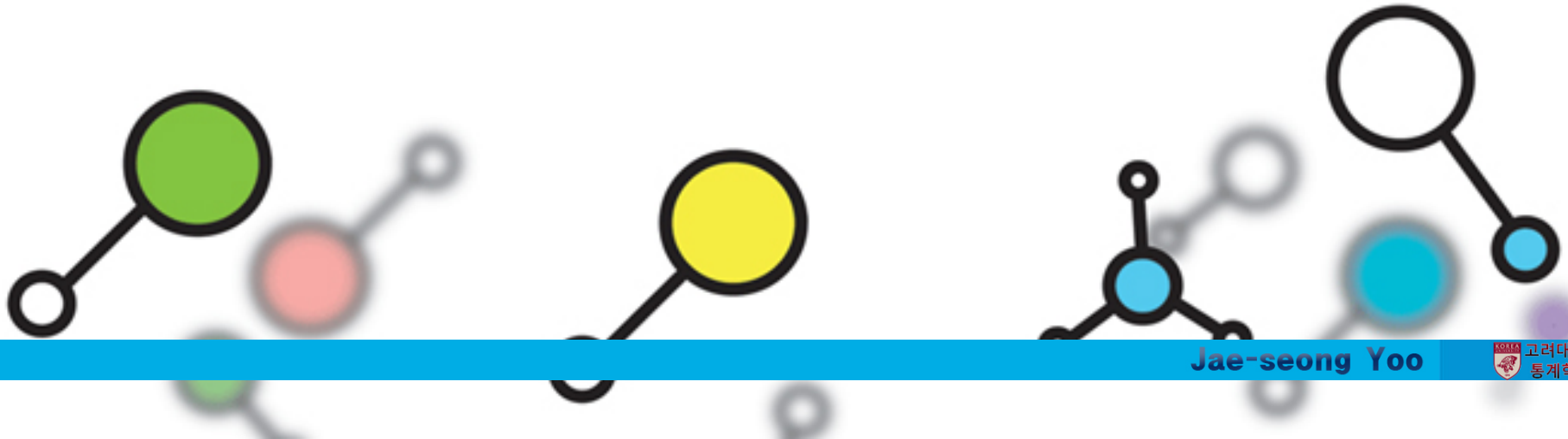
A Study on Comparison of Bayesian Network Structure Learning Algorithm for Selecting Appropriate Model

유재성

Dept. of Statistics

November 14, 2014



Title

Outline

점수 비교

BDe, Log Likelihood, AIC, BIC

모형 형태 비교*

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

* 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

Outline

선행연구

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

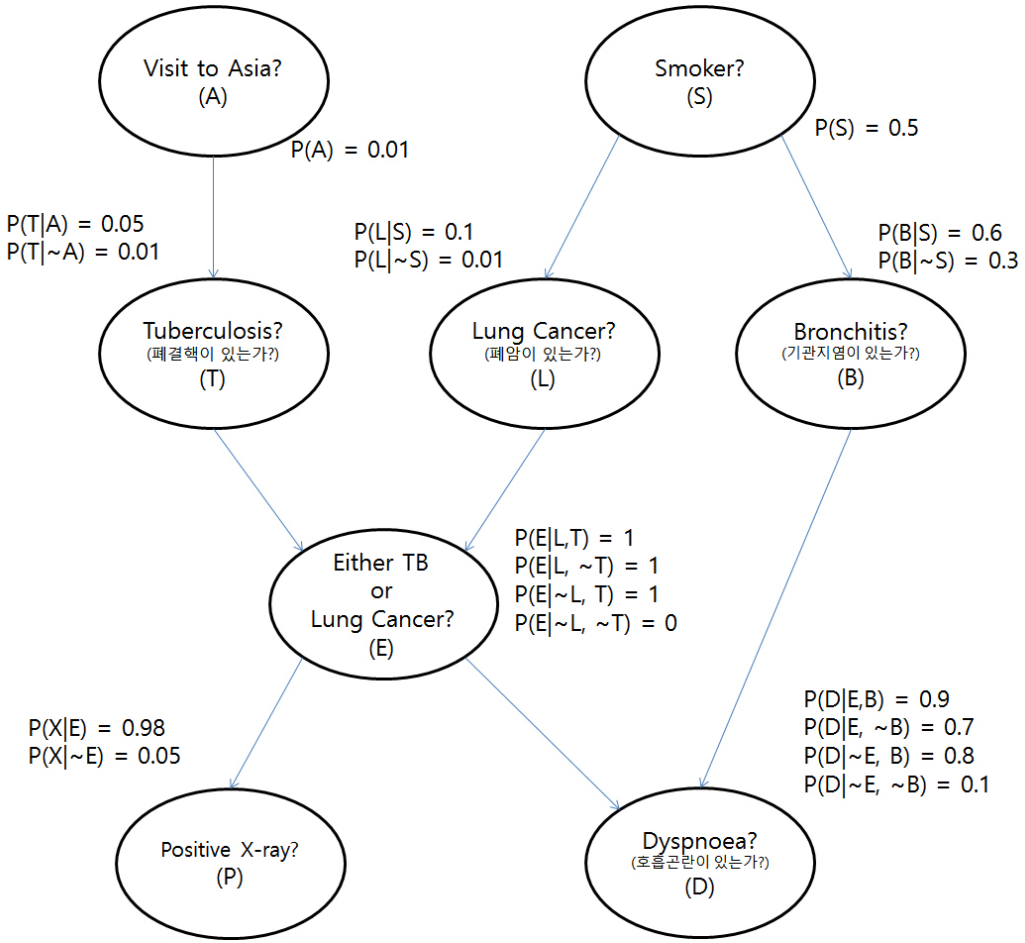
"Learning Bounded Tree-width Bayesian Networks using Integer Linear Programming"

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

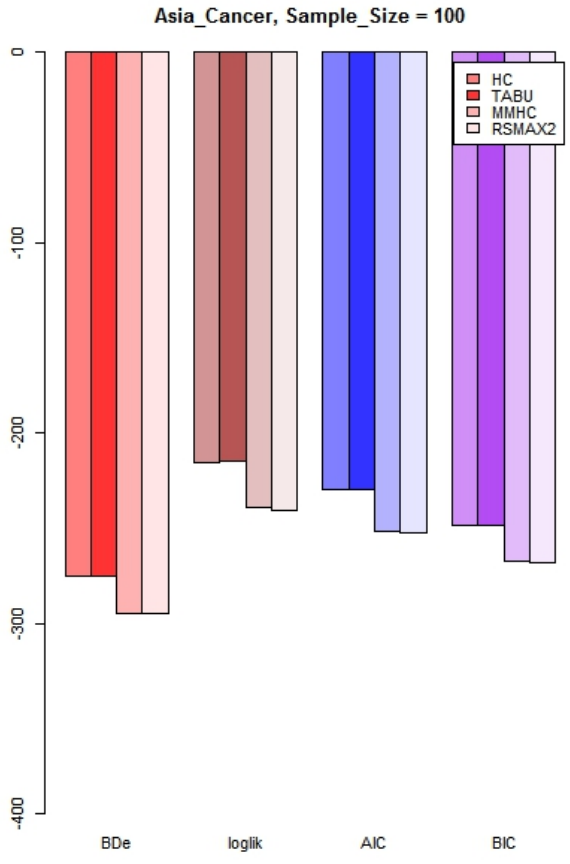
선행연구와 결과 비교 : Asia DataSet

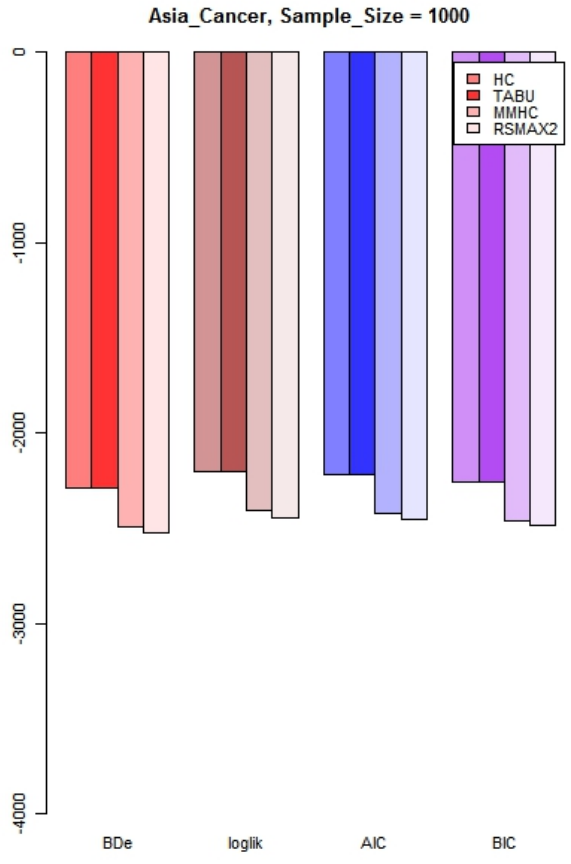
Description	Small synthetic data set from Lauritzen and Spiegelhalter (1988) about lung diseases (tuberculosis, lung cancer or bronchitis) and visits to Asia.
Number of nodes	8
Number of arcs	8
Number of parameters	18
Source	Lauritzen S, Spiegelhalter D (1988). "Local Computation with Probabilities on Graphical Structures and their Application to Expert Systems (with discussion)". Journal of the Royal Statistical Society: Series B (Statistical Methodology), 50(2), 157-224.

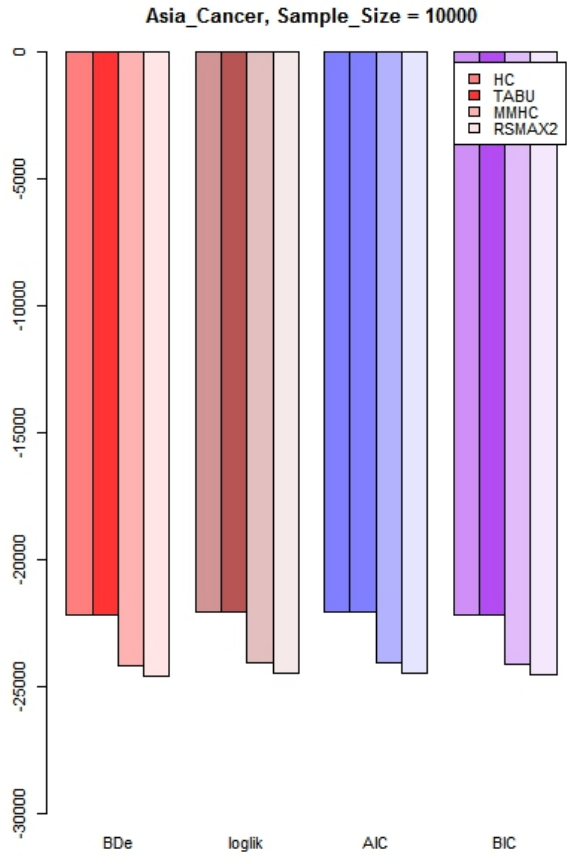
선행연구와 결과 비교 : Asia DataSet



선행연구와 결과 비교 : Asia DataSet

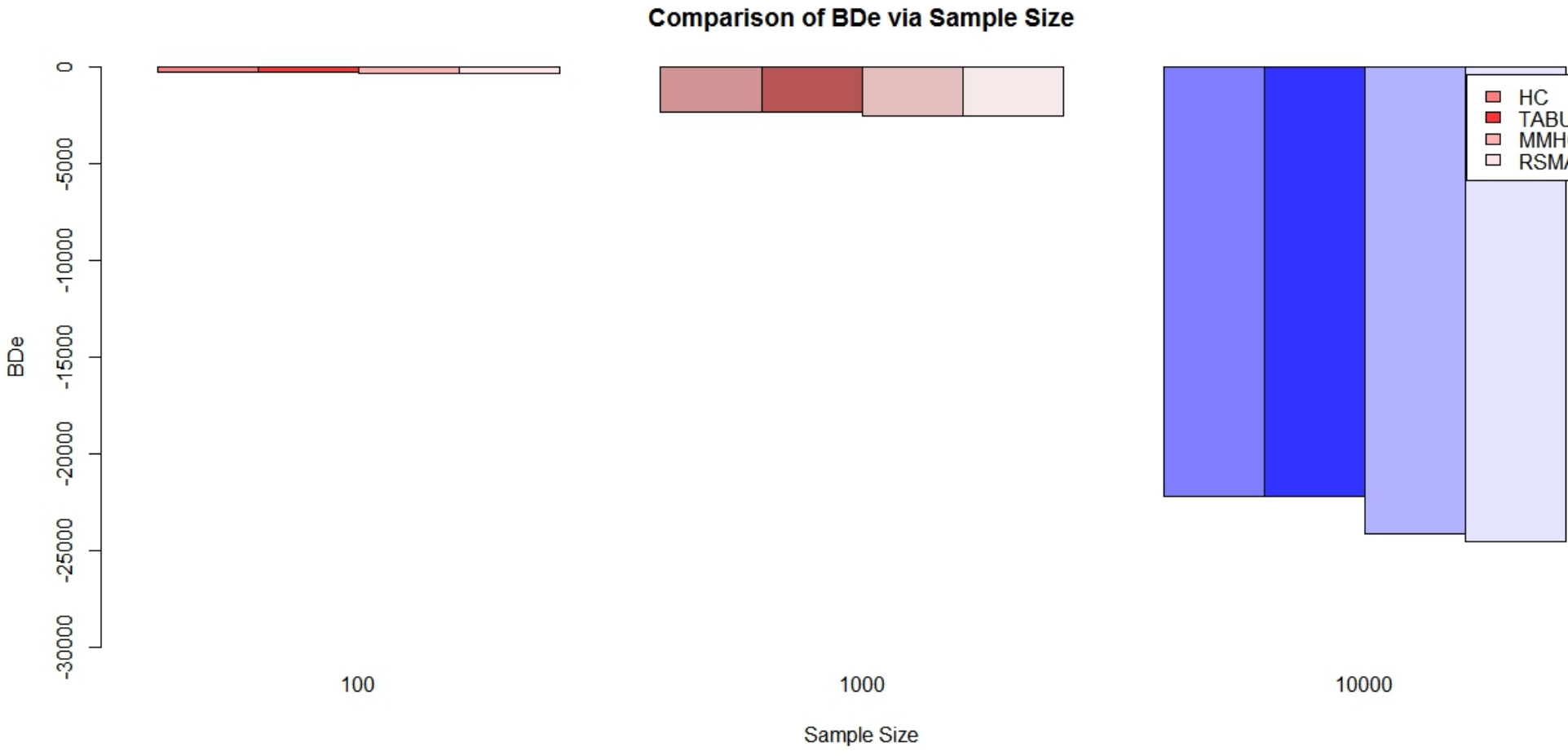




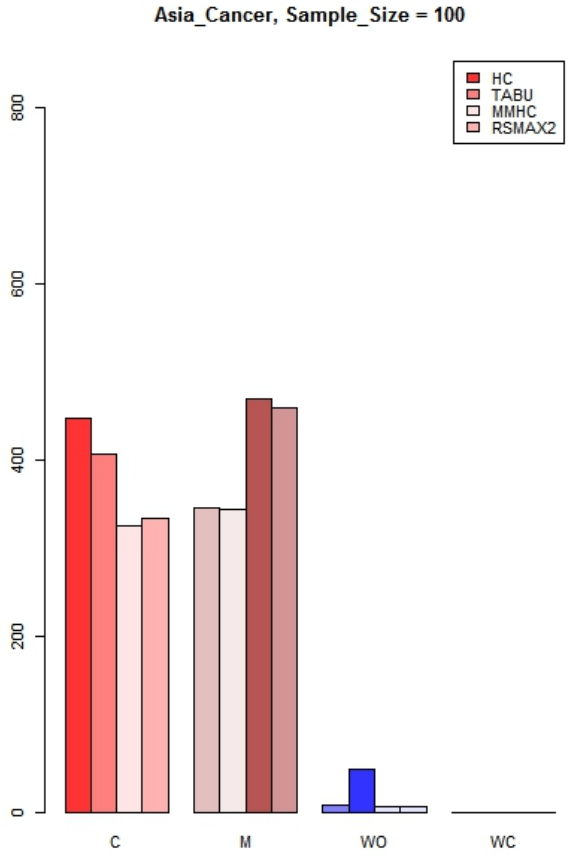


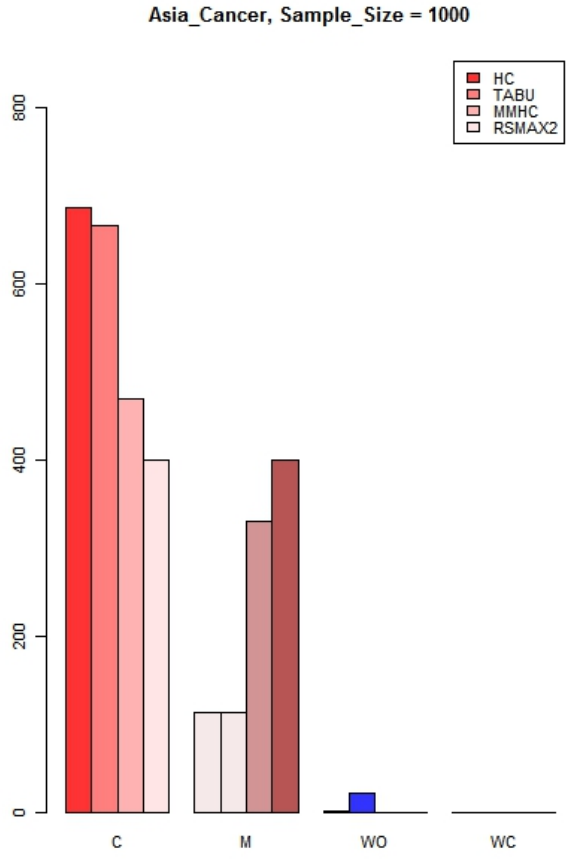
색상이 진해질수록 score가 낮다.

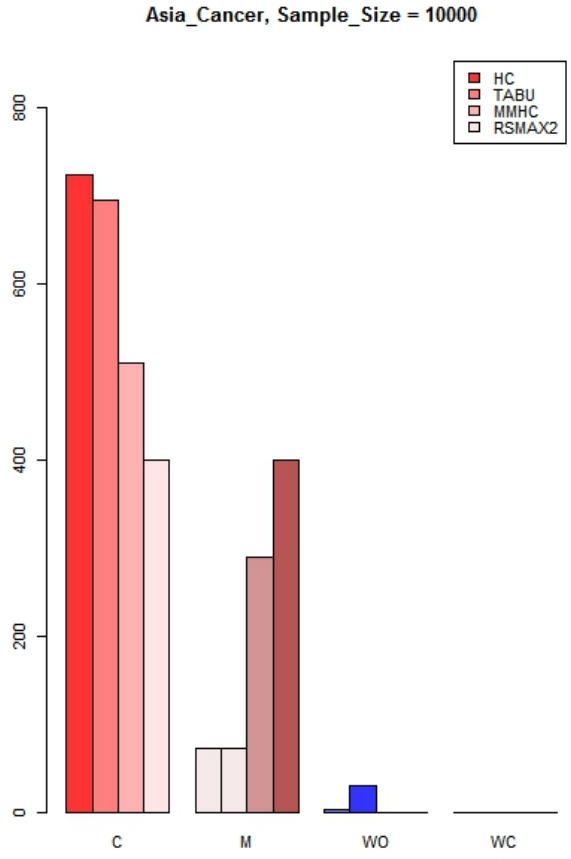
선행연구와 결과 비교 : Asia DataSet



선행연구와 결과 비교 : Asia DataSet



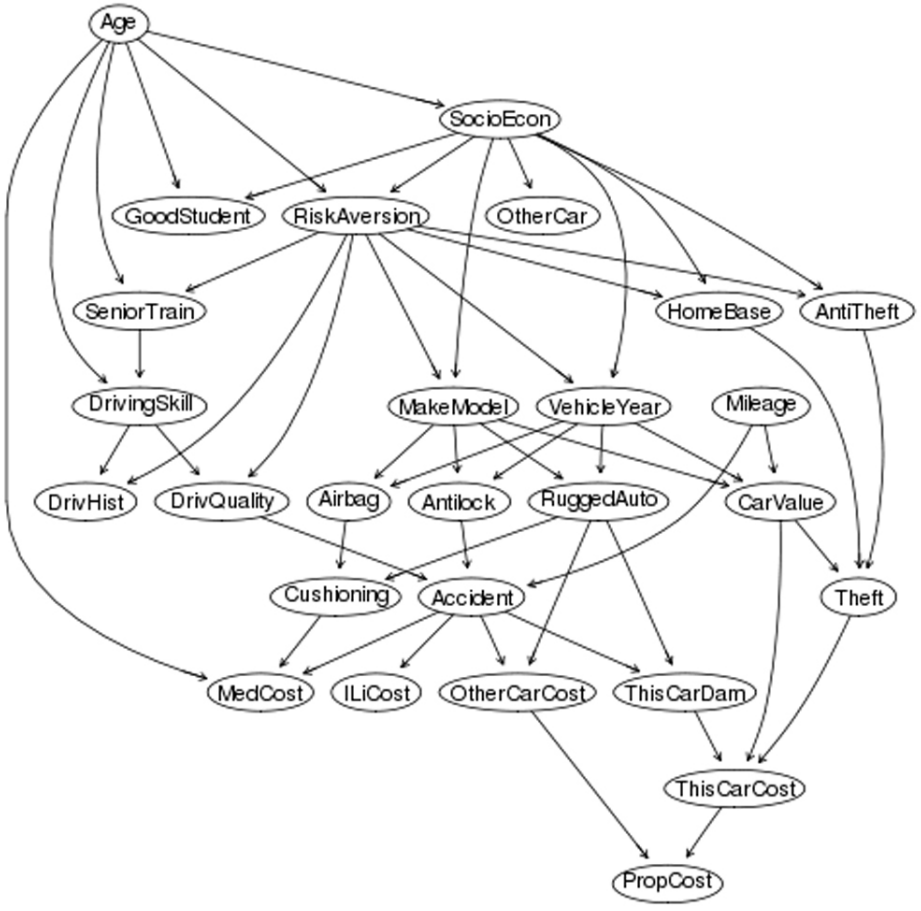




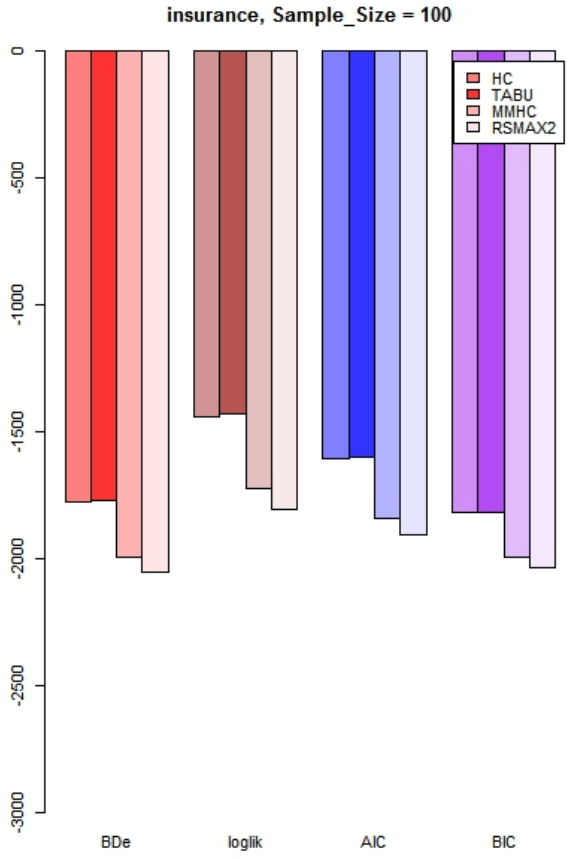
선행연구와 결과 비교 : Insurance DataSet

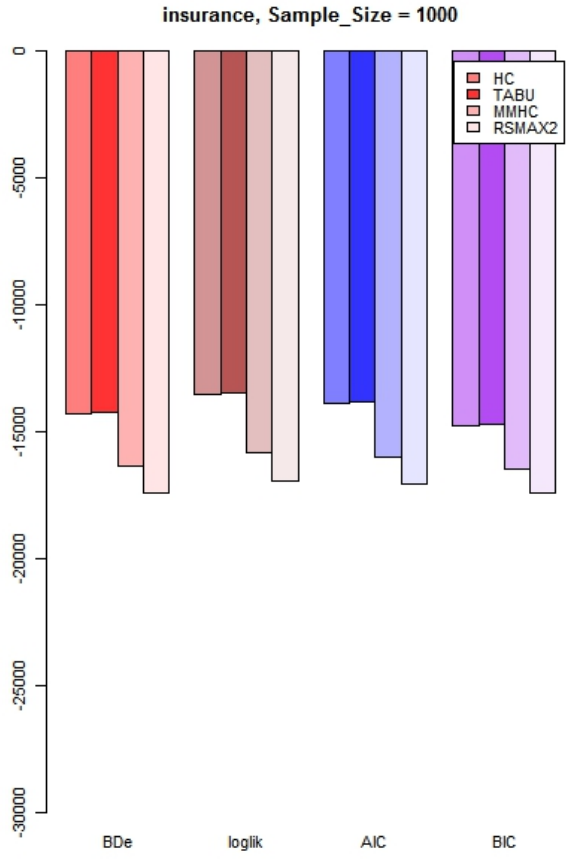
Description	Insurance is a network for evaluating car insurance risks.
Number of nodes	27
Number of arcs	52
Number of parameters	984
Source	Binder J, Koller D, Russell S, Kanazawa K (1997). "Adaptive Probabilistic Networks with Hidden Variables". Machine Learning, 29(2-3), 213-244.

