BIOINFORMATICS

Vol. 00 no. 00 2012
Pages 1–2

Supplementary Data for Inferring Time-Delayed Causal Gene Network using Time-series Expression Data

Leung-Yau Lo*, Kwong-Sak Leung and Kin-Hong Lee

Department of Computer Science & Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Availability:

Contact: lylo@cse.cuhk.edu.hk

Supplementary Information:

1 SUPPLEMENTARY DATA

1.1 Algorithms

Algorithm 1 is the main inference algorithm. Algorithm 2 is the algorithm for randomly generating acyclic gene network for the analysis of synthetic data. Algorithm 3 is the algorithm for simulating expression data from synthetic network.x

2 FIGURES

2.1 Comparison of Stage 1 and Stage 2

Figures 1 and 2 show respectively the comparison of the median links and delays performance of 100 trials after stage 1 and stage 2 for partial correlation and conditional mutual information. Figure 6 shows the comparison of effects performance for partial correlation.

2.2 Effects of Different Number of Time Points Provided

Figures 3, 5 and 7 show the performances of Delays, Links and Effects against the number of time points for $n = 30$ genes, and the maximum number of neighbors conditioned on is $N_0 = 4$.

As expected, as the number of time points increases, the performance improves. If the score threshold is more strict, the recall is lower, but the precision (PPV) is higher.

2.3 Effects of Varying the Maximum Number of Neighbors to Condition on, N_0

For completeness, figure 4 shows the median F-measures for links for Partial Correlation (PCor) and Conditional Mutual Information (MI) under different number of maximum number of neighbors conditioned on (N_0). As can be seen, N_0 has little effect on the median F-measure attained, except for a little improvement of going from $N_0 = 1$ to $N_0 = 2$. The graphs for recall and precision are similar (not shown). This is presumably because the conditioning rarely goes to a very high order.

Algorithm 1 Infer a time-delayed GRN given the expression data e, gene number n, time steps m, maximum time-delay T, maximum neighbors N_0, score thresholds s_1 and s_2, pair-wise test function $Test$, and conditional test function $CTest$

Proc CLINDE(e,n,m,T,N_0,s_1,s_2,$Test$,$CTest$)
{Stage 1: Initial Pair-wise Links}
$R \leftarrow \emptyset$
for $1 \leq i, j \leq n$ and $i \neq j$ **do**
  for $1 \leq \tau \leq T$ **do**
    {c is the correlation or mutual information, s is the score of the test}
    $(c, s) \leftarrow Test((e_{i,1}, \ldots, e_{i,m-\tau}), (e_{j,1+\tau}, \ldots, e_{j,m}))$
    if $c \geq s_1$ **then**
      $R \leftarrow R \cup \{(\tau, c, i, j)\}$
    end if
  end for
end for
{Stage 2: Pruning}
$h \leftarrow 1$
while $h \leq N_0$ and $\exists (i, j)$ with $\geq h$ neighbors **do**
  for $(\tau, c, i, j) \in R$ **do**
    {Note that R may be changed}
    $N \leftarrow$ neighbors of (τ, c, i, j) in R
    if $|N| \geq h$ **then**
      for $N' \subseteq N$ and $|N'| = h$ **do**
        for $L' \in$ set of combinations of neighbor links of each member of N' **do**
          Calculate Rel_i and Adj for members of L'
          Get the adjusted time-series for i, j and L'
          $(c, s) \leftarrow CTest$ applied on the adjusted time-series of i and j conditioned on L'
          if $s < s_2$ **then**
            $R \leftarrow R \setminus \{(\tau, c, i, j)\}$
          end if
        end for
      end for
    end if
  end for
  $h \leftarrow h + 1$
end while
return R

2.4 IRMA On

Figure 8 shows the links recall-precision graphs for IRMA On dataset (after stage 2) with different inference parameters, using Conditional Mutual Information (MI) and Partial Correlation

*to whom correspondence should be addressed

Algorithm 2 Simulate a GRN given the gene number n and maximum predecessor M for each gene

Proc SIMGRN(n,M)
Require: $n > 0$ and $M > 0$
  $R \leftarrow \emptyset$
  for $1 \leq i \leq n$ **do**
    $N_i \leftarrow$ the set of M randomly chosen numbers from $\{1, \ldots, n\}$
    $N_i' \leftarrow N_i \setminus \{i, \ldots, n\}$
    for $k \in N_i'$ **do**
      $\tau_{ki} \sim U(1, 10)$,
      $c_{ki} \sim (1 - 2 * Bern(0.5)) * U(0.5, 1.5)$
      $R \leftarrow R \cup \{(\tau_{ki}, c_{ki}, k, i)\}$
    end for
  end for
  $\theta \leftarrow$ random permutation of $\{1, \ldots, n\}$
  return R with θ applied

Algorithm 3 Simulate the expression data from a GRN, given the GRN R, the gene number n, time step size dt and number of time steps m

Proc SIMEXPRESSION(R,n,dt,m)
  $e \leftarrow$ empty n by m matrix
  for $1 \leq i \leq m$ **do**
    for $1 \leq j \leq n$ **do**
      $x \leftarrow N(0, 0.01)$
      for $(\tau_k, c_k, k, j) \in R$ **do**
        {Predecessors of j}
        $\tau \leftarrow \lceil \frac{\tau_k}{dt} \rceil$
        if $\tau < i$ **then**
          $x \leftarrow x + c_k e_{k,i-\tau}$
        end if
      end for
      $e_{j,i} \leftarrow x$
    end for
  end for
  return e

(PCor), different score thresholds, maximum time delay T_0 and different maximum number of neighbors conditioned on (N_0). We show the performance on (directed) links only, because the regulatory effects for the IRMA network are not clearly stated by Zoppoli *et al.* (2010). Since our algorithm may output more than one link for each gene pair, so for the precision, we calculate the proportion of correct prediction among the predictions. There are a few things to note. Using Conditional Mutual Information can yield better results than using Partial Correlation. The recall after stage 2 is generally higher when a lower (less strict) threshold is used: for mutual information it can reach 0.75 by using the threshold 0.01, and for partial correlation it can reach 0.5 by using a threshold of 1 or 1.30103 (corresponding to p-values 0.1 and 0.05 respectively). And the precision can reach 1 by using thresholds 0.4 for MI, and 3.30103 and 4 (corresponding to p-values 0.0005 and 0.0001 respectively) for PCor. In figure 8, for the same score threshold, there may be multiple points which correspond to different values of T_0 and N_0.

