Monitoring System DataSet
The HailFinder Weather Forecast System DataSet

Summary

Sample Size 1000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
Asia	2	1	3	4	1	2	3	4	4	4	2	1	4	1	4	4	3	2	1	4
Insurance	2	1	3	4	2	1	3	4	3	4	2	1	2	1	4	3	1	2	4	3
Alarm	2	1	3	4	2	1	3	4	3	4	2	1	1	2	3	4	1	2	3	4
HaliFinder	2	1	3	4	2	1	3	4	4	4	2	1	1	2	3	4	1	2	3	4
Sample Size 5000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
Asia	2	1	3	4	1	2	3	4	4	4	2	1	4	1	4	4	3	2	1	4
Insurance	2	1	3	4	2	1	3	4	3	4	2	1	1	2	3	4	1	2	3	4
Alarm	1	2	3	4	2	1	3	4	4	3	2	1	1	3	2	4	2	3	1	4
HaliFinder	1	1	3	4	1	1	3	4	4	4	2	1	4	4	4	4	2	2	1	4
Sample Size 10000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
Asia	2	1	3	4	1	2	3	4	4	4	2	1	4	1	4	4	3	1	2	4
Insurance	2	1	3	4	2	1	3	4	3	4	2	1	1	3	2	4	1	2	3	4
Alarm	2	1	3	4	2	1	3	4	4	4	2	1	1	2	3	4	1	2	3	4
HaliFinder	2	1	3	4	1	2	3	4	4	3	2	1	2	1	4	4	3	2	1	4

Outline

Data Generator 성능 평가

BN_Data_Generator {User-Defined Function}

베이지안 네트워크 데이터 생성기

Description 베이지안 네트워크 모형을 기반으로 Synthetic Data를 생성하여 준다.

Usage BN_Data_Generator (arcs, input_Probs, n, node_names)

Arguments

arcs	(matrix)	모형 내 arc 유무와 방향을 결정하는 행렬
input_Probs	(list)	arc가 이어져있는 node간 관계의 조건부 확률
n	(constant)	Sample Size
node_names	(vector)	Node의 이름

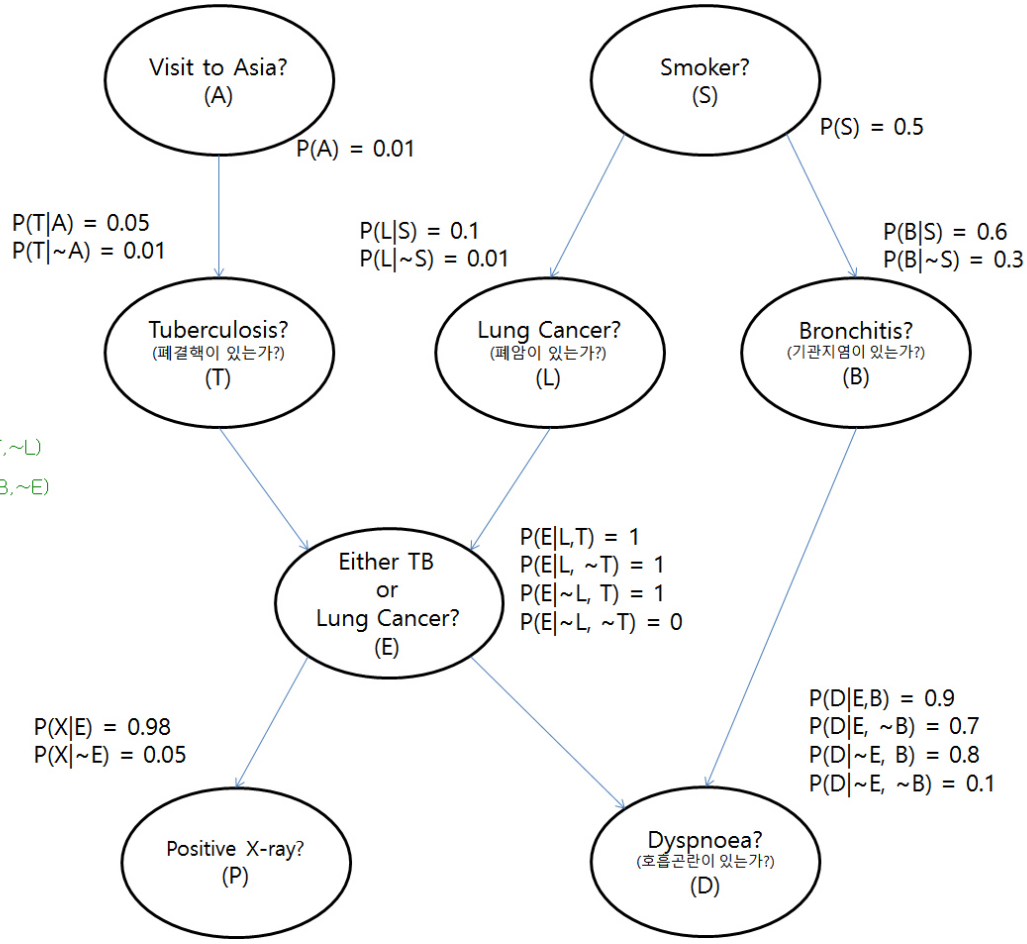
Data Generator 성능 평가

run_multirun.R | graph.R | BN_Data_Generator.R

```
BN_Data_Generator = function (arcs, input_Probs, n, node_names)
{
    # Node 개수
    num_of_nodes = dim(arcs)[1];

    # 각 Node의 Parent Node 개수
    num_of_parent_nodes = apply(arcs, 2, sum);

    list_parent_nodes = list();
    for(i in 1:num_of_nodes)
    {
        if (length(which(arcs[,i]==1)) == 0)
        {
            list_parent_nodes[[i]] = NULL;
        } else {
            list_parent_nodes[[i]] = which(arcs[,i]==1);
        }
    }

    # Root node의 개수
    root_nodes = sum(num_of_parent_nodes == 0);

    # 결과는 여기에 저장이 된다.
    result_mat = matrix(0, n, num_of_nodes);
    dimnames(result_mat)[[2]] = node_names;
    # result_mat

    # 지정해야할 조건부 확률 개수
    num_of_probs = t(as.matrix(2^num_of_parent_nodes));
    dimnames(num_of_probs)[[2]] = node_names;
    num_of_probs


    # 지정해야할 조건부 확률 개수만큼 input이 맞는지 확인. 만일 false이면 프로그램 종료
    input_prob_len = length(input_Probs);
    for (i in 1:input_prob_len)
    {
        if (length(input_Probs[[i]]) != num_of_probs[i])
        {
            cat("Error");
            return(break);
        }
    }


    # Root Node Initialization
    for(i in 1:root_nodes)
    {
        p = input_Probs[[i]][1];
        result_mat[,i] = sample(c("Y", "N"), n, prob=c(p, 1-p), rep=T);
    }


    # Generator
    init = 0
```

Data Generator 성능 평가

```
# Asia
arcs = rbind(
      #  A  S  T  L  B  E  X  D
       c(0,  0, 1, 0, 0, 0, 0, 0), #A
       c(0,  0, 0, 1, 1, 0, 0, 0), #S
       c(0,  0, 0, 0, 0, 1, 0, 0), #T
       c(0,  0, 0, 0, 0, 1, 0, 0), #L
       c(0,  0, 0, 0, 0, 0, 0, 1), #B
       c(0,  0, 0, 0, 0, 0, 1, 1), #E
       c(0,  0, 0, 0, 0, 0, 0, 0), #X
       c(0,  0, 0, 0, 0, 0, 0, 0) #D
       )
arc_name = c("A", "S", "T", "L", "B", "E", "X", "D")
dimnames(arcs)[[1]] = arc_name
dimnames(arcs)[[2]] = arc_name

Probs = list(
    c(0.01),                  # P(A)
    c(0.5),                   # P(S)
    c(0.05, 0.01),            # P(T|A), P(T|~A)
    c(0.1, 0.01),             # P(L|S), P(L|~S)
    c(0.6, 0.3),            # P(B|S), P(B|~S)
    c(1, 1, 1, 0),            # P(E|T,L), P(E|~T,L), P(E|T,~L), P(E|~T,~L)
    c(0.98, 0.05),            # P(X|E), P(X|~E)
    c(0.9, 0.7, 0.8, 0.1)     # P(D|B,E), P(D|~B,E), P(D|B,~E), P(D|~B,~E)
)
```