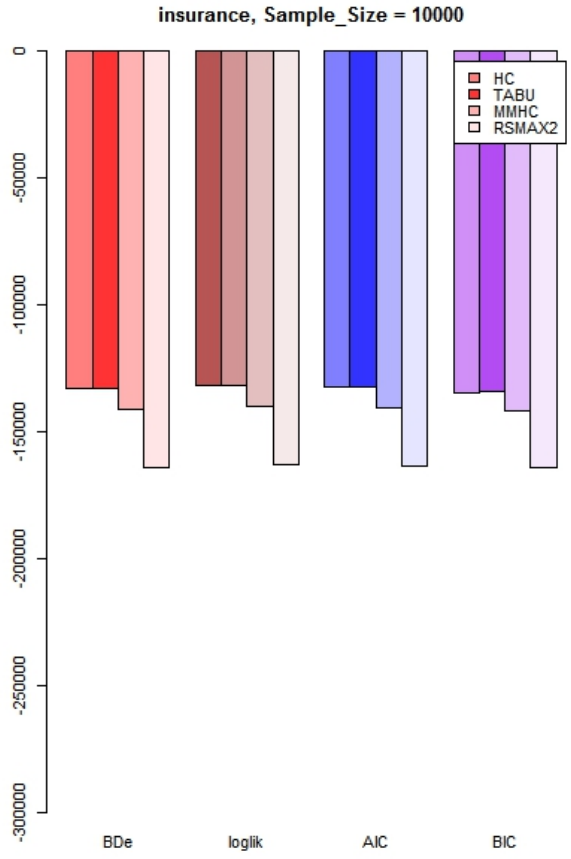
선행연구와 결과 비교 : Insurance DataSet



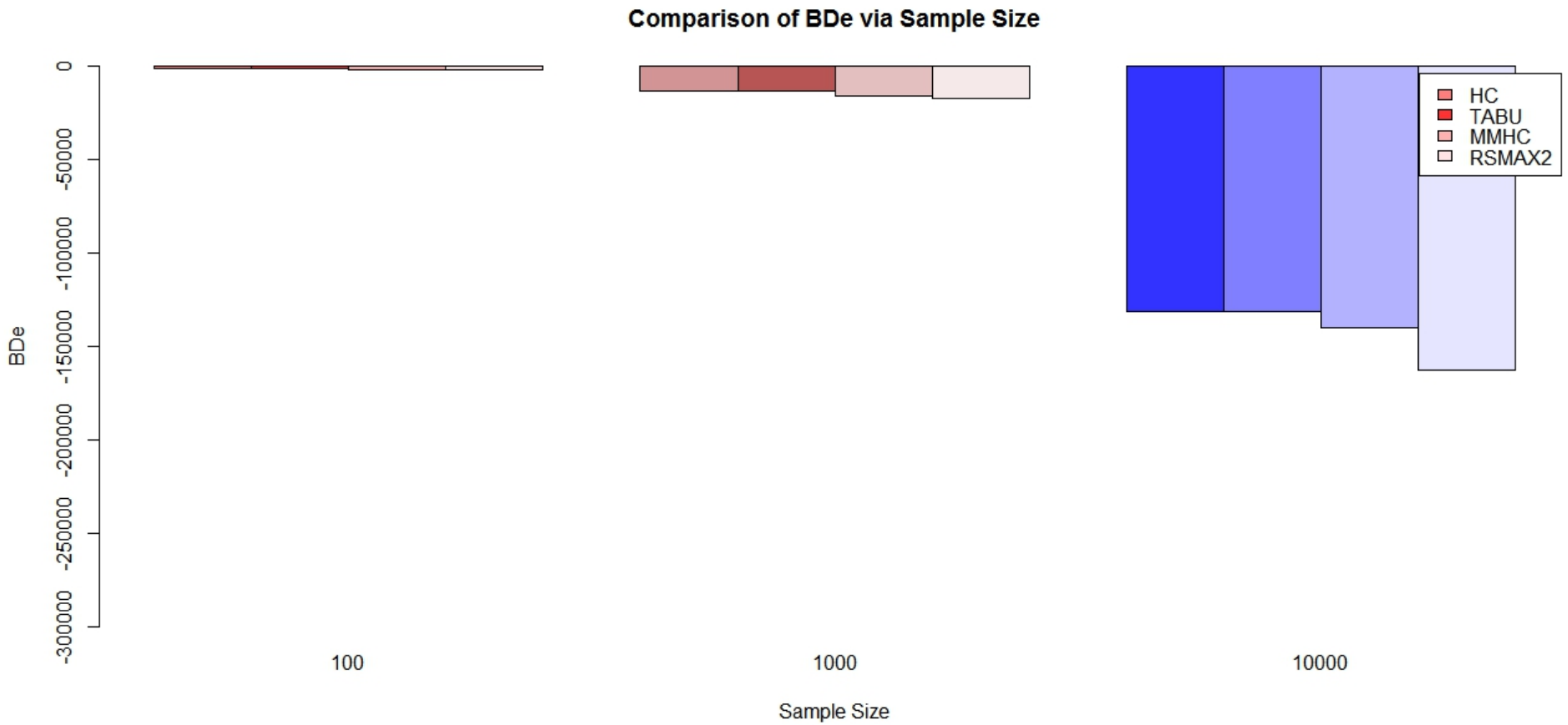
선행연구와 결과 비교 : Insurance DataSet



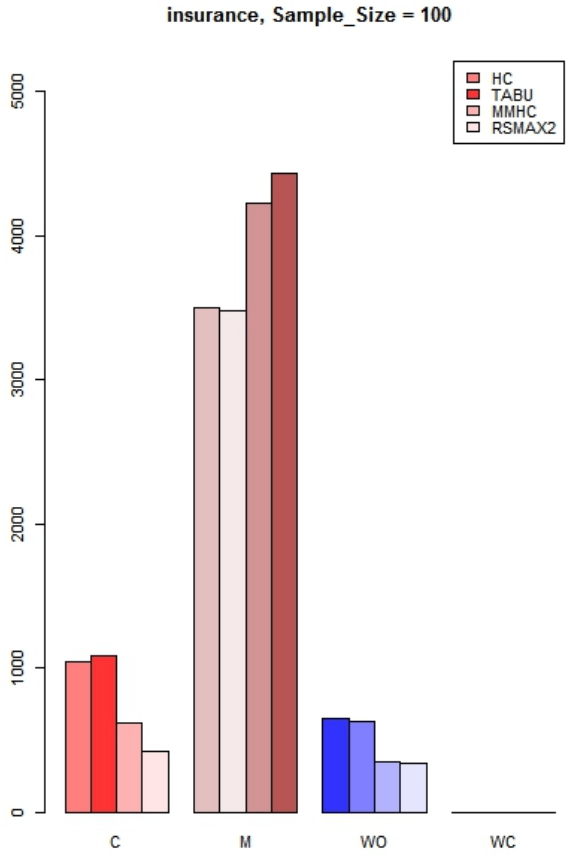


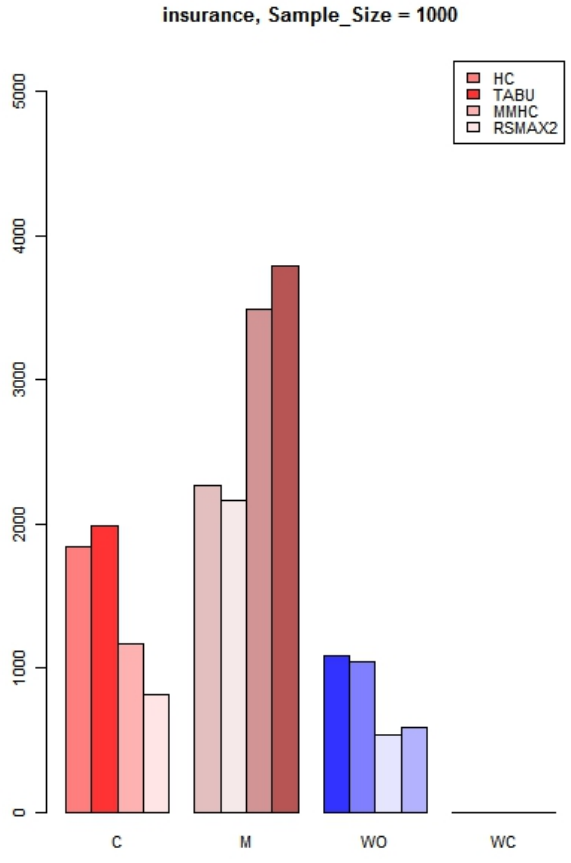


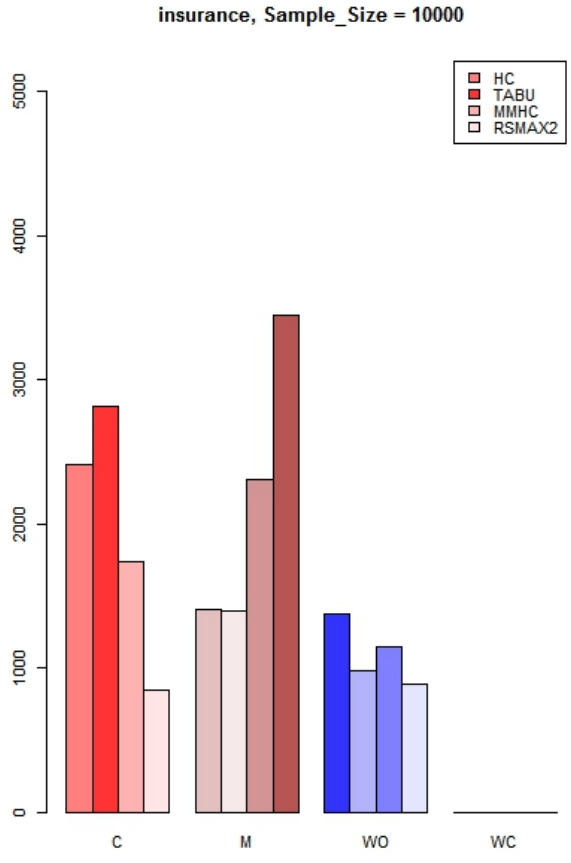
선행연구와 결과 비교 : Insurance DataSet



선행연구와 결과 비교 : Insurance DataSet



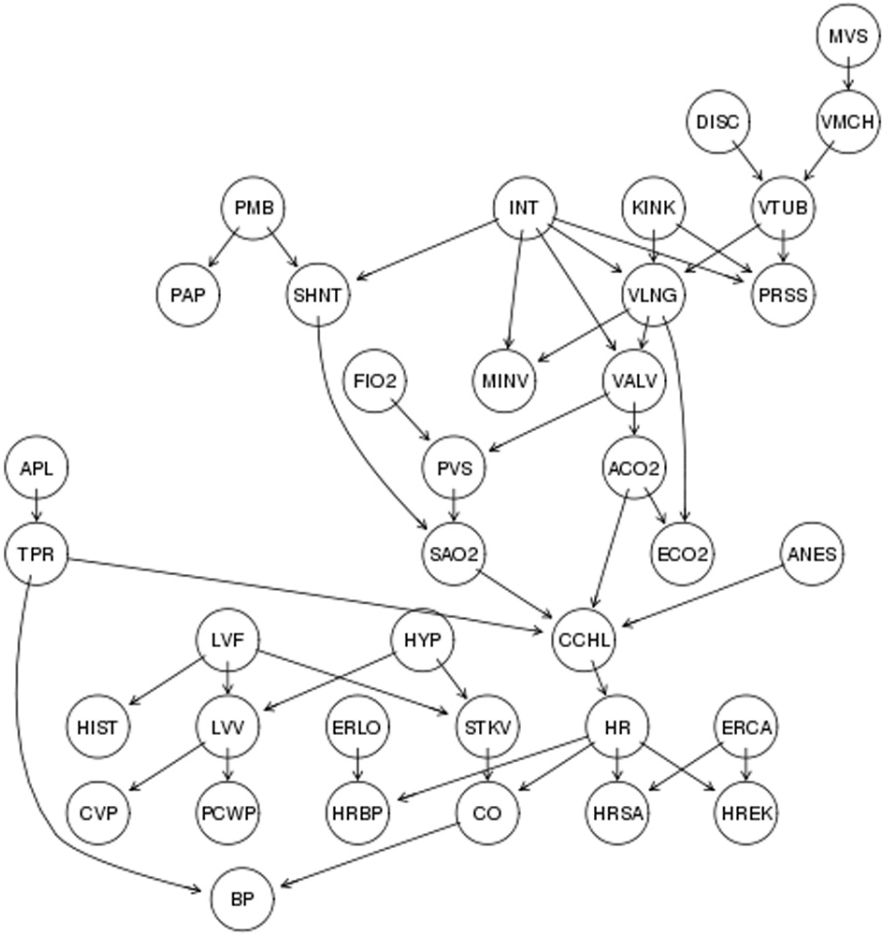




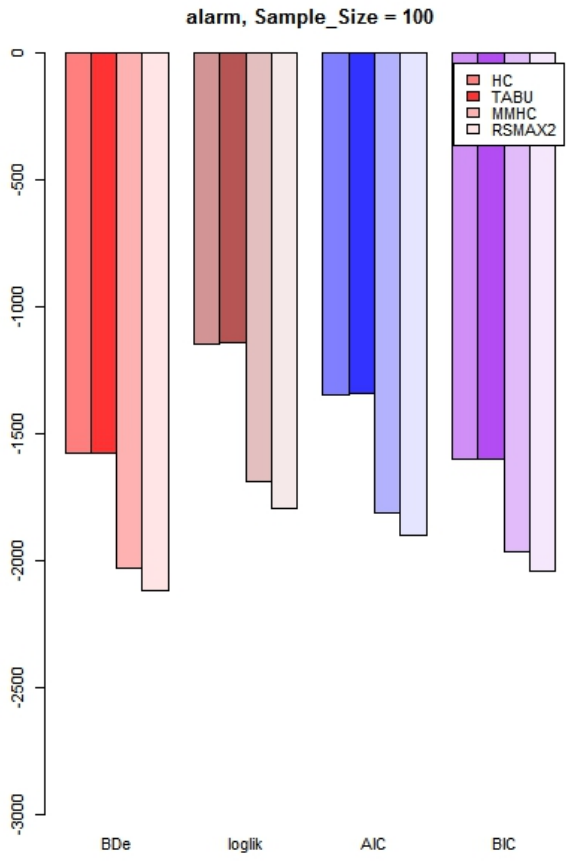
선행연구와 결과 비교 : Alarm DataSet

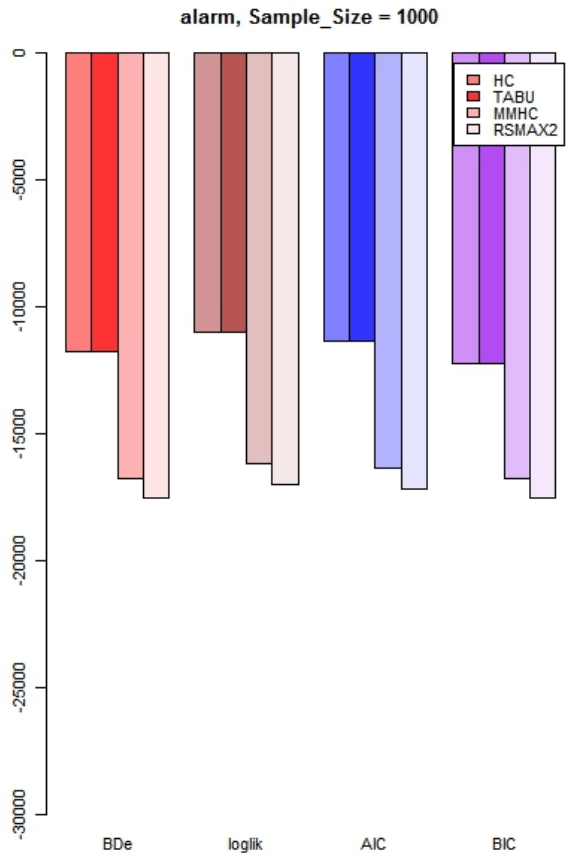
Description	The ALARM ("A Logical Alarm Reduction Mechanism") is a Bayesian network designed to provide an alarm message system for patient monitoring.
Number of nodes	37
Number of arcs	46
Number of parameters	509
Source	Beinlich I, Suermondt HJ, Chavez RM, Cooper GF (1989). "The ALARM Monitoring System: 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.

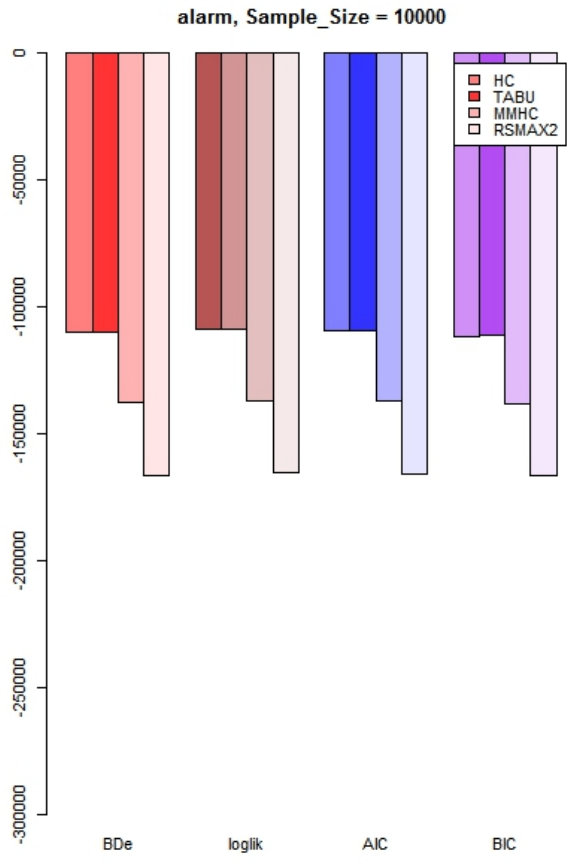
선행연구와 결과 비교 : Alarm DataSet



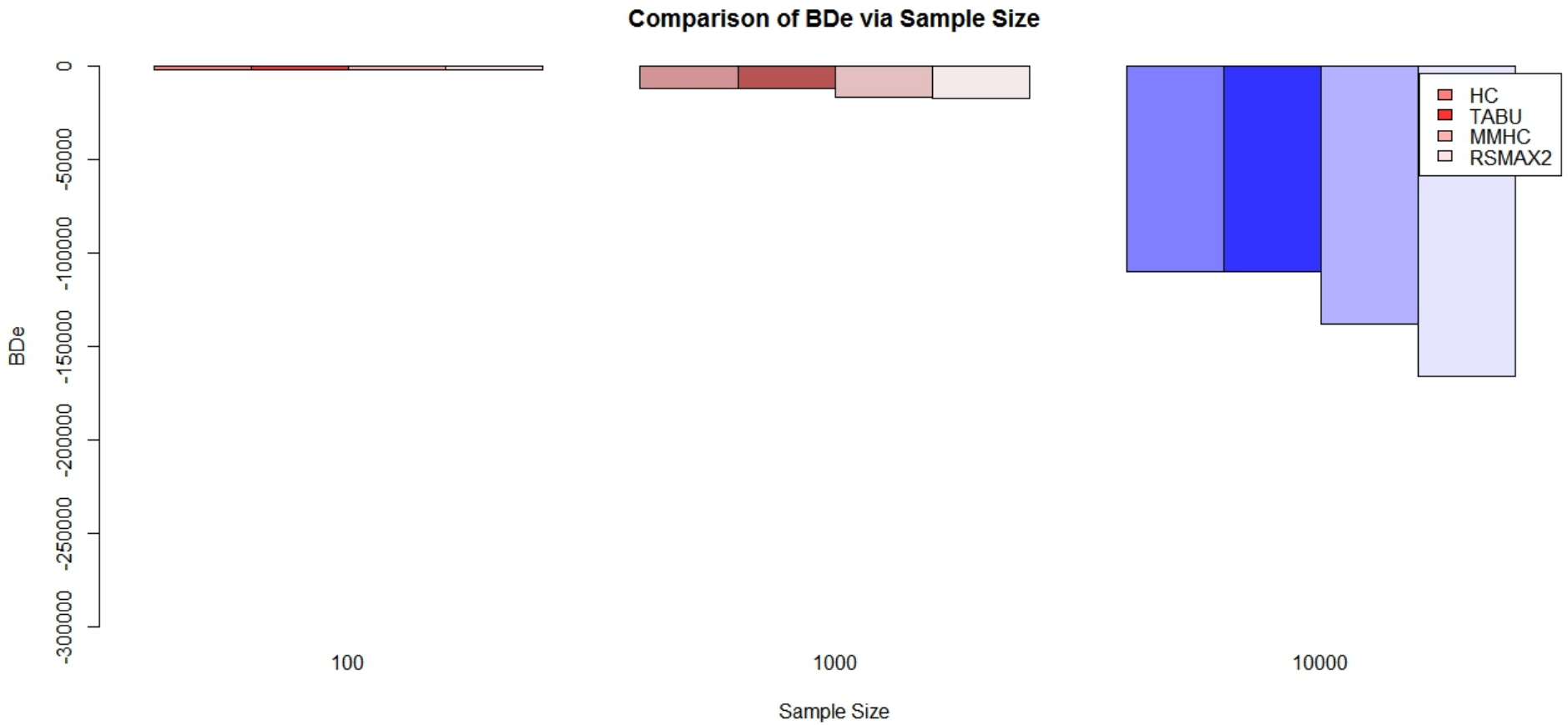
선행연구와 결과 비교 : Alarm DataSet



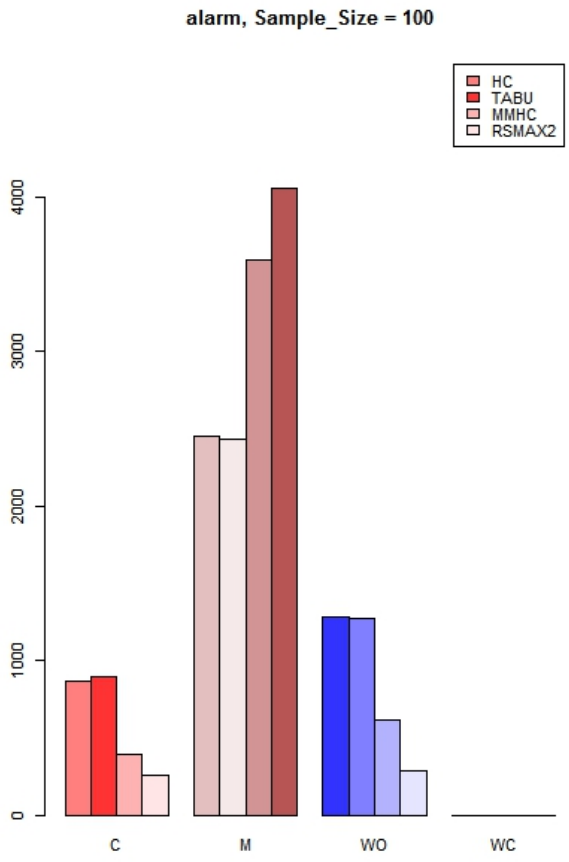


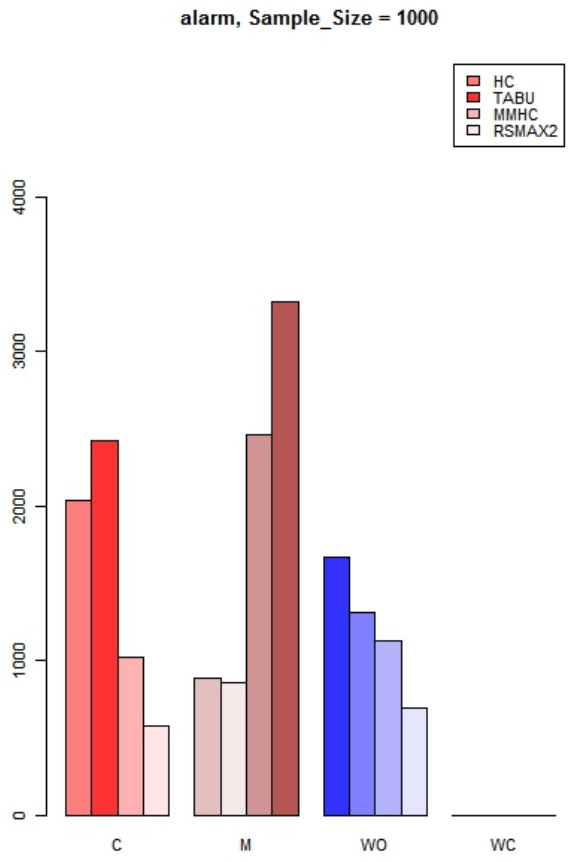


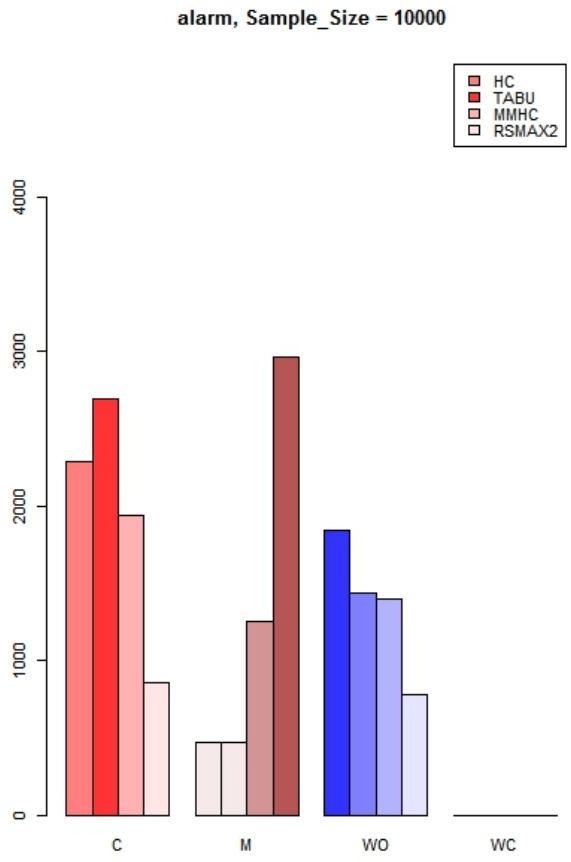
선행연구와 결과 비교 : Alarm DataSet



선행연구와 결과 비교 : Alarm DataSet



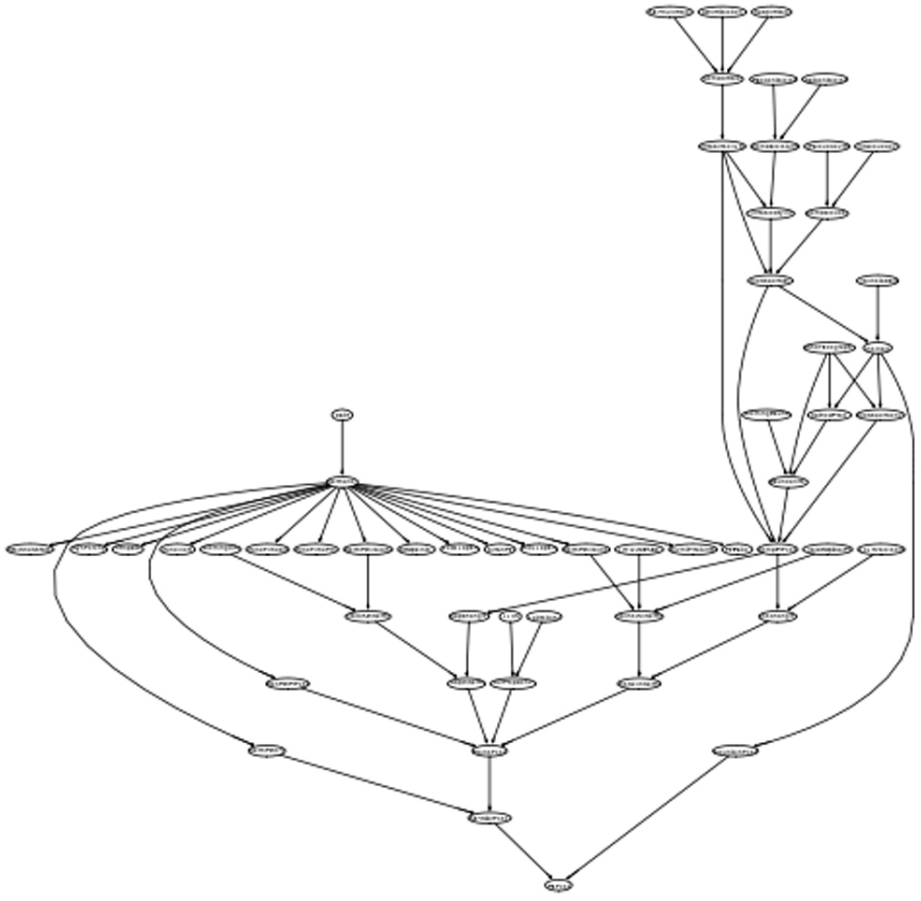




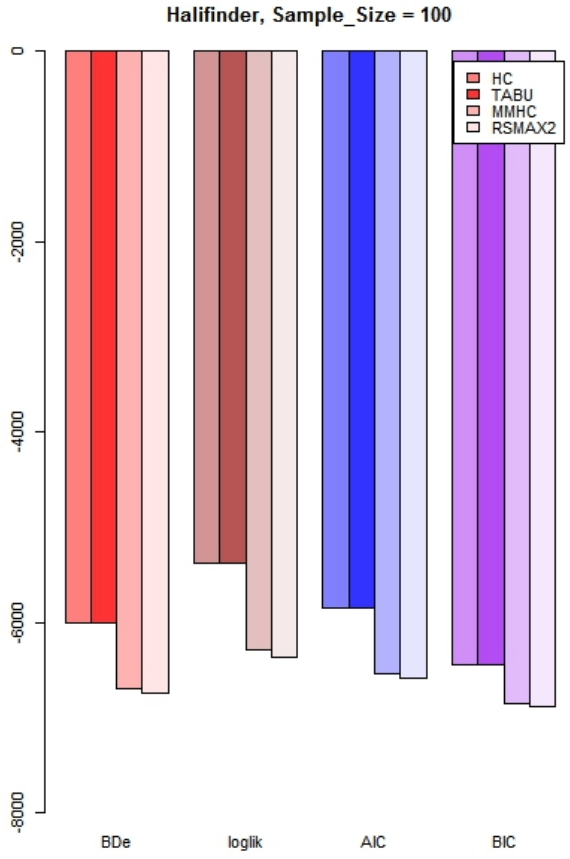
선행연구와 결과 비교 : HailFinder DataSet

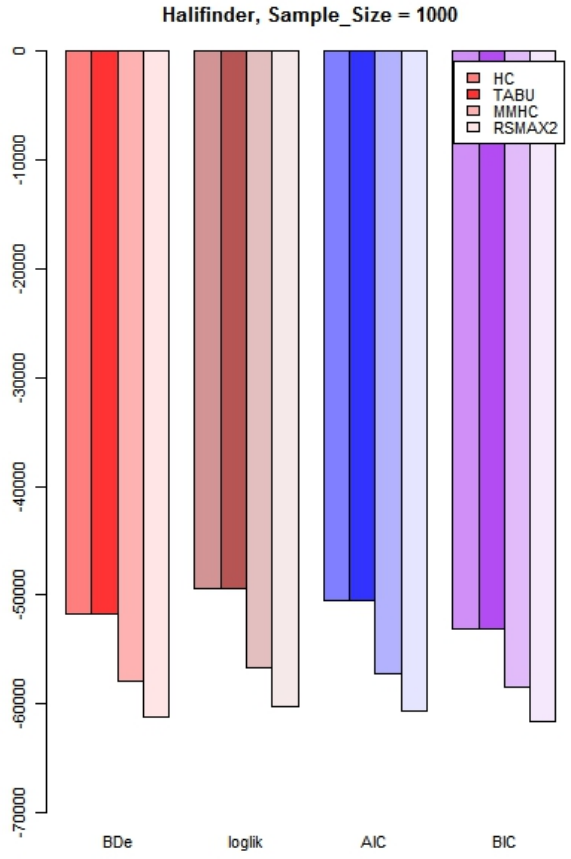
Description	Hailfinder is a Bayesian network designed to forecast severe summer hail in northeastern Colorado.
Number of nodes	56
Number of arcs	66
Number of parameters	2656
Source	Abramson B, Brown J, Edwards W, Murphy A, Winkler RL (1996). "Hailfinder: A Bayesian system for forecasting severe weather". International Journal of Forecasting, 12(1), 57-71.