2.5 IRMA Off

Figure 9 shows the links recall-precision graphs for IRMA Off dataset (after stage 2) with different inference parameters. Qualitatively, the effects of varying the score threshold also apply to this case, which is not surprising. But it should be noted that the performance for the IRMA Off dataset is worse than that for IRMA On dataset, although the underlying network is the same, and the number of time points for the IRMA Off dataset is higher (21 points in IRMA Off versus 16 points in IRMA On). This observation is consistent with Zoppoli *et al.* (2010). Presumably the worse performance is because the stimulus in IRMA Off case is much weaker (Zoppoli *et al.*, 2010). Compared to the IRMA On dataset, the performance difference for Conditional Mutual Information (MI) and Partial Correlation (PCor) is more prominent for IRMA Off dataset when there is no interpolation, it maybe because mutual information is more robust when the time points are scarce. Another possible reason is that the regulatory relationships between genes are not linear enough for partial correlation to capture well. When interpolation is used, the performance of using Partial Correlation can be improved quite a bit. This is quite surprising, because although doing interpolation increases the number of time points, but *no new* information is put into the dataset.

3 FEATURE COMPARISON

Table 1 lists the comparison of the features of our algorithm with other GRN inference algorithms compared in this paper. ARACNE estimates undirected links only. TSNI, Banjo and ARACNE does not have time delays in the links. TimeDelay-ARACNE, while has time delays in the inference, does not explicitly estimate the regulatory effect of the links because it uses mutual information. The main strength of our algorithm is that it infers the time delays in the links, as well as the regulatory effect (activatory or repressive) in the links, which are all very important information in understanding gene regulations. We did not compare with (Hyvärinen *et al.*, 2008) because they assume non-gaussian noise, and in our inference, we assume gaussian noise.

ACKNOWLEDGEMENT

Funding:

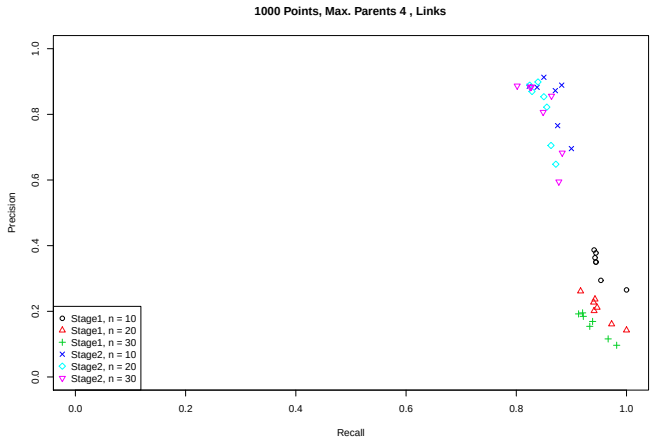
REFERENCES

Hyvärinen, A., Shimizu, S., and Hoyer, P. O. (2008). Causal modelling combining instantaneous and lagged effects: an identifiable model based on non-gaussianity. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 424–431, New York, NY, USA. ACM.

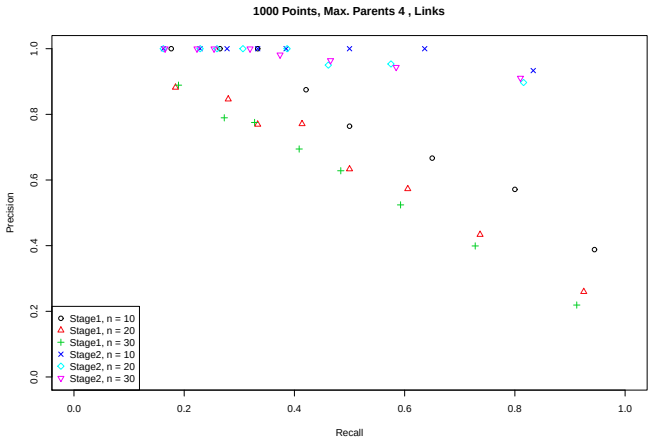
Zoppoli, P., Morganella, S., and Ceccarelli, M. (2010). Timedelay-aracne: Reverse engineering of gene networks from time-course data by an information theoretic approach. *BMC Bioinformatics*, **11**(1), 154.

Fig. 1: Comparison of Links Median Performance of 100 Trials after Stage 1 and Stage 2 for Partial Correlation (PCor), Conditional Mutual Information (MI), different number of genes (n=10,20,30), different number of time points m=1000, 100, 20. Maximum number of neighbors conditioned on (N_0) is 4. The different points correspond to different score thresholds.