Data Generator 성능 평가

RStudio

File Edit Code View Plots Session Build Debug Tools Help

Console D:/Dropbox/_베이지안 네트워크/Scores/R/

```
> ##### 시작
> # set.seed(1234)
> require(bnlearn)
필요한 패키지를 로딩중입니다: bnlearn
>
>
> n = 1000
>
> # Asia
> arcs = rbind(
+     # A S T L B E X D
+     c(0, 0, 1, 0, 0, 0, 0, 0), #A
+     c(0, 0, 0, 1, 1, 0, 0, 0), #S
+     c(0, 0, 0, 0, 0, 1, 0, 0), #T
+     c(0, 0, 0, 0, 0, 1, 0, 0), #L
+     c(0, 0, 0, 0, 0, 0, 0, 1), #B
+     c(0, 0, 0, 0, 0, 0, 1, 1), #E
+     c(0, 0, 0, 0, 0, 0, 0, 0), #X
+     c(0, 0, 0, 0, 0, 0, 0, 0) #D
+ )
> arc_name = c("A", "S", "T", "L", "B", "E", "X", "D")
> dimnames(arcs)[[1]] = arc_name
> dimnames(arcs)[[2]] = arc_name
>
> Probs = list(
+     c(0.01),      # P(A)
+     c(0.5),       # P(S)
+     c(0.05, 0.01),    # P(T|A), P(T|~A)
+     c(0.1, 0.01),     # P(L|S), P(L|~S)
+     c(0.6, 0.3),     # P(B|S), P(B|~S)
+     c(1, 1, 1, 0),    # P(E|T,L), P(E|~T,L), P(E|T,~L), P(E|~T,~L)
+     c(0.98, 0.05),    # P(X|E), P(X|~E)
+     c(0.9, 0.7, 0.8, 0.1)  # P(D|B,E), P(D|~B,E), P(D|B,~E), P(D|~B,~E)
+ )
>
> res = BN_Data_Generator(arcs, Probs, n, arc_name)
> data = res$data
> head(data)
  A S T L B E X D
1 N N N N N N N N
2 N N N N N N N N
3 N Y N N Y N N Y
4 N N N N N N N N
5 N Y N N Y N N N
6 N Y N N N N N N
> dim(data)
[1] 1000    8
>
> # Constraint-based algorithms
> bn_gs = gs(data)       # the Grow-Shrink(GS)
> bn_gs
```

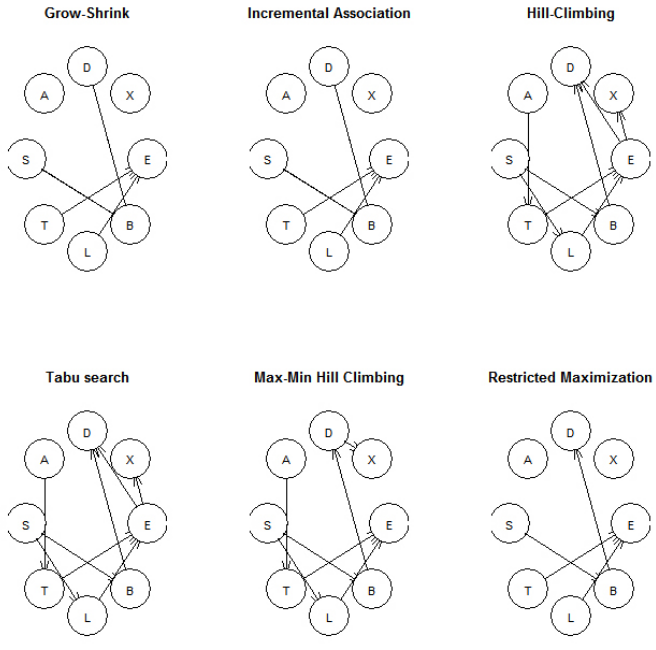
Environment History

Import Dataset Clear List

Global Environment

Files Plots Packages Help Viewer

Zoom Export Clear All



성능 평가를 위한 선행 연구

Pekka Parviainen, Hossein Shahrabi Farahani, and Jens Lagergren (2014).

"Learning Bounded Tree-width Bayesian Networks using Integer Linear Programming"

Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)

Data Generator 성능 평가 via Hill Climbing (HC)

Dataset	Num. of Nodes	Sample Size	ILP	Score via HC			
			BDe	BDe	loglik	AIC	BIC
Asia	8	100	-245.64	-251.07	-194.08	-210.08	-230.92
real		1000	-2317.41	-2281.98	-2188.63	-2205.63	-2247.35
dataset		10000	-22466.40	-21937.24	-21812.37	-21832.37	-21904.47
Gen.		100		-271.30	-209.69	-222.69	-239.62
by		1000		-2350.21	-2262.16	-2279.16	-2320.87
WEKA		10000		-22529.57	-22419.24	-22437.24	-22502.14
Gen.		100		-	-	-	-
by		1000		-2289.26	-2197.43	-2214.43	-2256.15
Generator		10000		-22521.67	-22403.53	-22420.53	-22481.82

Data Generator 성능 평가 via TABU Search

Dataset	Num. of Nodes	Sample Size	ILP	Score via TABU			
			BDe	BDe	loglik	AIC	BIC
Asia	8	100	-245.64	-270.99	-206.14	-219.14	-236.08
real		1000	-2317.41	-2342.17	-2249.24	-2266.24	-2307.96
dataset		10000	-22466.40	-21996.98	-21870.55	-21889.55	-21958.05
Gen.		100		-270.74	-208.81	-221.81	-238.74
by		1000		-2350.21	-2262.16	-2279.16	-2320.87
WEKA		10000		-22529.57	-22419.24	-22437.24	-22502.14
Gen.		100		-	-	-	-
by		1000		-2289.26	-2197.43	-2214.43	-2256.15
Generator		10000		-22521.67	-22403.53	-22420.53	-22481.82