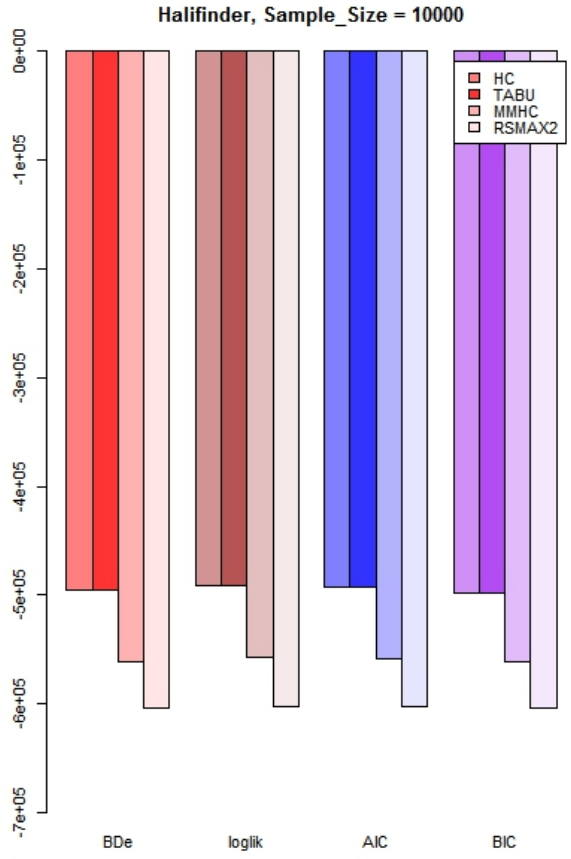
선행연구와 결과 비교 : HailFinder DataSet



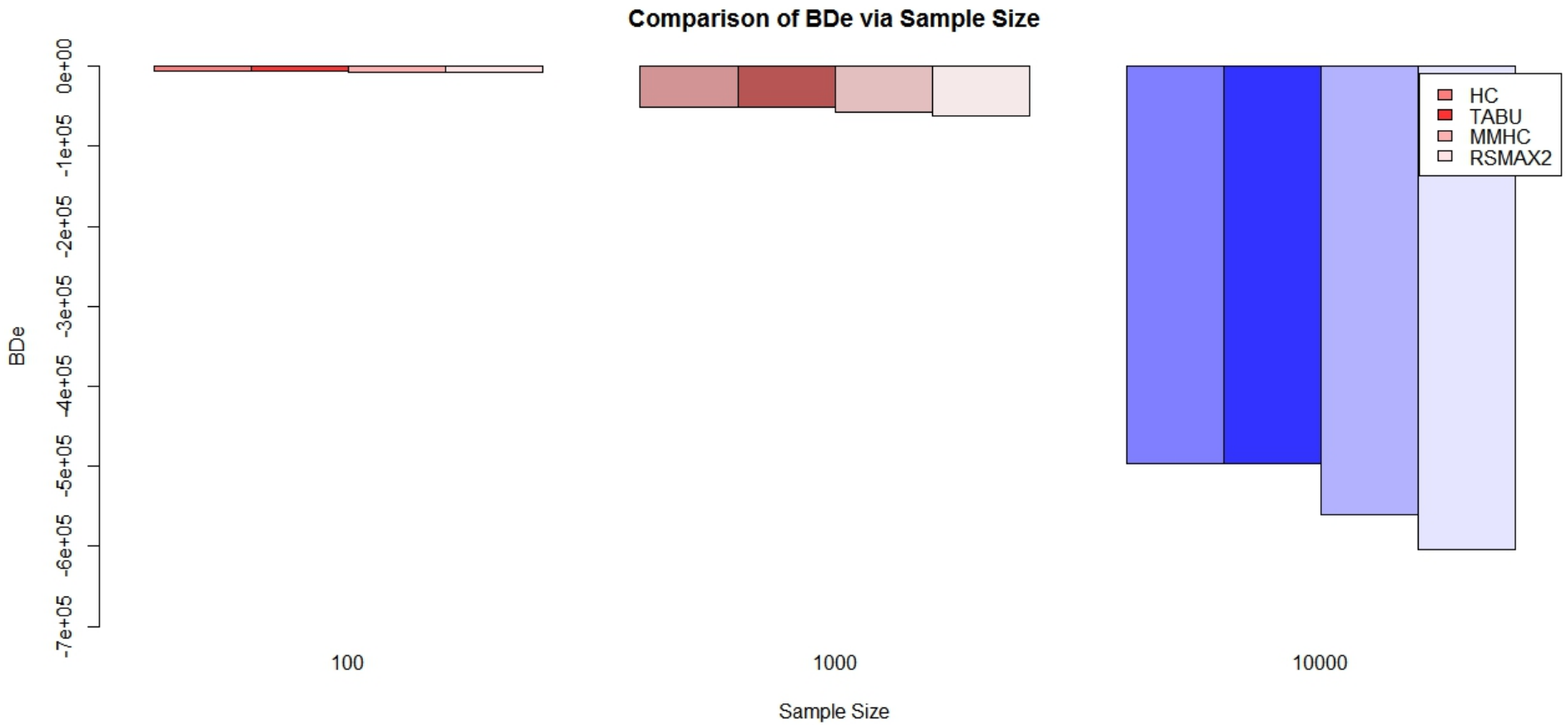
선행연구와 결과 비교 : HailFinder DataSet



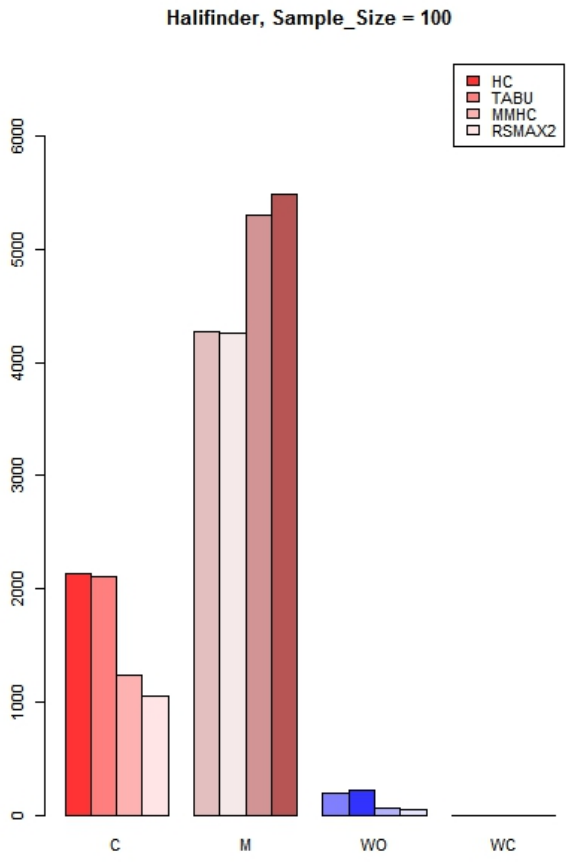


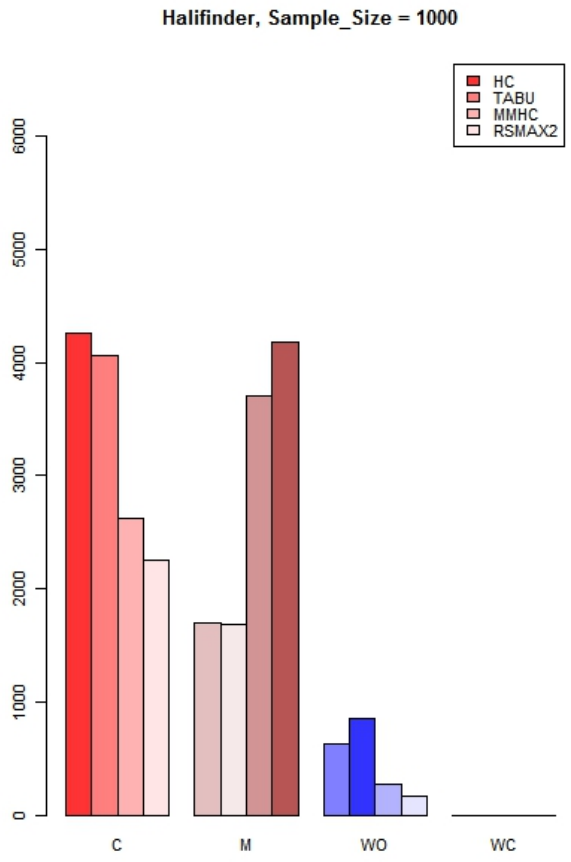


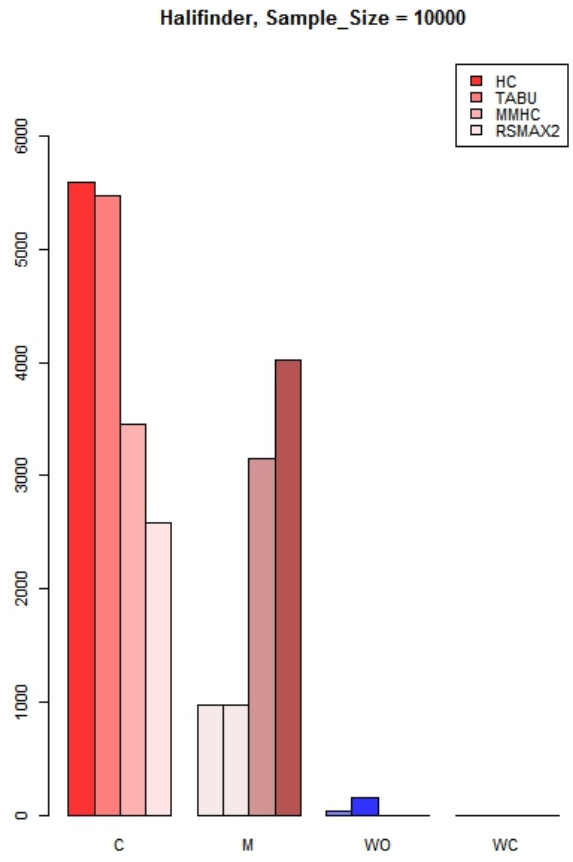
선행연구와 결과 비교 : HailFinder DataSet



선행연구와 결과 비교 : HailFinder DataSet







Summary

By Score	Best			Worst		
	100	1000	10000	100	1000	10000
Asia	TABU	TABU	TABU	RSMAX2	RSMAX2	RSMAX2
Insurance	TABU	TABU	TABU	RSMAX2	RSMAX2	RSMAX2
Alarm	TABU	TABU	TABU	RSMAX2	RSMAX2	RSMAX2
HaliFinder	TABU	TABU	TABU	RSMAX2	RSMAX2	RSMAX2

HC TABU MMHC RSMAX2

By Model	Best			Worst		
	100	1000	10000	100	1000	10000
Asia	HC	HC	HC	MMHC	RSMAX2	RSMAX2
Insurance	TABU	HC	TABU	RSMAX2	RSMAX2	RSMAX2
Alarm	TABU	TABU	TABU	RSMAX2	RSMAX2	RSMAX2
HaliFinder	HC	HC	HC	RSMAX2	RSMAX2	RSMAX2

HC TABU MMHC RSMAX2

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)	aa
input_Probs	(list)	bb
n	(constant)	Sample Size
node_names	(vector)	dd

