Granger Causality vs. Dynamic Bayesian Network Inference： A Comparative Study

Cunlu Zou┼[2]　　Jianfeng Feng*[1,2]

[1,2]Centre for Computational System Biology, Fudan University,
Shanghai, PR China
[2]Department of Computer Science and Mathematics, University of
Warwick

July 2008

┼ email address: csrcbh@dcs.warwick.ac.uk

*Corresponding author:　jianfeng.feng@warwick.ac.uk

Abstract

In computational biology, one often faces the problem of deriving the causal relationship among different elements such as genes, proteins, metabolites, neurons and so on, based upon multi-dimensional temporal data. Currently, there are two common approaches used to explore the network structure among elements. One is the Granger causality approach, and the other is the dynamic Bayesian network inference approach. Both have at least a few thousand publications reported in the literature. A key issue is to choose which approach is used to tackle the data, in particular when they give rise to contradictory results. In this paper, we provide an answer by focusing on a systematic and computationally intensive comparison between the two approaches on both synthesized and experimental data. For synthesized data, a critical point of the data length is found: the dynamic Bayesian network outperforms the Granger causality approach when the data length is short, and vice versa. We then test our results in experimental data of short length which is a common scenario in current biological experiments: it is again confirmed that the dynamic Bayesian network works better. A software package is available at supplement material.

1 Introduction

Based upon high throughput data, to reliably and accurately explore the network structure of elements (genes, proteins, metabolites, neurons etc.) is one of the most important issues in computational biology [1,2,3,4,5,6]. Currently, there are two main approaches which are often used to infer causal relationships [7] or interactions among a set of elements [8,9]. One is the Granger causality approach [10,11], and the other is the Bayesian network inference approach [12,13]. The latter is often applied to static data. However, one can employ the dynamic Bayesian networks to deal with time series data for which the Granger causality has been solely developed. The Granger causality has the advantage of having a corresponding frequency domain decomposition so that one can clearly find at which frequencies two elements interact with each other.

Giving a multi-variable time series dataset, the Granger causality and dynamic Bayesian networks [14] can both be applied. The Granger causality notation, which was firstly introduced by Wiener and Granger [15,16], proposed that we can determine a causal influence of one time series on another: the prediction of one time series can be improved by incorporating the knowledge of the second one. On the other hand, The Bayesian network [17] is a special case of a diagrammatic representation of probability distributions, called probabilistic graphical models [18,19,20]. The Bayesian network graph model comprises nodes (also called vertices) connected by directed links (also called edges or arcs) and there is no cycle in the graph. To learn the structure and the parameters for the Bayesian networks from a set of data, we should search the space(s) of all possible graph representations, and find out which structure is most likely to produce our data. If we have a scoring function (or likelihood function) which can determine the structure and parameter likelihood from the data, then the problem is to find the highest score (maximum likelihood) structure among all the possible representations.

The causal relationship derived from these two approaches could be different, in particular when we face the data obtained from experiments. Therefore it is of vital importance to compare these two causal inferring approaches before we could confidently apply them to biological data. By doing the comparison, one expects to find the advantages, performances and stabilities for each technique.

Adopting the most common existing methods to find the coefficients of the time series in both approaches in the literature, we compare the dynamic Bayesian network with the Granger causality both in the linear and nonlinear model.

Interestingly, a critical point of the data length is found. When the data length is shorter than the critical point, the dynamic Bayesian network approach outperforms the Granger causality approach. But when the data length is longer, the Granger causality is more reliable. The conclusion is obtained via intensive computations (more than 100 computers over a few weeks). A biological data set of gene microarray is analyzed using both approaches, which indicates that for a data set with a short sampling length the dynamic Bayesian network produces more reliable results. In summary, we would argue that the dynamic Bayesian network is more suitable for dealing with experimental data.

2. Results

To illustrate and compare the differences between the dynamic Bayesian network inference and the conditional Granger causality, a simple multivariate model with fixed coefficients, which has been discussed in many papers to test the Granger causality, is tested first. We then extend our comparisons to the more general case of the model with random coefficients, which requires considerable computational resources. More than 100 networked computers are used to perform the comparisons for more than a week. Both the Granger causality and the dynamic Bayesian network are applied to nonlinear models. Finally, we test our approach on a set of microarray data recently acquired from a comparison of mock and infected Arabidopsis leaf.

2.1 Synthesized Data: Linear Case

Example 1 Suppose we have 5 simultaneously recorded time series generated according to the equations:

$$\begin{cases} X_1(n) = 0.95\sqrt{2}X_1(n-1) - 0.9025X_1(n-2) + \varepsilon_1 \\ X_2(n) = 0.5X_1(n-2) + \varepsilon_2 \\ X_3(n) = -0.4X_1(n-3) + \varepsilon_3 \\ X_4(n) = -0.5X_1(n-1) + 0.25\sqrt{2}X_4(n-1) + 0.25\sqrt{2}X_5(n-1) + \varepsilon_4 \\ X_5(n) = -0.25\sqrt{2}X_4(n-1) + 0.25\sqrt{2}X_5(n-1) + \varepsilon_5 \end{cases} \quad (1)$$

where n is the time, and $[\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5]$ are independent Gaussian white noise processes with zero means and unit variances. From the equations, we see that $X_1(n)$ is a cause of $X_2(n)$, $X_3(n)$ and $X_4(n)$, and $X_4(n)$ and $X_5(n)$ share a feedback loop with each other, as depicted in **Fig. 1 II. Fig. 1 I** shows an example of the time trace of 5 time series. For the Granger causality approach, we simulated the fitted vector autoregressive (VAR) model to generate a data set of 100 realizations of 1000 time points, and applied the bootstrap approach to construct the 95% confidence intervals (**Fig. 1 III**). For Granger causality, we assume the causality value is Gaussian distributed. Then the confidence intervals can be obtained by calculating the mean and standard derivation values [21][22]. According to the confidence intervals, one can derive the network structure as shown in **Fig.1 II,** which correctly recovers the pattern of the connectivity in our toy model. For the dynamic Bayesian network inference approach, we can infer a network structure (**Fig. 1 IVa**) for each realization of 1000 time points. The final resulting causal network model was inferred with high-confidence causal arcs (the arcs occur more than 95% of the time in the whole population) between various variables [13]. This complex network contains the information of different time-lags for each variable. It fits exactly the pattern of connectivity in our VAR model. In order to compare it with the Granger causality approach, we can further simplify the network by hiding the information of time-lags, and then we infer the exactly same structure as the Granger causality approach (**Fig. 1 IVd**). From this simple example, we can find that both approaches can reveal correct network structures for the data with a large sample size (1000 here).

Most, if not all, experimental data has a very limited time step due to various experimental restrictions. Hence one of the key quantities to test the reliability of an approach is the data length (sample size). In the next setup, we reduce the sample size to a smaller value and check its impact. **Fig. 2I** shows the case of the sample size of 80: we find both approaches start failing to detect some interactions (false negative). By reducing the sample size to 20, we can see that the Bayesian network inference can derive more true positive connections than the Granger causality. This is certainly an interesting phenomenon and we intend to explore whether it is true for a more general case.