Data Generator 성능 평가 via MMHC

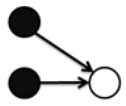
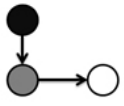
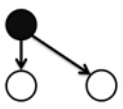
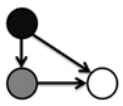
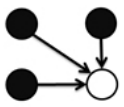
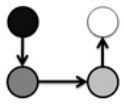
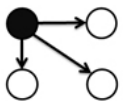
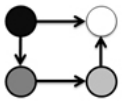
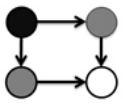
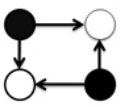
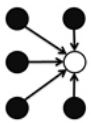
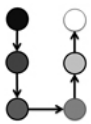
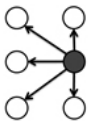
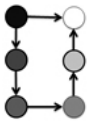
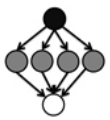
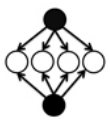
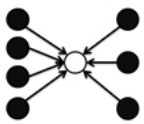
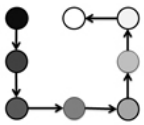
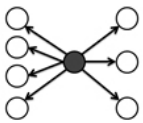
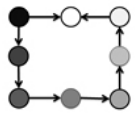
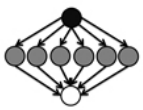
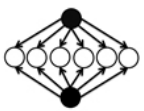
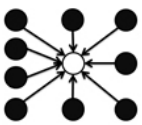
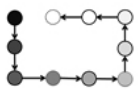
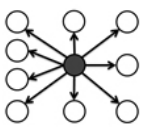
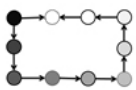
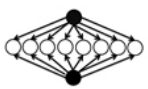
Dataset	Num. of Nodes	Sample Size	ILP	Score via MMHC			
			BDe	BDe	loglik	AIC	BIC
Asia	8	100	-245.64	-301.29	-246.32	-257.32	-271.65
real		1000	-2317.41	-2504.74	-2421.80	-2436.80	-2473.61
dataset		10000	-22466.40	-24346.08	-24232.89	-24249.89	-24311.18
Gen.		100		-272.31	-213.08	-225.08	-240.71
by		1000		-2508.05	-2423.51	-2439.51	-2478.77
WEKA		10000		-22815.07	-22709.96	-22725.96	-22783.64
Gen.		100		-	-	-	-
by		1000		-2446.05	-2360.82	-2376.82	-2416.08
Generator		10000		-24137.75	-24026.34	-24042.34	-24100.02

Data Generator 성능 평가 via RSMAX2

Dataset	Num. of Nodes	Sample Size	ILP	Score via RSMAX2			
			BDe	BDe	loglik	AIC	BIC
Asia	8	100	-245.64	-299.97	-246.07	-260.07	-278.31
real		1000	-2317.41	-2531.66	-2451.91	-2464.91	-2496.81
dataset		10000	-22466.40	-24295.92	-24194.06	-24207.06	-24253.93
Gen.		100		-272.31	-213.08	-225.08	-240.71
by		1000		-2534.36	-2456.36	-2469.36	-2501.26
WEKA		10000		-22823.55	-22715.80	-22730.80	-22784.88
Gen.		100		-	-	-	-
by		1000		-2479.98	-2401.17	-2414.17	-2446.07
Generator		10000		-24504.43	-24403.62	-24416.62	-24463.49

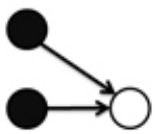
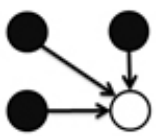
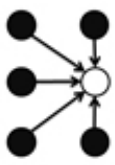
Outline

Topology에 따른 비교

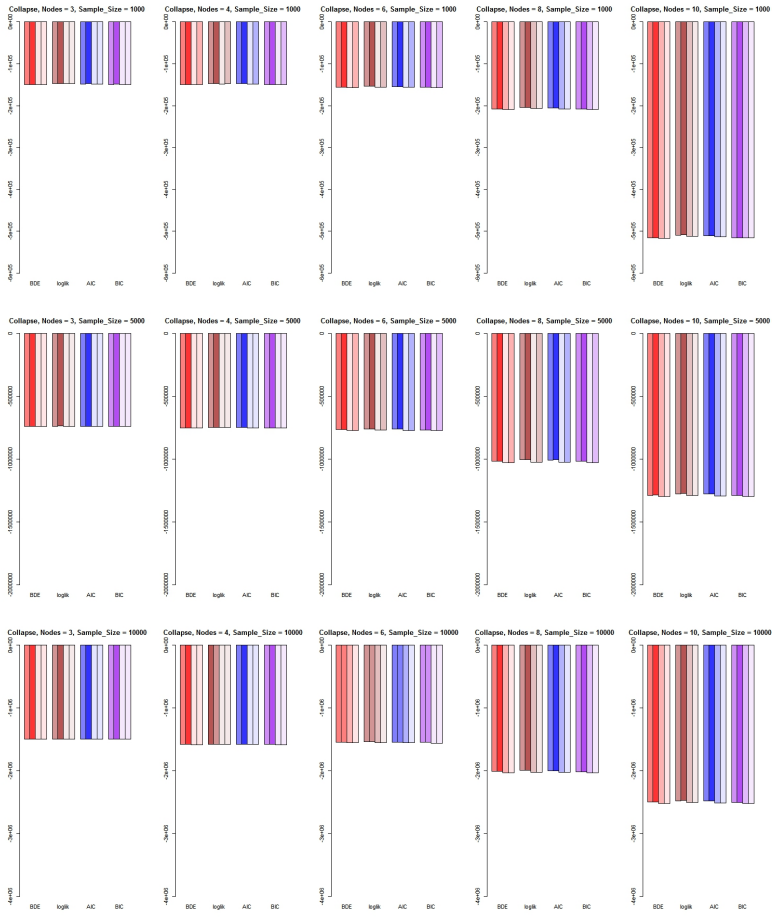
	Collapse	Line	Star	Pseudo Loop	Diamond	Rhombus
3						
4						
6						
8						
10						

Eitel J. M. Lauría,
"An Information-Geometric Approach to Learning Bayesian Network Topologies from Data",
Innovations in Bayesian Networks Studies in Computational Intelligence Volume 156, 2008, pp 187-217

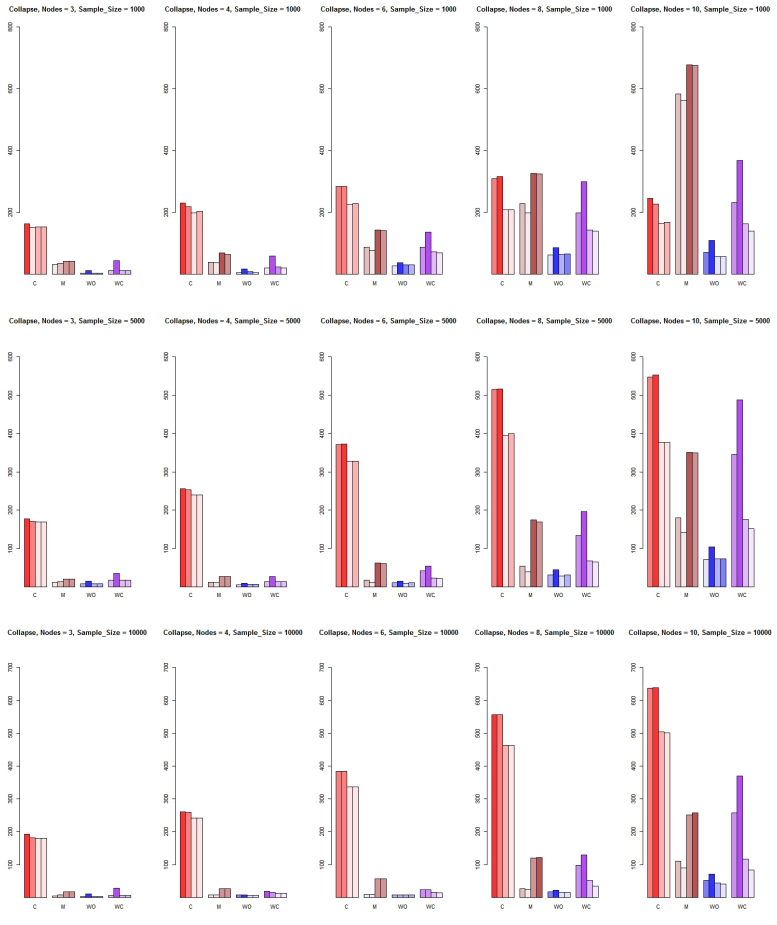
Topology에 따른 비교 분석 : Collapse

	3	4	6	8	10
Collapse					

Topology에 따른 비교 : Collapse (Score)