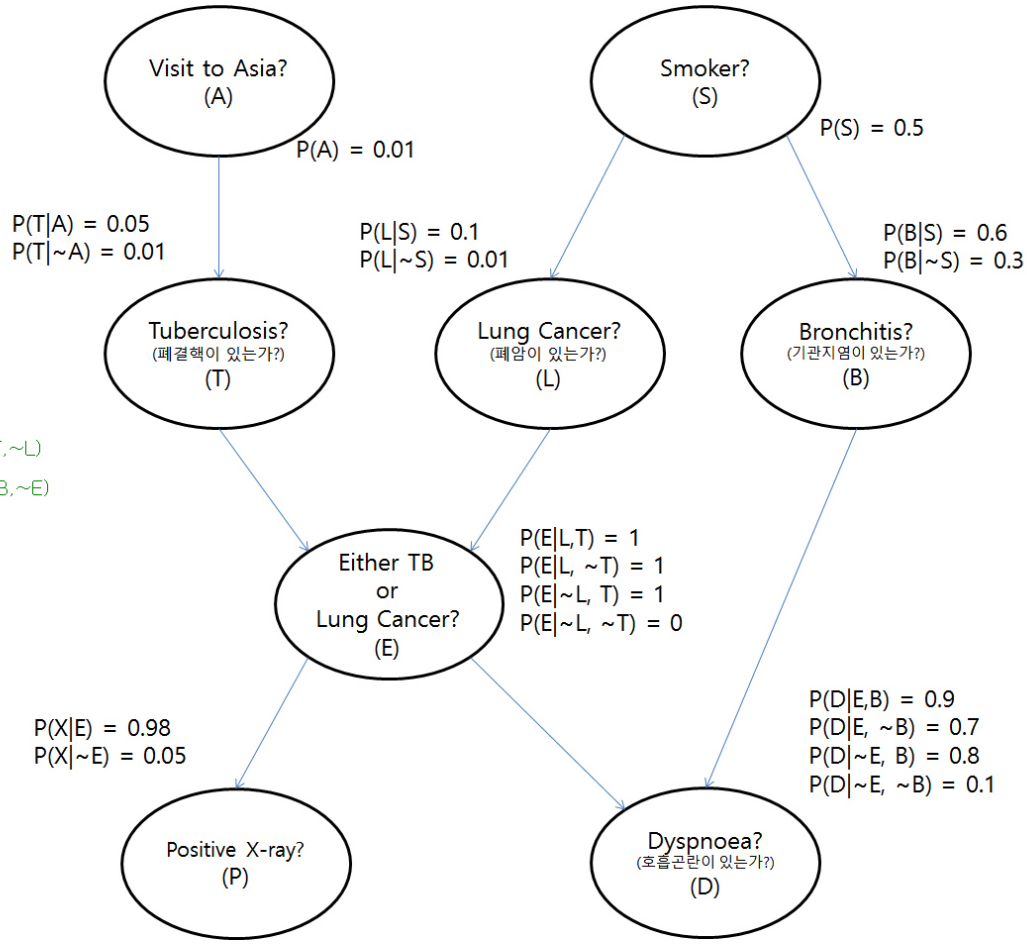
Data Generator 성능 평가

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
```

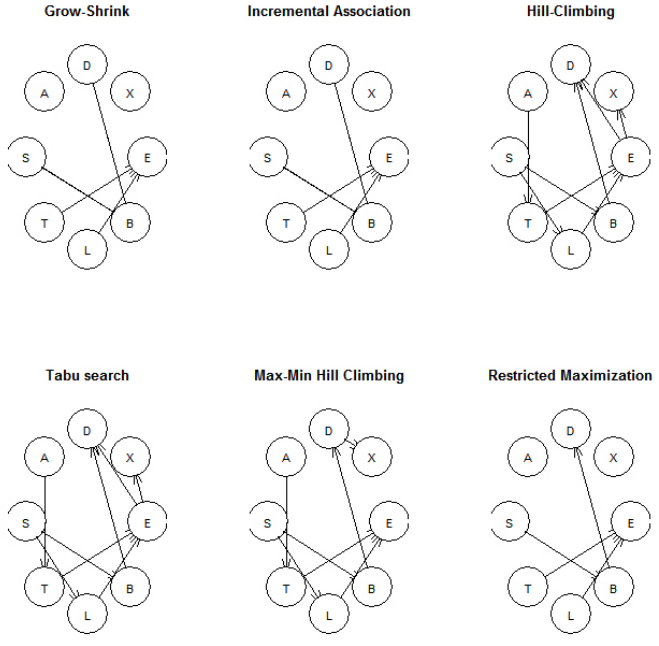
Environment History

Import Dataset Clear List

Global Environment

Files Plots Packages Help Viewer

Zoom Export Clear All



Data Generator 성능 평가

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

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

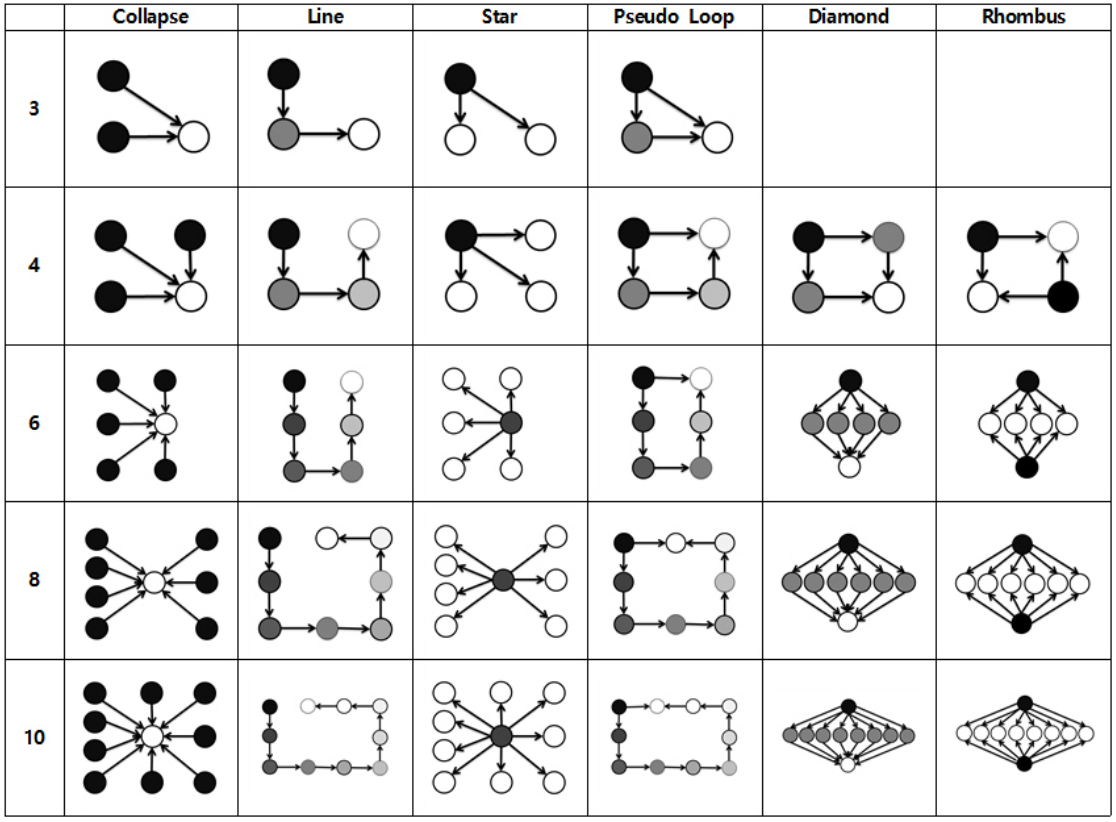
Dataset	Num. of Nodes	Sample Size	ILP	Score via MMHC			
			BDe	BDe	loglik	AIC	BIC
Asia	8	100	-245.64	-301.29	-246.32	-257.32	-271.65
real		1000	-2317.41	-2504.74	-2421.80	-2436.80	-2473.61
dataset		10000	-22466.40	-24346.08	-24232.89	-24249.89	-24311.18
Gen.		100		-272.31	-213.08	-225.08	-240.71
by		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

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

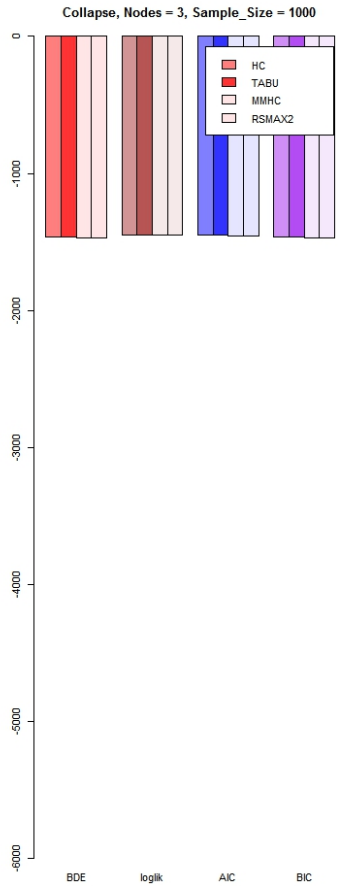
Outline

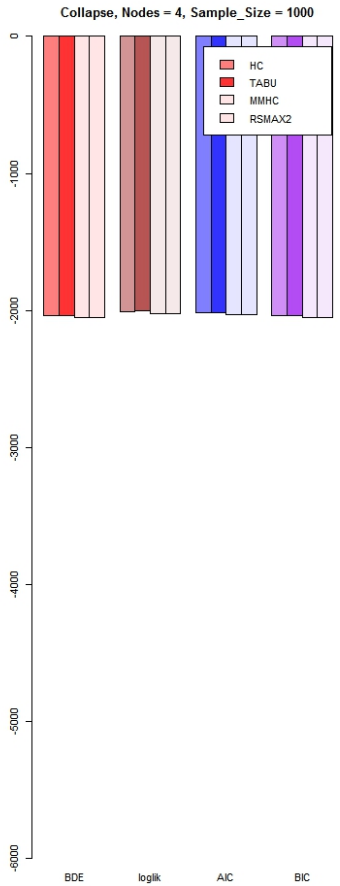
Synthetic Data 유형에 따른 비교 분석

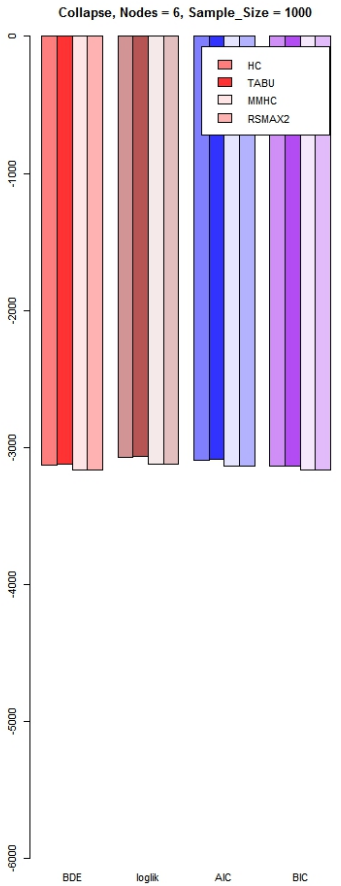


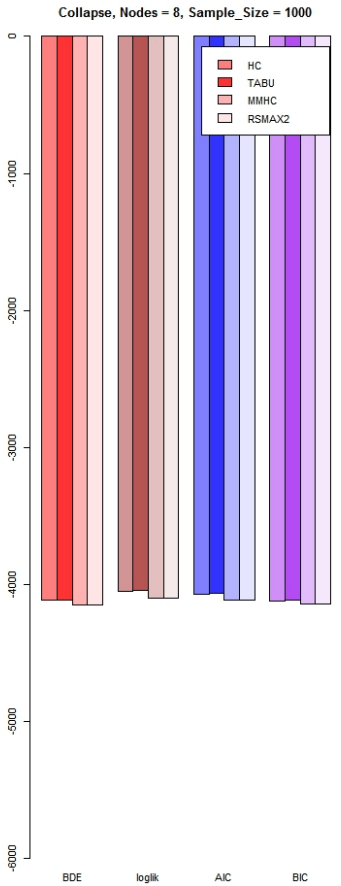
Eitel J. M. Lauría,
"An Information-Geometric Approach to Learning Bayesian Network Topologies from Data",
Innovations in Bayesian Networks Studies in Computational Intelligence Volume 156, 2008, pp 187-217

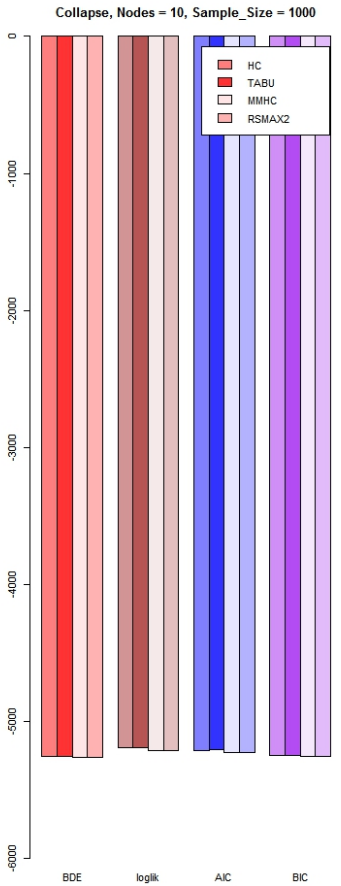
Synthetic Data 유형에 따른 비교 분석



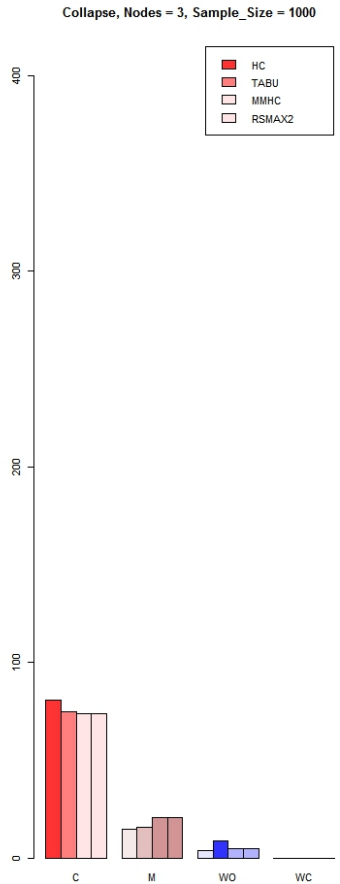


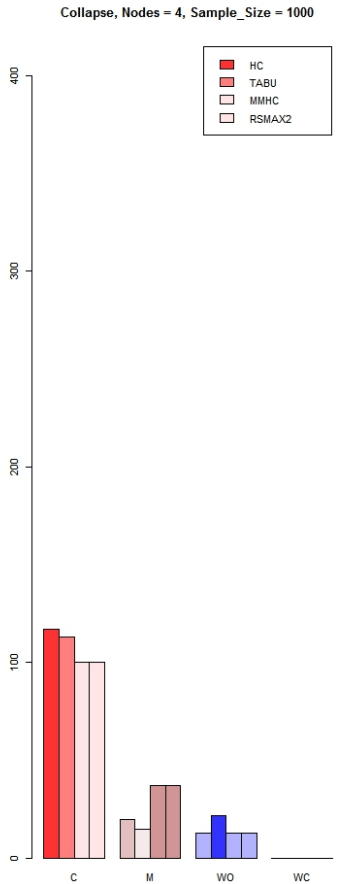


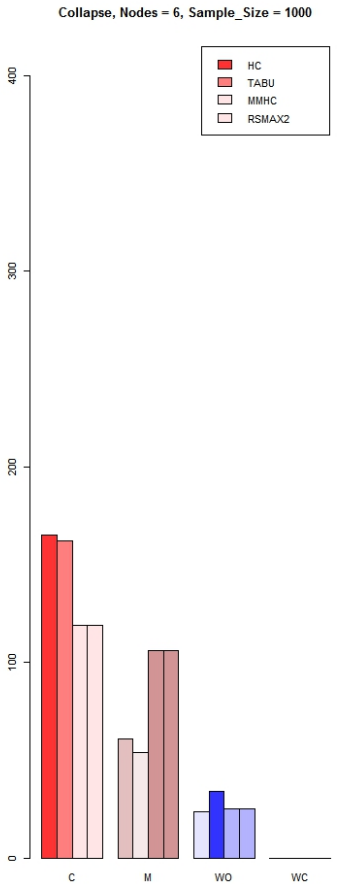


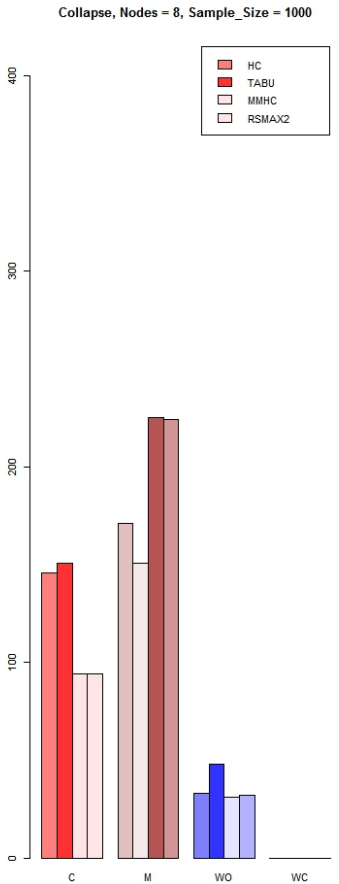


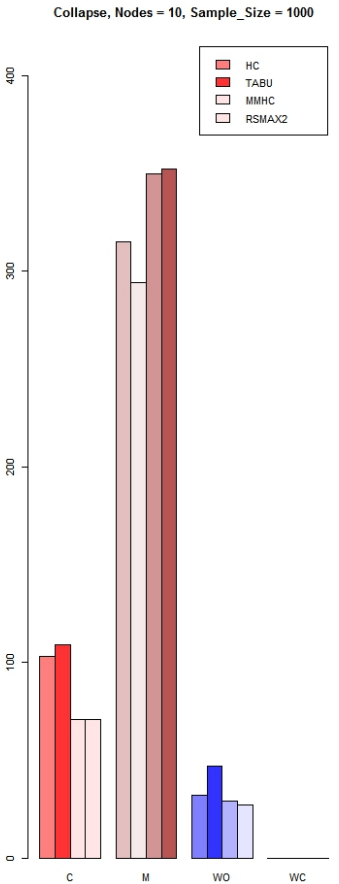
Synthetic Data 유형에 따른 비교 분석



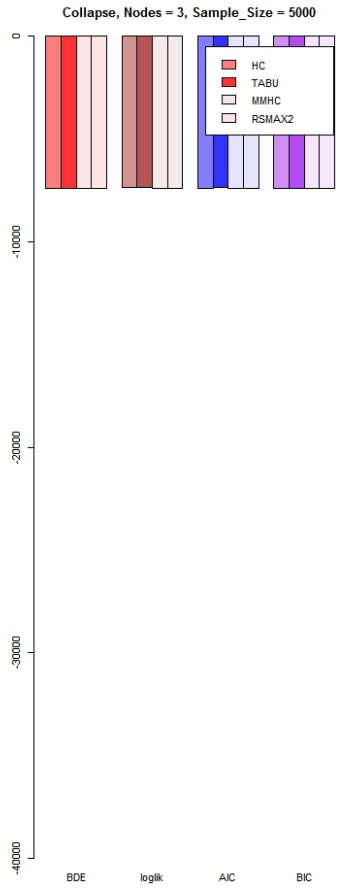


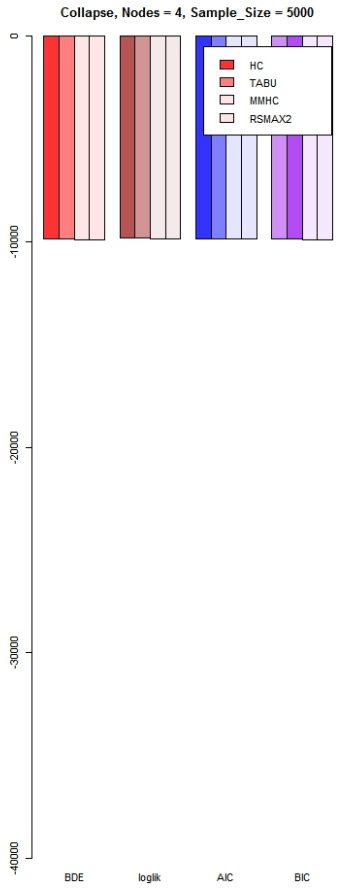


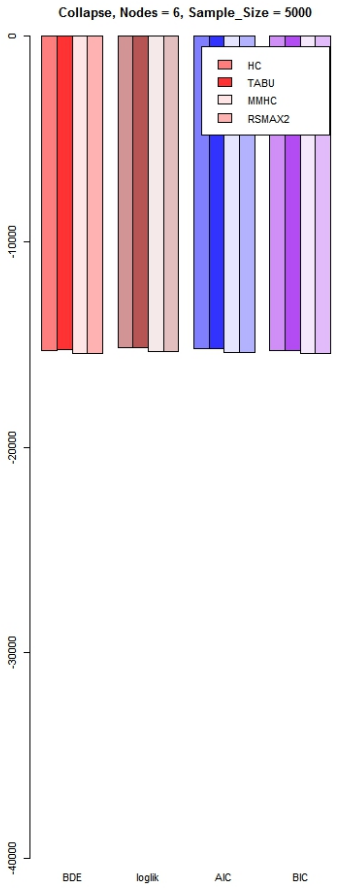


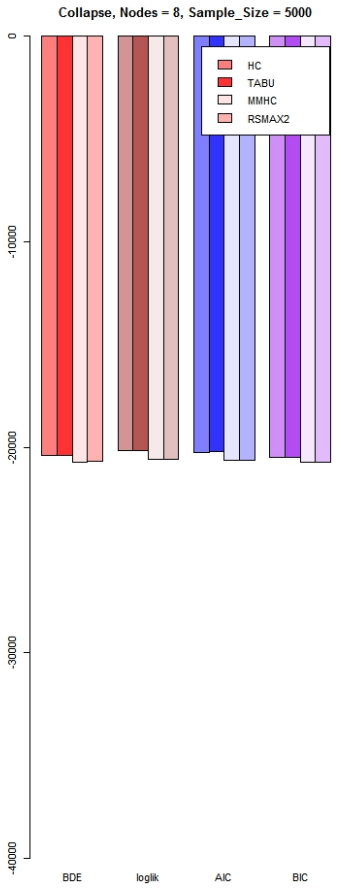