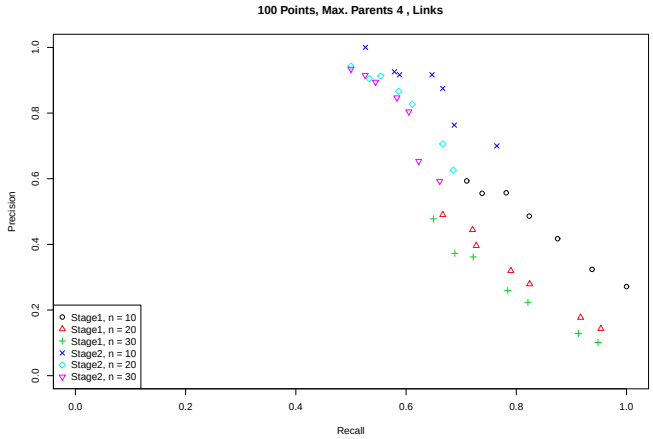
(a) PCor, $m = 1000$



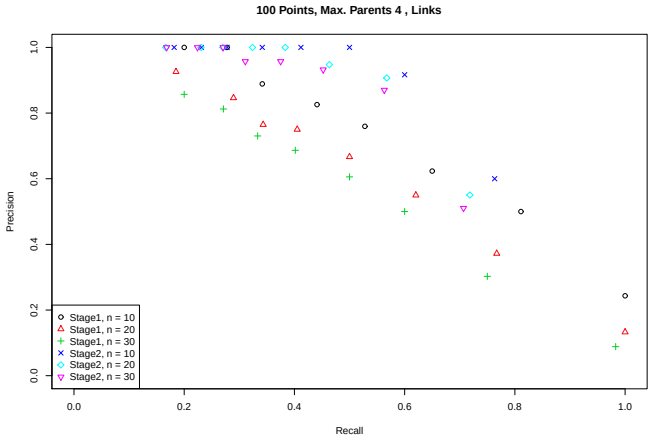
(b) MI, $m = 1000$



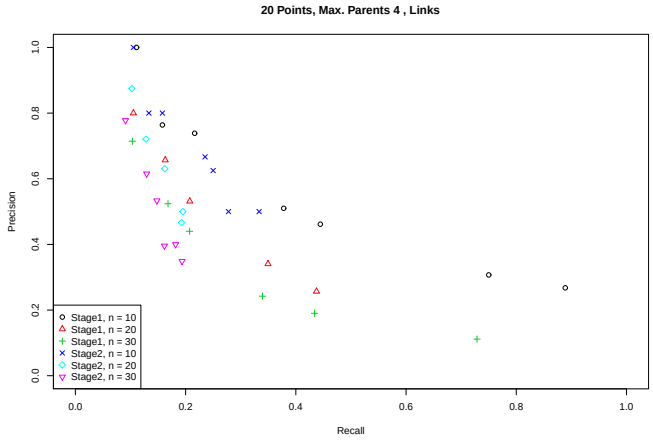
(c) PCor, $m = 100$



(d) MI, $m = 100$



(e) PCor, $m = 20$



(f) MI, $m = 20$

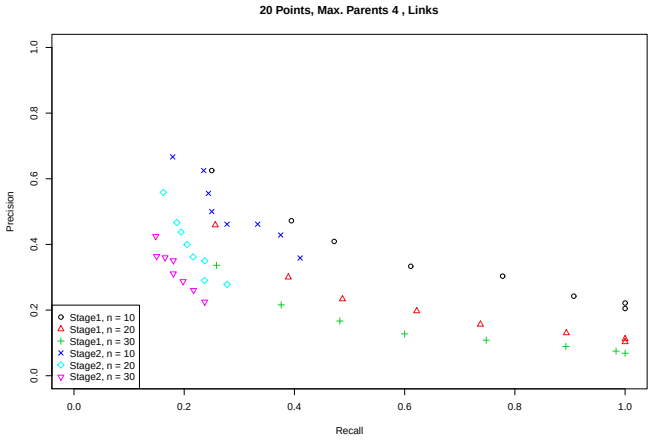
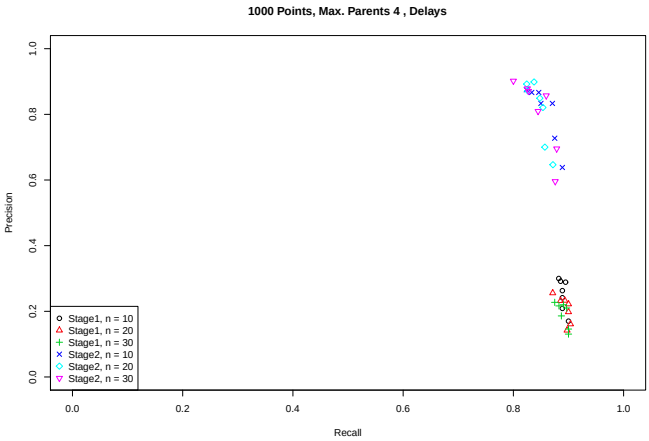
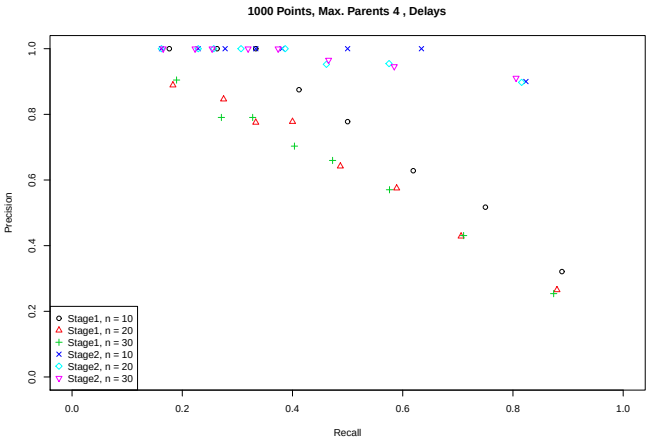


Fig. 2: Comparison of Delays Median Performance of 100 Trials after Stage 1 and Stage 2 for Partial Correlation (PCor), Conditional Mutual Information (MI), different number of genes (n=10,20,30), different number of time points m=1000, 100, 20. Maximum number of neighbors conditioned on (N_0) is 4. The different points correspond to different score thresholds.