Example 2 we considered a more general toy model; the coefficients in the equations (1) of Example 1 are randomly generated. This toy model aims to test the causality sensitivity for the two approaches. Suppose 5 simultaneously generated time series according to the equations:

$$\begin{cases} X_1(n) = w_1 X_1(n-1) + w_2 X_1(n-2) + \varepsilon_1 \\ X_2(n) = w_3 X_1(n-2) + \varepsilon_2 \\ X_3(n) = w_4 X_1(n-3) + \varepsilon_3 \\ X_4(n) = w_5 X_1(n-1) + w_6 X_4(n-1) + w_7 X_5(n-1) \\ X_5(n) = w_8 X_4(n-1) + w_9 X_5(n-1) \end{cases} \tag{2}$$

where $w_1, w_2, \cdots, w_9$ are uniformly distributed random variables in the interval [-1,1]. The randomly generated coefficients are also required to make the system stable. The stability can be tested by using the z-plane pole-zero method, which states if the outermost poles of the z-transfer function describing the time series are inside the unit circle on the z-plane pole-zero plot, then the system is stable.

The above toy model is then used to test the two different causality approaches: Bayesian network inference and Granger causality. They are applied with different sample sizes. For each sample size, we randomly generated 100 different coefficient vectors [$w_1, w_2, \cdots, w_9$], which corresponds to100 different toy models in Example 1. For each different coefficient vectors model, we applied the same approach as in Example 1, using Monte Carlos method to construct 95% confidence interval for the Granger causality approach and chose high-confidence arcs (appearing in at least 95% of all samplings) for the Bayesian network inference approach. The total number of arcs (or causalities) is 500 (5 interactions for each realization) for each sample size. However we cannot expect to detect the maximum number of arcs in our system, since the coefficients are randomly generated, which could be significantly small.

Fig. 3 Ia shows the comparison result of the percentage of true positive connections derived from these two methods. In general, the Granger causality approach can infer slightly more true positive causalities compared to the Bayesian network inference approach when the data length is long. It is interesting to see that there is a critical point at around 30 in Fig. 3 Ia. If the sample size is larger than 30, then the Bayesian network recovers less positive connections. However, if the sample size is smaller than 30, the Bayesian network performs better. From Fig. 3 Ib, we see that computing time for the Bayesian network inference is much larger than the Granger causality.

Now we are in the position to find out why the dynamic Bayesian network is better than the Granger causality when the data length is short, and vise verse. In Fig. 3 II, we compare the performances on different coefficients (strength of interaction) for a fixed sample size of 900 (super-critical case). The x-axis shows the absolute value of coefficients, and y shows the corresponding causality (1 indicates positive causality and 0 indicates no causality). For visualization purposes, the figure for the Granger causality is shifted upward.

From the five graphs, we can see that there is no difference between these two approaches if the coefficients are significant large (strong interactions with an absolute value of coefficients being greater than 0.15): both approaches can infer the correct connections. For most cases, the Granger causality approach performs with more stability when the coefficients are larger than 0.15, and the Bayesian network inference approach shows slightly more oscillations around this point. Hence we conclude that the Granger causality is less sensitive to the small value of the connection when the data length is large (see also the nonlinear case below).

We now compare the fitting accuracy of the two approaches, as shown in **Fig. 3 III**. We use the average mean-square error as a measurement of the fitting. Not surprisingly, the dynamic Bayesian network approach considerably outperforms the simple fitting algorithm in the Granger approach ([15,16]), in particular when the data length is short.

In conclusion, when the data is a reasonable fit to the original model, the Granger causality works better. This is due to the fact that the Granger causality approach is more sensitive to a small value of the interactions. When the data length is short, the Bayesian approach can fit the data much more reliably and it outperforms the Granger approach.

2.2 Synthesized Data: Non-linear Case

In real situations, all data should be nonlinear and a linear relationship as described above is only an approximation. To address the nonlinear issue, we turn our attention to kernel models. As we all know, any nonlinear relationship can be approximated by a series of kernel functions.

Example 3 we modify the model in example 1 to a series of nonlinear equations as follows:

$$\begin{cases} X_1(n) = 0.125\sqrt{2}\exp(-\frac{X_1(n-1)^2}{2}) + \varepsilon_1 \\ X_2(n) = 1.2\exp(-\frac{X_1(n-1)^2}{2}) + \varepsilon_2 \\ X_3(n) = -1.05\exp(-\frac{X_1(n-1)^2}{2}) + \varepsilon_3 \\ X_4(n) = -1.15\exp(-\frac{X_1(n-1)^2}{2}) \\ \qquad +0.2\sqrt{2}\exp(-\frac{X_4(n-1)^2}{2}) + 1.35\exp(-\frac{X_5(n-1)^2}{2}) + \varepsilon_4 \\ X_5(n) = -0.5\sqrt{2}\exp(-\frac{X_4(n-1)^2}{2}) + 0.25\sqrt{2}\exp(-\frac{X_5(n-1)^2}{2}) + \varepsilon_5 \end{cases} \tag{3}$$

In this example, the center and variance of each time series is chosen as the center and variance in the kernel function. We use the fuzzy c-mean method to find the center of each time series and then applied the same approach as in **Example 1**. For the Granger causality approach, we simulated the fitted VAR model to generate a data set of 100 realizations of 1000 time points, and applied the bootstrap approach to construct the 95% confidence intervals (Fig. 4 IIIa). According to the confidence interval, one can derive the network structure (Fig. 4 IIIb) which correctly recovers the pattern of connectivity in our non-linear model. For the Bayesian network inference approach, we can infer a network structure (Fig. 4 IVa) for each realization of 1000 time points. We can then obtain a simplified network structure (Fig. 4 IV).

For a small sample size (see Fig. 5), worse results are obtained for both approaches comparing to the previous linear model. Both approaches start to miss interactions when the sample size is smaller than 300. When the sample size is 150, the Bayesian network inference approach can detect one more true positive interaction than the Granger causality. However, when the sample size is 50, both approaches fail to detect all the interactions.

In the next step, we extend our non-linear model to a more general setting in which the coefficients in the equations are randomly generated. Fig. 6 Ia shows the comparison result of the percentage of true positive connections derived from these two methods. It is very interesting to see that a critical point around 500 exists in the non-linear model, similar to the linear model before. From Fig. 6 Ib, the computing time required for the Bayesian network inference is still much larger than the Granger causality. In Fig. 6 II, we compare the performances on different coefficients (strength of interaction) for a fixed sample size of 900. From the five graphs, we can see that in general the Granger approach is more sensitive to a small value of the coefficients (see **Fig. 3**. X_5 -> X_4 and X_4 -> X_5).

Therefore, all conclusions in the linear case are confirmed in the nonlinear model. In the literature [23], the result they obtained shows the same direction as we did here, which finds that the Granger causality performs better than the dynamic Bayesian network inference concerning a nonlinear kernel model of genetic regulatory pathways and for a sufficiently large sample size (2000 data points).

2. 3. Experimental Data

Finally we carry out a study on experimental data of microarray experiments. The gene data were collected from two cases of Arabidopsis Leaf: the mock (normal) case and the infected case with the plant pathogen *Botrytis cinerea*. A total of 31,000 genes were measured with a time interval of two hours, with a total of 24 sampling points (two days) and four replicates. We test the Granger causality approach and dynamic Bayesian network inference approach on a well-known circadian circuit. This circuit contains 7 genes: PRR7, GI, PRR9, ELF4, LHY, CCA1 and TOC1. **Fig. 7 I** shows the time traces of the 7 genes. From the time traces figure, it is clearly to see that they exhibit a 24 hour rhythm. Note that the total number of time points is only 24. Compared to our previous toy model case, this sample size is quite small. We therefore expect the Bayesian network inference to be more reliable.