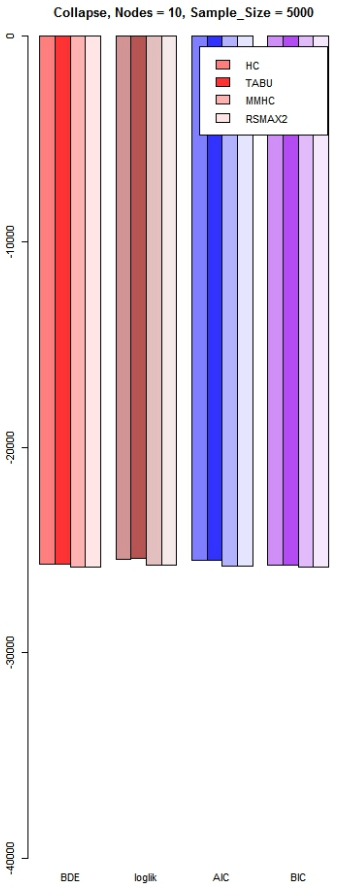
Synthetic Data 유형에 따른 비교 분석



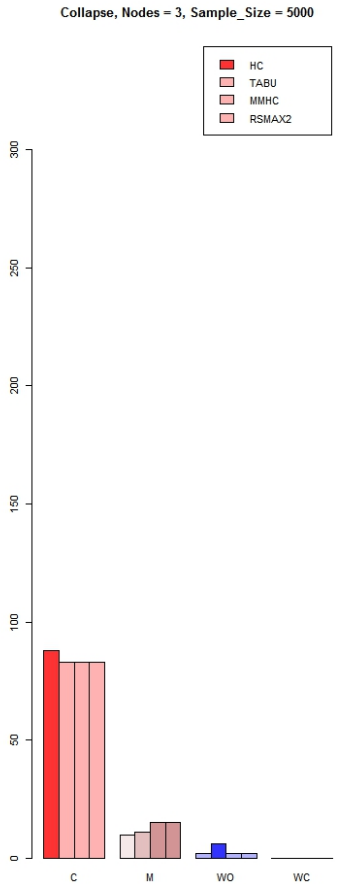


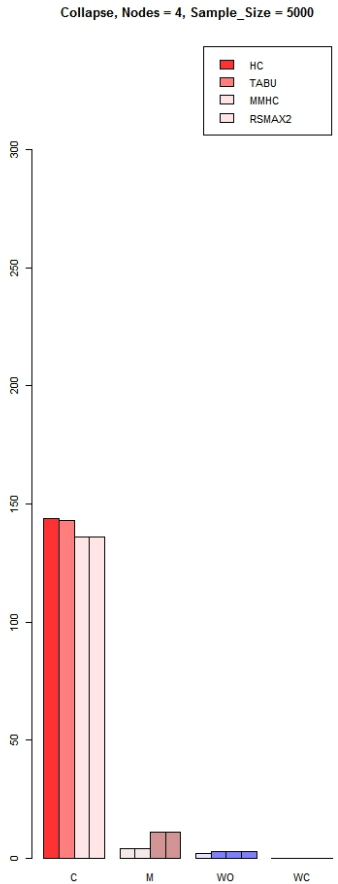


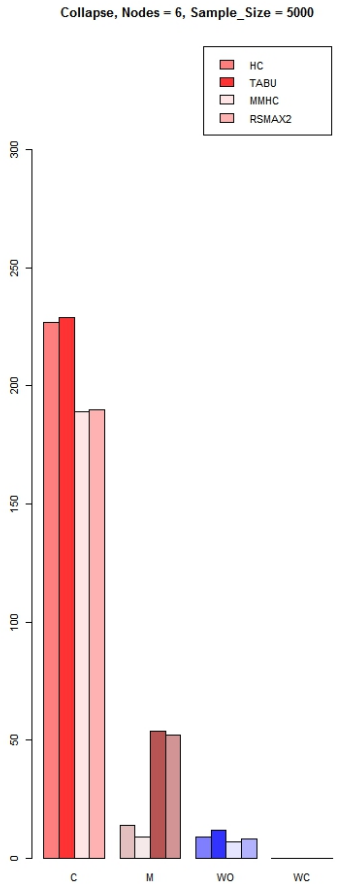


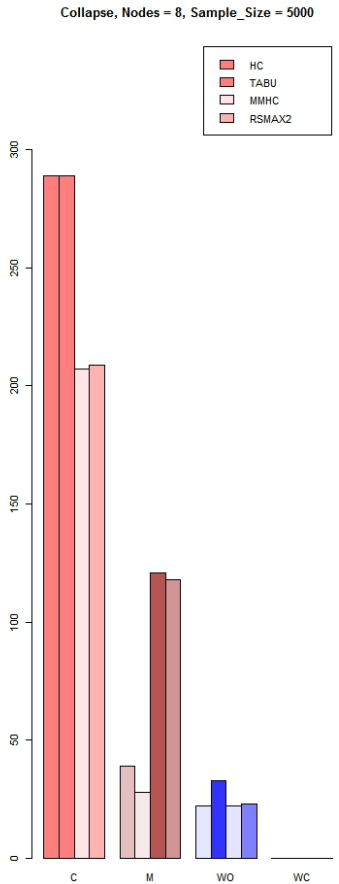


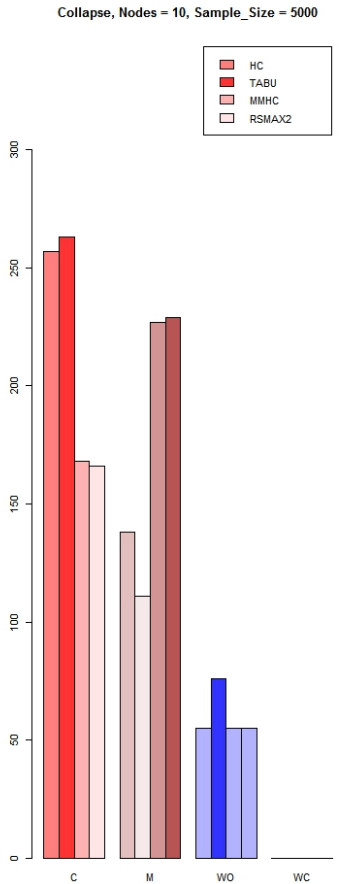
Synthetic Data 유형에 따른 비교 분석



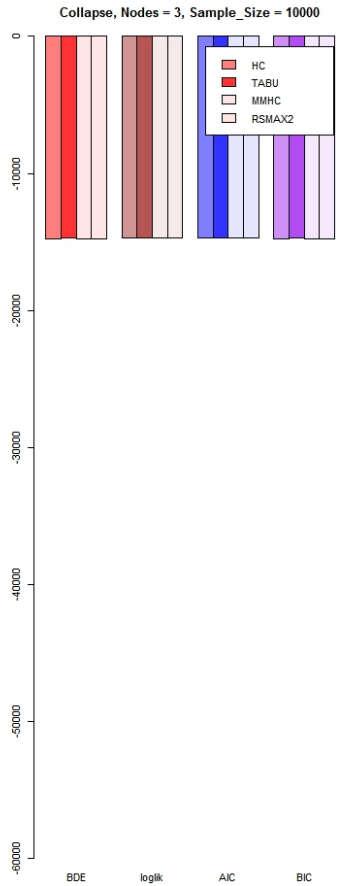


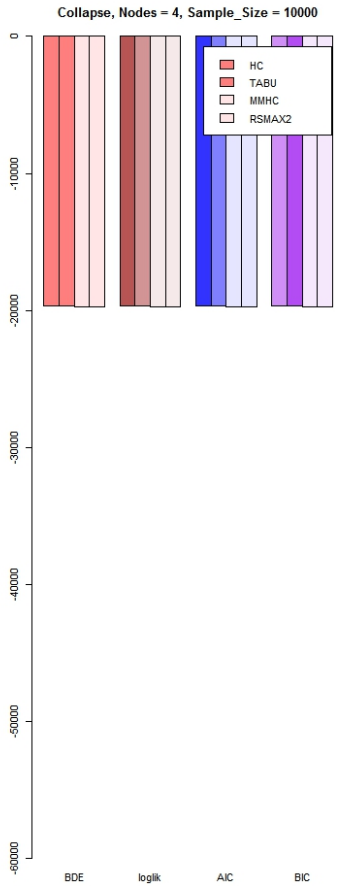


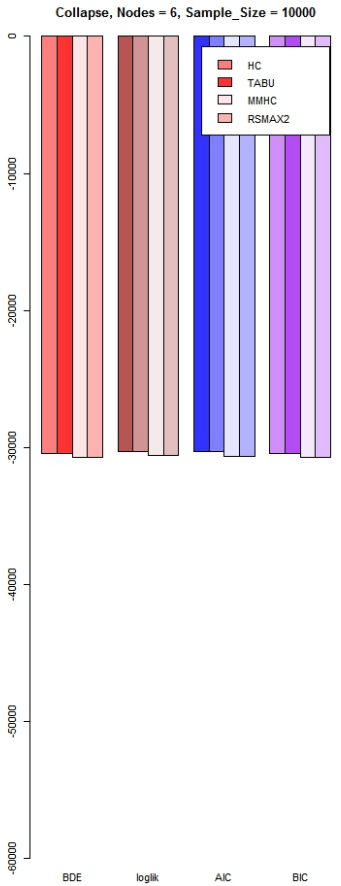


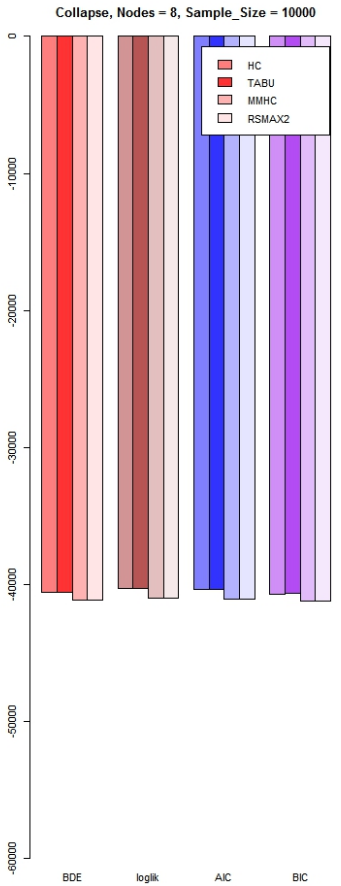


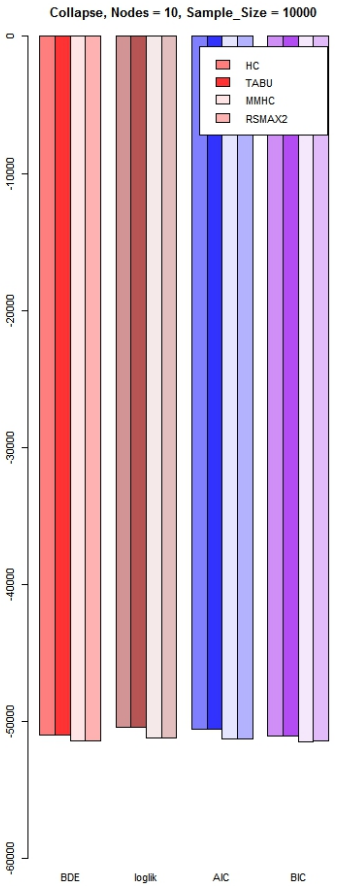
Synthetic Data 유형에 따른 비교 분석



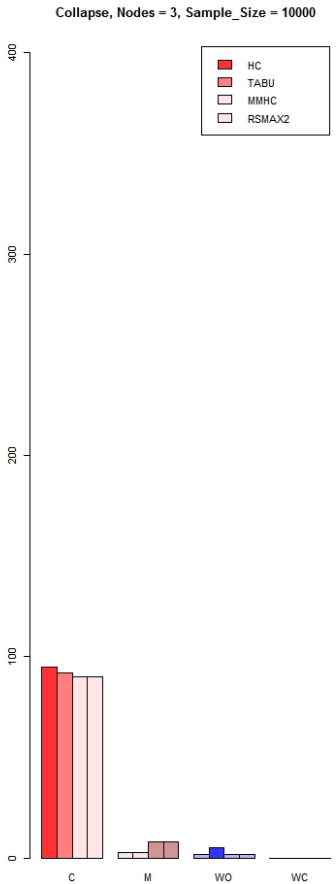


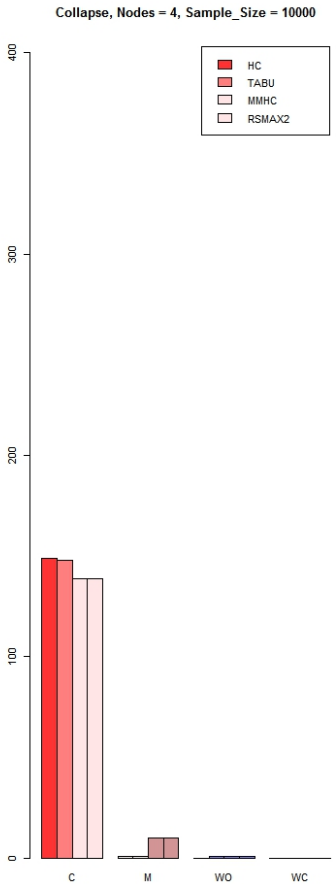


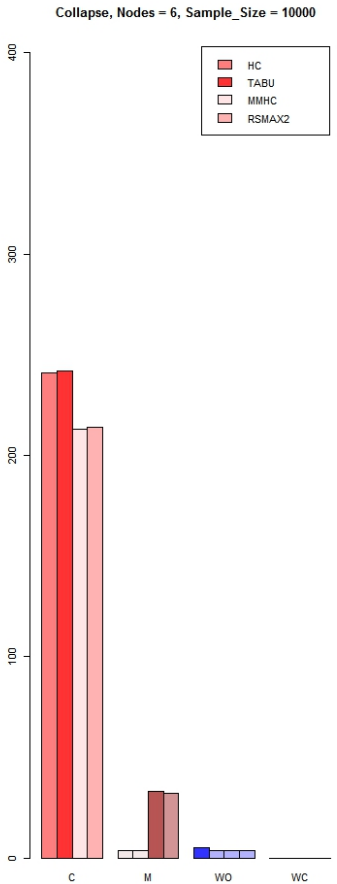


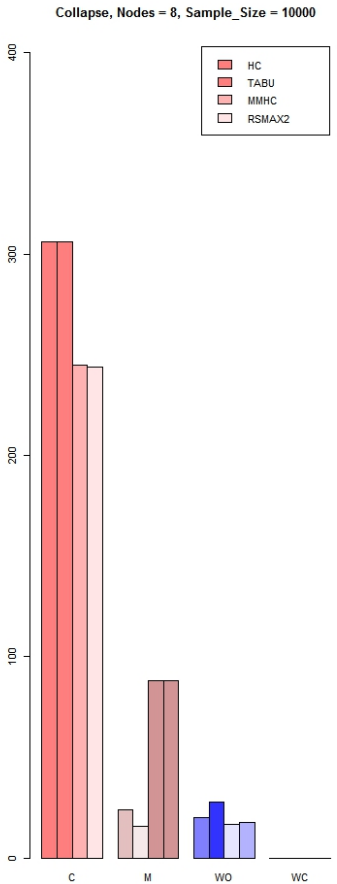


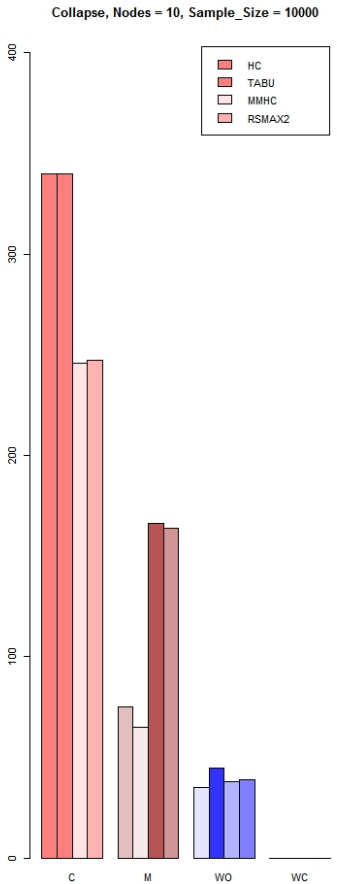
Synthetic Data 유형에 따른 비교 분석



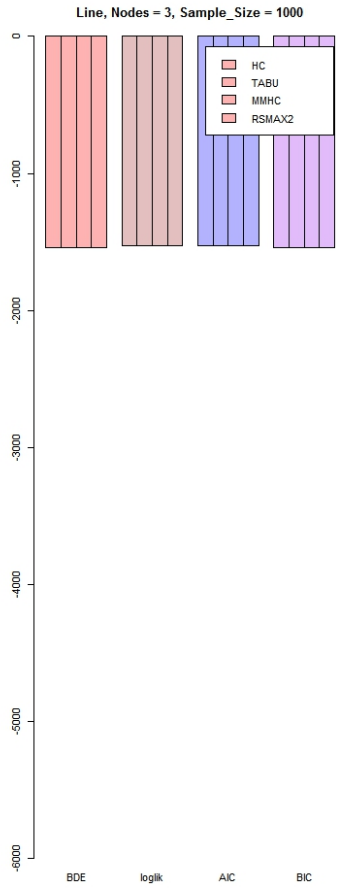


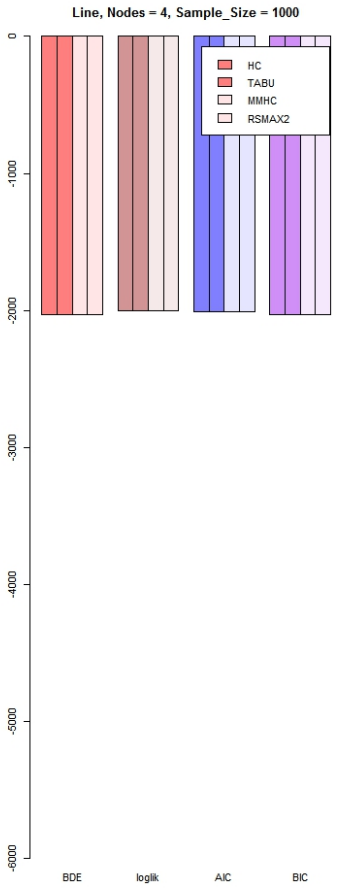


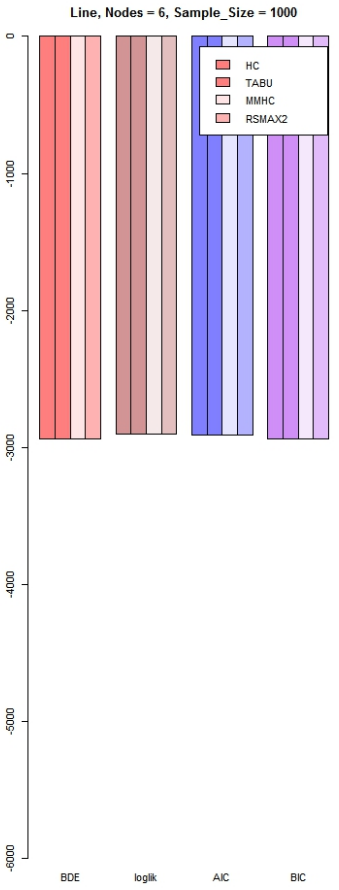


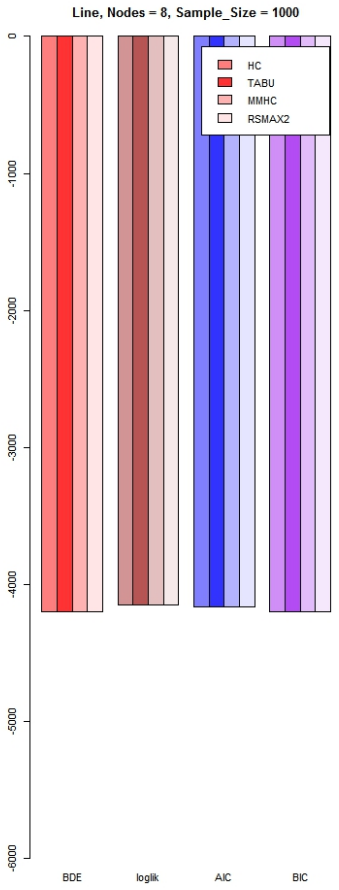


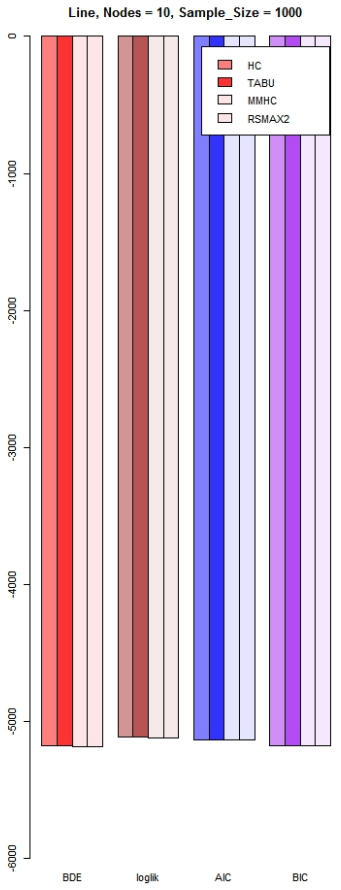
Synthetic Data 유형에 따른 비교 분석



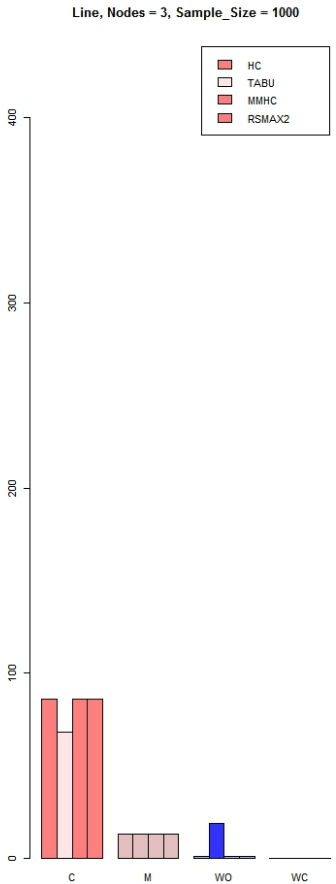


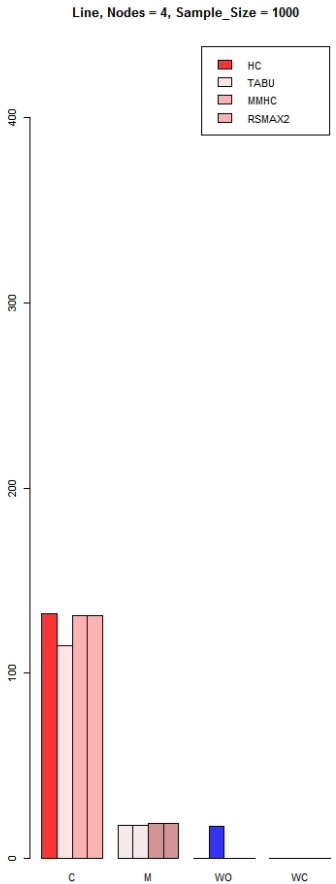


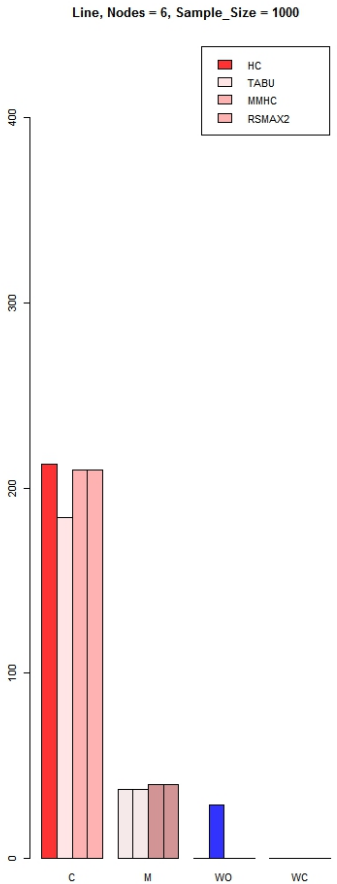


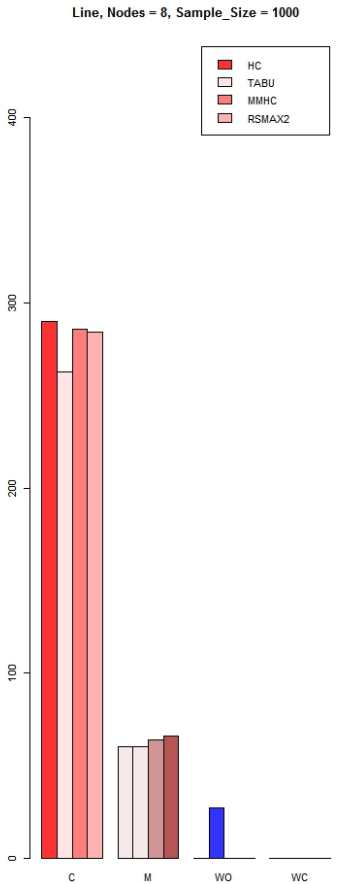


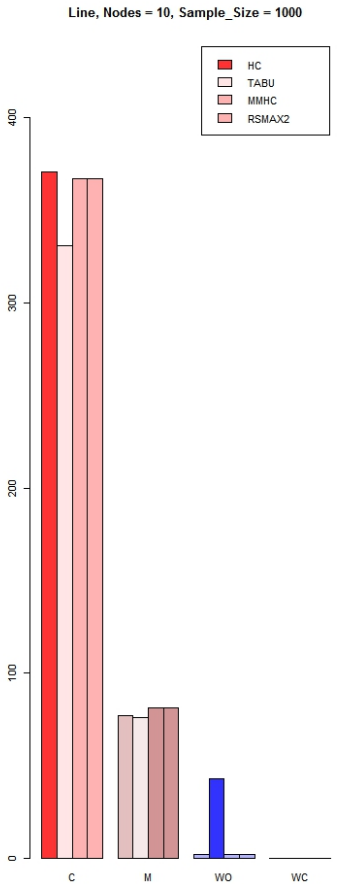
Synthetic Data 유형에 따른 비교 분석



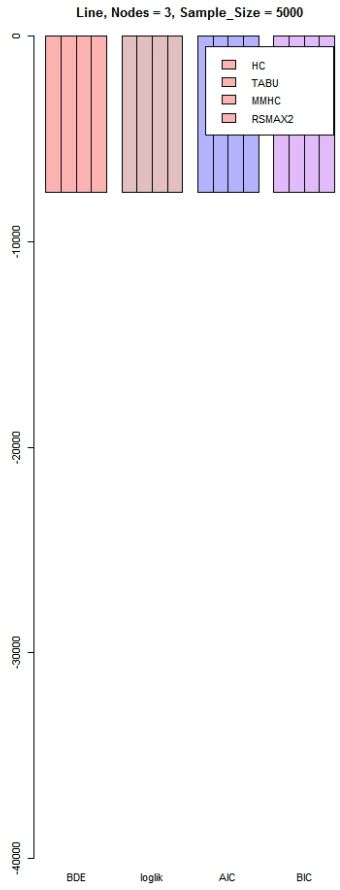


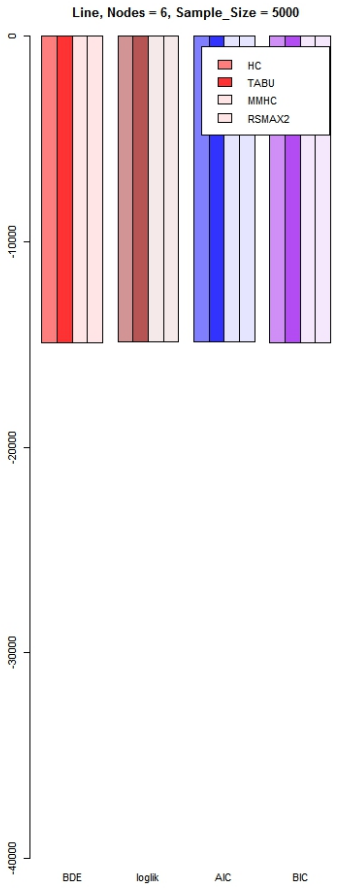


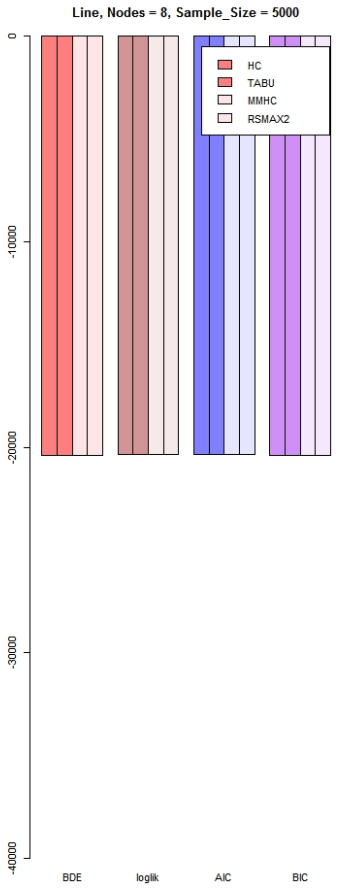


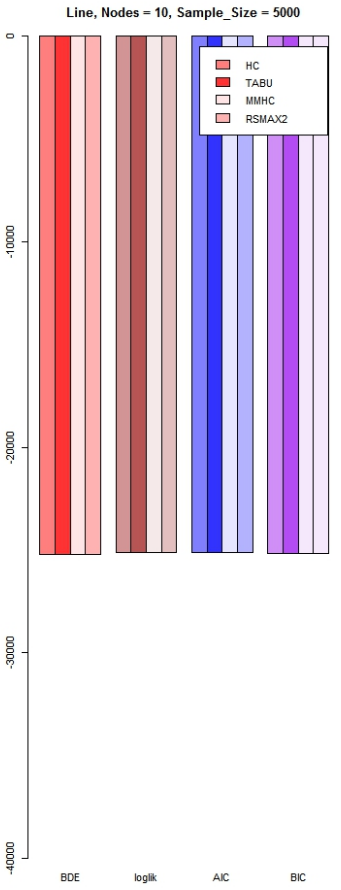


Synthetic Data 유형에 따른 비교 분석

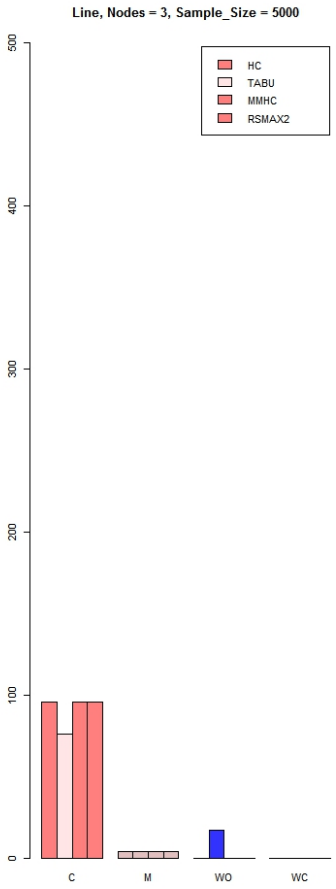


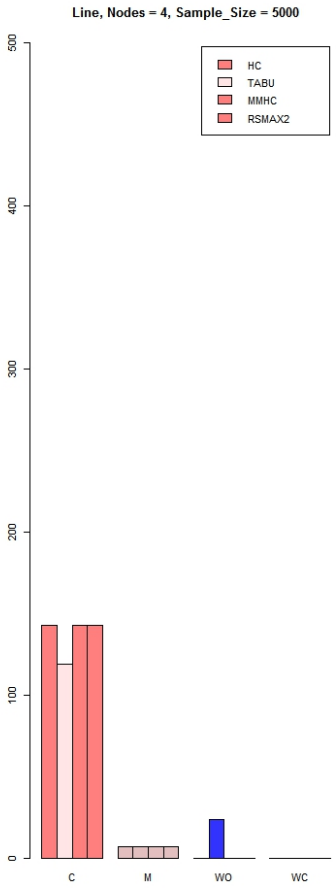


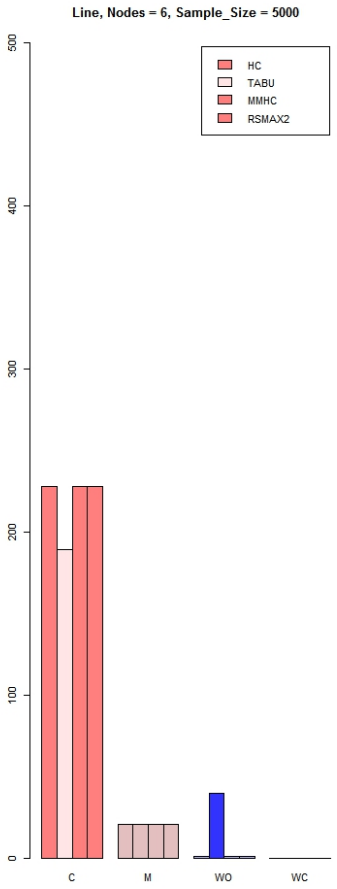


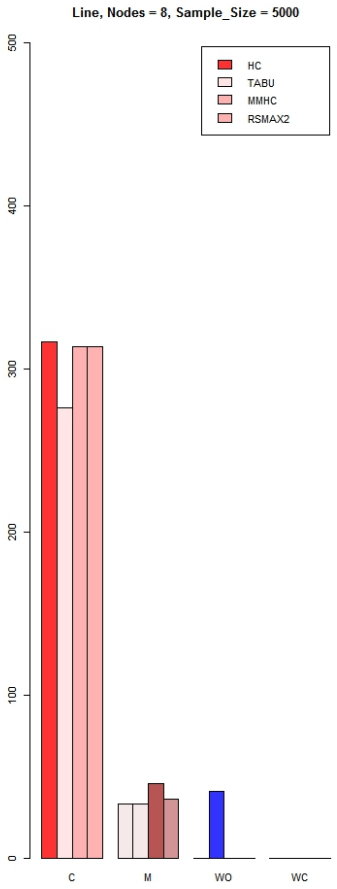


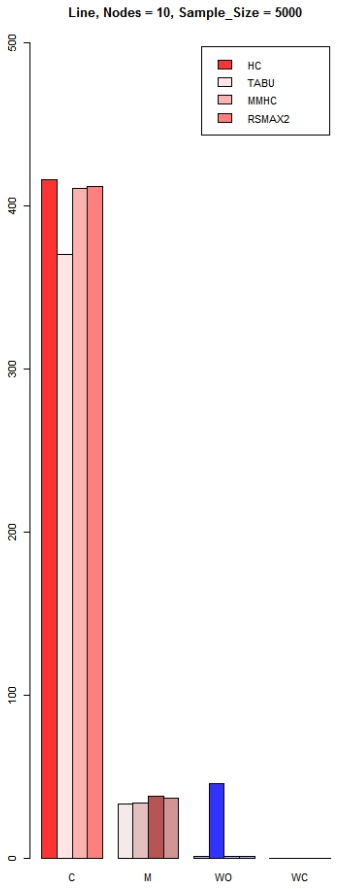
Synthetic Data 유형에 따른 비교 분석



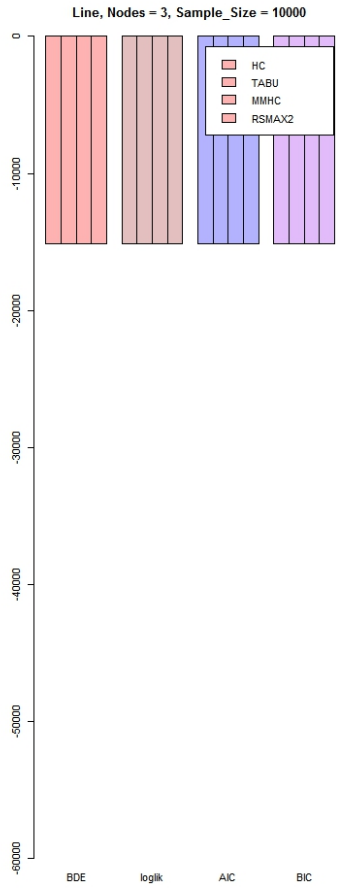


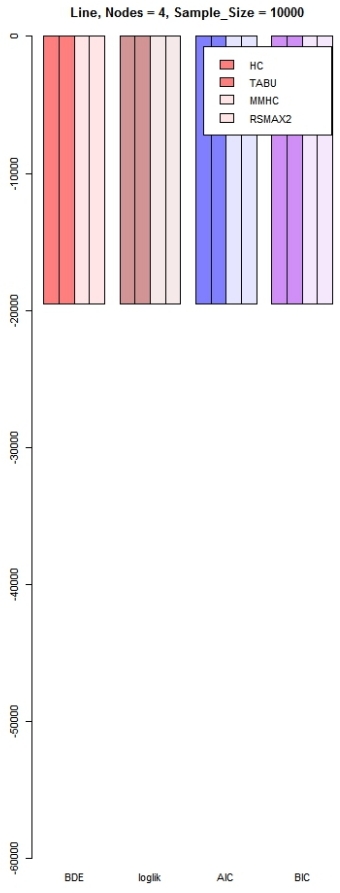