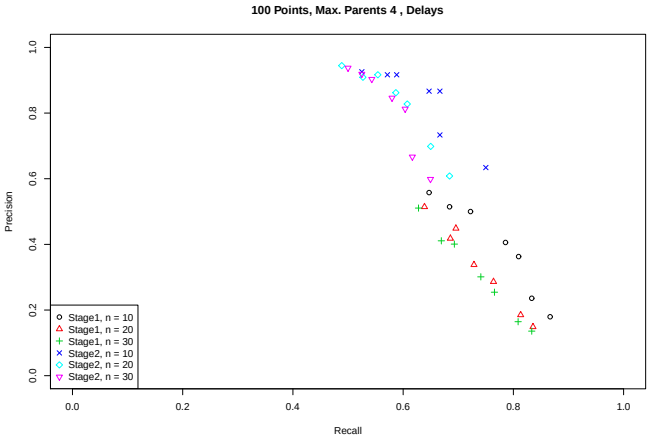
(a) PCor, $m = 1000$



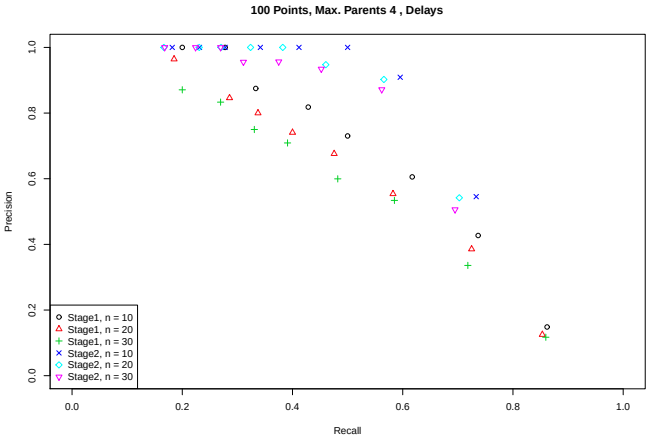
(b) MI, $m = 1000$



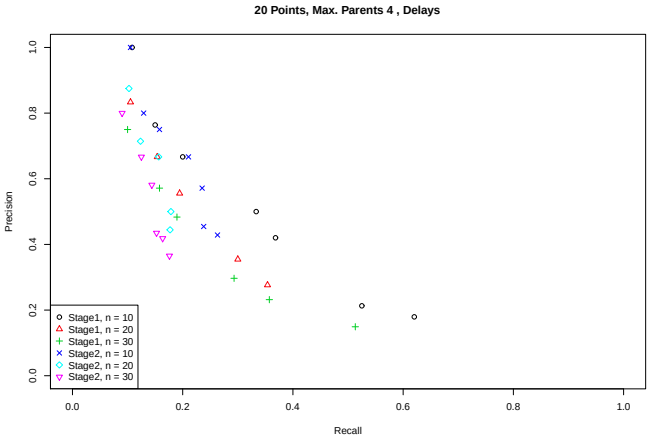
(c) PCor, $m = 100$



(d) MI, $m = 100$



(e) PCor, $m = 20$



(f) MI, $m = 20$

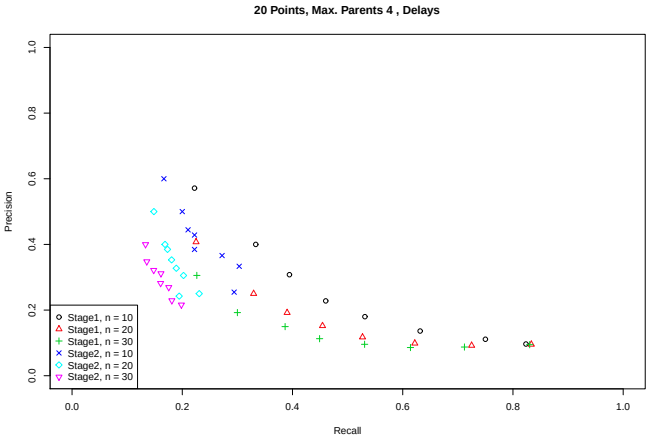
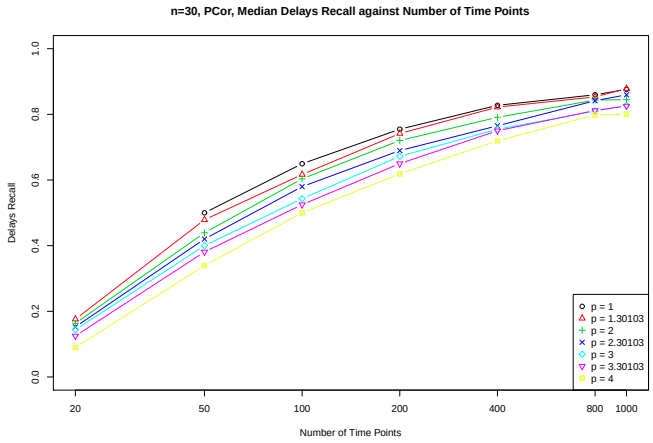
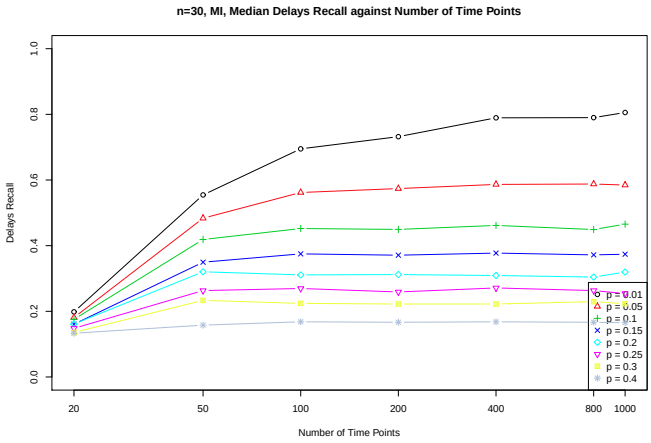


Fig. 3: Delays Median Stage 2 Performance (Recall, PPV, F-measure) of 100 Trials for Partial Correlation (PCor), Conditional Mutual Information (MI) against different number of time points. Number of genes is $n = 30$. Maximum number of neighbors conditioned on (N_0) is 4. The different lines correspond to different score thresholds p.