We first apply the dynamic Bayesian network inference approach on these two data sets. The two network structures for two cases are shown in **Fig. 7 II**. In the next step, the conditional Granger causality approach is applied. By using the bootstrapping method, we construct 95% confidence intervals as shown in **Fig. 7 IIIc**. Finally, we can also obtain two network structures for two different cases shown in **Fig. 7 IIIa** and **Fig. 7 IIIb**. It is clearly seen that the globe patterns for the mock case and the infected case are different.

From the literature, there are three well known connections among the whole structure for the mock case. (1) It is known that GI alone is independent of the remaining six genes in the circuit. There should be no connection to and from the GI node (node 2 in the **Fig. 7**) in our derived network. From **Fig. 7 IIa** and **Fig. 7 IIIa**, we find that the dynamic Bayesian network inference method clearly picks this up, but the conditional Granger causality approach fails to detect this property. The Granger causality approach derived two false positive arcs which were connected to a GI node as shown in **Fig. 7 IIIa**. (2) It is known that PRR7 and LHY share a feedback loop. In other words, there should be two directed arcs connected from node 1 (PRR7) to node 5 (LHY) and from node 5 to node 1. The network structures derived from both approaches are in agreement with this known relationship. (3) It is known that ELF4 has some interactions with both LHY and CCA1. There should be some connections between node 4 (ELF4) to node 5 (LHY), and between node 4 (ELF4) and node 6 (CCA1). From our derived structures, both approaches can detect these connections, which are in agreement with the known structure in the literature [24,25].

According these three known relationships in the structure, we can find that the Bayesian network structure is in agreement with all three rules, but the network structure derived from the conditional Granger causality is not: two more false positive interactions are found. Again for a small sample size, the Bayesian network inference approach could be more reliable than the conditional Granger causality approach.

3. Discussion

3.1 A fair comparison

In our results presented here, one of the key issues which is the cause of the critical point of the sampling size between the dynamic Bayesian approach and the Granger causality lies in the fact that a batch fitting approach is used in the Granger causality approach. One might argue that we could use the sequential fitting approach as in the Bayesian network to improve the performance of the Granger causality approach. This is certainly the case. However, due to the thousand publications in both topics [26], we simply adopted the most common approaches in the dynamic Bayesian network approach and the Granger causality. Developing one of the approaches, for example the Granger causality, so that it could always outperform the other is an interesting future research topic.

3.2 How long is long enough?

Although we have found the critical point of the two approaches, in practical applications, we have no certain idea where the critical point is. Hence, we still have to choose one of them to tackle the data. In molecular biology, we have to deal with a very limited data size; but in physiology, for example, neurophysiology, the data we record is usually very long. Hence one could argue that we use the dynamic Bayesian network in gene, protein or metabolite data, and apply the Granger causality to physiology data. The dynamic Bayesian network is more often reported in molecular biology, but the Granger causality has been very successfully applied in neurophysiological data [27] and fMRI. The result we chose to use was always chosen through experimental validation, as we did here for the plant data.

3.3 Frequency decomposition

As we emphasized at the beginning, the advantage of the Granger causality over the dynamic Bayesian network is the frequency decomposition, which is usually informative when we deal with temporal data. For example, in neurophysiology data, we know the brain employs different frequency bands to communicate between neurons and brain areas [28,29]. We would expect a similar situation to arise in genes, proteins and metabolites, although we lack a detailed analysis due to the limited data length. To this end, we have also presented frequency decomposition results in Appendix 1 for the dynamic Bayesian network.

3.4 False positive

In our synthesized data, for both approaches, we did not find any false positive links in our experiments. However, there were a few false positive links found when we applied the conditional Granger causality and also partial Granger causality (data not shown, [21]) on the gene data. One might ask why this is the case; there are several different reasons. Firstly, the experimental data is not strictly stationary: it is a natural process and evolves with time. As a first approximation, we treat it as stationary. Of course, we could use ARIMA rather than ARMA model to fit the data in the Granger causality. Secondly, the seven gene network is only a small network embedded in complete and large network, so there are latent variables. Using the partial Granger causality [21] which was originally developed for eliminating latent variables, GI still has links with the other six genes. Whether the dynamic Bayesian network could do a better job in the presence of latent variables is another research topic.

3.5 The meaning of the found motifs

Two circuits are found: one with the mock plant and one with the infected plant. The plant rewires its circadian circuit after infection. Ignoring the issue of identifying the molecular mechanisms which control circuit rewiring, which is itself an interesting and challenging problem, we intend to discuss the functional meaning of the two circuits. To this end, we could assign a dynamics to the network and try to decipher the implications of the rewiring. Interestingly, we found that GI is recruited to save the network: if we leave GI as it is in the mock case, the whole network will rapidly converge to a fixed point state (a dead state). We will publish the results elsewhere.

3.6 Reasons for short size data?

In our synthesized data, we test both short and long data samples and come to the conclusion that there is a critical size, at which the two approaches behave differently. In our experimental data, we only tested it for the short data set. Of course, as we mentioned above, in neurophysiological data, we have recordings of long time traces and the Granger causality is widely used there. However, we have to realize that all *in vivo* recordings are very dynamic and stationary of data will become a key issue once we apply both approaches to a long dataset. Furthermore, when the dataset is long, both approaches could do well and it is more difficult to find the difference between the two. Hence we have only compared the results for short data length in the experimental setup.

3.7 Reasons for small size of variables

In our synthesized data, we only used 5 variables to simulate a small interacting network; the number of variables could affect the result we derived. As expected, see also [23], the estimation of the Granger causality becomes unfeasible when the number of variables is large and the amount of the data sets is small. Hence, all results in the literature on estimating Granger causality are exclusive for small networks (around the order of 10), as we considered here. This is more or less true for dynamic Bayesian network inference as well. Extending the Granger causality and the dynamic Bayesian network inference to large networks is a challenging problem, even before we carry out the same comparison study on these two approaches as we did here.

4 Methods

4.1 Granger causality

Causal influence measurement notation for time series was firstly proposed by Wiener-Granger. We can determine a causal influence of one time series on another, if the predication of one time series can be improved by incorporating the knowledge of the second one. Granger applied this notation by using the context of linear vector auto-regression (VAR) model of stochastic processes [30,31,32,33]. In the AR model, the variance of the prediction error is used to test the perdition improvement. For instance, assume two time series; if the variance of the autoregressive prediction error of the first time series at the present time is reduced by inclusion of past measurements from the second time series, then one can conclude that the second time series have a causal influence on the first one. Geweke [15,16] decomposed the VAR process into the frequency domain, it converted the causality measurement into a spectral representation and made the interpretation more appealing.

The pairwise analysis introduced above can only be applied to bivairate time series. For more than two time series, a time series can have a direct or indirect causal influence to other time series. In this case, pairwise analysis is not sufficient or misleading for revealing whether the causal interaction between a pair is direct or indirect. In order to distinguish the direct and indirect causal affect, one introduces the conditional causality which takes account of the other time series' effect in a multivariate time series. In this paper, we used conditional causality to compare with the Bayesian network inference introduced before.