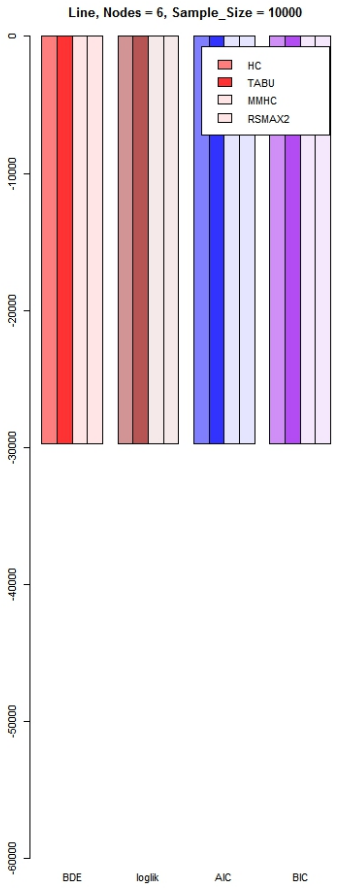


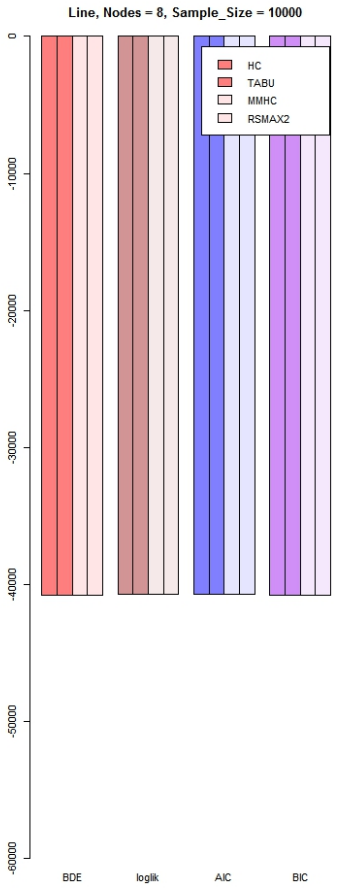


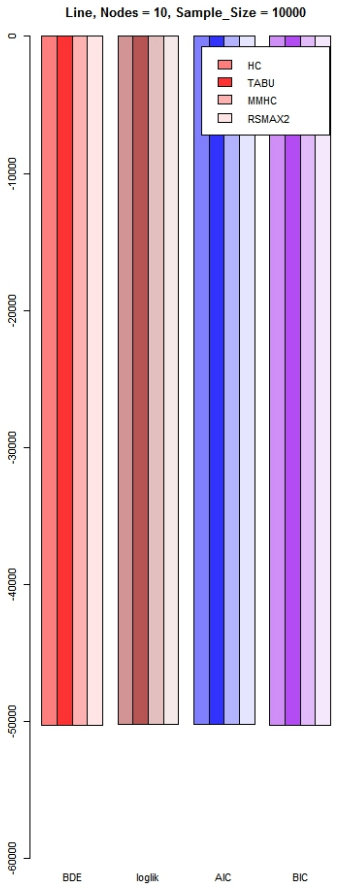
Synthetic Data 유형에 따른 비교 분석



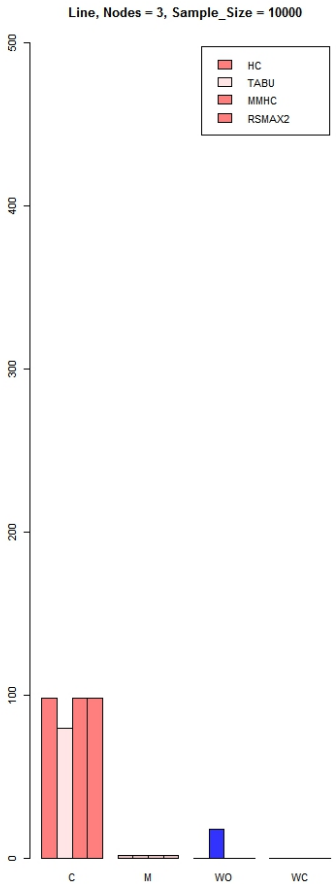


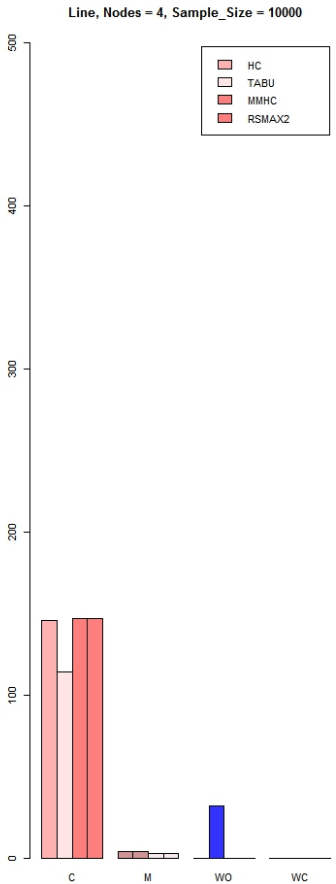


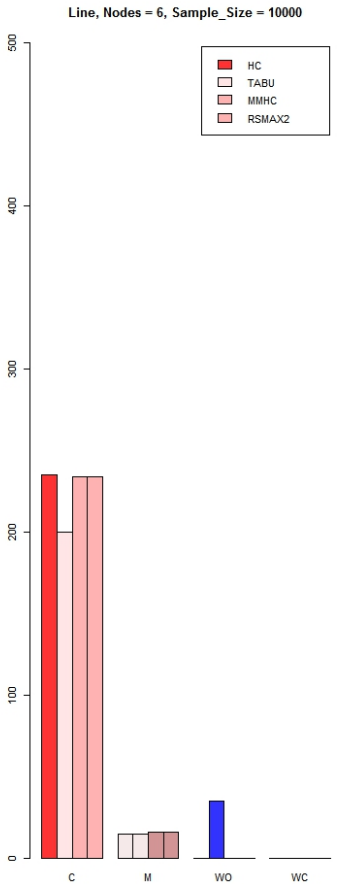


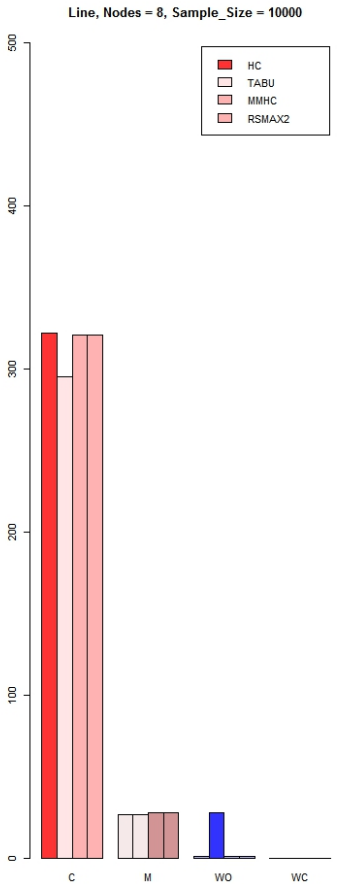


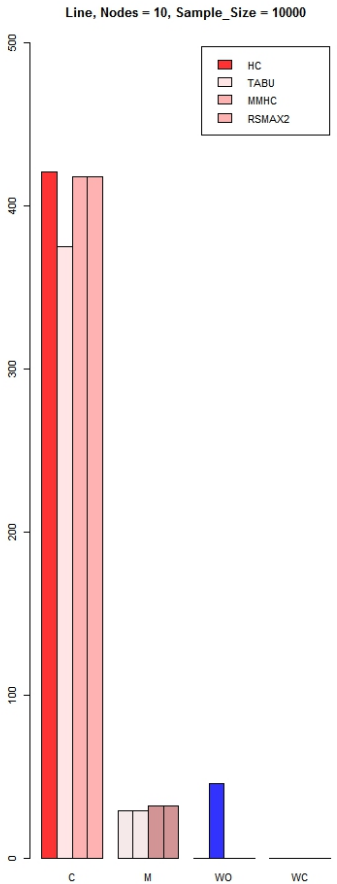
Synthetic Data 유형에 따른 비교 분석



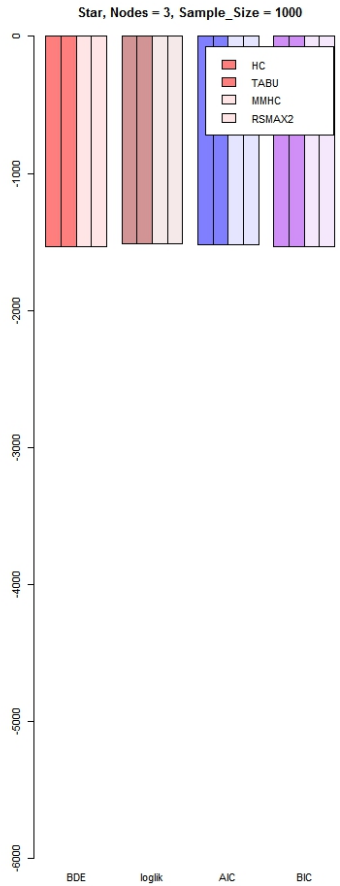


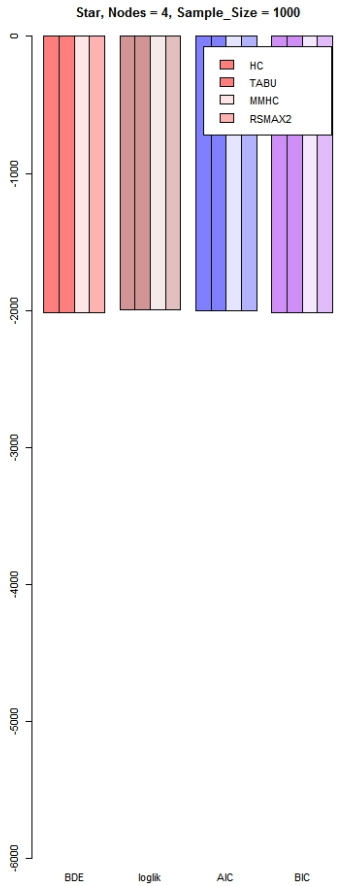


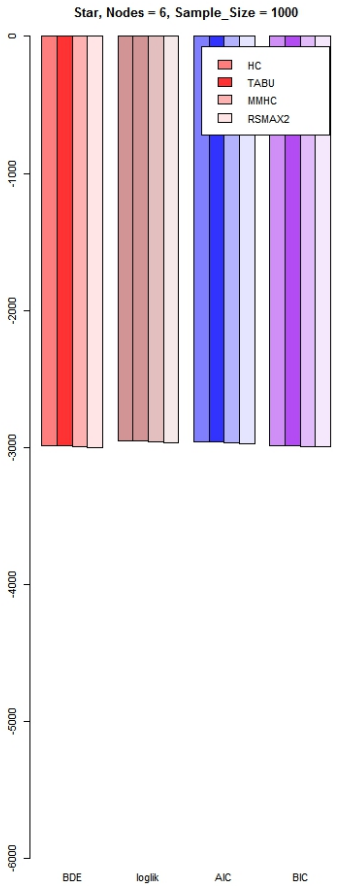


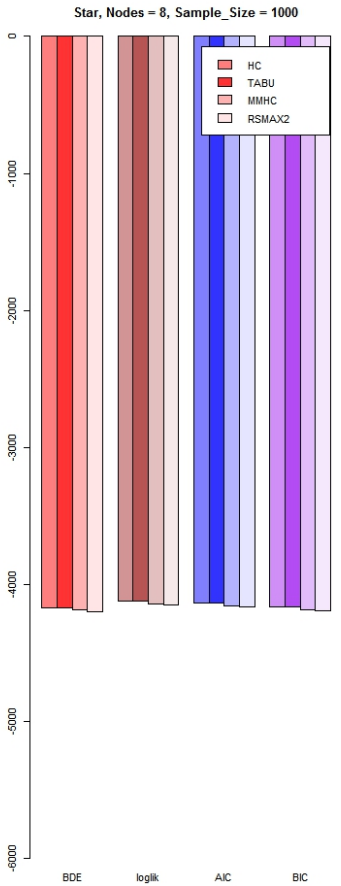


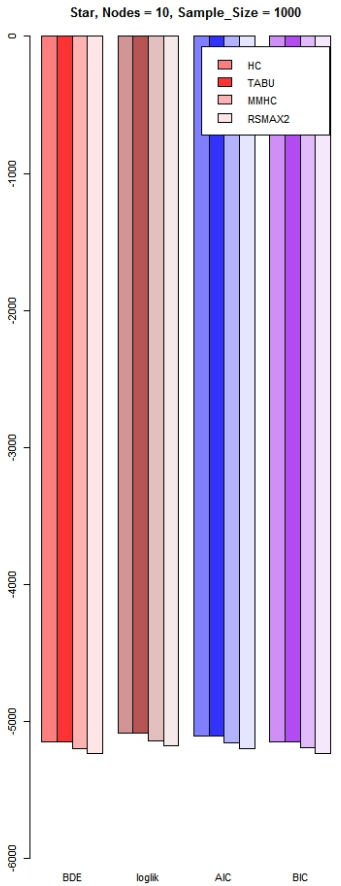
Synthetic Data 유형에 따른 비교 분석



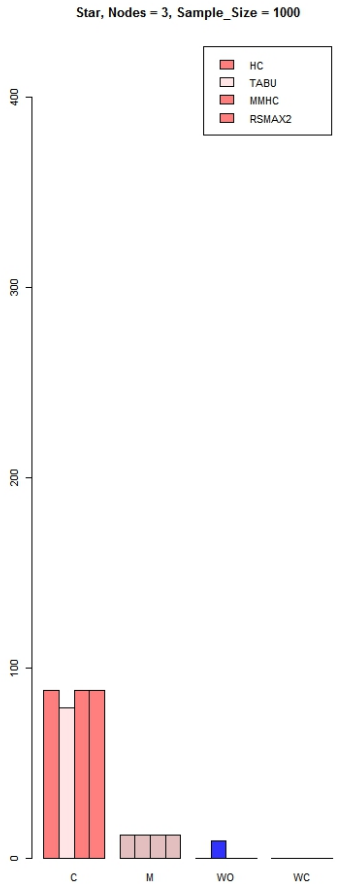


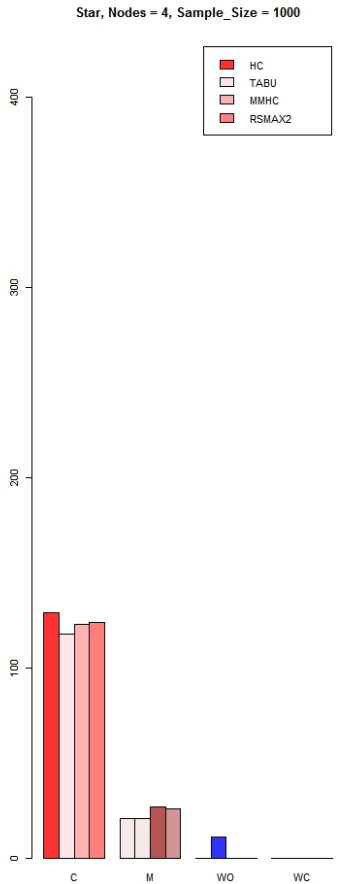


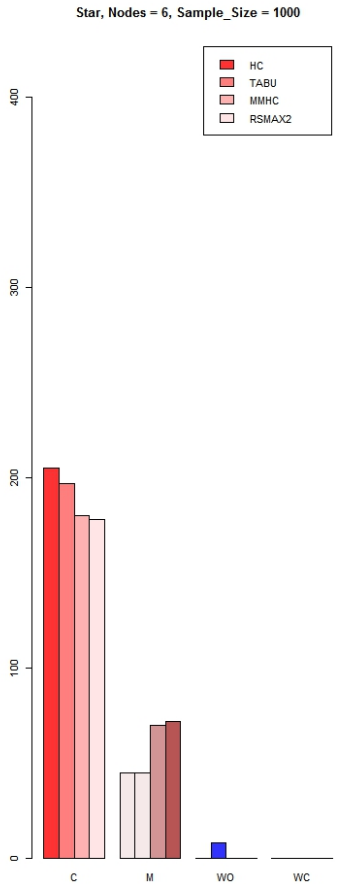


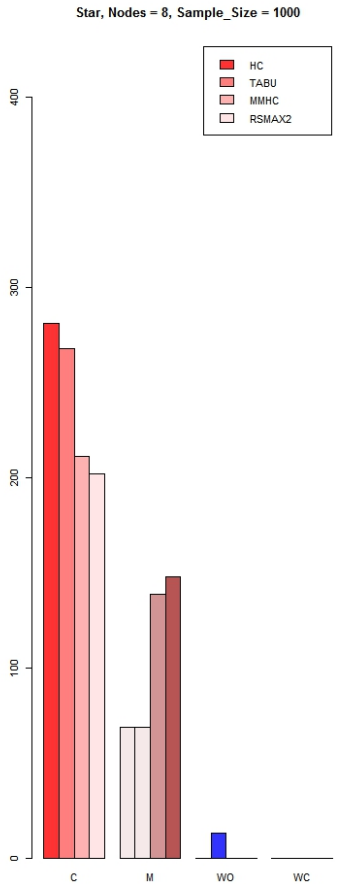


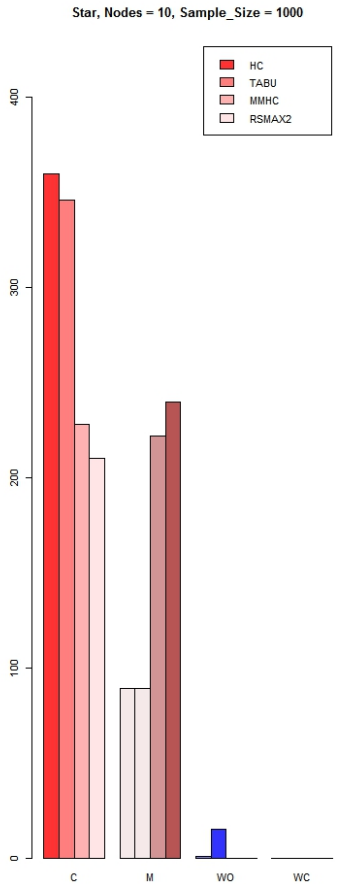
Synthetic Data 유형에 따른 비교 분석



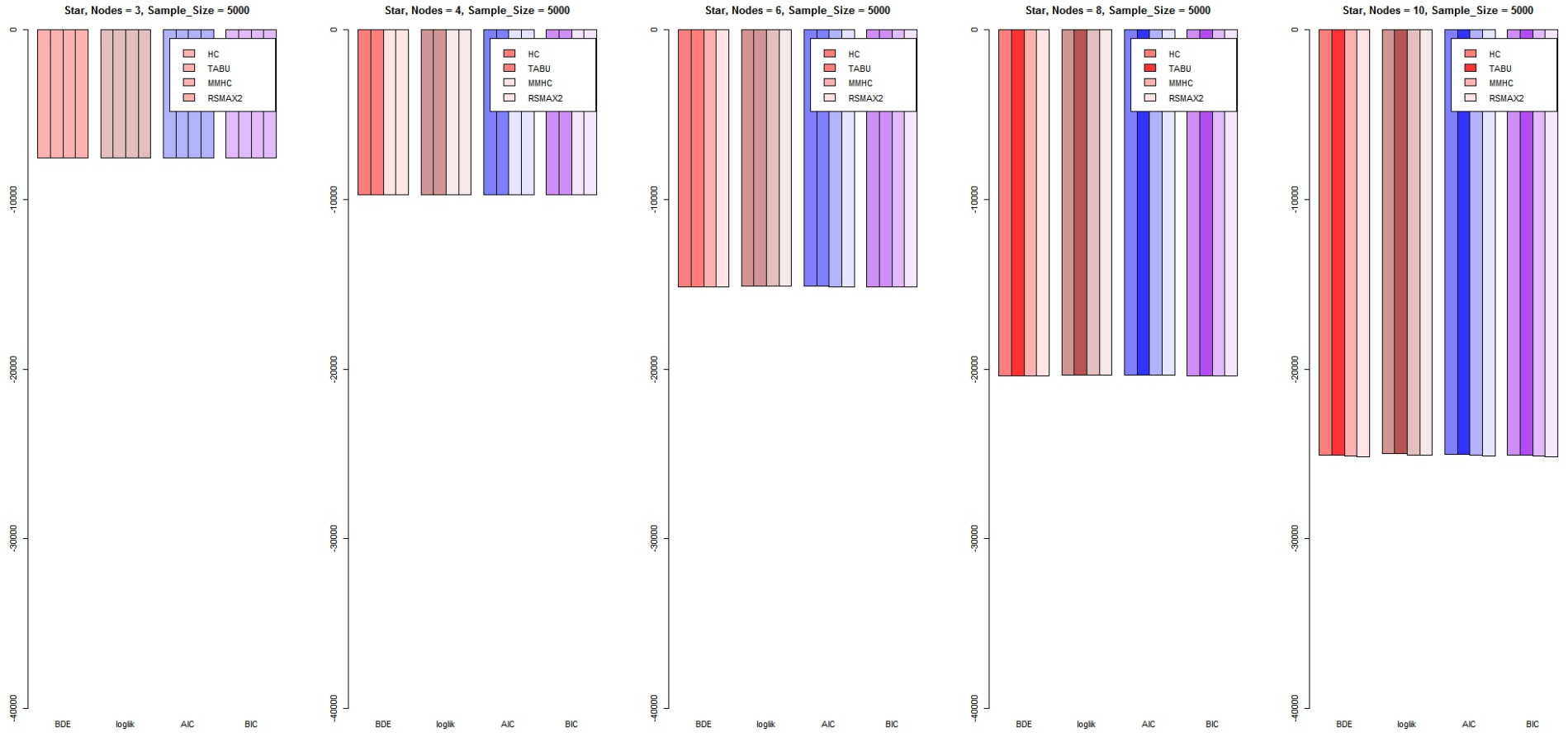




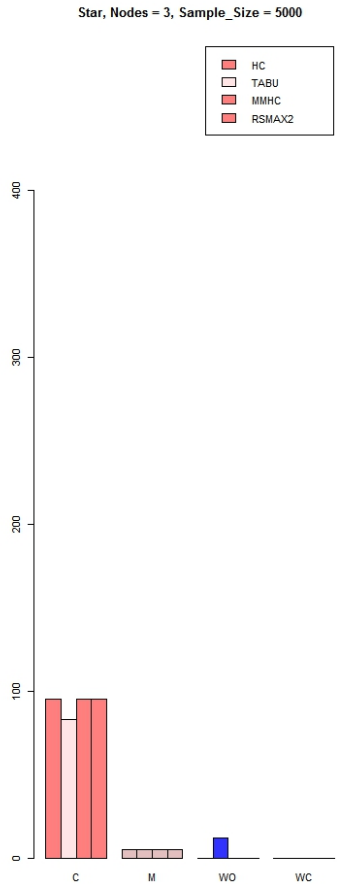


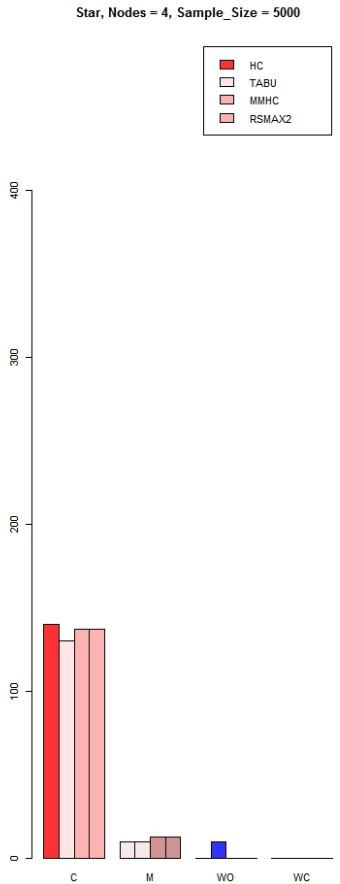


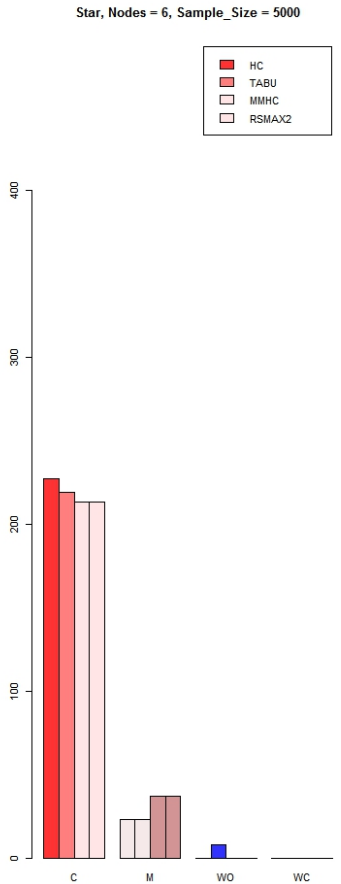
Synthetic Data 유형에 따른 비교 분석

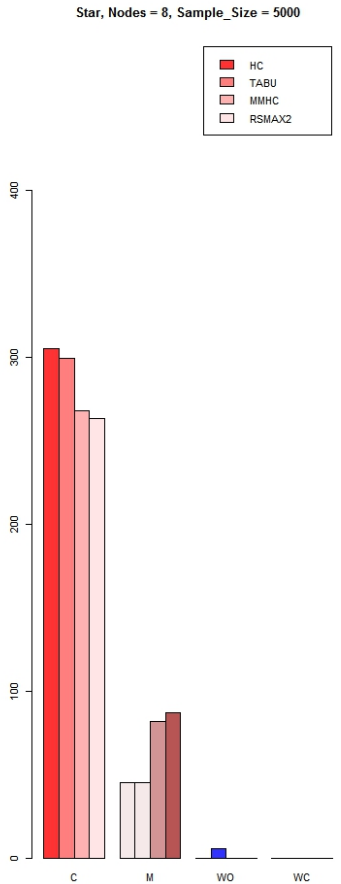


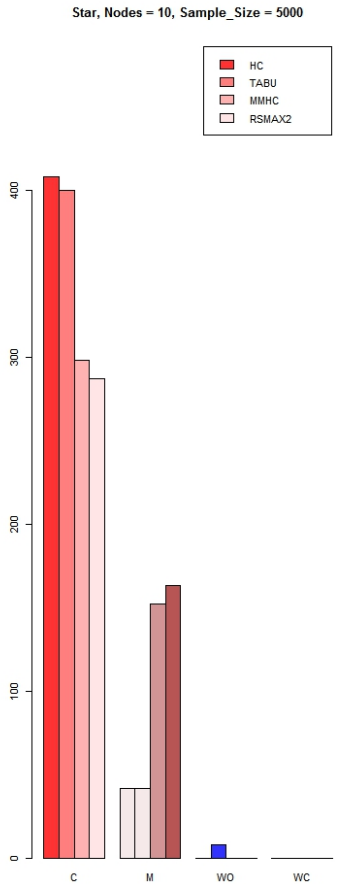
Synthetic Data 유형에 따른 비교 분석



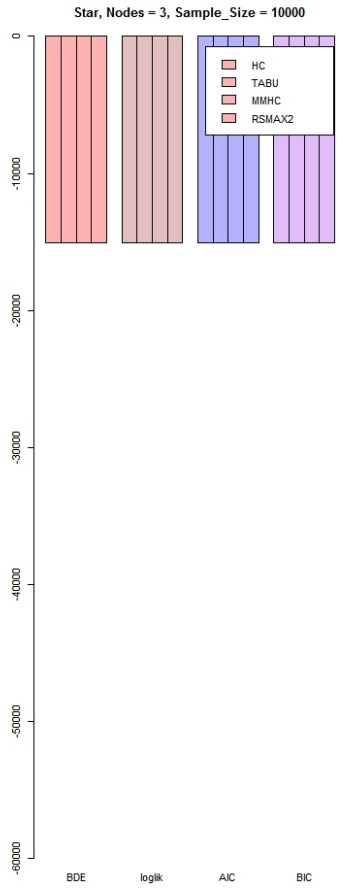


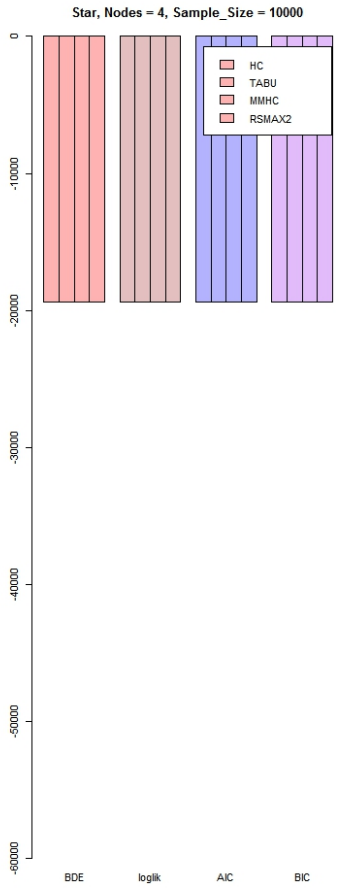


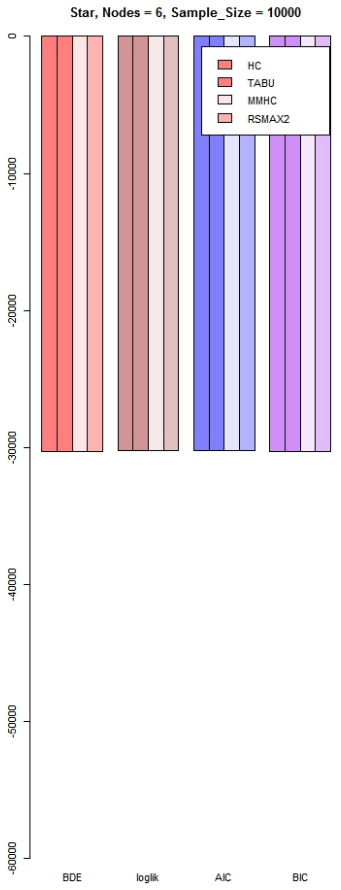


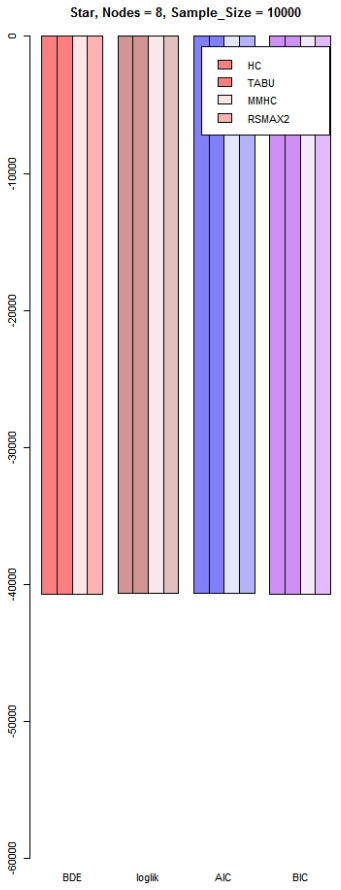


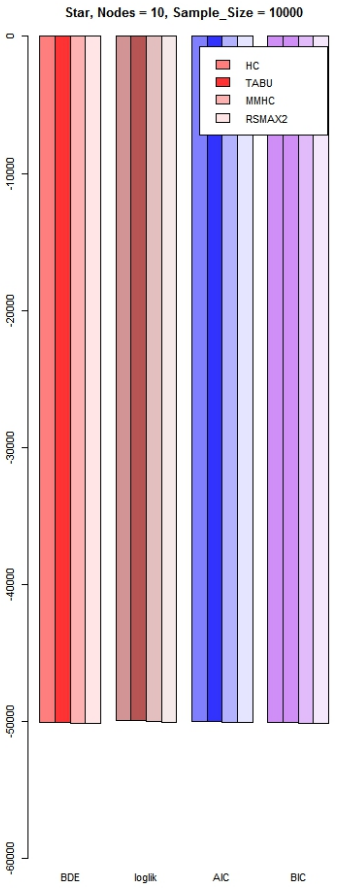
Synthetic Data 유형에 따른 비교 분석



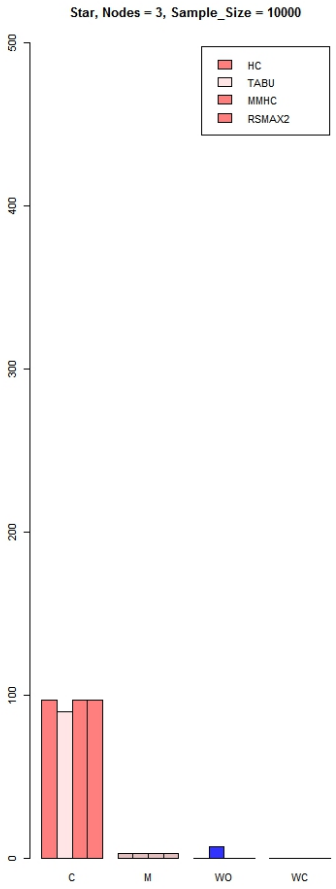


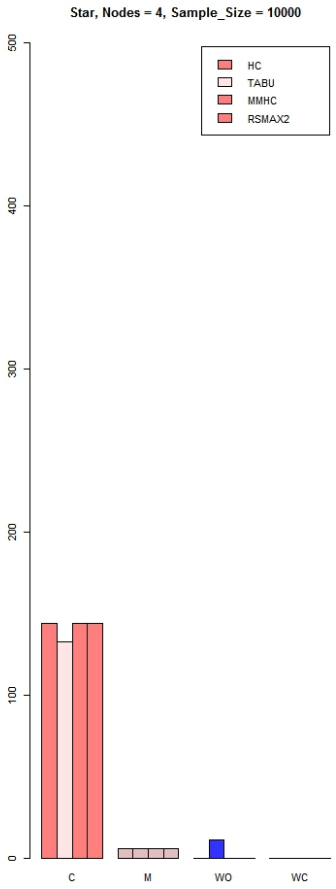


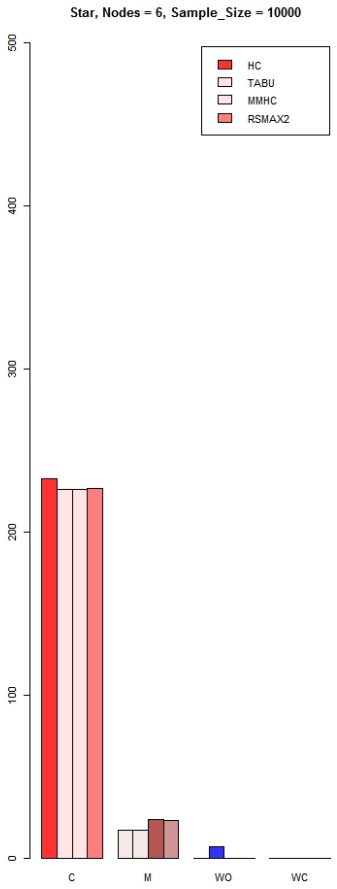


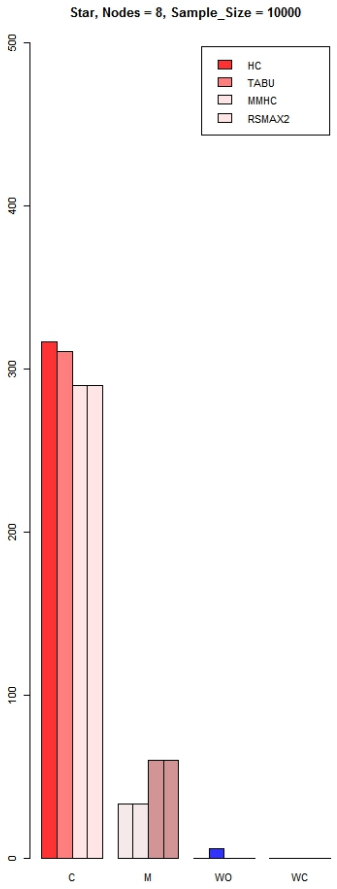


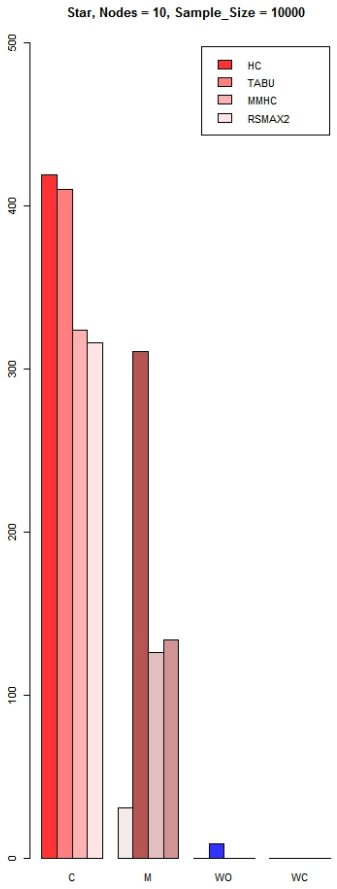
Synthetic Data 유형에 따른 비교 분석



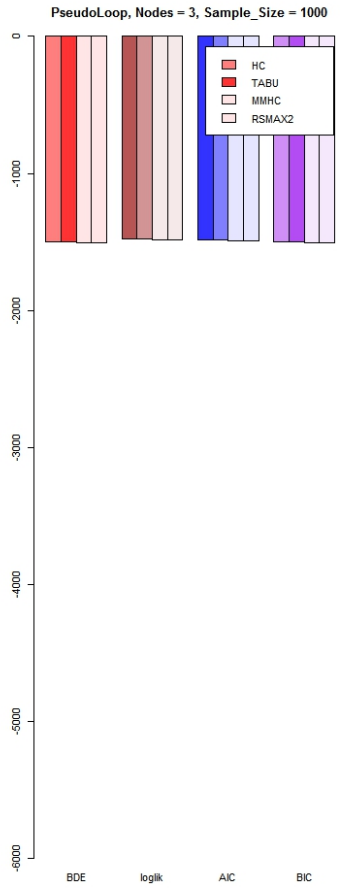


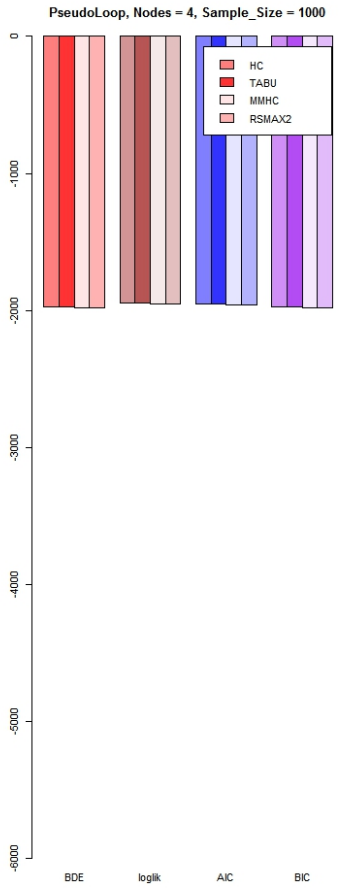