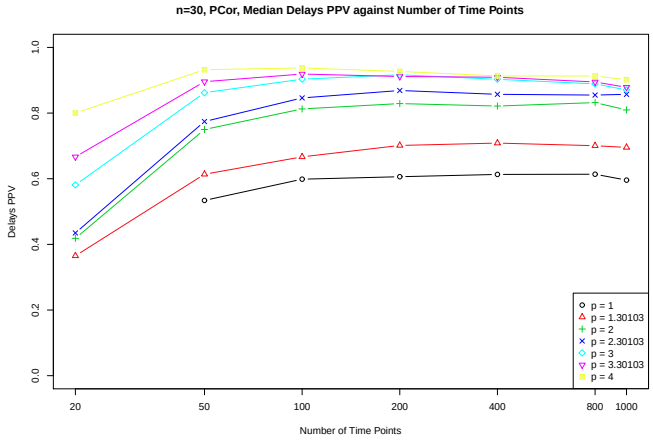
(a) PCor, Recall



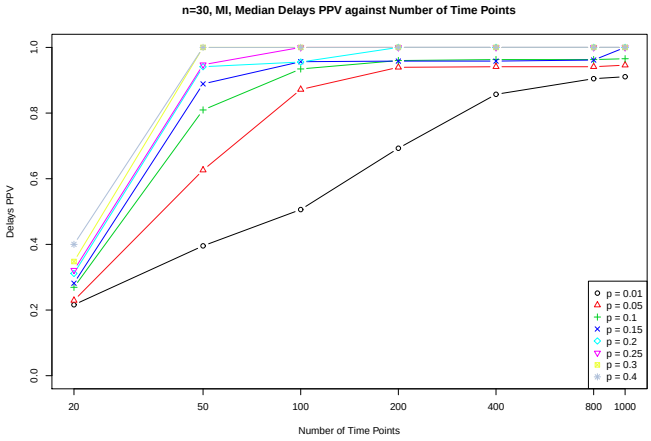
(b) MI, Recall



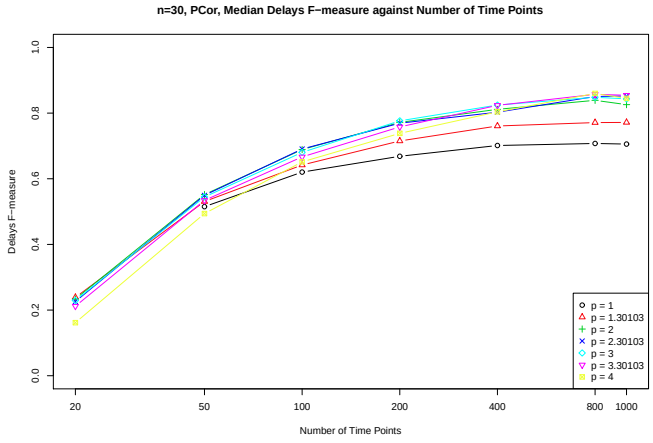
(c) PCor, PPV



(d) MI, PPV



(e) PCor, F-measure



(f) MI, F-measure

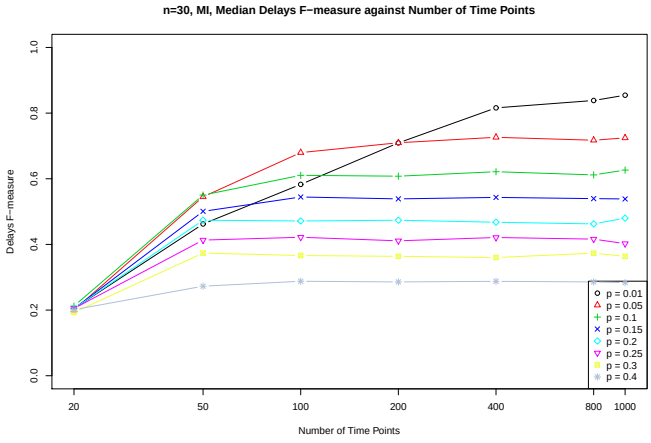
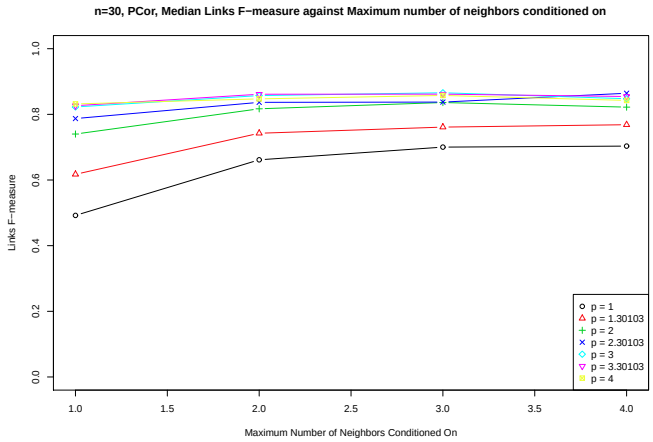
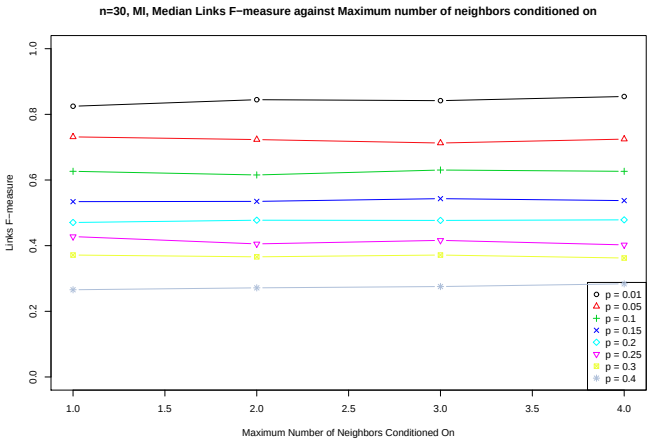


Fig. 4: Links Median Stage 2 F-measure of 100 Trials for Partial Correlation (PCor), Conditional Mutual Information (MI) against different maximum number of neighbors conditioned on (N_0). Number of genes is $n = 30$. The different lines correspond to different score thresholds p.