4.1.1 Linear conditional Granger causality

The conditional Granger causality was defined by Granger. It can be explained as following. Giving Two time series $\mathbf{X}_t$ and $\mathbf{Z}_t$, the joint autoregressive representation for $\mathbf{X}_t$ and $\mathbf{Z}_t$ by using the knowledge of their past measurement can be expressed as

$$\begin{cases} \mathbf{X}_t = \sum_{i=1}^{\infty} a_{1i}\mathbf{X}_{t-i} + \sum_{i=1}^{\infty} c_{1i}\mathbf{Z}_{t-i} + \boldsymbol{\varepsilon}_{1t} \\ \mathbf{Z}_t = \sum_{i=1}^{\infty} b_{1i}\mathbf{Z}_{t-i} + \sum_{i=1}^{\infty} d_{1i}\mathbf{X}_{t-i} + \boldsymbol{\varepsilon}_{2t} \end{cases} \tag{4}$$

and the noise covariance matrix for the system can be represented as

$$\mathbf{S} = \begin{bmatrix} \mathrm{var}(\boldsymbol{\varepsilon}_{1t}) & \mathrm{cov}(\boldsymbol{\varepsilon}_{1t}, \boldsymbol{\varepsilon}_{2t}) \\ \mathrm{cov}(\boldsymbol{\varepsilon}_{2t}, \varepsilon_{1t}) & \mathrm{var}(\boldsymbol{\varepsilon}_{2t}) \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix} \tag{5}$$

where var and cov represent variance and co-variance respectively. Incorporating the knowledge of third time series, the vector autoregressive mode can be represented involving three time series $\mathbf{X}_t$, $\mathbf{Y}_t$ and $\mathbf{Z}_t$ can be represented as

$$\begin{cases} \mathbf{X}_t = \sum_{i=1}^{\infty} a_{2i}\mathbf{X}_{t-i} + \sum_{i=1}^{\infty} b_{2i}\mathbf{Y}_{t-i} + \sum_{i=1}^{\infty} c_{2i}\mathbf{Z}_{t-i} + \boldsymbol{\varepsilon}_{3t} \\ \mathbf{Y}_t = \sum_{i=1}^{\infty} d_{2i}\mathbf{X}_{t-i} + \sum_{i=1}^{\infty} e_{2i}\mathbf{Y}_{t-i} + \sum_{i=1}^{\infty} f_{2i}\mathbf{Z}_{t-i} + \boldsymbol{\varepsilon}_{4t} \\ \mathbf{Z}_t = \sum_{i=1}^{\infty} g_{2i}\mathbf{X}_{t-i} + \sum_{i=1}^{\infty} h_{2i}\mathbf{Y}_{t-i} + \sum_{i=1}^{\infty} k_{2i}\mathbf{Z}_{t-i} + \boldsymbol{\varepsilon}_{5t} \end{cases} \tag{6}$$

And the noise covariance matrix for the above system is

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathrm{var}(\boldsymbol{\varepsilon}_{3t}) & \mathrm{cov}(\boldsymbol{\varepsilon}_{3t}, \boldsymbol{\varepsilon}_{4t}) & \mathrm{cov}(\boldsymbol{\varepsilon}_{3t}, \boldsymbol{\varepsilon}_{5t}) \\ \mathrm{var}(\boldsymbol{\varepsilon}_{4t}, \boldsymbol{\varepsilon}_{3t}) & \mathrm{var}(\boldsymbol{\varepsilon}_{4t}) & \mathrm{cov}(\boldsymbol{\varepsilon}_{4t}, \boldsymbol{\varepsilon}_{5t}) \\ \mathrm{var}(\boldsymbol{\varepsilon}_{5t}, \boldsymbol{\varepsilon}_{3t}) & \mathrm{var}(\boldsymbol{\varepsilon}_{5t}, \boldsymbol{\varepsilon}_{4t}) & \mathrm{var}(\boldsymbol{\varepsilon}_{5t}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}} & \boldsymbol{\Sigma}_{\mathbf{xz}} \\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}} & \boldsymbol{\Sigma}_{\mathbf{yz}} \\ \boldsymbol{\Sigma}_{\mathbf{zx}} & \boldsymbol{\Sigma}_{\mathbf{zy}} & \boldsymbol{\Sigma}_{\mathbf{zz}} \end{bmatrix} \tag{7}$$

where $\boldsymbol{\varepsilon}_{it}, i = 1, 2, \cdots, 5$ are the prediction error, which are uncorrelated over time. From above two sets of equations, the conditional Granger causality form $\mathbf{Y}$ to $\mathbf{X}$ conditional on $\mathbf{Z}$ can be defined as

$$F_{\mathbf{Y}\to\mathbf{X}|\mathbf{Z}} = \ln(\frac{|\mathrm{var}(\boldsymbol{\varepsilon}_{1t})|}{|\mathrm{var}(\boldsymbol{\varepsilon}_{3t})|}) \tag{8}$$

When the causal influence from $\mathbf{Y}$ to $\mathbf{X}$ is entirely mediated by $\mathbf{Z}$, the coefficient b_{2i} is uniformly zero, and the two autoregression models for two time series and three time series will be exactly same, thus we can get $\mathrm{var}(\boldsymbol{\varepsilon}_{1t}) = \mathrm{var}(\boldsymbol{\varepsilon}_{3t})$. We then can deduce $F_{\mathbf{Y}\to\mathbf{X}|\mathbf{Z}} = 0$, which means $\mathbf{Y}$ can not futher improve the prediction of $\mathbf{X}$ including past measurements of $\mathbf{Y}$ conditional on $\mathbf{Z}$. For $\mathrm{var}(\boldsymbol{\varepsilon}_{1t}) > \mathrm{var}(\boldsymbol{\varepsilon}_{3t})$ and $F_{\mathbf{Y}\to\mathbf{X}|\mathbf{Z}} > 0$, we can say that there is still a direct influence from $\mathbf{Y}$ to $\mathbf{X}$ conditional on the past measurements of $\mathbf{Z}$.

4.1.2 non-linear conditional Granger causality

We can extend our Granger causality to a non-linear model by using a series kernel functions [22,34]. Let $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ be three time series of n simultaneously measured quantities, which are assumed to be stationary. We are supposed to quantify how much $\mathbf{Y}$ cause $\mathbf{X}$ conditional on $\mathbf{Z}$. The general expression for the nonlinear model is:

$$\begin{cases} \mathbf{X}_t = \sum_j \mathbf{w}_{1j}\Phi_j(\mathbf{X}_{t-j}) + \sum_j \mathbf{w}_{2j}\Phi_j(\mathbf{Z}_{t-j}) + \boldsymbol{\varepsilon}_6 \\ \mathbf{Z}_t = \sum_j \mathbf{w}_{3j}\Phi_j(\mathbf{X}_{t-j}) + \sum_j \mathbf{w}_{4j}\Phi_j(\mathbf{Z}_{t-j}) + \boldsymbol{\varepsilon}_7 \end{cases} \tag{9}$$