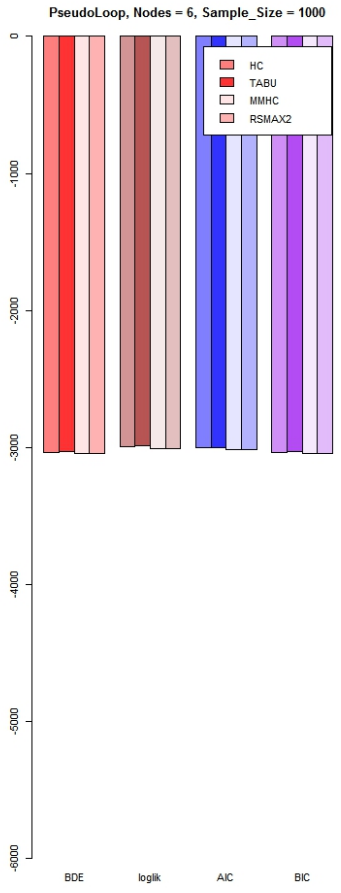


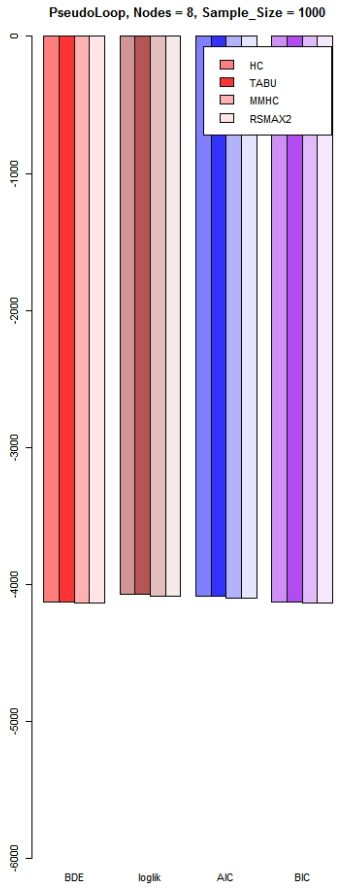


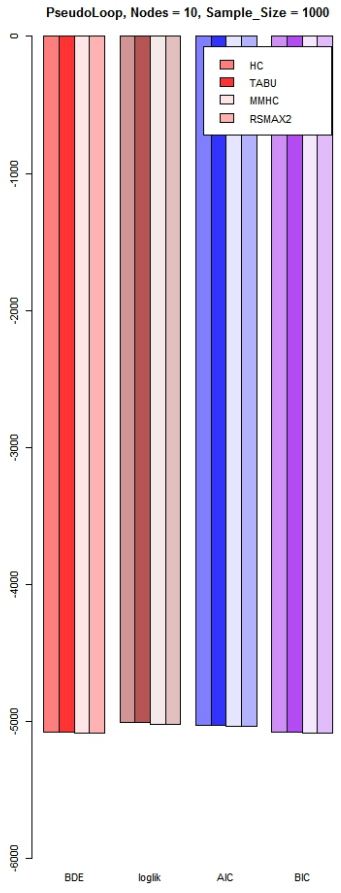
Synthetic Data 유형에 따른 비교 분석



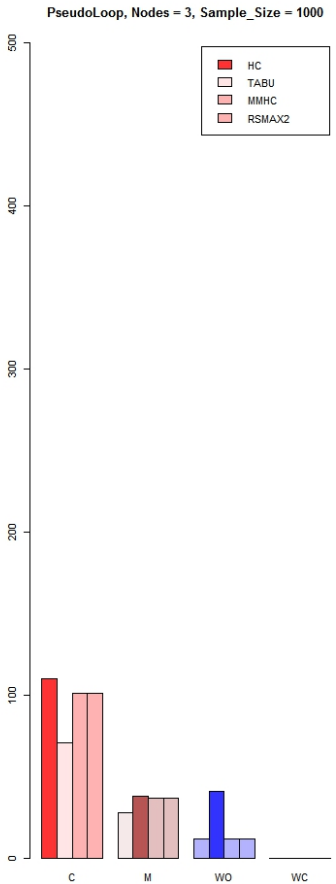


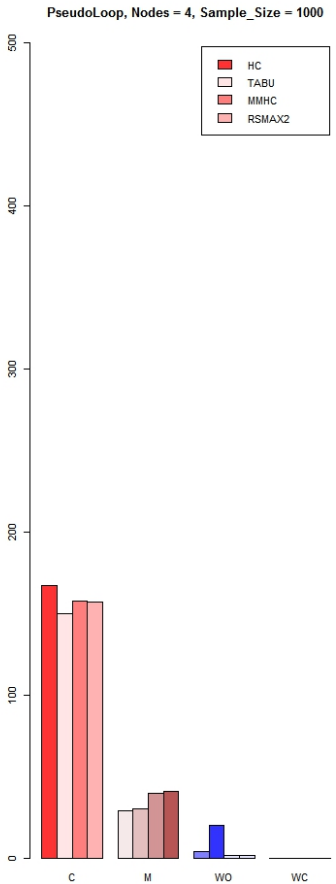


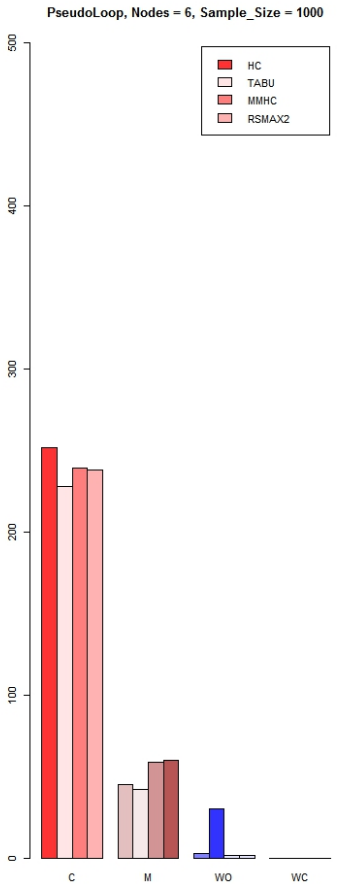


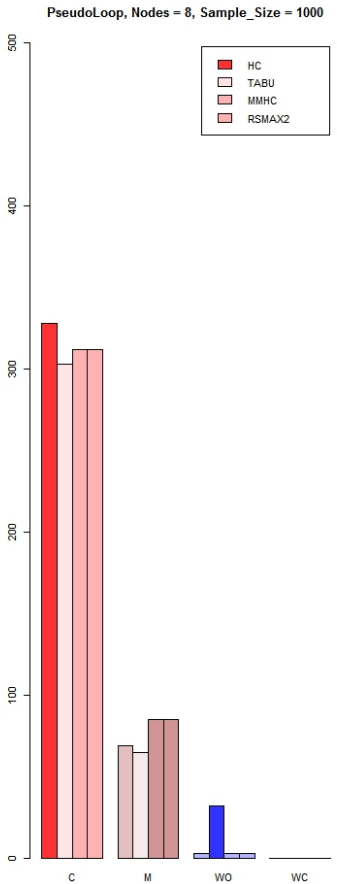


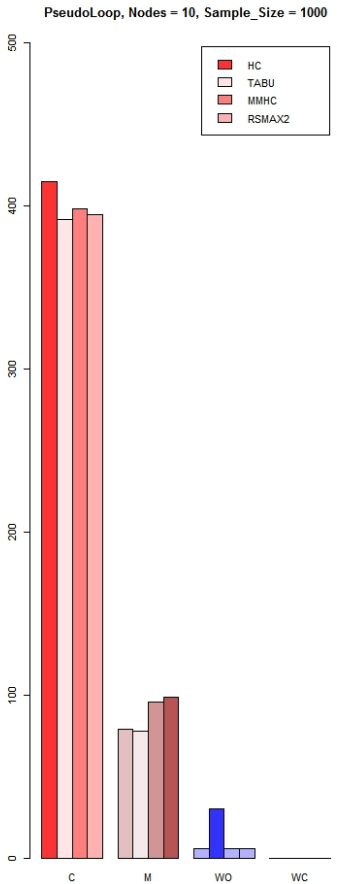
Synthetic Data 유형에 따른 비교 분석



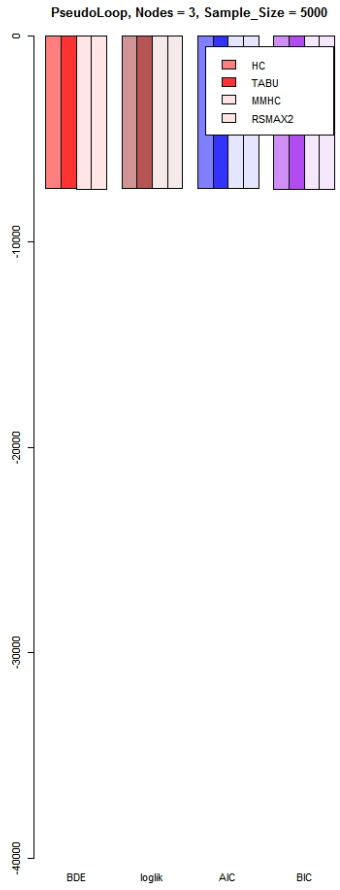


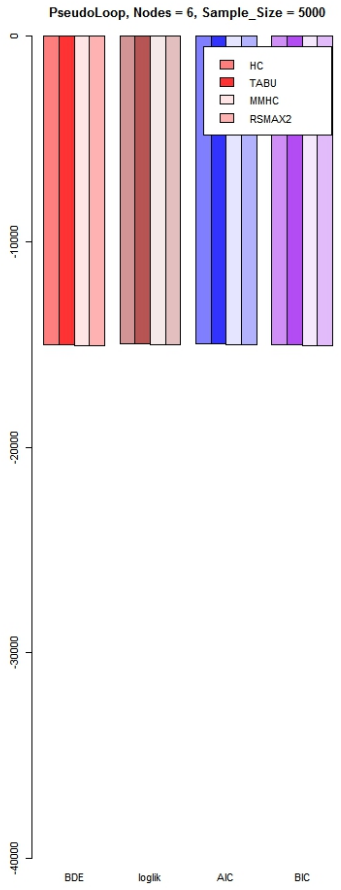


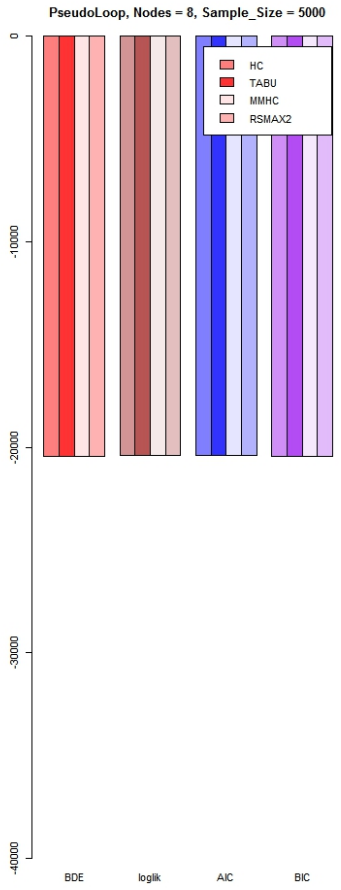


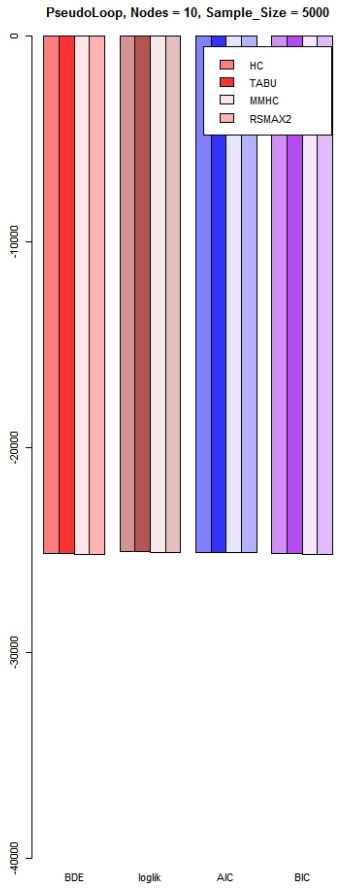


Synthetic Data 유형에 따른 비교 분석

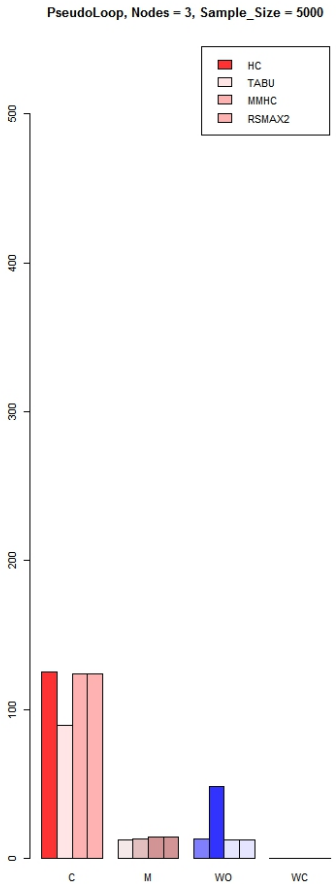


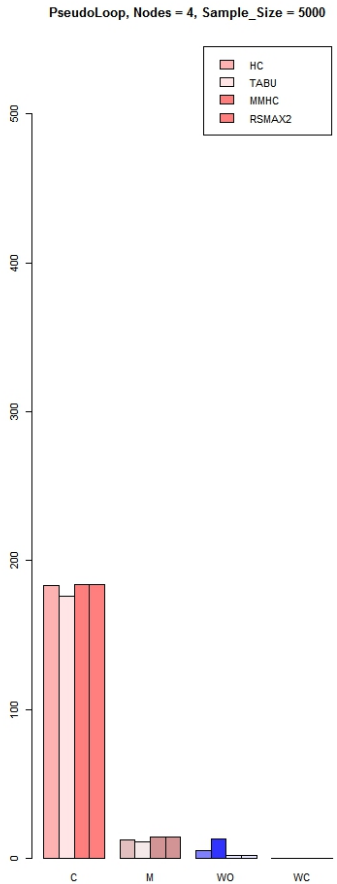


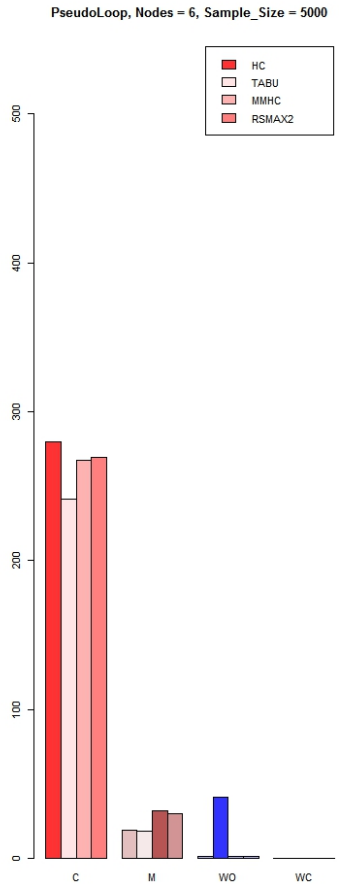


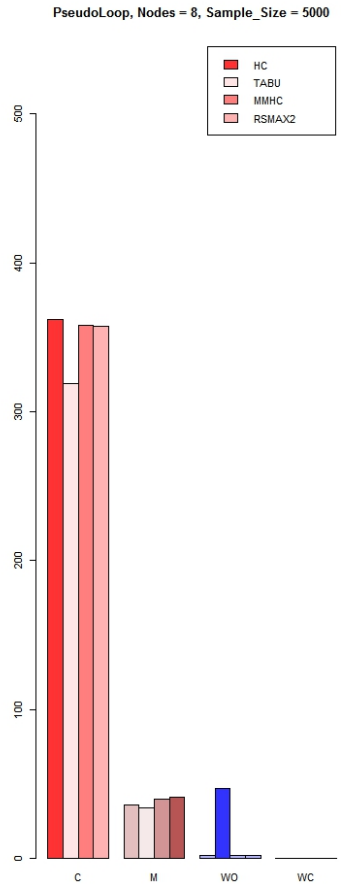


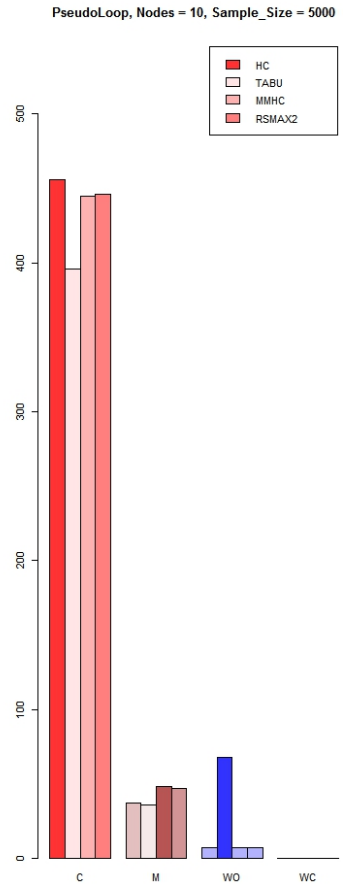
Synthetic Data 유형에 따른 비교 분석



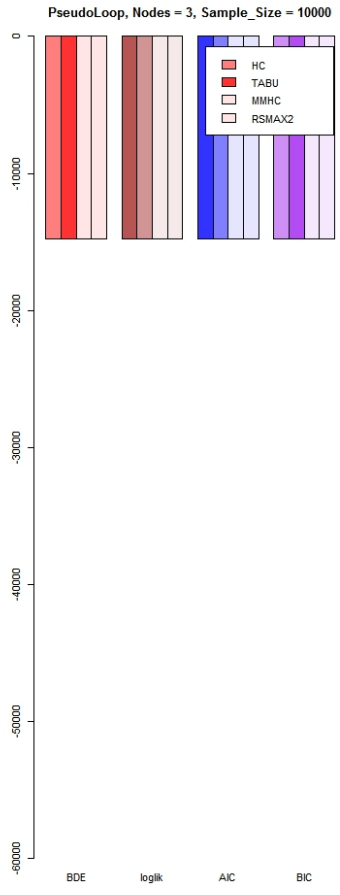


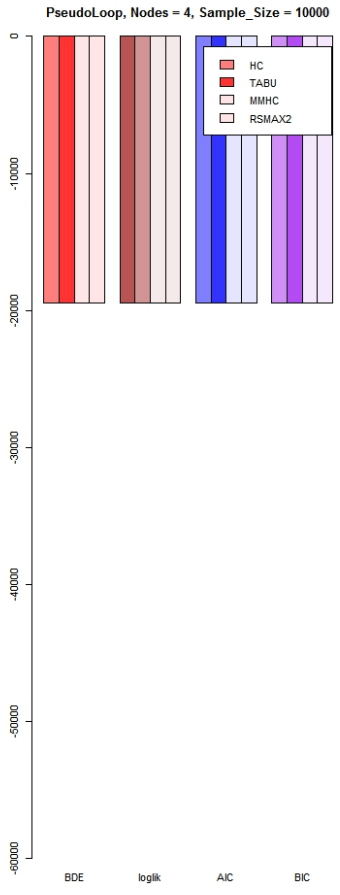


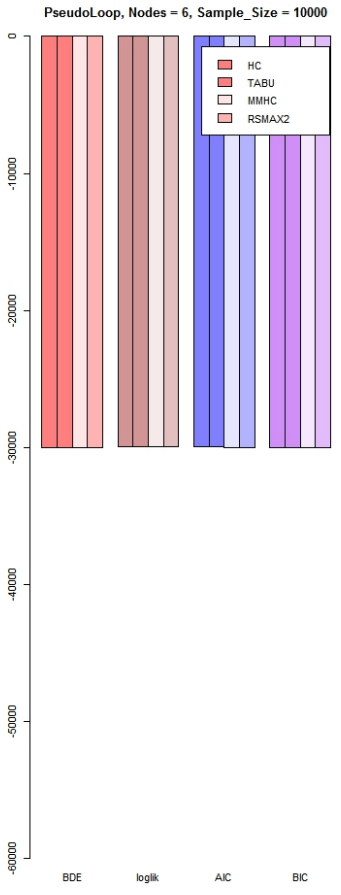


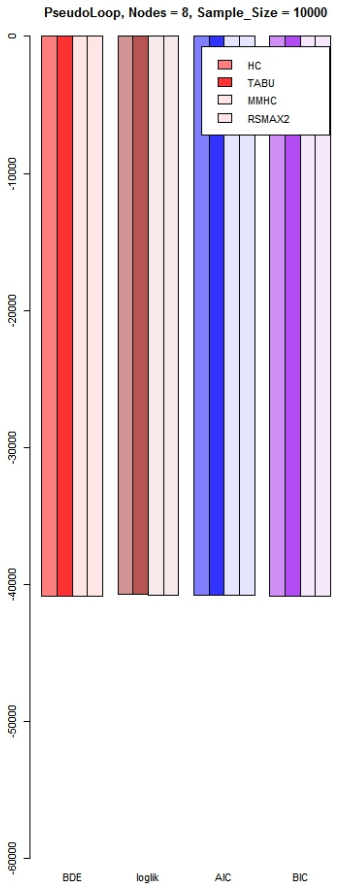


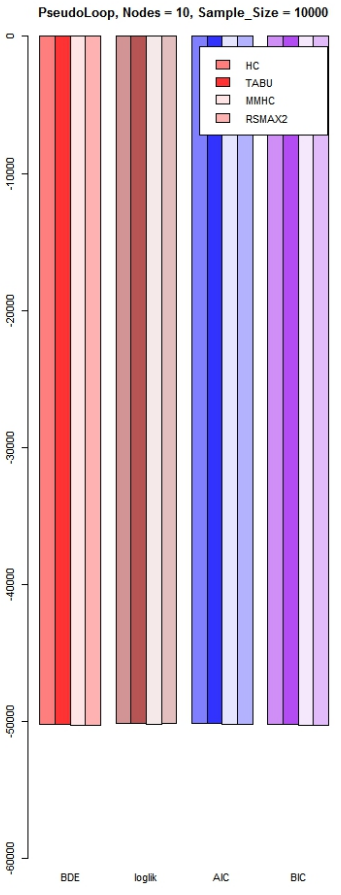
Synthetic Data 유형에 따른 비교 분석



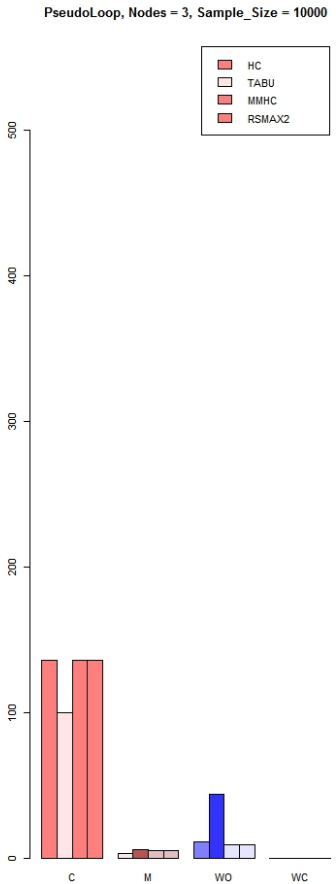


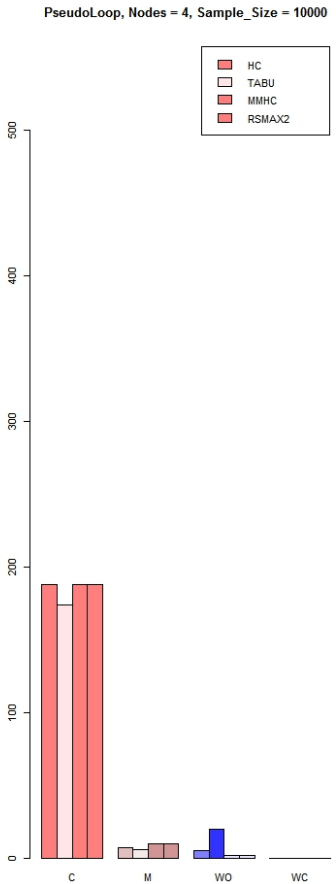


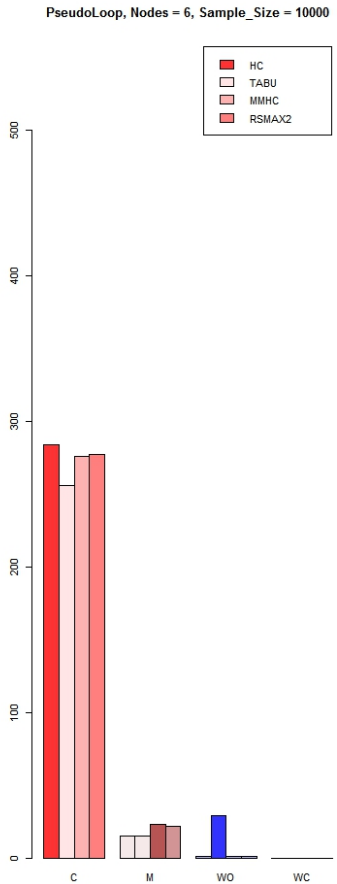


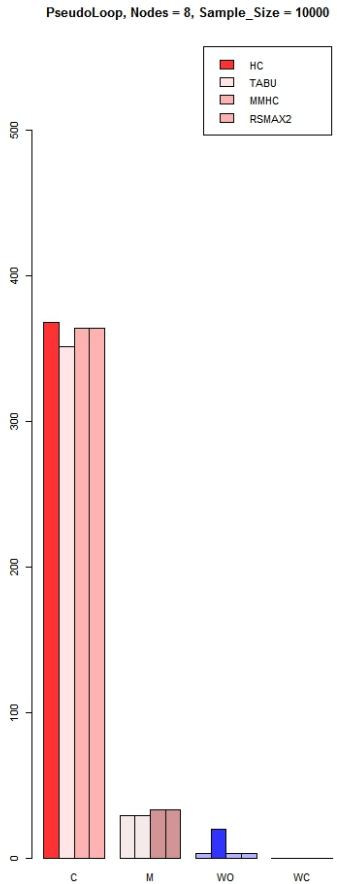


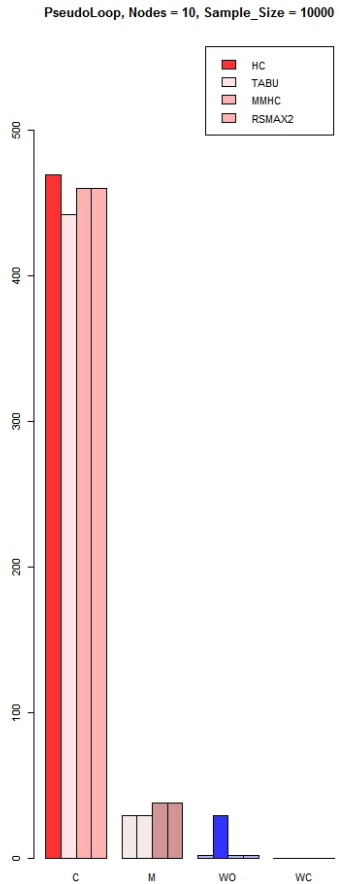
Synthetic Data 유형에 따른 비교 분석



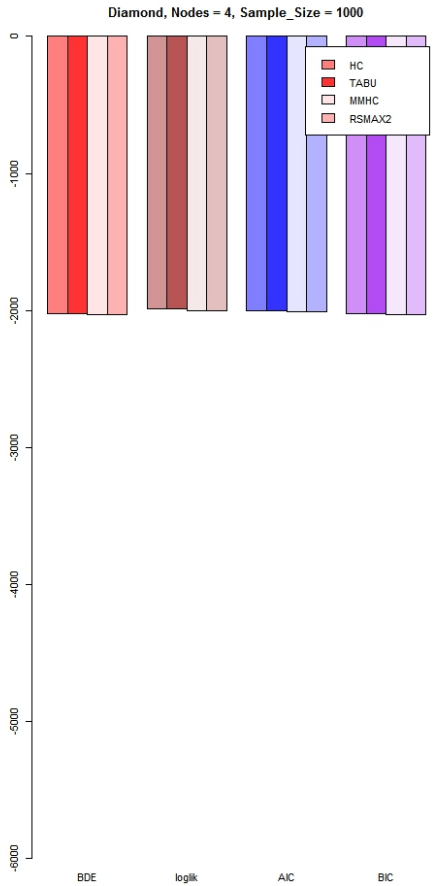


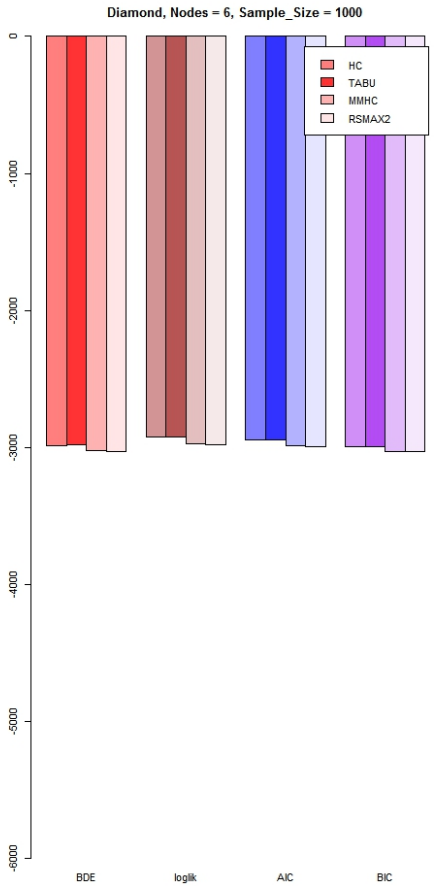


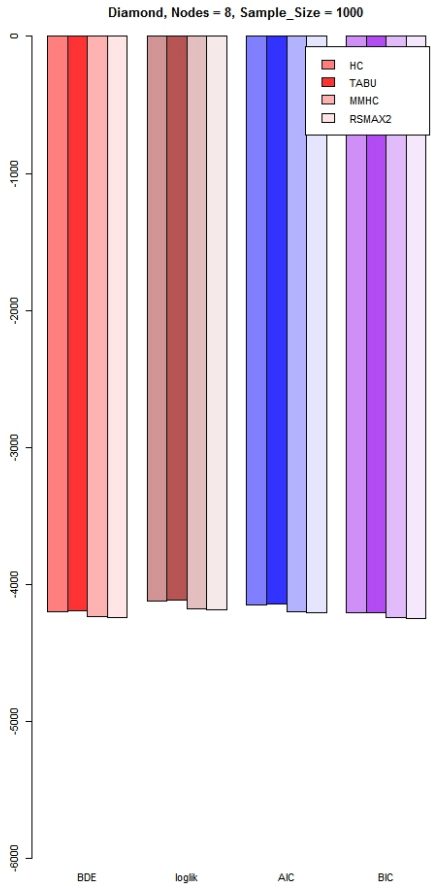


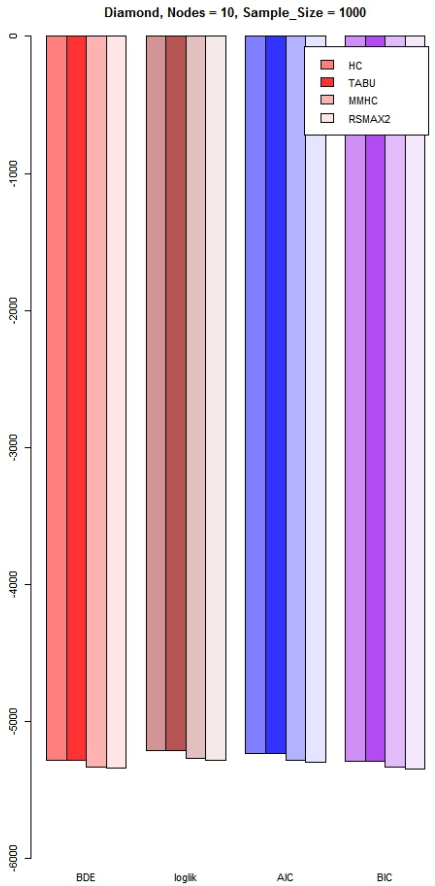


Synthetic Data 유형에 따른 비교 분석

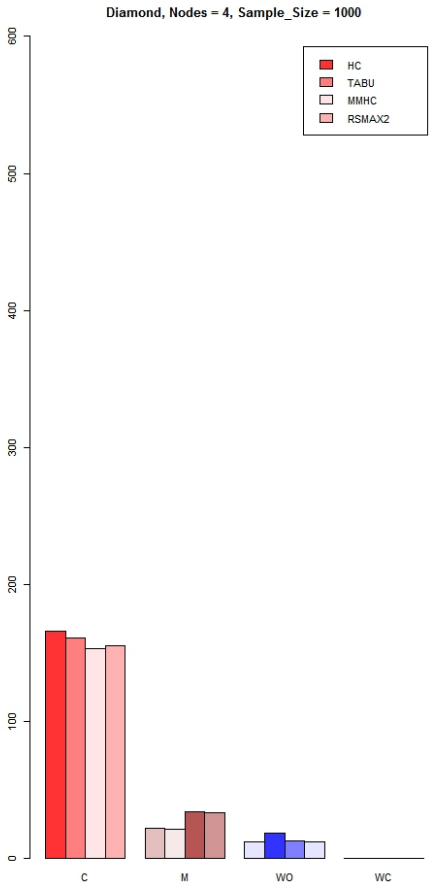


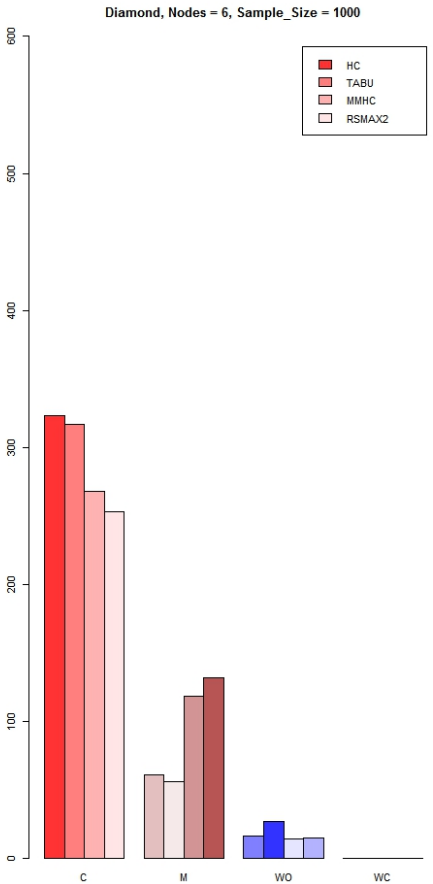


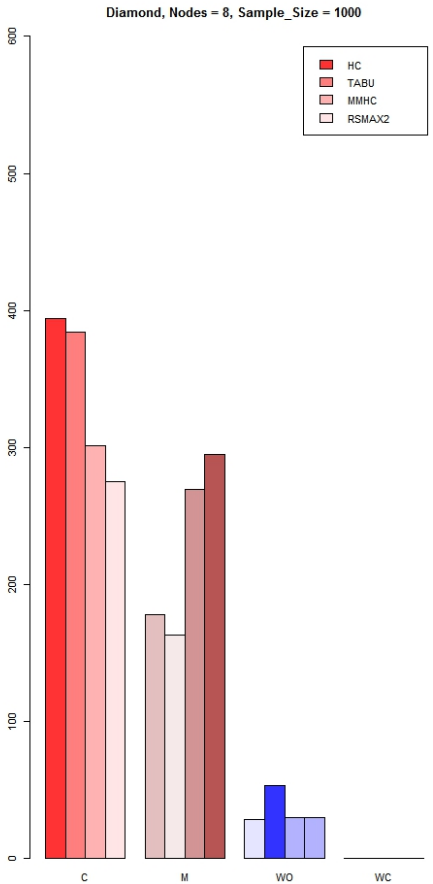


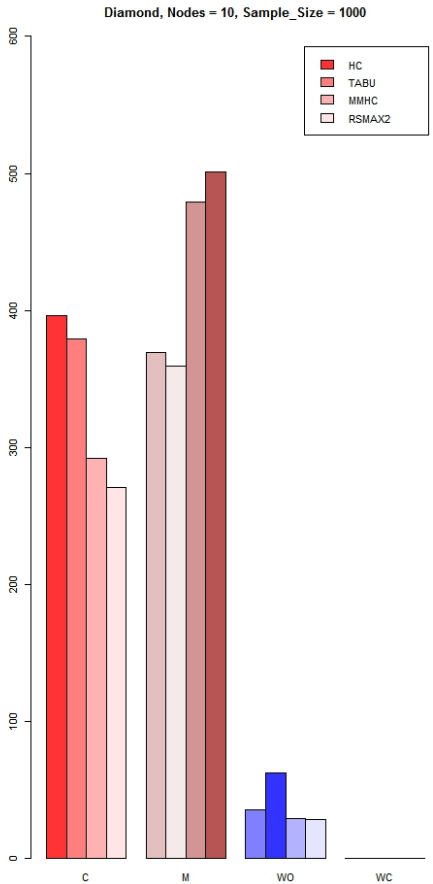


Synthetic Data 유형에 따른 비교 분석

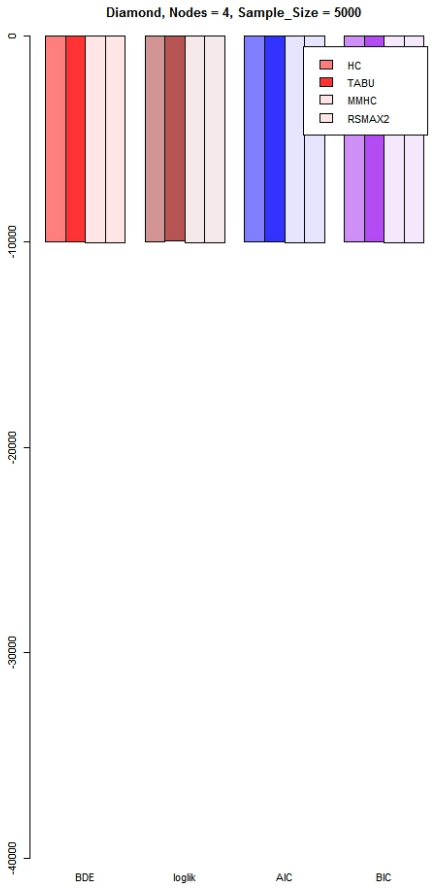