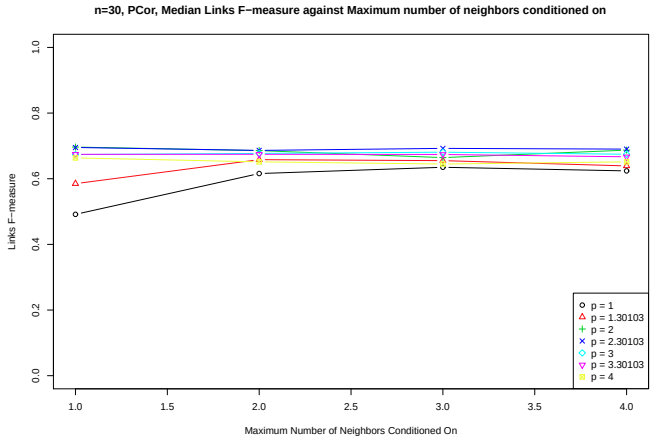
(a) PCor, $m = 1000$



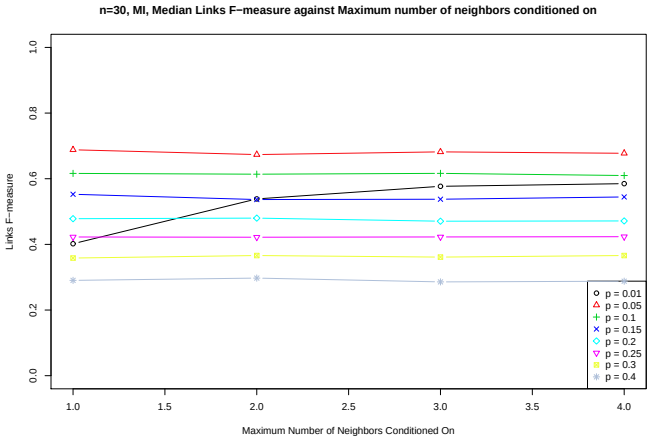
(b) MI, $m = 1000$



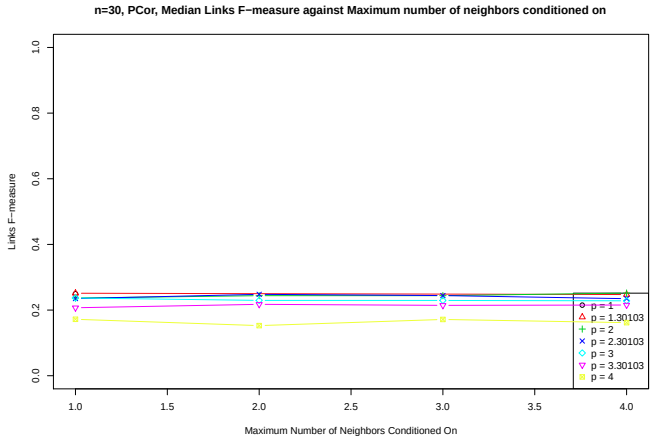
(c) PCor, $m = 100$



(d) MI, $m = 100$



(e) PCor, $m = 20$



(f) MI, $m = 20$

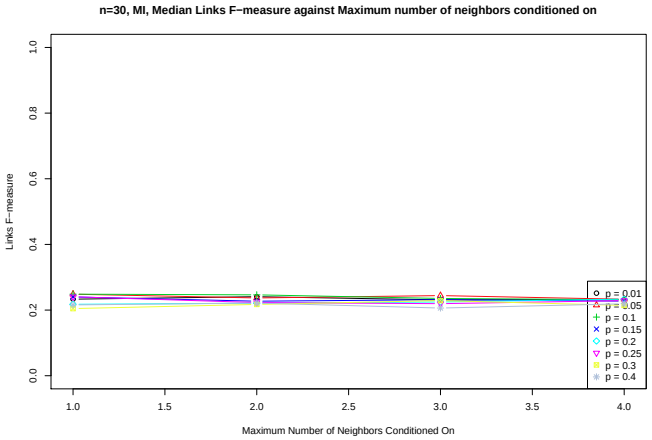
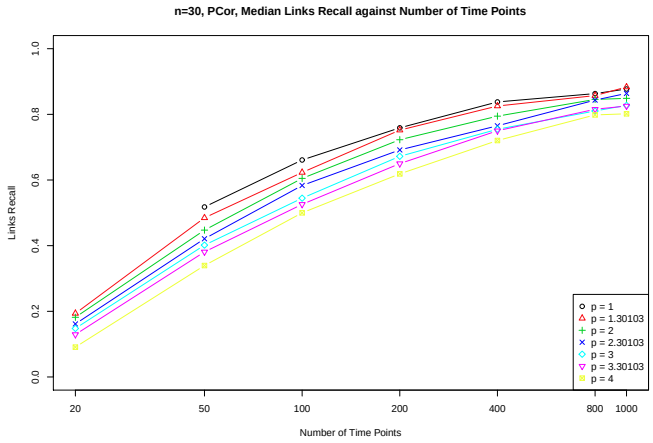
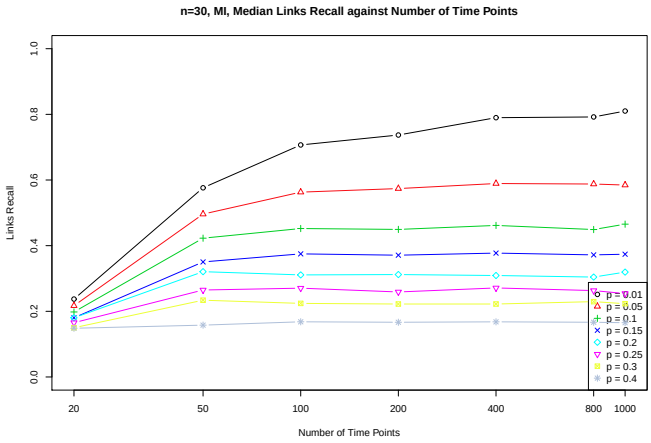


Fig. 5: Links Median Stage 2 Performance (Recall, PPV, F-measure) of 100 Trials for Partial Correlation (PCor), Conditional Mutual Information (MI) against different number of time points. Number of genes is $n = 30$. Maximum number of neighbors conditioned on (N_0) is 4. The different lines correspond to different score thresholds p.