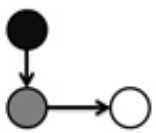
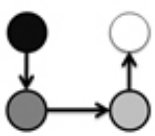
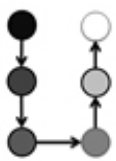
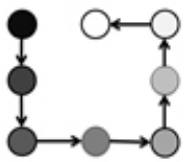
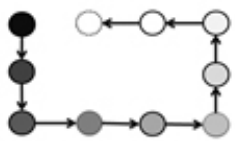
Topology에 따른 비교 : Collapse (Arcs)



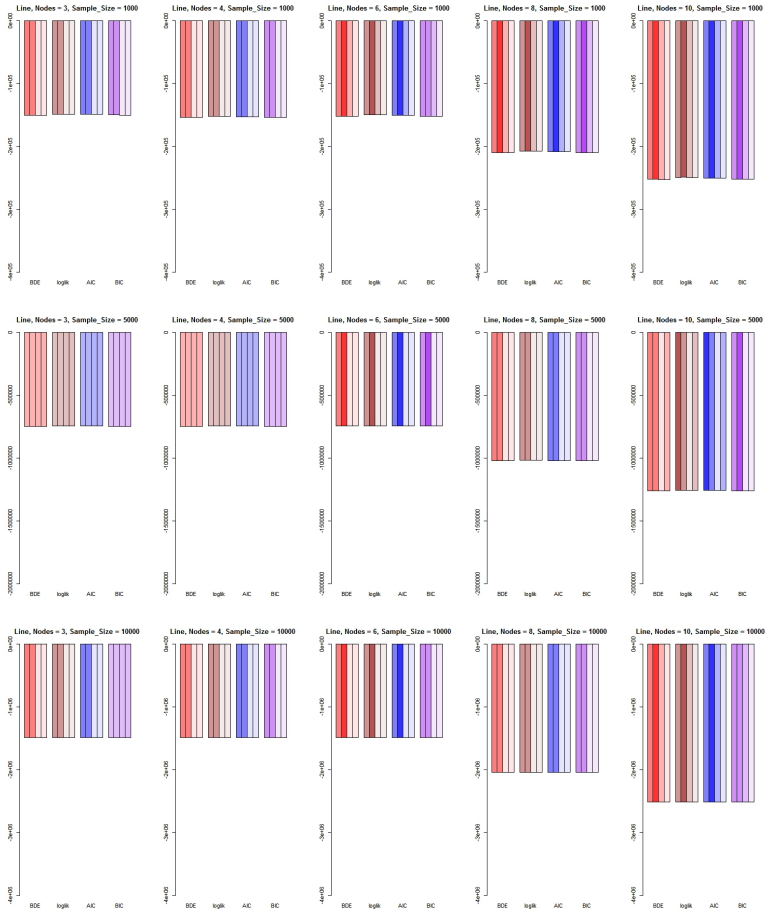
Topology에 따른 비교 : Collapse

Sample Size 1000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
3	2	1	4	4	1	4	2	2	4	3	1	1	4	1	4	4	4	1	4	4
4	2	1	4	3	1	2	4	3	3	4	1	2	4	1	2	4	4	1	2	4
6	2	1	4	3	1	1	4	3	3	4	1	2	4	1	2	3	2	1	3	4
8	2	1	4	3	2	1	4	4	3	4	1	2	4	1	3	2	2	1	3	4
10	2	1	3	4	1	2	4	3	3	4	1	2	2	1	4	4	2	1	3	4
Sample Size 5000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
3	2	1	4	4	1	2	4	4	4	3	1	1	4	1	4	4	4	1	4	4
4	2	1	4	4	1	2	4	4	3	4	1	1	4	1	2	2	4	1	4	4
6	2	1	4	3	2	1	4	3	3	4	1	2	2	1	4	2	2	1	3	4
8	2	1	4	3	2	1	4	3	3	4	1	2	2	1	4	2	2	1	3	4
10	2	1	3	4	2	1	4	4	3	4	1	2	2	1	4	2	2	1	3	4
Sample Size 10000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
3	2	1	4	4	1	2	4	4	4	3	1	1	4	1	4	4	4	1	4	4
4	2	1	4	4	1	2	4	4	4	4	1	1	2	1	4	4	1	2	4	4
6	2	1	3	4	1	1	4	4	4	4	1	1	1	1	1	1	1	1	3	4
8	2	1	3	4	1	2	3	4	3	4	2	1	2	1	4	4	2	1	3	4
10	2	1	3	4	2	1	3	4	3	4	2	1	2	1	3	4	2	1	3	4

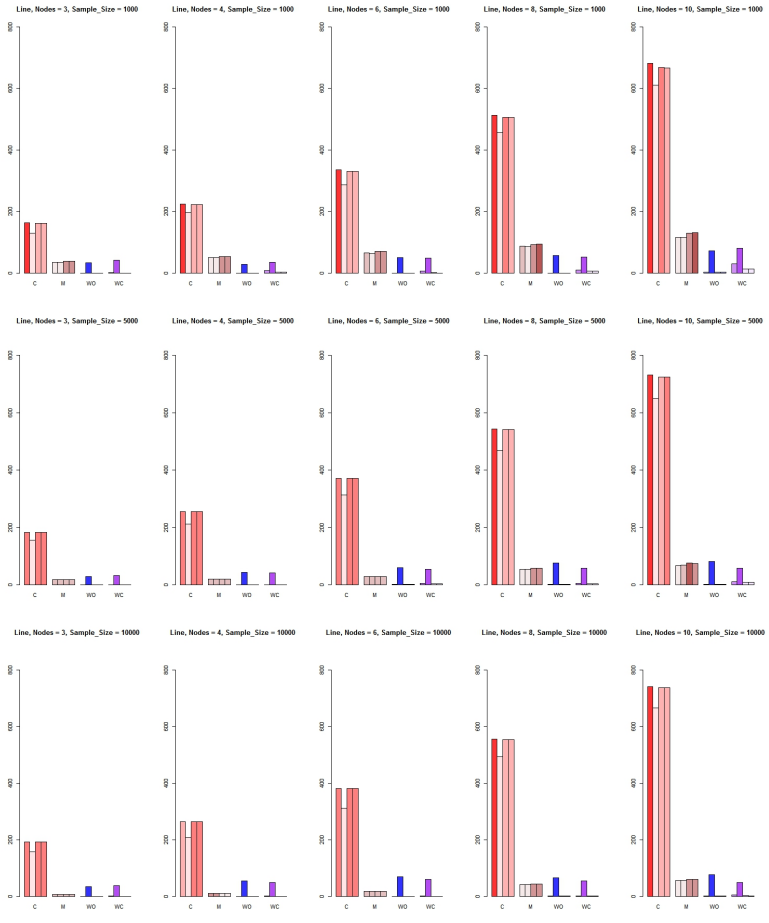
Topology에 따른 비교 분석 : Line

	3	4	6	8	10
Line					

Topology에 따른 비교 : Line (Score)



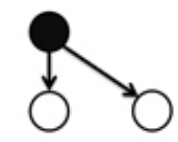
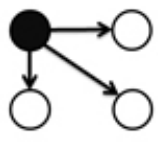
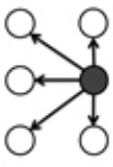
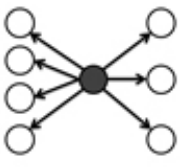
Topology에 따른 비교 : Line (Arcs)