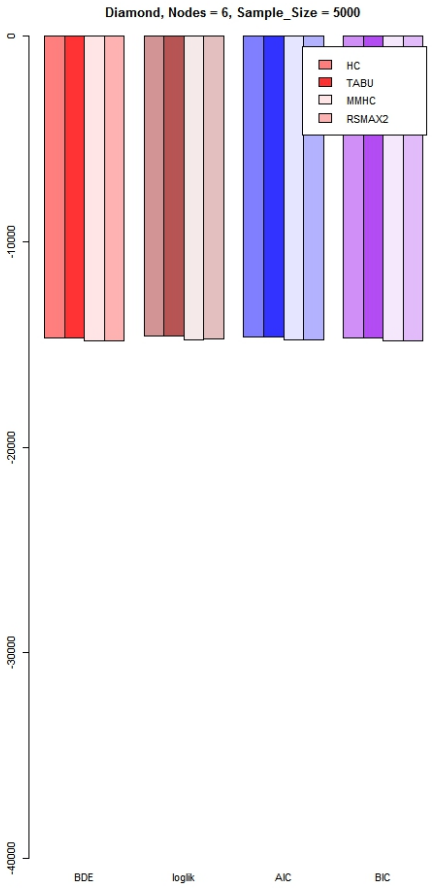


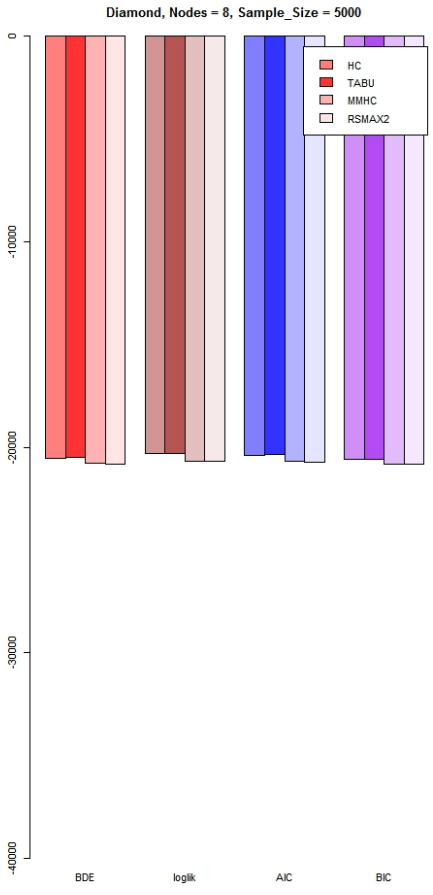


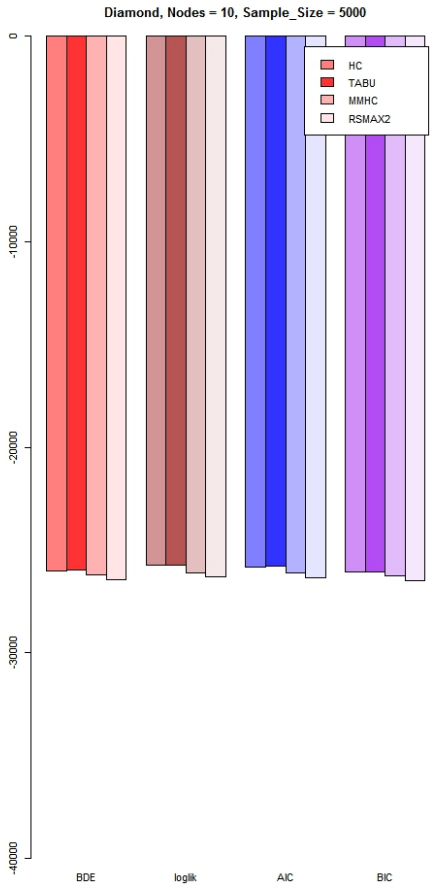


Synthetic Data 유형에 따른 비교 분석

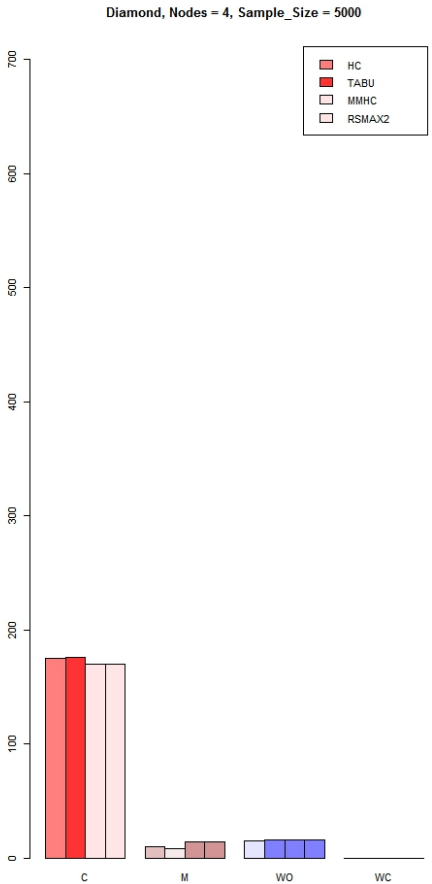


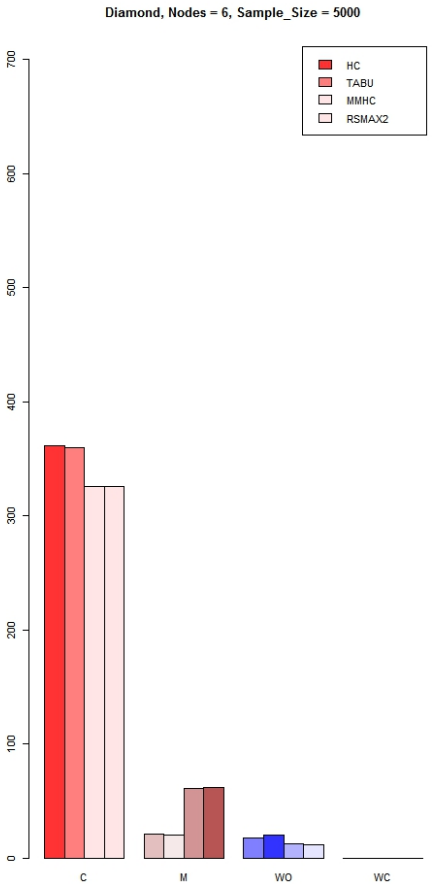


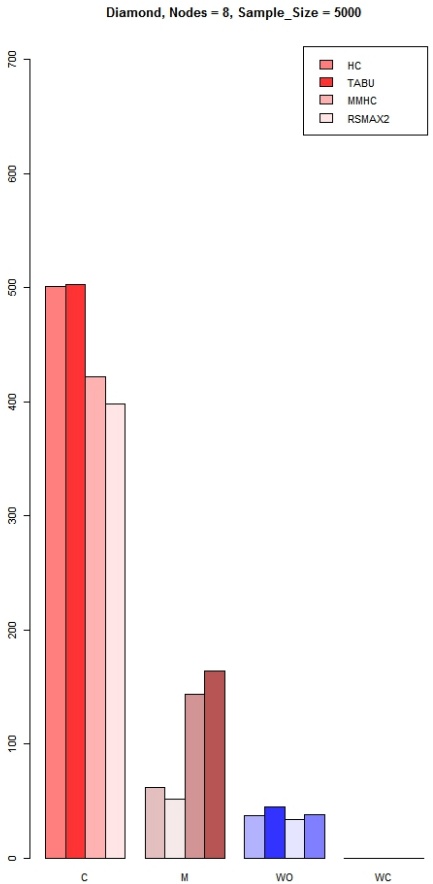


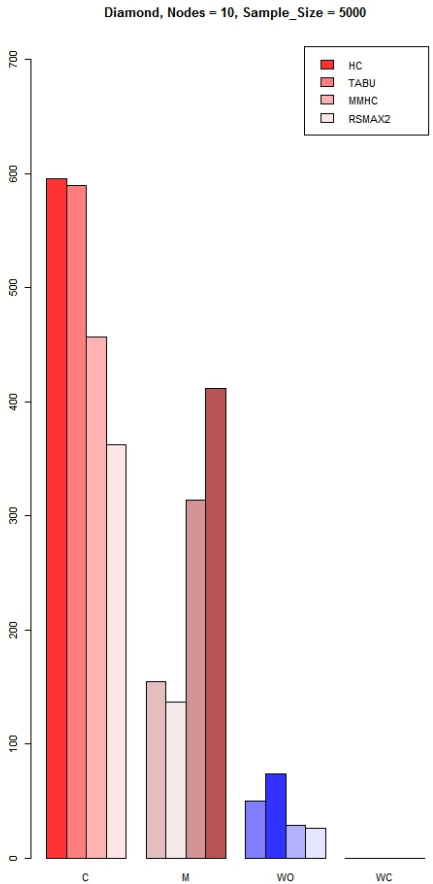


Synthetic Data 유형에 따른 비교 분석

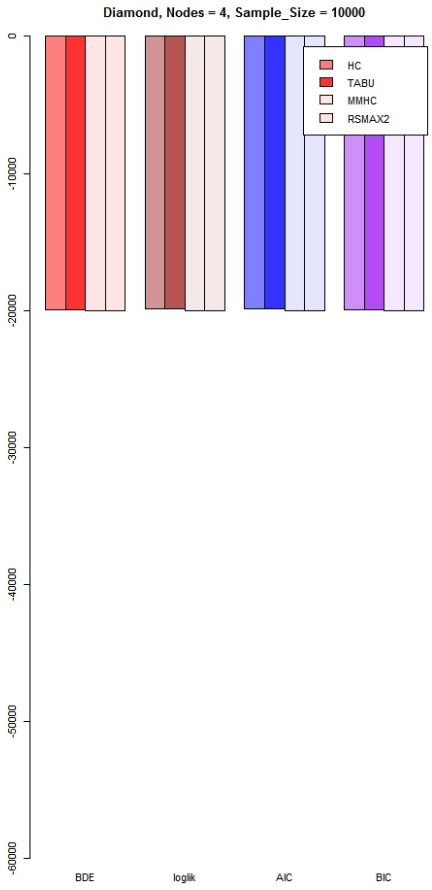


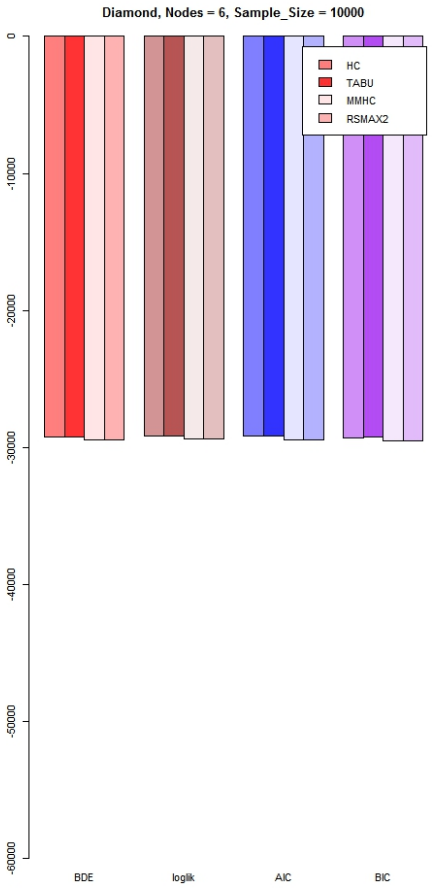


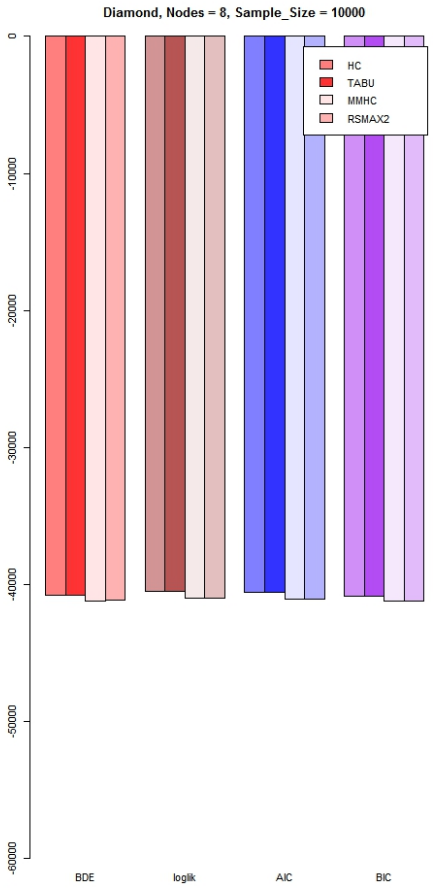


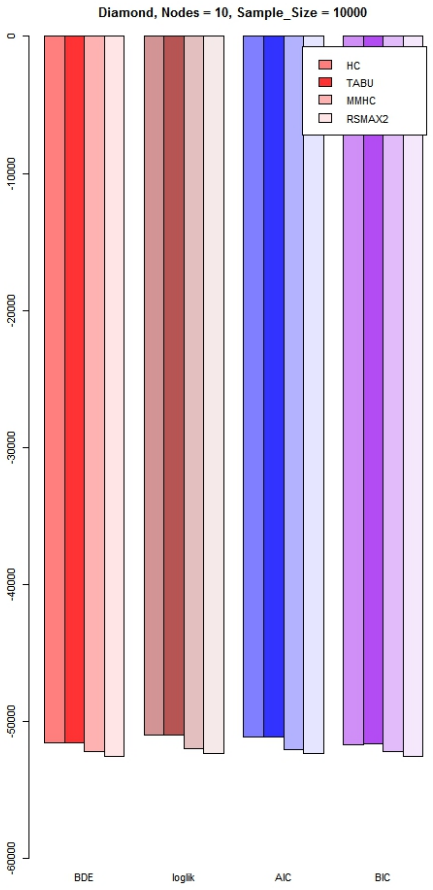


Synthetic Data 유형에 따른 비교 분석

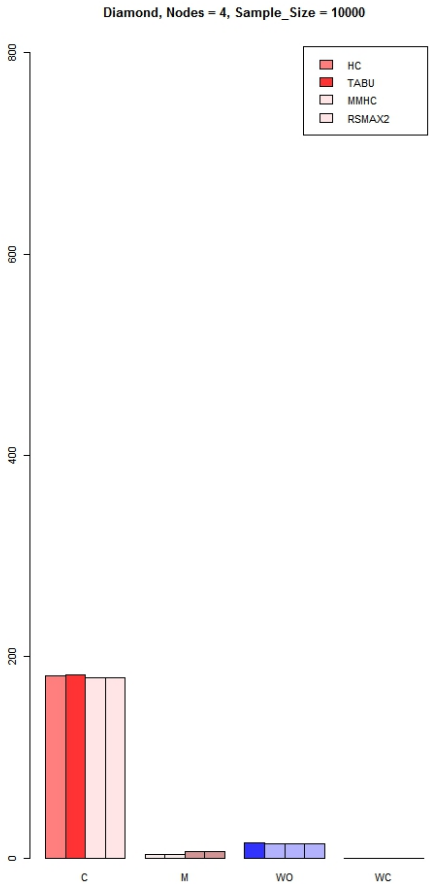


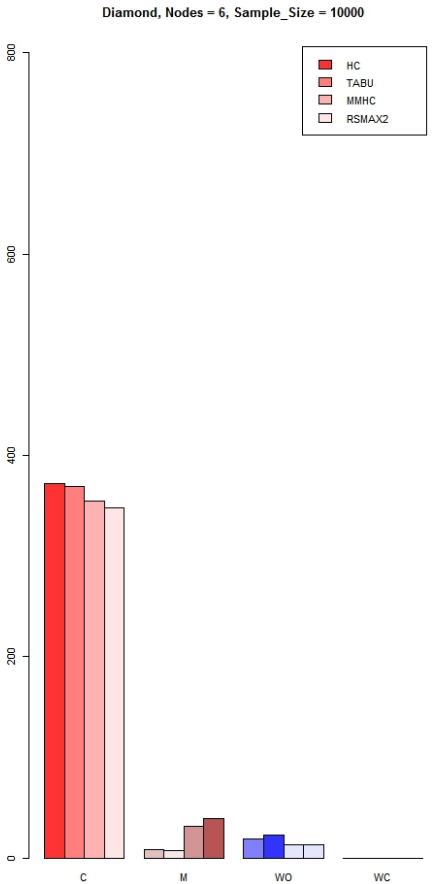


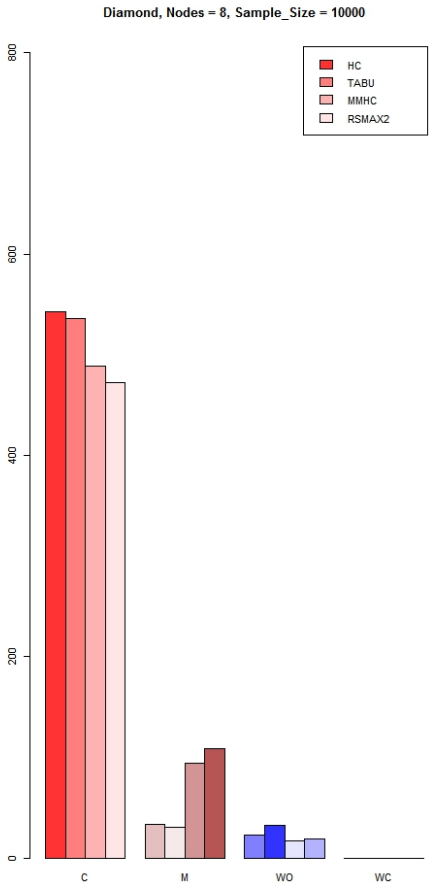


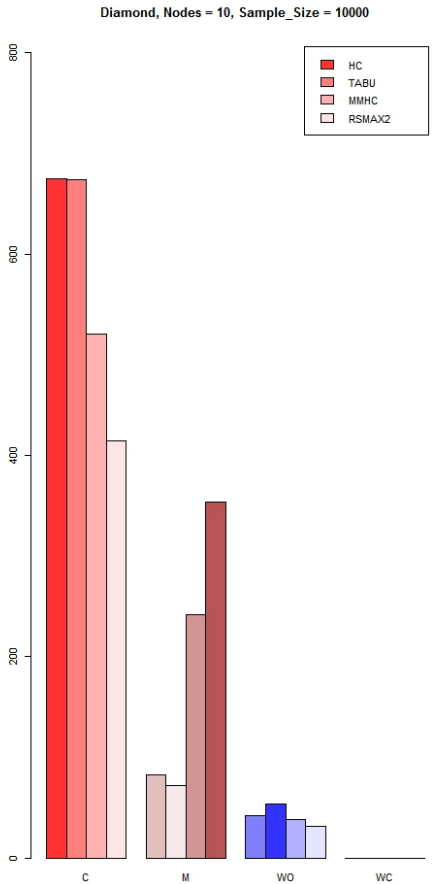


Synthetic Data 유형에 따른 비교 분석

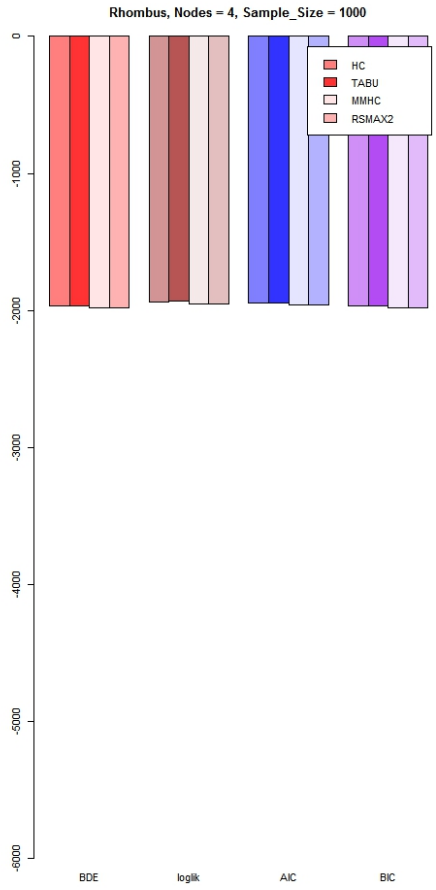


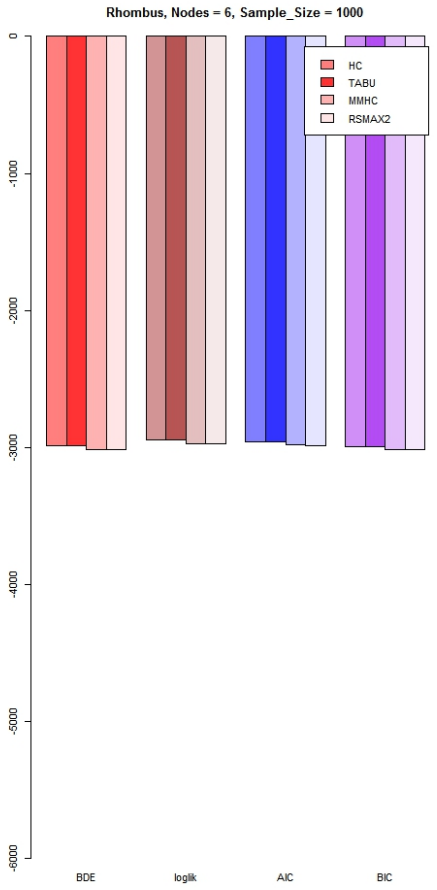


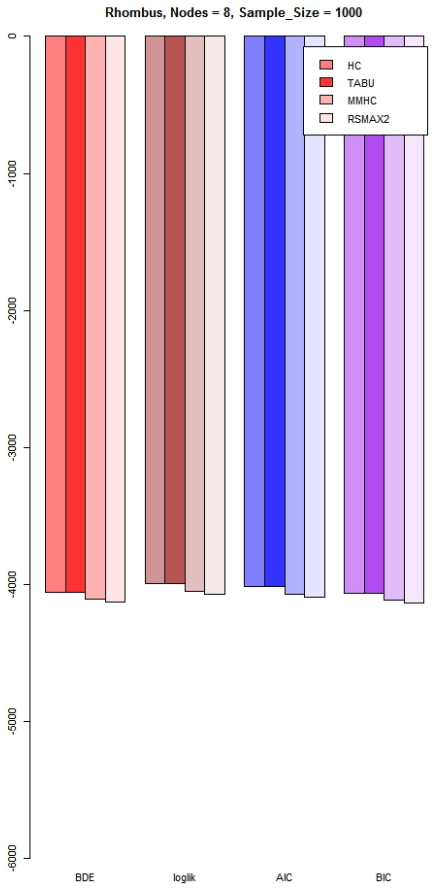


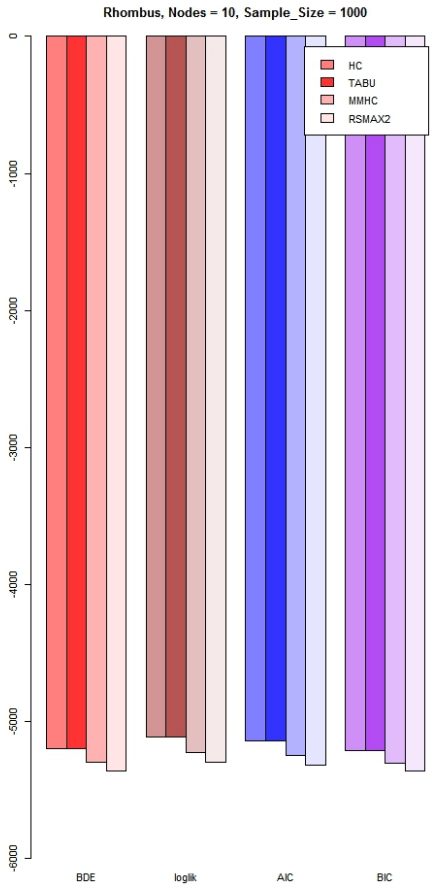


Synthetic Data 유형에 따른 비교 분석

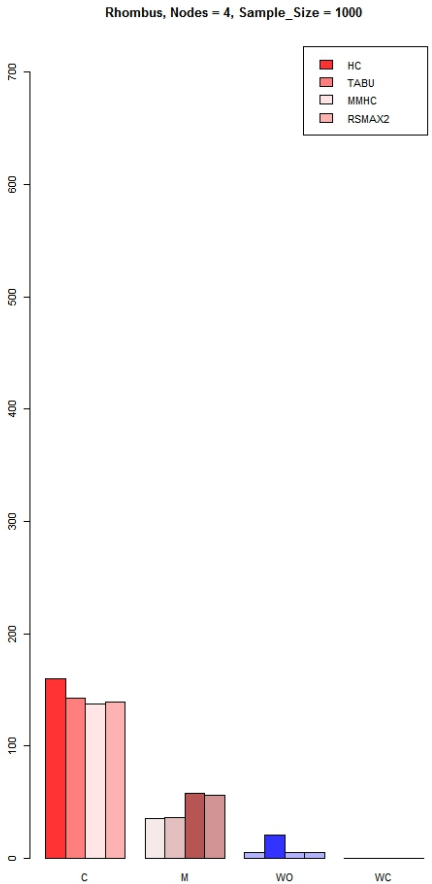


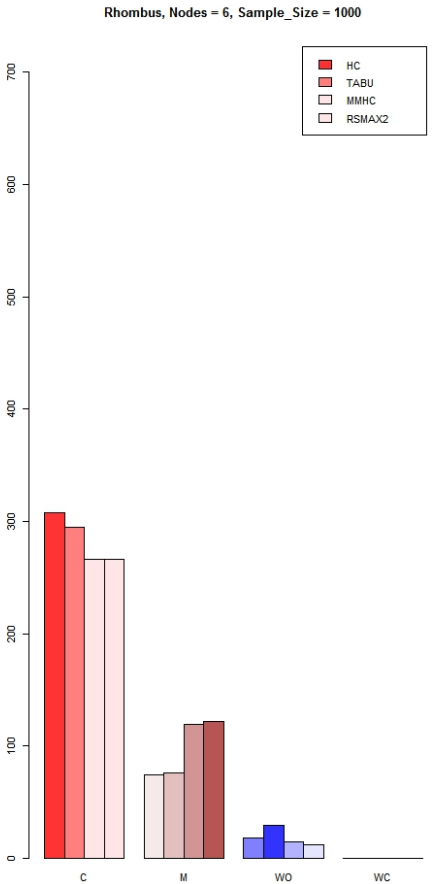


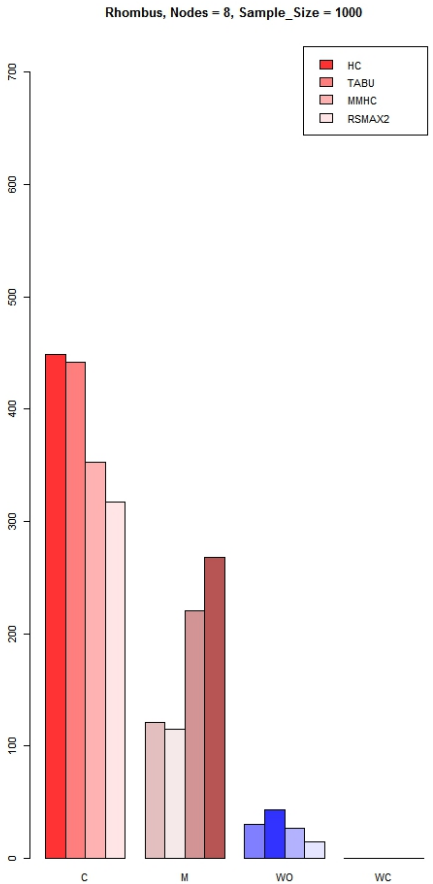


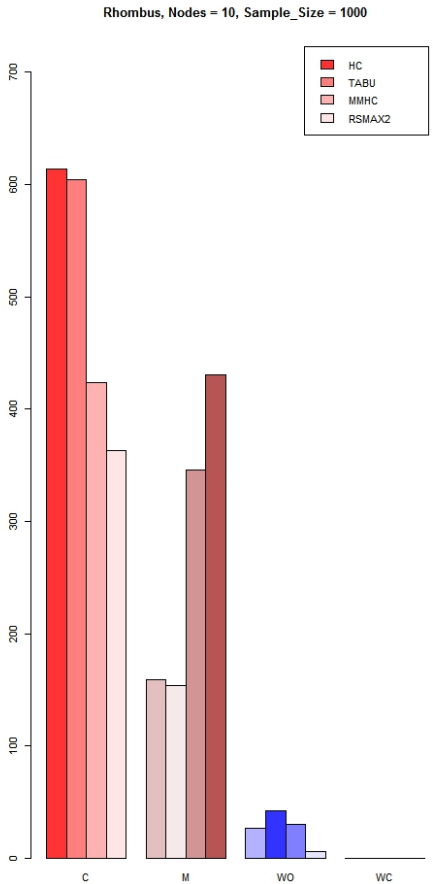


Synthetic Data 유형에 따른 비교 분석

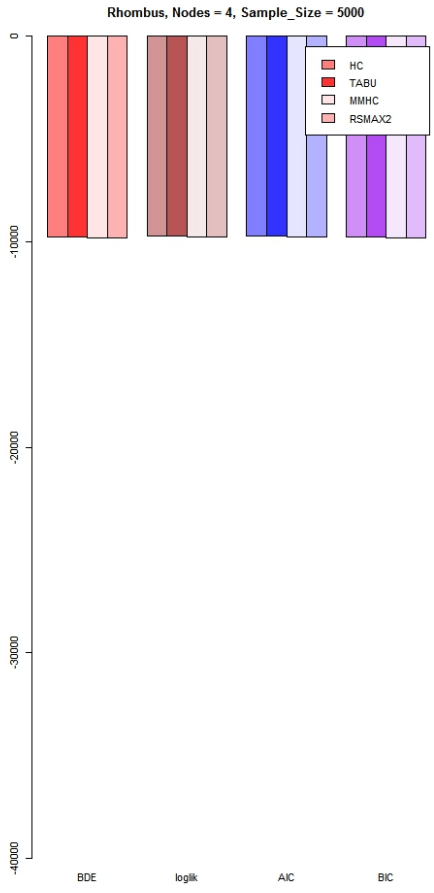


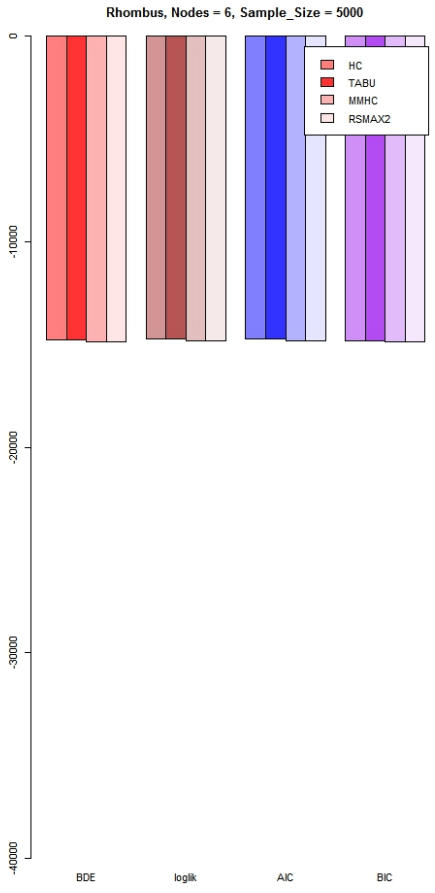


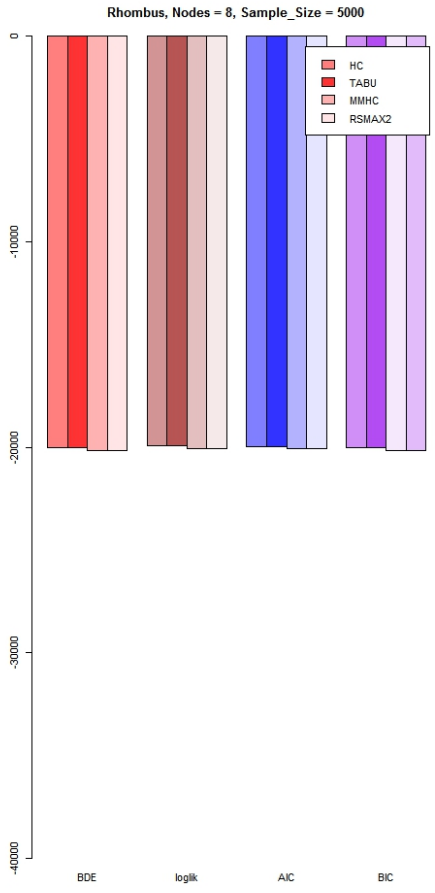


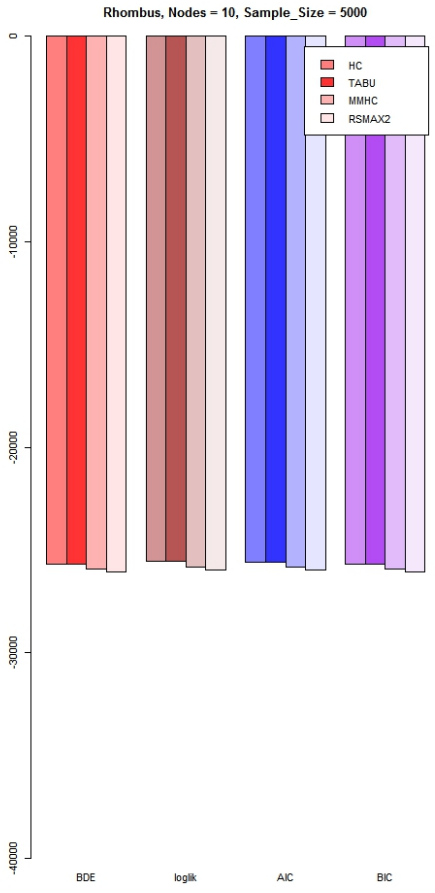


Synthetic Data 유형에 따른 비교 분석

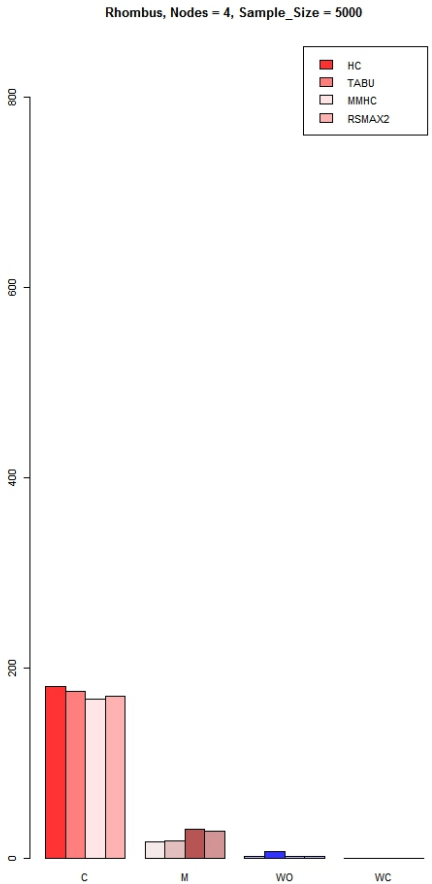


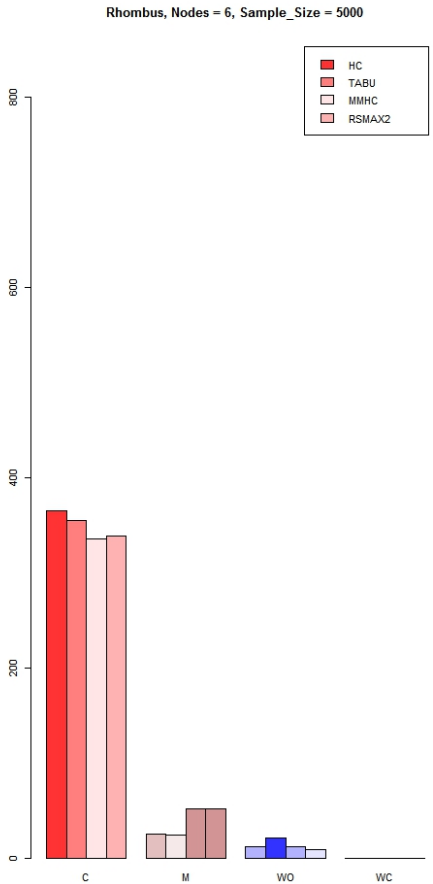


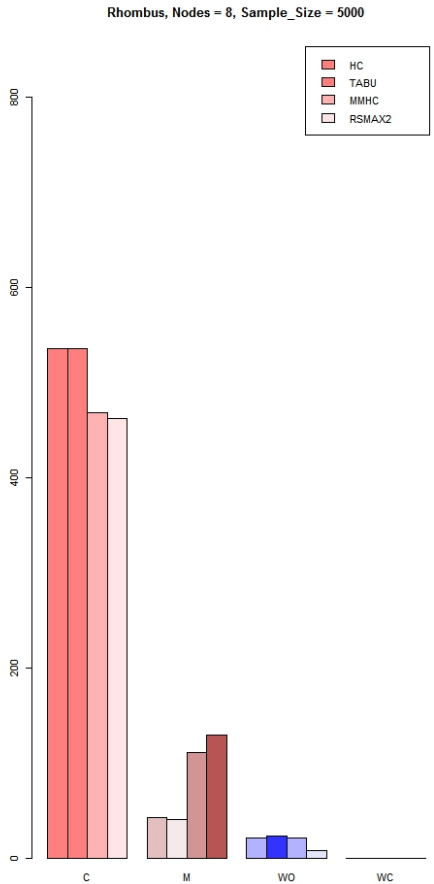


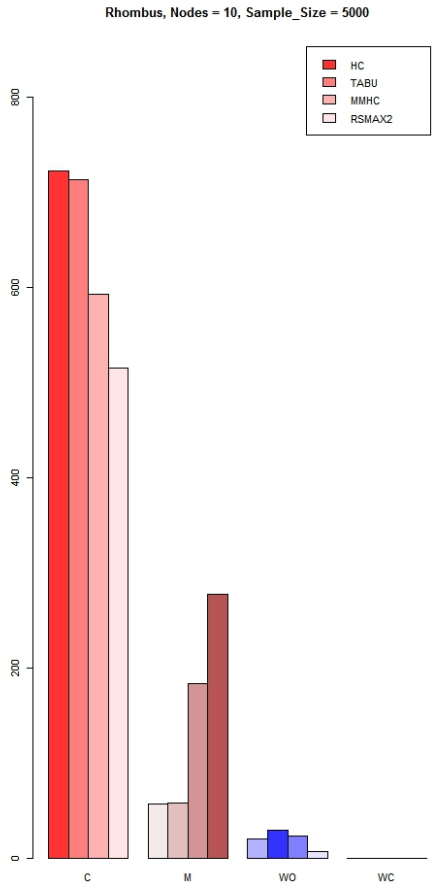


Synthetic Data 유형에 따른 비교 분석

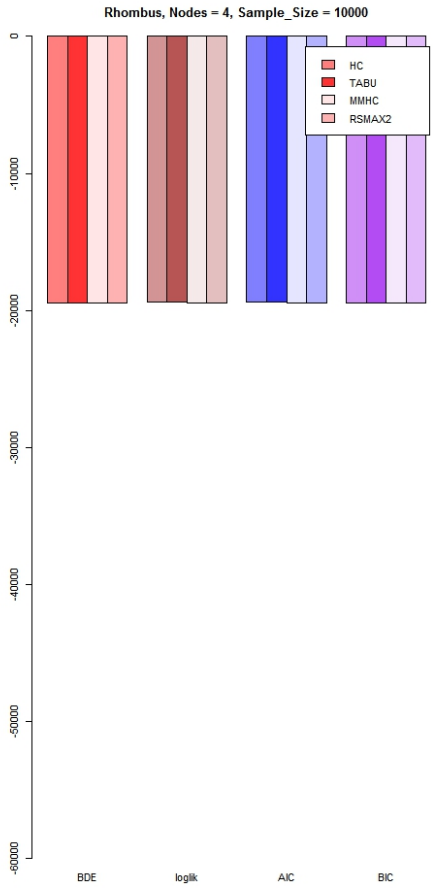


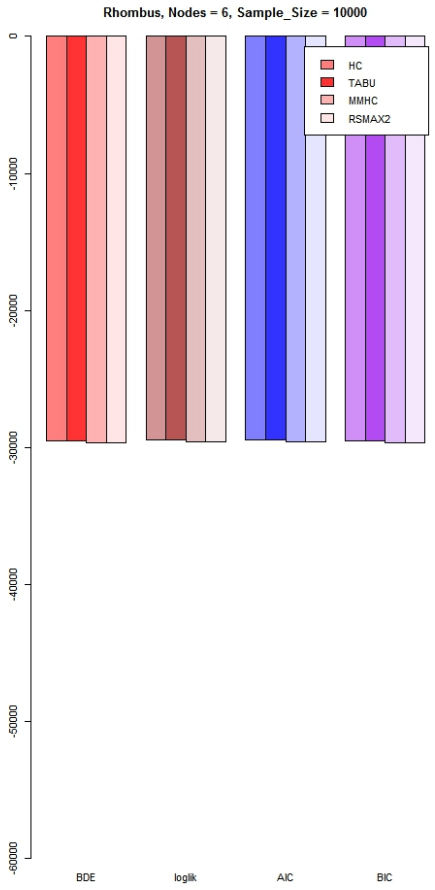


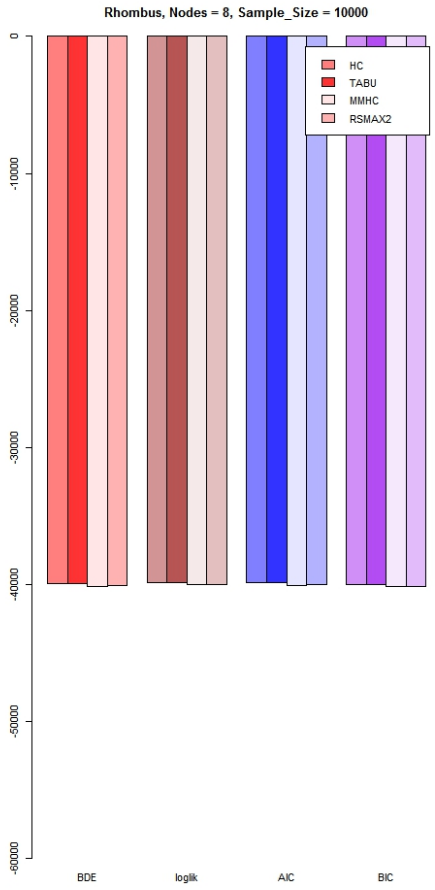


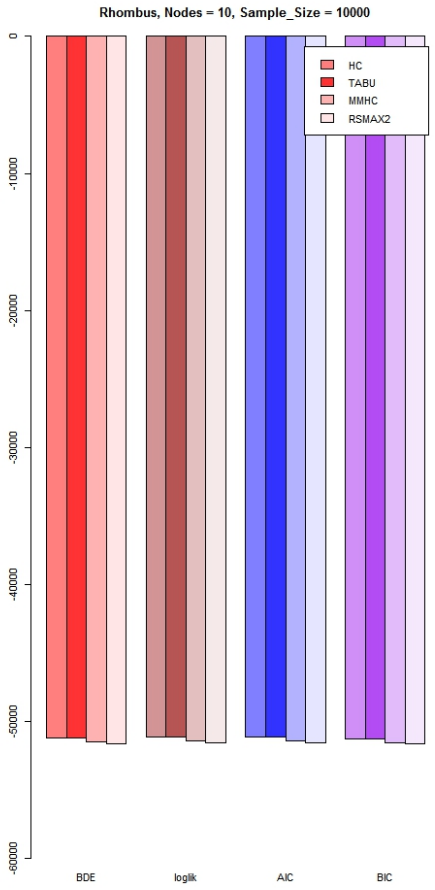


Synthetic Data 유형에 따른 비교 분석









Synthetic Data 유형에 따른 비교 분석