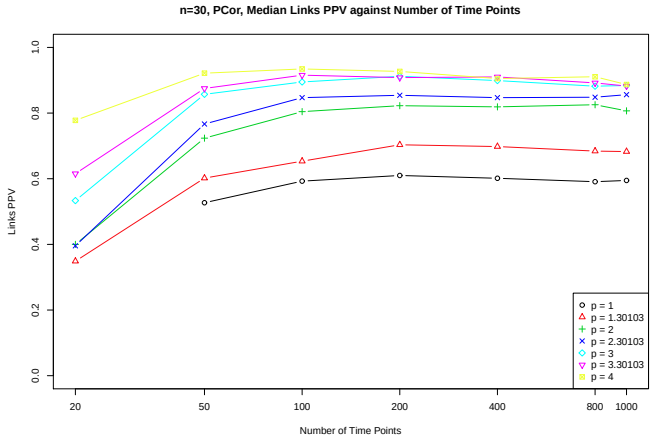
(a) PCor, Recall



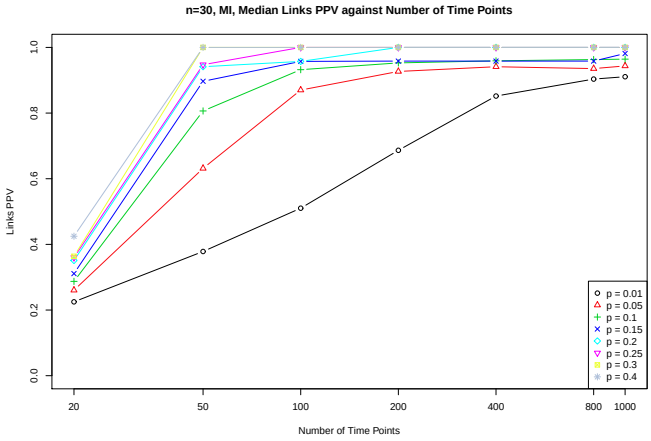
(b) MI, Recall



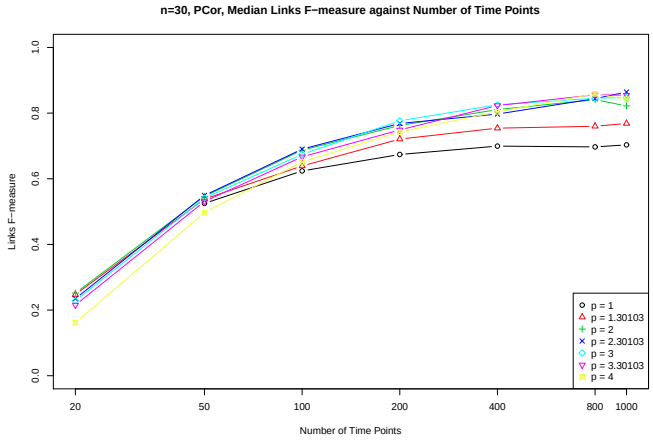
(c) PCor, PPV



(d) MI, PPV



(e) PCor, F-measure



(f) MI, F-measure

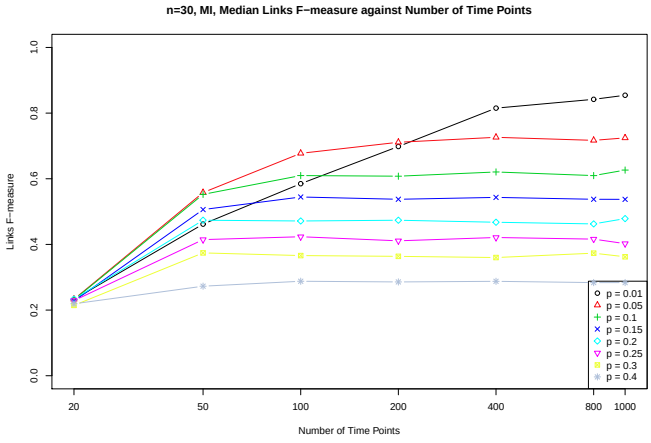
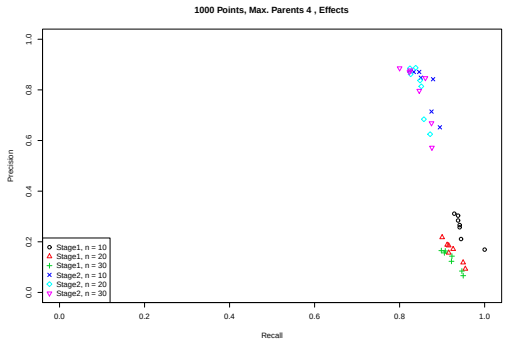
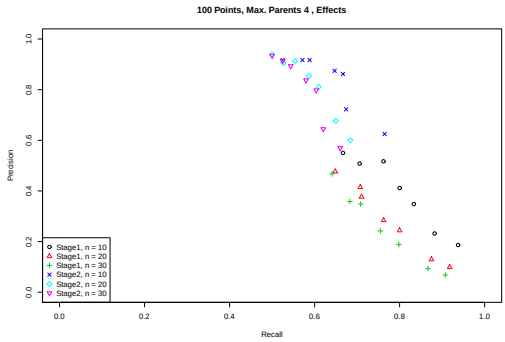


Fig. 6: Comparison of Effects Median Performance of 100 Trials after Stage 1 and Stage 2 for Partial Correlation (PCor), different number of genes (n=10,20,30), different number of time points m=1000, 100, 20. Maximum number of neighbors conditioned on (N_0) is 4. The different points correspond to different score thresholds.

(a) PCor, $m = 1000$



(b) PCor, $m = 100$



(c) PCor, $m = 20$

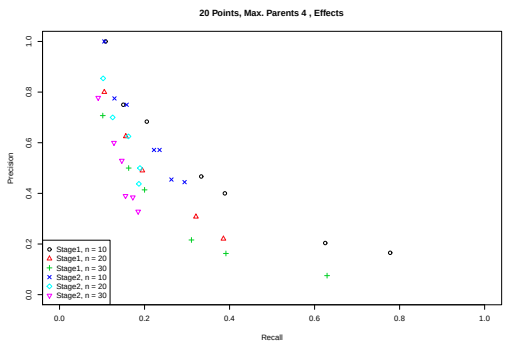
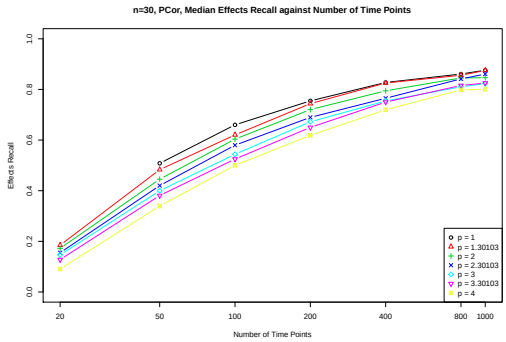
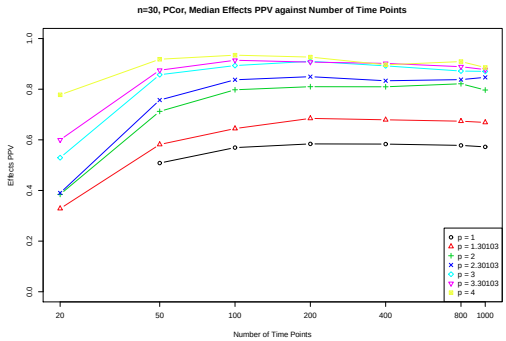


Fig. 7: Effects Median Stage 2 Performance (Recall, PPV, F-measure) of 100 Trials for Partial Correlation (PCor) against different number of time points. Number of genes is $n = 30$. Maximum number of neighbors conditioned on (N_0) is 4. The different lines correspond to different score thresholds p.