Function Φ can be selected as the kernel function of $\mathbf{X}$ and $\mathbf{Z}$ which has the following expression:

$$\Phi_j(\mathbf{X}) = \exp(-\left\|\mathbf{X} - \bar{X}_j\right\|^2 / 2\sigma_{\mathbf{X}}^2) \tag{10}$$

$$\Phi_j(\mathbf{Z}) = \exp(-\left\|\mathbf{Z} - \bar{Z}_j\right\|^2 / 2\sigma_{\mathbf{Z}}^2) \tag{11}$$

where $\bar{X}$, $\bar{Z}$ are centers of $\mathbf{X}$ and $\mathbf{Z}$, $\sigma_{\mathbf{X}}^2$, $\sigma_{\mathbf{Z}}^2$ are variances of $\mathbf{X}$ and $\mathbf{Z}$. The covariance matrix of prediction error can be expressed as

$$\mathbf{S} = \begin{bmatrix} \mathrm{var}(\boldsymbol{\varepsilon}_6) & \mathrm{cov}(\boldsymbol{\varepsilon}_6, \boldsymbol{\varepsilon}_7) \\ \mathrm{cov}(\boldsymbol{\varepsilon}_7, \boldsymbol{\varepsilon}_6) & \mathrm{var}(\boldsymbol{\varepsilon}_7) \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{11}^2 & \mathbf{S}_{12}^2 \\ \mathbf{S}_{21}^2 & \mathbf{S}_{22}^2 \end{bmatrix} \tag{12}$$

A joint autoregressive representation has the following expression:

$$\begin{cases} \mathbf{X}_t = \sum_j w_{5j}\Phi_j(\mathbf{X}_{t-j}) + \sum_j w_{6j}\Phi_j(\mathbf{Y}_{t-j}) + \sum_j w_{7j}\Phi_j(\mathbf{Z}_{t-j}) + \boldsymbol{\varepsilon}_8 \\ \mathbf{Y}_t = \sum_j w_{8j}\Phi_j(\mathbf{X}_{t-j}) + \sum_j w_{9j}\Phi_j(\mathbf{Y}_{t-j}) + \sum_j w_{10j}\Phi_j(\mathbf{Z}_{t-j}) + \boldsymbol{\varepsilon}_9 \\ \mathbf{Z}_t = \sum_j w_{11j}\Phi_j(\mathbf{X}_{t-j}) + \sum_j w_{12j}\Phi_j(\mathbf{Y}_{t-j}) + \sum_j w_{13j}\Phi_j(\mathbf{Z}_{t-j}) + \boldsymbol{\varepsilon}_{10} \end{cases} \tag{13}$$

The covariance matrix of prediction error can be expressed as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathrm{var}(\boldsymbol{\varepsilon}_8) & \mathrm{cov}(\boldsymbol{\varepsilon}_8, \boldsymbol{\varepsilon}_9) & \mathrm{cov}(\boldsymbol{\varepsilon}_8, \boldsymbol{\varepsilon}_{10}) \\ \mathrm{cov}(\boldsymbol{\varepsilon}_9, \boldsymbol{\varepsilon}_8) & \mathrm{var}(\boldsymbol{\varepsilon}_9) & \mathrm{cov}(\boldsymbol{\varepsilon}_9, \boldsymbol{\varepsilon}_{10}) \\ \mathrm{cov}(\boldsymbol{\varepsilon}_{10}, \boldsymbol{\varepsilon}_8) & \mathrm{cov}(\boldsymbol{\varepsilon}_{10}, \boldsymbol{\varepsilon}_8) & \mathrm{var}(\boldsymbol{\varepsilon}_{10}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{xx}}^2 & \boldsymbol{\Sigma}_{\mathbf{xy}}^2 & \boldsymbol{\Sigma}_{\mathbf{xz}}^2 \\ \boldsymbol{\Sigma}_{\mathbf{yx}}^2 & \boldsymbol{\Sigma}_{\mathbf{yy}}^2 & \boldsymbol{\Sigma}_{\mathbf{yz}}^2 \\ \boldsymbol{\Sigma}_{\mathbf{zx}}^2 & \boldsymbol{\Sigma}_{\mathbf{zy}}^2 & \boldsymbol{\Sigma}_{\mathbf{zz}}^2 \end{bmatrix} \quad (14)$$

Similarly, we can define the conditional causality as

$$F_{\mathbf{Y}\to\mathbf{X}|\mathbf{Z}} = \ln\left(\frac{|\mathrm{var}(\boldsymbol{\varepsilon}_{6t})|}{|\mathrm{var}(\boldsymbol{\varepsilon}_{8t})|}\right) \quad (15)$$

4.2 Bayesian network

Bayesian networks are probabilistic graphical models initially introduced by [Kim & Pearl, 1987]. A Bayesian network is the specific type of graphical model which is directed acyclic graph [35,36]. Each arc in the model is directed and there is no way to start from any nodes and travel along a set of directed edges and get back at the initial node. The set of nodes represent a set of random variables $[\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n]$, and the arcs express statistical dependence between the downstream variables and the upstream variables. The upstream variables are also called the parent variables of the downstream variables. Bayesian network inference yields the most concise model, automatically excluding arcs based on dependencies already explained by the model, which means the arcs in the network can be interpreted as a conditional causality. The edges in the Bayesian network encode a particular factorization of the joint distribution. The joint probability distribution can be decomposed as following:

$$\mathrm{P}(\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n) = \prod_{i=1}^{n} \mathrm{P}(\mathbf{X}_i \mid \mathit{parents}(\mathbf{X}_i)) \tag{16}$$

That is, the joint probability distribution is the product of the local distributions of each node and its parents. If node $\mathbf{X}_i$ has no parents, its local probability distribution is said to be unconditional, otherwise it is conditional. This decomposition is useful for Bayesian networks inference algorithm to deal with the uncertain situation and incomplete data.

To learn the parameter of the Bayesian network is to essentially estimate two kinds of probability distributions: the probability $\mathrm{P}(\mathbf{X})$ and the conditional probability $\mathrm{P}(\mathbf{X} \mid \mathbf{Y})$. There are two kinds of approaches to density estimation; the nonparametric method and the parametric method. The easiest estimation for nonparametric method is to use the histogram approach. The distribution can then be a tabular conditional probability distribution, which is represented as a table. However this approach requires a much larger sample size to give us an accurate estimation, which is not suitable for general experimental data. For parametric method, one needs to make some assumptions about the form of the distribution such as widely used Gaussian distribution. For a D-dimensional vector $\mathbf{X}$, the multivariate Gaussian distribution is in the form

$$\mathrm{N}(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\mathbf{X} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right\} \tag{17}$$

Where $\boldsymbol{\mu}$ is a D-dimensional mean vector, $\boldsymbol{\Sigma}$ is a $\mathbf{D} \times \mathbf{D}$ covariance matrix, and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$. In this paper, we first consider every node's conditional probability distribution as a conditional Gaussian distribution for the following inferences. The distribution on a node $\mathbf{X}$ can be defined as follows:

-no parents: $\mathrm{P}(\mathbf{X}) \sim \mathrm{N}(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\sigma})$ (18)

-continuous parents $\mathbf{Y}$: $\mathrm{P}(\mathbf{X} \mid \mathbf{Y} = \mathbf{y}) \sim \mathrm{N}(\mathbf{X} \mid \boldsymbol{\mu} + \mathbf{W}^{\mathrm{T}}\mathbf{y}, \boldsymbol{\sigma})$ (19)

Where the T is the matrix transposition. $\mathbf{W}$ is the connection weight vector between node $\mathbf{X}$ and its parents $\mathbf{Y}$. It can be represented by using the covariance matrix as following:

$$\mathbf{W} = \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}} \tag{20}$$

The detailed inductions of parameter estimations are given in the next chapter.