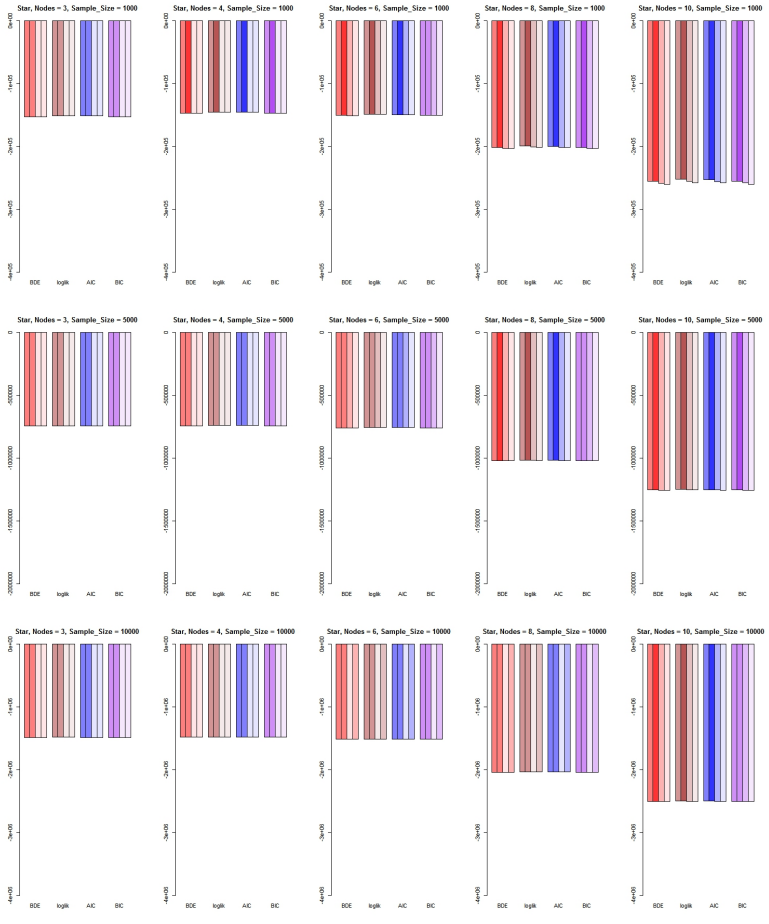
Topology에 따른 비교 : Line

Sample Size 1000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
3	1	1	4	4	1	4	2	2	4	4	1	1	4	1	4	4	2	1	4	4
4	1	1	4	4	1	4	2	2	4	4	1	1	4	1	4	4	2	1	3	4
6	2	1	3	4	1	4	2	2	3	4	1	1	4	1	4	4	2	1	3	4
8	2	1	3	4	1	4	2	3	3	4	2	1	4	1	4	4	2	1	4	4
10	2	1	3	4	1	4	2	3	4	4	2	1	4	1	4	4	2	1	4	4
Sample Size 5000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
3	1	1	1	1	1	4	1	1	1	1	1	1	4	1	4	4	4	1	4	4
4	1	1	1	1	1	4	1	1	1	1	1	1	4	1	4	4	4	1	4	4
6	2	1	4	4	1	4	1	1	1	1	1	1	4	1	4	4	2	1	4	4
8	1	1	4	4	1	4	2	2	4	4	1	1	4	1	4	4	2	1	4	4
10	1	1	4	3	1	4	3	2	4	3	1	2	4	1	4	4	2	1	4	4
Sample Size 10000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
3	1	1	4	4	1	4	1	1	1	1	1	1	4	1	4	4	2	1	4	4
4	1	1	4	4	3	4	1	1	1	1	4	4	4	1	4	4	2	1	4	4
6	2	1	4	4	1	4	1	1	1	1	1	1	4	1	4	4	2	1	4	4
8	1	1	4	4	1	4	2	2	4	4	1	1	4	1	4	4	4	1	4	4
10	2	1	3	4	1	4	2	2	4	4	1	1	4	1	4	4	2	1	3	4

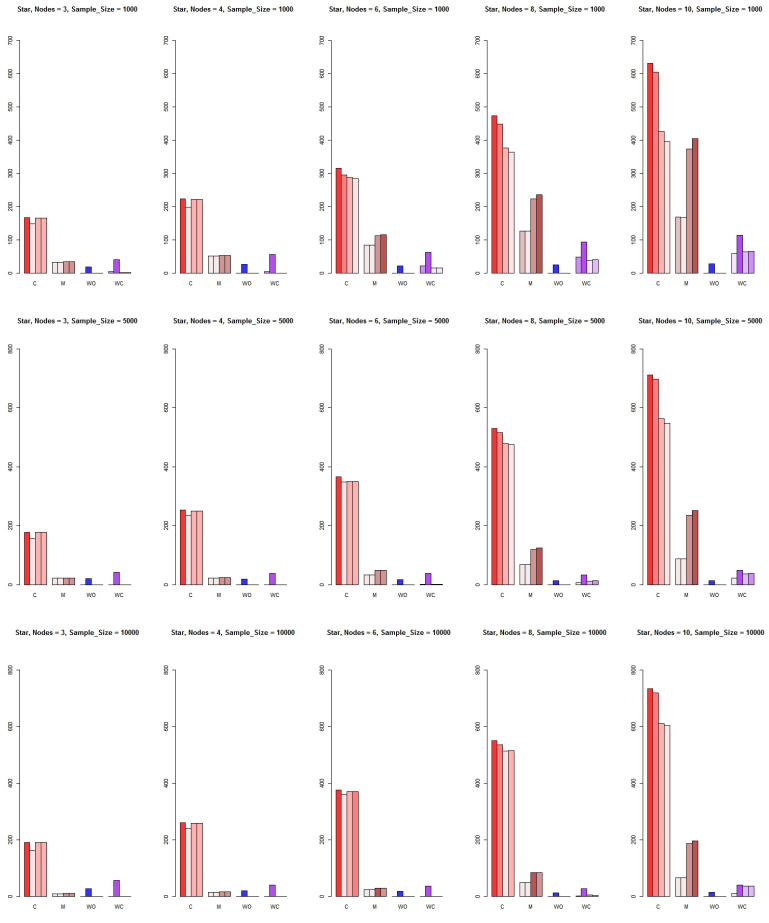
Topology에 따른 비교 분석 : Star

	3	4	6	8	10
Star					

Topology에 따른 비교 : Star (Score)



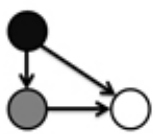
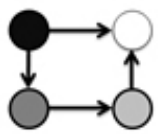
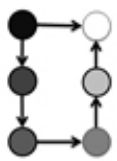
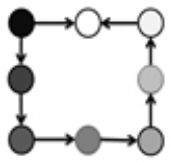
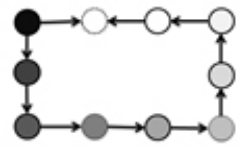
Topology에 따른 비교 : Star (Arcs)