(a) PCor, Recall



(b) PCor, PPV



(c) PCor, F-measure

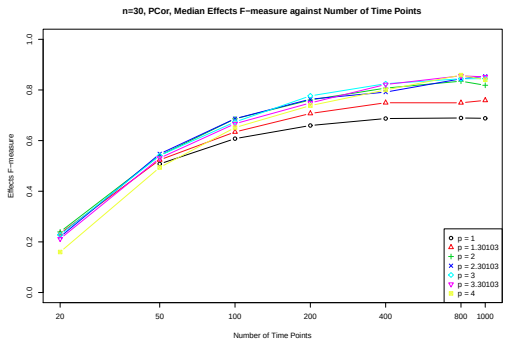
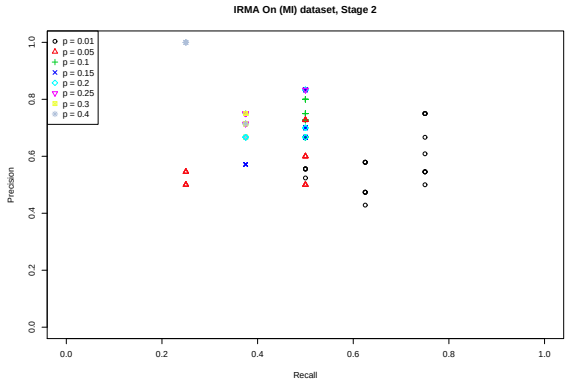


Fig. 8: Links Recall-Precision graphs for IRMA On Dataset (after Stage 2) with different inference parameters, without interpolation. For partial correlation, the score threshold is $-\log_{10}(p)$ for p-value p. For the same score threshold, there are multiple points, corresponding to different maximum time delay T_0 and maximum number of neighbors conditioned N_0. (a) IRMA On using Mutual Information (MI). (b) IRMA On using Partial Correlation (PCor).

(a) IRMA On, MI



(b) IRMA On, PCor

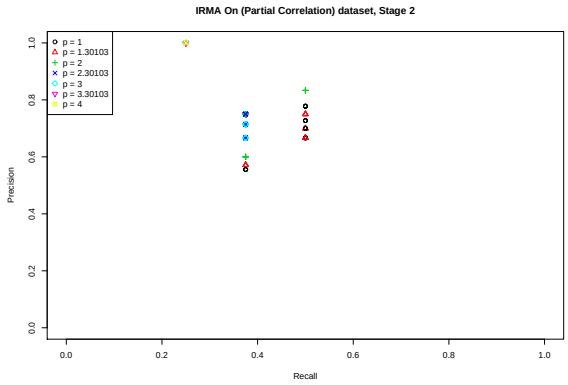
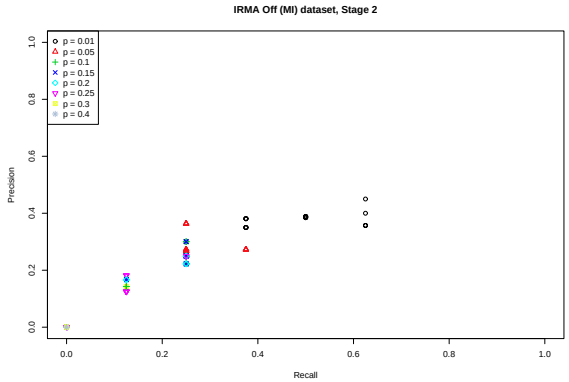


Fig. 9: Links Recall-Precision graphs for IRMA Off Dataset (after Stage 2) with different inference parameters, without interpolation. For partial correlation, the score threshold is $-\log_{10}(p)$ for p-value p. For the same score threshold, there are multiple points, corresponding to different maximum time delay T_0 and maximum number of neighbors conditioned N_0. (a) IRMA Off using Mutual Information (MI). (b) IRMA Off using Partial Correlation (PCor).

(a) IRMA Off, MI



(b) IRMA Off, PCor

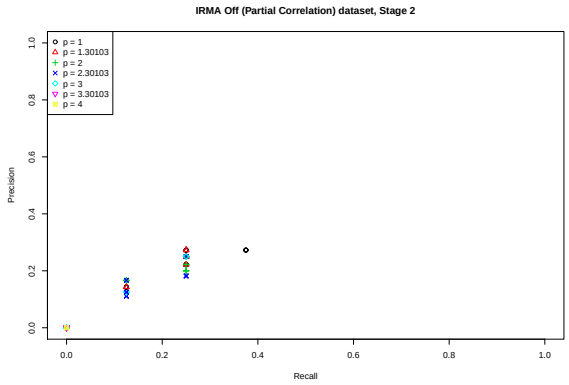


Table 1. Feature Comparison of some GRN Inference Algorithms. PCor: Partial Correlation. MI: Mutual Information. DPI: Data Processing Inequality.

Algorithm	Approach	Directed Links?	Delays?	Effects in the Links?	Type of Data	Allows Directed Cycles?
CLINDE	PCor / Conditional MI for link estimation and pruning	Yes	Yes	Yes for Partial Correlation	Time-series	Yes
TimeDelay-ARACNE	MI for link estimation with DPI for pruning	Yes	Yes	No	Time-series	Yes
TSNI	Solving discretized linear ODE model	Yes	No	Yes	Time-series (with perturbation data)	Yes
Banjo	Search networks with high score	Yes	No	Yes	Steady-state /Time-series	Not for steady-state data
ARACNE	MI for link estimation with DPI for pruning	No	No	No	Steady-state /Time-series	—