For learning the structure of the Bayesian network, one needs to search the space of all the possible structures and find out the best one which can be used to describe the input data, which is to maximum the conditional probability $\mathrm{P}(Data \mid \theta, M)$ of data ($Data$) by give the parameters (θ) and the network structure (M). In order to balance the complex and concise of the structure, we can use BIC (Bayesian Information Criterion) as a scoring function, which includes an extra penalty term.

4.2.1 Parameter learning for linear Gaussian

The parameter learning part can be approached by fitting a linear-Gaussian model. The goal of learning parameter in Bayesian network is to estimate the mean and covariance of the conditional Gaussian distribution, thus we can deduce the parameters of $\boldsymbol{\mu}, \boldsymbol{\sigma}$ and $\mathbf{W}$ in the equation (19).

Suppose $\mathbf{X}$ is a D-dimensional vector with Gaussian distribution $\mathrm{N}(\mathbf{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$, and one partition $\mathbf{X}$ into two disjoint subsets $\mathbf{X}_a$ and $\mathbf{X}_b$, To be easily illustration, one takes the first M components of $\mathbf{X}$ to form $\mathbf{X}_a$, and the remaining components to form $\mathbf{X}_b$, so that

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_a \\ \mathbf{X}_b \end{pmatrix} \tag{21}$$

We also define mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ given by

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \tag{22}$$

Considering the quadratic form in the exponent of the Gaussian distribution, we can get following equation by a transformation.

$$\begin{aligned} &-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X} + \mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const} \\ &= -\frac{1}{2}(\mathbf{X}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Sigma}_{aa}{}^{-1}(\mathbf{X}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{X}_a-\boldsymbol{\mu}_a)^{\mathrm{T}}\boldsymbol{\Sigma}_{ab}{}^{-1}(\mathbf{X}_b-\boldsymbol{\mu}_b) \\ &-\frac{1}{2}(\mathbf{X}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Sigma}_{ba}{}^{-1}(\mathbf{X}_a-\boldsymbol{\mu}_a) - \frac{1}{2}(\mathbf{X}_b-\boldsymbol{\mu}_b)^{\mathrm{T}}\boldsymbol{\Sigma}_{bb}{}^{-1}(\mathbf{X}_b-\boldsymbol{\mu}_b) \end{aligned} \tag{23}$$

From these and regard $\mathbf{X}_b$ as a constant, we obtain the following expressions for the mean and covariance of the conditional probability distribution $\mathrm{P}(\mathbf{X}_a \mid \mathbf{X}_b)$.

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}{}^{-1}(\mathbf{X}_b - \boldsymbol{\mu}_b) \tag{24}$$

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}{}^{-1}\boldsymbol{\Sigma}_{ba} \tag{25}$$

Thus the parameters in the Bayesian network can be learned from above two equations.

4.2.2 Parameter learning for non-linear Gaussian

We can also extend our linear model to a non-linear model like that for the Granger causality case. Suppose we have two variables which can be expressed as in equation (9). The kernel function is also chosen as described in equation (10) and equation (11).

Our non-linear model, the probability distribution for $\mathbf{X}_t$ is no longer a Gaussian distribution. From the expression in equation (9), we can find that the probability distribution for $\mathbf{X}_t$ is a combined distribution of kernel function distribution for the past measured values of $\mathbf{X}$ and $\mathbf{Z}$, and a Gaussian distribution for the noise term. The kernel distribution is very difficult to derive, so one can use a mixture of Gaussian models to approximate the real distribution of kernel function. The mixture Gaussian model is in the form:

$$\mathrm{P}(\mathbf{X}) = \sum_{k=1}^{K} \pi_k \, \mathrm{N}(\mathbf{X} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{26}$$

Each Gaussian density $\mathrm{N}(\mathbf{X} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is called a component of the mixture and has its own mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$. The parameter π_k are called mixing coefficients which satisfies:

$$\sum_{k=1}^{K} \pi_k = 1 \tag{27}$$

The conditional probability distribution for $\mathbf{X}_t$ conditional on the past observation of $\mathbf{X}$ and $\mathbf{Y}$ in the nonlinear model is still a Gaussian distribution which can be easily obtained as following:

$$\mathrm{P}(\mathbf{X} \mid \mathbf{Y} = \mathbf{y}) = \mathrm{N}(\mathbf{X} \mid \boldsymbol{\mu} + \sum_i w_i \Phi(y_i), \sigma) \tag{28}$$

where w is the connection weights between node $\mathbf{X}$ and it parents. It can be estimated by using the simple linear regression method.

4.2.3 Structure learning

There are two very different approaches to structure learning: one is constraint-based and the other is search and score algorithm. For the constraint-based algorithm, we start with a fully connected network and then remove the arcs, which are conditional independent. This has the disadvantage that repeated independence tests lose statistical power. For the latter algorithm, we perform a search on all possible graphs and select one graph which best describes the statistical dependence relationship in the observed data.

Unfortunately, the number of possible graphs increases super-exponentially with the number of nodes, so some search algorithms are required for overcoming this kind of complex problem rather than doing a exhaustive search in the space. There are several searching algorithms that can be applied; such as annealing search, genetic algorithm search and so on. The question could become easier if we know the total order of the nodes. The K2 algorithm allows us to find the best structure by selecting the best set of parents for each node independently. In the dynamic Bayesian networks, the order of nodes can be interpreted as the sequence of time lags represented for each node, so the K2 algorithm is applied for Bayesian network structure learning in this paper (see appendix 2 for more details descriptions). The K2 algorithm tests parent insertion according to the order. The first node cannot have any parent, for other nodes, we can only choose the parent nodes which are behind it in this order. Then the scoring function can be applied to determine the best parent set, i.e. the one which gives the best score.

In addition to the search algorithm, a scoring function must be defined in order to decide which structure is the best (a high scoring network). There are two popular choices. One is the Bayesian score metric which is the marginal likelihood of the model, and the other is BIC (Bayesian Information Criterion) defined as following:

$$LogP(Data \mid \theta) - \frac{d}{2} Log(N) \tag{29}$$

Where $Data$ is the observed data, θ is the estimated value of the parameters, d is the number of parameters and N is the number of data cases. The term of $\frac{d}{2} Log(N)$ is regarded as a penalty term in order to balance both simple and accurate structure representation.

Suppose we observed a set of independent and identically distributed data $Data = \{\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^N\}$, each of which can be a case of multi-dimensional data. Then the log likelihood of the data set can be defined as

$$\begin{aligned} LogP(Data \mid \theta) &= \sum_{i=1}^{N} \log P(\mathbf{Y}^i \mid \theta) = \sum_{i=1}^{N} \log \prod_j P(\mathbf{Y}_j^i \mid \mathbf{Y}_{pa(j)}^i, \theta_j) \\ &= \sum_{i=1}^{N} \sum_j \log P(\mathbf{Y}_j^i \mid \mathbf{Y}_{pa(j)}^i, \theta_j) \end{aligned} \tag{30}$$

Where j is the index of the nodes or variables in the Bayesian network, $pa(j)$ is the set of parents of node j, and θ_j are the parameters that define the conditional probability of Y_j giving its parents.