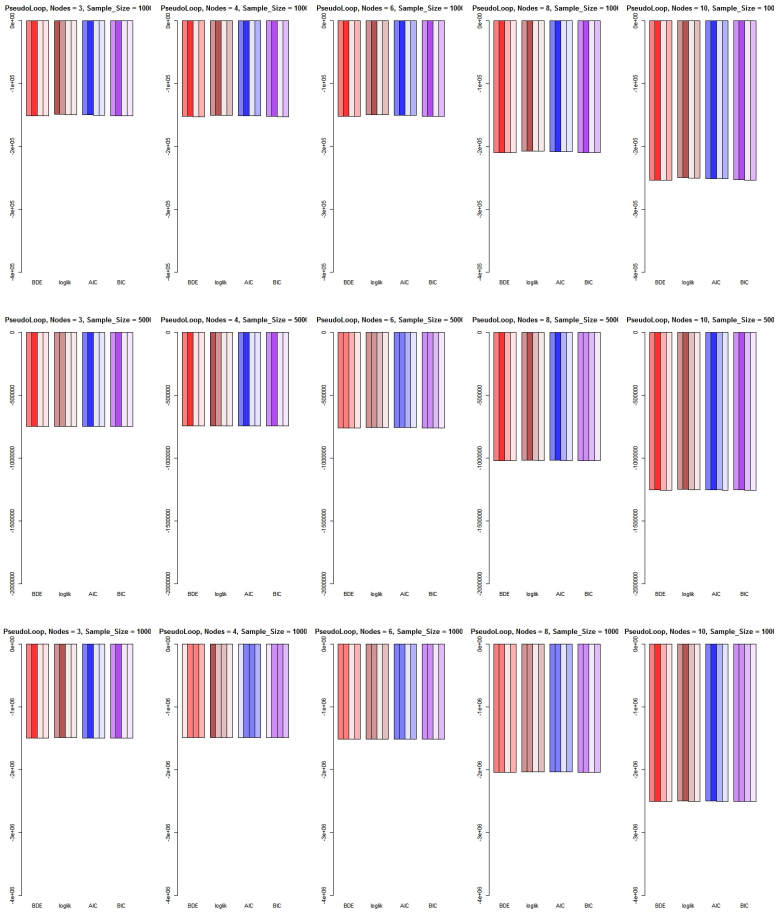
Topology에 따른 비교 분석 : Star

Sample Size 1000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
3	1	1	4	4	1	4	2	2	4	4	1	1	4	1	4	4	2	1	4	4
4	2	1	4	4	1	4	2	2	4	4	1	1	4	1	4	4	2	1	4	4
6	2	1	3	4	1	2	3	4	4	4	2	1	4	1	4	4	2	1	4	4
8	2	1	3	4	1	2	3	4	3	4	2	1	4	1	4	4	2	1	4	3
10	2	1	3	4	1	2	3	4	3	4	2	1	4	1	4	4	4	1	3	2
Sample Size 5000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
3	1	1	4	4	1	4	2	2	4	4	1	1	4	1	4	4	4	1	4	4
4	1	1	4	4	1	4	2	2	4	4	1	1	4	1	4	4	4	1	4	4
6	1	1	3	4	1	4	2	2	4	4	1	1	4	1	4	4	4	1	4	4
8	2	1	3	4	1	2	3	4	4	4	2	1	4	1	4	4	4	1	3	2
10	2	1	3	4	1	2	3	4	4	4	2	1	4	1	4	4	4	1	3	2
Sample Size 10000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
3	1	1	4	4	1	4	2	2	4	4	1	1	4	1	4	4	4	1	4	4
4	1	1	4	4	1	4	2	2	4	4	1	1	4	1	4	4	4	1	4	4
6	1	1	4	3	1	2	4	3	4	4	1	2	4	1	4	4	4	1	4	4
8	1	1	4	3	1	2	4	3	4	4	1	2	4	1	4	4	4	1	2	3
10	2	1	3	4	1	2	3	4	4	4	2	1	4	1	4	4	4	1	2	2

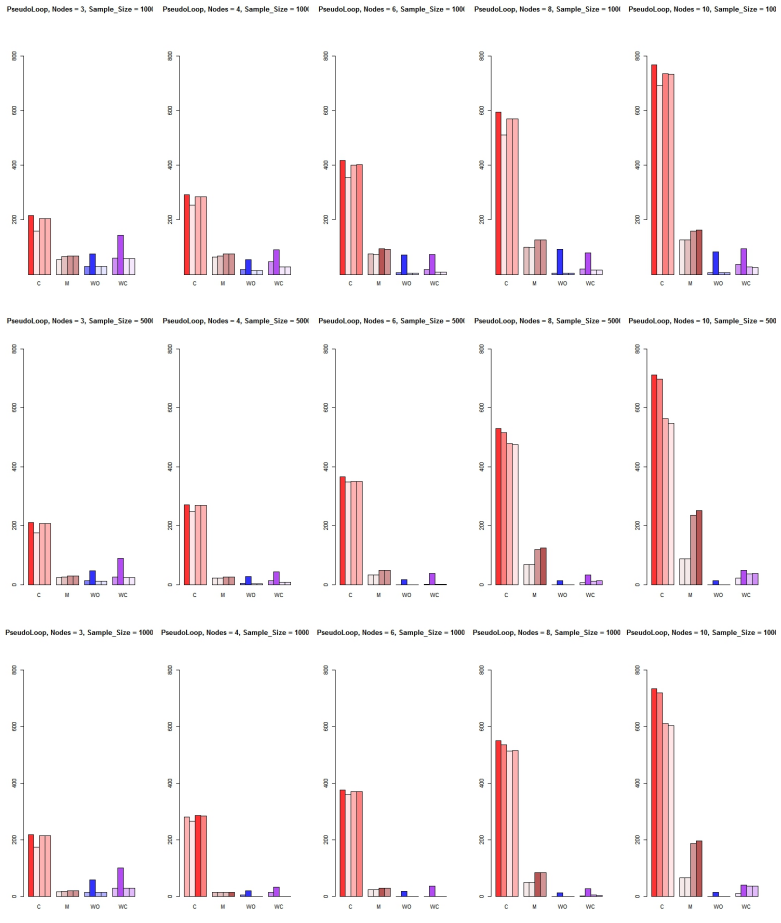
Topology에 따른 비교 분석 : PseudoLoop

	3	4	6	8	10
Pseudo Loop					

Topology에 따른 비교 분석 : PseudoLoop (Score)



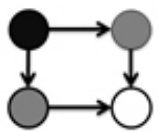
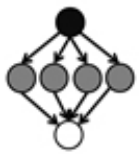
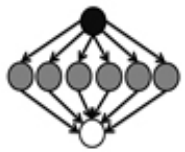
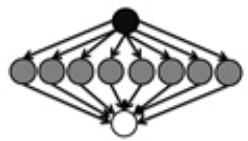
Topology에 따른 비교 분석 : PseudoLoop (Arc)



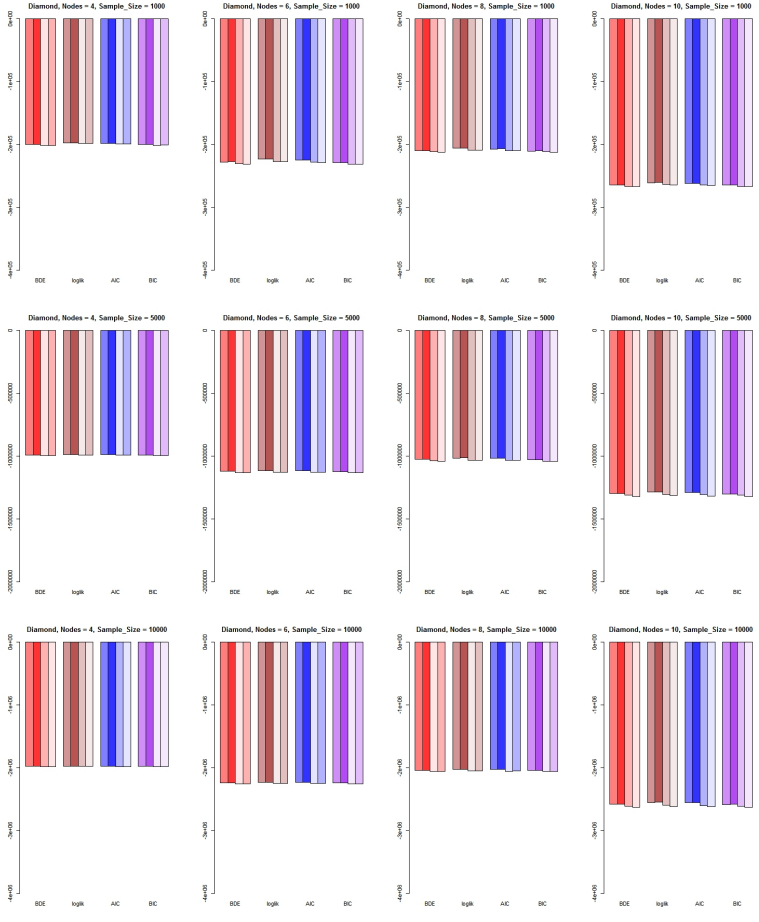
Topology에 따른 비교 분석 : PseudoLoop

Sample Size 1000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
3	2	1	4	4	1	4	2	2	4	3	1	1	2	1	4	4	2	1	4	4
4	2	1	4	3	1	4	2	2	4	3	1	1	2	1	4	4	2	1	4	4
6	2	1	4	3	1	4	3	2	3	4	1	2	2	1	4	4	2	1	4	4
8	2	1	3	4	1	4	2	2	3	4	1	1	4	1	4	4	2	1	4	4
10	2	1	4	3	1	4	2	3	4	3	2	1	4	1	4	4	2	1	3	4
Sample Size 5000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
3	2	1	4	4	1	4	2	2	3	2	1	1	2	1	4	4	2	1	4	4
4	2	1	4	4	1	4	2	2	4	4	1	1	2	1	4	4	2	1	4	4
6	1	1	3	4	1	4	2	2	4	4	1	1	4	1	4	4	4	1	4	4
8	2	1	3	4	1	2	3	4	4	4	2	1	4	1	4	4	4	1	3	2
10	2	1	3	4	1	2	3	4	4	4	2	1	4	1	4	4	4	1	3	2
Sample Size 10000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
3	2	1	4	4	1	4	2	2	4	3	1	1	4	1	4	4	4	1	4	4
4	4	1	1	3	2	4	1	2	4	4	4	1	2	1	4	4	2	1	4	4
6	1	1	4	3	1	4	3	2	4	4	1	2	4	1	4	4	4	1	4	4
8	1	1	4	3	1	2	4	3	4	4	1	2	4	1	4	4	4	1	2	3
10	2	1	3	4	1	2	3	4	4	4	2	1	4	1	4	4	4	1	2	2

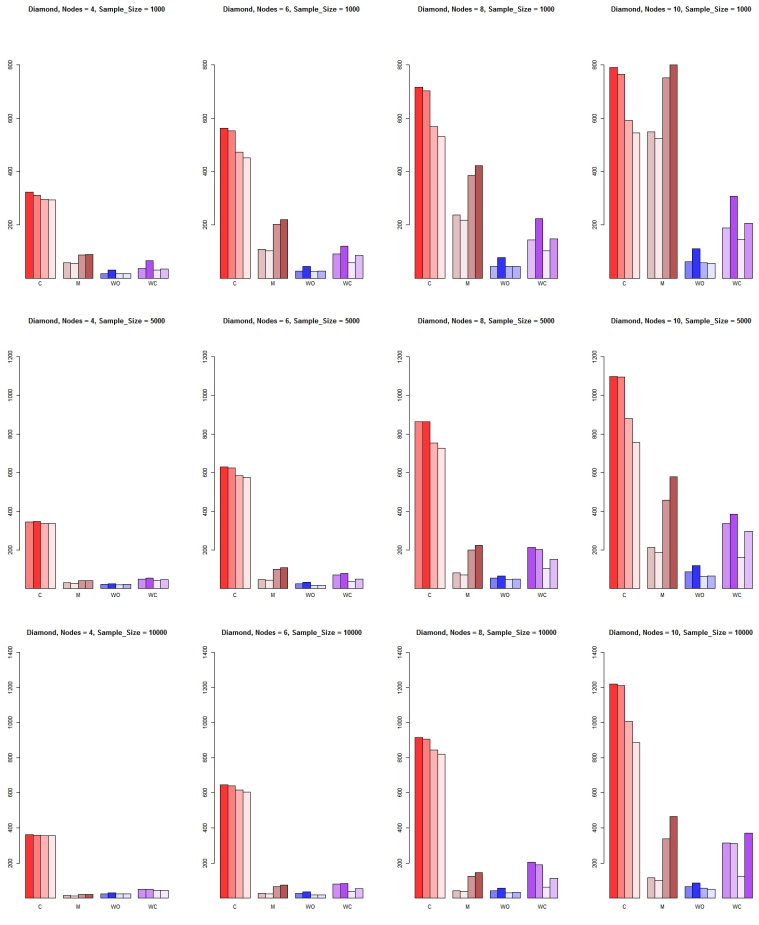
Topology에 따른 비교 분석 : Diamond

	3	4	6	8	10
Diamond					

Topology에 따른 비교 분석 : Diamond (Score)



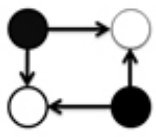
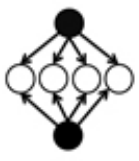
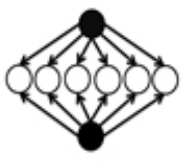
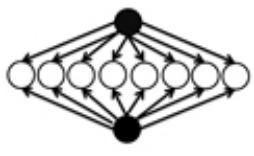
Topology에 따른 비교 분석 : Diamond (Arc)



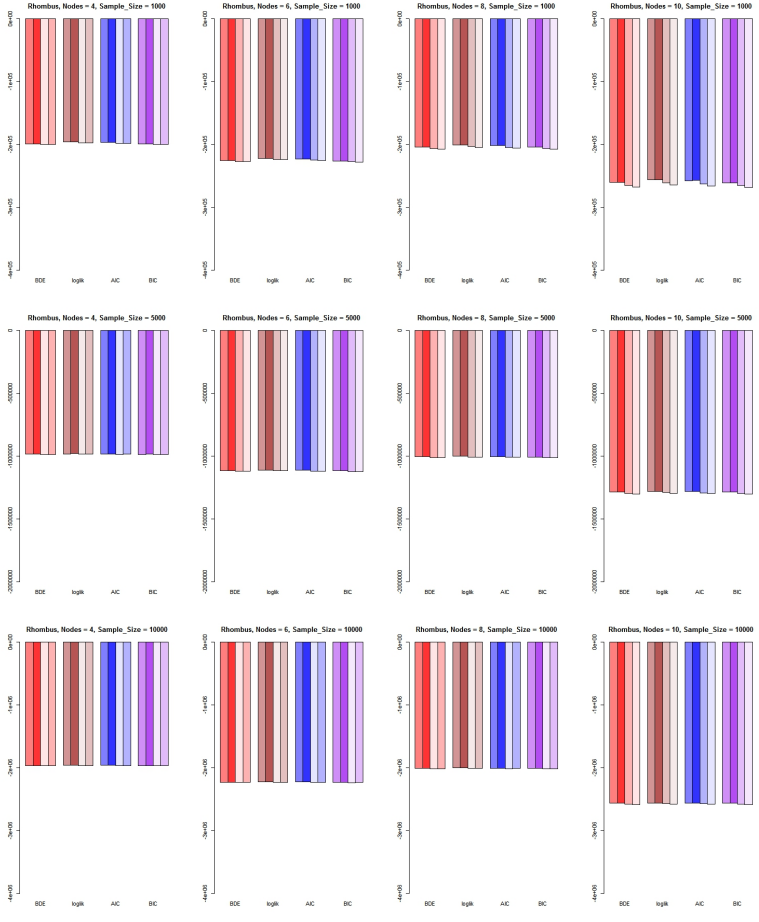
Topology에 따른 비교 분석 : Diamond

Sample Size 1000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
4	2	1	4	3	1	2	3	4	3	4	2	1	2	1	4	4	2	1	4	3
6	2	1	3	4	1	2	3	4	3	4	2	1	2	1	4	2	2	1	4	2
8	2	1	3	4	1	2	3	4	3	4	2	1	4	1	4	4	3	1	4	2
10	2	1	3	4	1	2	3	4	3	4	2	1	2	1	3	4	3	1	4	2
Sample Size 5000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
4	2	1	4	3	2	1	3	4	3	4	1	1	2	1	4	3	2	1	4	3
6	2	1	4	3	1	2	3	4	3	4	2	1	2	1	4	4	2	1	4	3
8	2	1	3	4	2	1	3	4	3	4	2	1	2	1	4	3	1	2	4	3
10	2	1	3	4	1	2	3	4	3	4	2	1	2	1	4	3	2	1	4	3
Sample Size 10000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
4	2	1	3	4	1	2	3	4	3	4	2	1	2	1	4	4	1	1	4	4
6	2	1	4	3	1	2	3	4	3	4	2	1	2	1	4	4	2	1	4	3
8	2	1	4	3	1	2	3	4	3	4	2	1	2	1	4	3	1	2	4	3
10	2	1	3	4	1	2	3	4	3	4	2	1	2	1	3	4	2	3	4	1