In this paper, the Bayesian networks inference can then be approached by following procedure: initially, K2 algorithm is applied to search the space of possible graphs. For each possible structure, we can use the parameter learning algorithm to estimate the parameters of the networks. The BIC scoring function assigns a corresponding score through the estimated parameters and observed data set. The best network we can get is the highest score structures among all the possible graphs [37].

Reference

1. Klipp E, Herwig R, Kowald A, Wierling C, Lehrach H: *Systems Biology in Practice: Concepts, Implementation and Application* Weinheim: Wiley-VCH Press; 2005.

2. Feng J, Jost J, Qian M: *Networks*: *From Biology to Theory* London: Springer Press; 2007.

3. Alon U: **Network motifs: theory and experimental approaches**, *Nature* 2007*,* **8**:450.

4. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Ménard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C: **Global Mapping of the Yeast Genetic Interaction Network**, *Science* 2004, **303**:808.

5. Tsai TY, Choi YS, Ma W, Pomerening JR, Tang C, Ferrell JE Jr: **Robust, Tunable Biological Oscillations from Interlinked Positive and Negative Feedback Loops**, *Science* 2008, **321**:126.

6. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings

EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA: Transcriptional Regulatory Networks in Saccharomyces cerevisiae, *Science* 2002, **298**:799.

7. Pearl J: *Causality: Models, Reasoning, and Inference* Cambridge: Cambridge Univ. Press; 2000.

8. Albo Z, Di Prisco GV, Chen Y, Rangarajan G, Truccolo W, Feng J, Vertes RP, Ding M: Is partial coherence a viable technique for identifying generators of neural oscillations, *Biological Cybernetics* 2004*,* **90**:318.

9. Horton PM, Bonny L, Nicol AU, Kendrick KM, Feng JF: Applications of multi-variate analysis of variances (MANOVA) to multi-electrode array data, *Journal of Neuroscience Methods* 2005, **146**:22.

10. Guo S, Wu J, Ding M, Feng J: Uncovering interactions in the frequence domain, *PLoS Computational Biology* 2008, **4** (5):e1000087.

11. Wu J, Liu X, Feng J: Detecting causality between different frequencies, *Journal of Neuroscience Methods* 2008, **167**:367.

12. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data, *Science* 2003, **302**:449.

13. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP: Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data, *Science* 2005, **308**:523.

14. Ghahramani Z: *Learning Dynamic Bayesian Networks* Berlin: Springer Press; 2004.

15. Geweke J: **Measurement of Conditional Linear Dependence and Feedback Between Time Series**, *Journal of the American Statistical Association* 1982, **79** (388):907.

16. Geweke J: **Measurement of Linear Dependence and Feedback Between Multiple Time Series**, *Journal of the American Statistical Association* 1982, **77** (378):304.

17. Jensen FV: *An introduction to Bayesian networks* London: UCL Press; 1996.

18. Bach FR, Jordan MI: **Learning Graphical Models for Stationary Time Series**, *IEEE transactions on signal processing* 2004, **52** (8):2189.

19. Buntine WL: **Operations for Learning with Graphical Models**, Journal of *Artificial Intelligence Research* 1994, **2**:159.

20. Friedman N: **Inferring Cellular Networks Using Probabilistic Graphical Models**, *Science* 2004, **303**:799.

21. Guo S, Seth AK, Kendrick KM, Zhou C, Feng J: **Partial Granger Causality-Eliminating Exogenous Inputs and latent Variables**, *Journal of neuroscience methods* 2008, **172** (1):79.

22. Chen Y, Rangarajan G, Feng J, Ding M: **Analyzing multiple nonlinear time series with extended Granger causality**, *Physics Letters A* 2004, **324**:26.

23. Marinazzo D, Pellicoro M, Stramaglia S: **Kernel-Granger causality and the**

analysis of dynamic networks, *Physical review E* 2008, **77**:056215.

24. Locke JC, Kozma-Bognár L, Gould PD, Fehér B, Kevei E, Nagy F, Turner MS, Hall A, Millar AJ: **Experimental Validation of a predicted feedback loop in the multi-oscillator clock of Arabidopsis thaliana**, *Molecular Systems Biology* 2006, **2**: 59.
25. Ueda HR: **Systems biology flowering in the plant clock field**, *Molecular Systems Biology* 2006, **2**: 60.
26. On 24th July, 2008, a search on ISI tells us that there are 3531 papers on Bayesian network in the area of Comp. Sci., Math., Math + Comp. Biol. and Business + Economics, and 1125 papers on Granger causality.
27. Wang S, Chen Y, Ding M, Feng J, Stein JF, Aziz TZ, Liu X: **Revealing the dynamic causal interdenpendence between neural and muscular signals in Parkinsonian tremor**, *Journal of Franklin Institute-Engineering and Applied Mathematics* 2007, **344** (3-4):180.
28. Wu J, Kendrick K, Feng J: **Detecting Hot-Spots in Multivariates Biological Data**, *BMC Bioinformatics* 2007, **8**: 331.
29. Zhan Y, Halliday D, Jiang Ping, Liu X, Feng J: **Detecting the time-dependent coherence between non-stationary electrophysiological signals-A combined statistical and time-frequency approach**, *Journal of Neuroscience Methods* 2006, **156**:322.
30. Akaike H: **Fitting Autoregressive Models for Regression**, *Annals of the Institute of Statistcal Mathmatics* 1969, **21**:243.

31. Beamish N, Priestley MB: **A Study of Autoregressive and Window Spectral Estimation** 1981, *Applied Statistics* **30** (1): 41.

32. Morettin PA: **Levinson Algorithm and Its Applications in Time Series Analysis**, *International Statistical Review* 1984, **52** (1):83.

33. Morf M, Vieira A, Lee DTL, Kailath T: **Recursive Multichannel Maximum Entropy Spectral Estimation**, *IEEE transactions on geosciences electronics* 1978, **GE-16** (2): 85.

34. Ancona N, Marinazzo D, Stramaglia S: **Radial Basis Function Approach to Nonlinear Granger Causality of Time Series**, *Physical Review. E* 2004, **70**:056221.

35. Bishop CM: *Neural Networks for Pattern Recognition* Oxford: Oxford Univ. Press; 1995.

36. Bishop CM: *Pattern Recognition and Machine Learning* New York: Springer Press; 2006.

37. Murphy K: Bayes Net Toolbox for Matlab. It can be downloaded from the website: http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html.

Fig. captions.

Fig. 1. Granger causality and Bayesian network inference approaches applied on a simple linear toy model.

I. Five time series are simultaneously generated, and the length of each time series is 1000. X_2, X_3, X_4 and X_5 are shifted upward for visualization purpose.

II. Granger causality results. (a) The network structure inferred from Granger causality approach. (b) The 95% confidence intervals graph for all the possible directed connections. (c) For visualization purpose, all directed edges (causalities) are sorted and enumerated into the table. The total number of edges is 20.