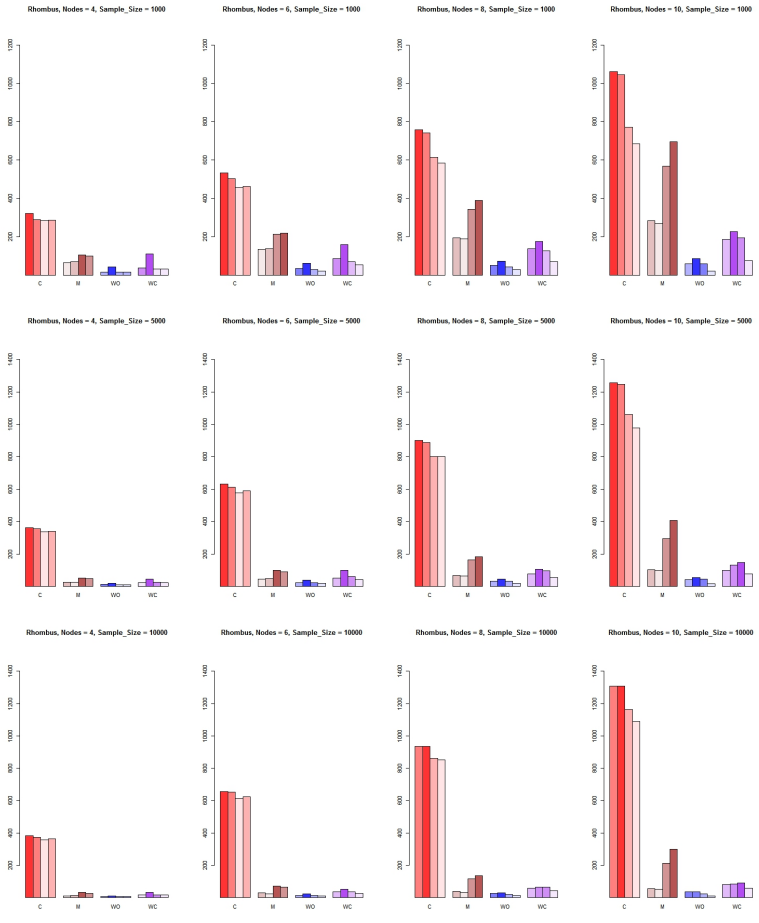
Topology에 따른 비교 분석 : Rhombus

	3	4	6	8	10
Rhombus					

Topology에 따른 비교 분석 : Rhombus (Score)



Topology에 따른 비교 분석 : Rhombus (Arc)



Topology에 따른 비교 분석 : Rhombus

Sample Size 1000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
4	2	1	4	3	1	2	4	3	4	3	1	2	4	1	4	4	2	1	4	4
6	2	1	3	4	1	2	4	3	4	3	2	1	2	1	3	4	2	1	3	4
8	2	1	3	4	1	2	3	4	3	4	2	1	2	1	3	4	2	1	3	4
10	2	1	3	4	1	2	3	4	3	4	2	1	3	1	2	4	3	1	2	4
Sample Size 5000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
4	2	1	4	2	1	2	4	3	3	4	1	2	2	1	4	4	4	1	2	4
6	2	1	3	4	1	2	4	3	4	3	1	2	3	1	2	4	3	1	2	4
8	2	1	3	4	1	2	3	4	3	4	2	1	2	1	2	4	3	1	2	4
10	2	1	3	4	1	2	3	4	3	4	2	1	3	1	2	4	3	2	1	4
Sample Size 10000	Score				C				M				WO				WC			
	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2	HC	TABU	MMHC	RSMAX2
4	2	1	4	3	1	2	4	3	4	3	1	2	4	1	4	4	4	1	2	4
6	2	1	4	3	1	2	4	3	3	4	1	2	2	1	2	4	2	1	2	4
8	2	1	4	3	2	1	3	4	3	4	2	1	2	1	3	4	3	1	1	4
10	2	1	3	4	2	1	3	4	3	4	2	1	1	1	3	4	3	2	1	4

결론

Algorithm User가 판단해야할 문제

- Score로 평가할 것인가, Model을 직접 비교하여 평가할 것인가?
- Model로 평가할 때, M, WO, WC에 대한 가중치는?

결론

bnlearn 패키지에서 제공하는, Score-Based Algorithm과 Hybrid Algorithm을 집중 비교해 본 결과 다음과 같은 결론을 내일 수 있었다.

- "Score가 무엇이 더 높은가?", "Correct Arcs가 무엇이 더 많은가?" 어떤 것으로 비교해도 Score-Based 알고리즘(HC, TABU)가 Hybrid 알고리즘(MMHC, RSMAX2)보다 좋게 나타났다.
- 그러나 Score-Based 알고리즘을 적용했을 때, Wrongly Oriented and Connected Arcs도 많이 나타났다.

  특히 HC와 TABU만을 놓고 보면, TABU가 score는 더 좋지만, 학습된 모형을 보면 WO, WC 비중이 압도적으로 높았다.

  만일 학습된 결과 WO, WC가 많은 것이 C가 적은 것보다 문제가 되는 분야라면, Hybrid Algorithm을 선택하는 것도 방법이 될 수 있다.

결론

bnlearn 패키지에서 제공하는, Score-Based Algorithm과 Hybrid Algorithm을 집중 비교해 본 결과 다음과 같은 결론을 내일 수 있었다.

- Topology별로 ???
- Sample Size별로 ??
- Node 개수별로 ??

Discussion

(컴퓨팅 시간을 기다릴 인내심만 있다면...)

- Topology의 Node 개수를 더 늘려볼 필요도 있다.
- 다른 Algorithm을 적용하여, Sample Size를 더 적게하여, Cardinality를 증가시켜 추가 실험을 진행해 볼 수 있다.
- 서로 다른 Topology를 결합하여 비교 분석 할 수도 있다.
- Bayesian Network Data Generator의 R 패키징 완성
- 확률 관계를 정의할 때 확률값을 U(0,1) 사이의 값에서 임의로 주었는데, 이 확률값을 "순차적"으로 준 것과의 관계를 알아볼 수도 있다.

다른 연구 과제

- Bayesian Network를 이용하여 Missing Value를 컨트롤
  (이 때 Bayesian Network Data Generator를 적극 활용할 수 있다.)
  1. 배경 지식을 바탕으로 만든 모형을 이용하여 데이터를 생성하는 방법
  2. Missing이 없는 부분을 이용하여 먼저 구조학습을 한 후, 학습된 모형으로 데이터를 생성하는 방법
  3. 조건부가 Missing인 부분에 대한 조건부 확률관계를 표현하는 Node를 추가로 생성하는 방법