III. Dynamic Bayesian network inference results. (a) The causal network structure learned from Bayesian network inference. (b) Each variable is represented by four nodes, representing different time-lags, we have a total of 20 nodes. They are numbered and enumerated in the table. (c) The simplified network structure: since we only care about the causality to the current time status, we can remove all the other edges and nodes that have no connection to the node 16 to node 20 (five variables with current time status).

(d). A further simplified network structure of causality.

Fig. 2. Granger causality and Bayesian network inference applied on data points of various sample sizes. The grey edges in the inferred network structures indicate undetected causalities in the toy model. For each sample size n, we simulated a data set of 100 realizations of n time points. The Bayesian network structure represents a model average from these 100 realizations. High-confidence arcs, appearing in at least 95% of the networks are shown. The Granger causality inferred the structure according to the 95% confidence interval constructed by using the bootstrap method. (Ⅰ) The sample size is 80. (Ⅱ) The sample size is 60. (Ⅲ) The sample size is 20.

Fig. 3. Granger causality and Bayesian network inference applied on a stochastic coefficients toy model: the parameters in polynomial equation are randomly generated in the interval [-1,1]. For each randomly generated coefficient vector, we applied the same approach as example 1: bootstrapping method and 95% confidence interval for Granger causality; 95% high confidence arcs are chosen from Bayesian network

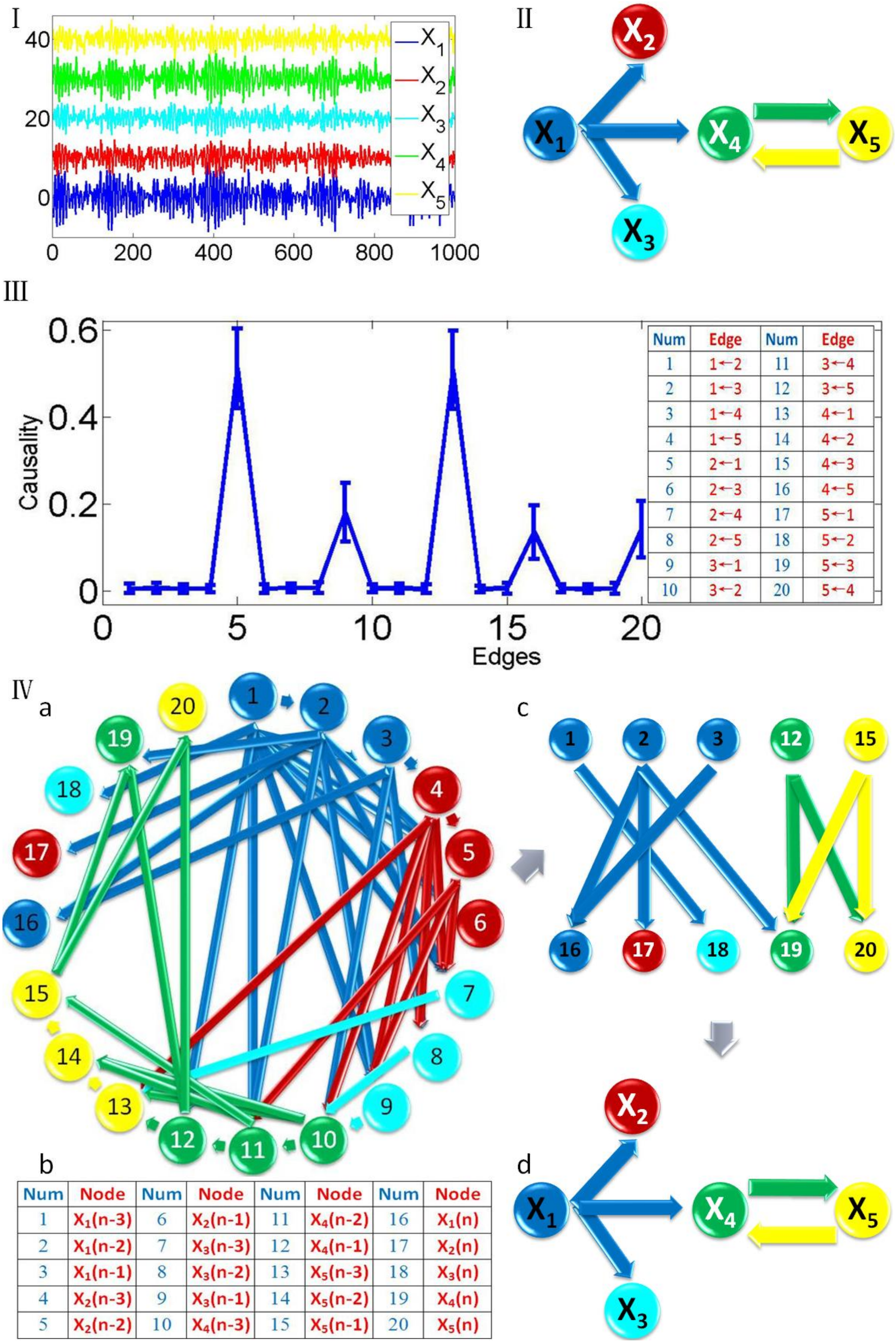
inference. (Ⅰ) We applied both approaches on different sample size (from 20 to 900). For each sample size, we generated 100 different coefficient vectors, so the total number of directed interactions for each sample size is 500. (a) The percentage of detected true positive causalities for both approaches. (b) Time cost for both approaches. (Ⅱ) For sample size 900, the derived causality (1 represents positive causality and 0 represents negative) is plotted with the absolute value of corresponding coefficients. For visualization purpose, the figure for Granger causality is shifted upward. （Ⅲ）Linear model fitting comparison for both Granger causality and Bayesian networks. Using a number of training data points to fit both linear models, one can calculate a corresponding predicted mean-square error by applying a set of test data. And we can find that Bayesian networks inference approach works much better than the Granger causality approach when the sample size is significant small (around 100). When the sample size is significant large, both approaches converge to the standard error which exactly fits the noise term in our toy model.

Fig. 4. Granger causality and Bayesian network inference approaches applied on a simple non-linear toy model. (Ⅰ) Five time series are simultaneously generated, and the length of each time series is 1000. They are assumed to be stationary. (Ⅱ) The five histogram graphs show the probability distribution for these five time series. (Ⅲ) Assuming no knowledge of MVAR toy model we fitted, we calculated Granger causality. Bootstrapping approach is used to construct the confidence intervals. The fitted MVAR model is simulated to generate a data set of 100 realizations of 1000 time points each. (a) For visualization purpose, all directed edges (causalities) are sorted and enumerated into the table. The total number of edges is 20. 95% confidence interval is chosen. (b) The network structure inferred from Granger causality method correctly recovers the pattern of connectivity in our MVAR toy model. (Ⅳ) Assuming no knowledge of MVAR toy model we fitted, we approach Bayesian network inference. (a) The causal network structure learned from Bayesian network inference for one realization of 1000 time points. (b) Each variable is represented by two nodes; each node represents different time statuses, so we have 10 nodes in total. They are numbered and enumerated into the table. (c) The simplified network structure: since we only care about the causality to the current time status, we can remove all the other edges and nodes that have no connection to the node 6 to node 10 (five variables with current time status). (d) A further simplified network structure: in order to compare with Granger causality approach, we hid the information of time status, and we obtained the same structure as Granger causality method had.

Fig. 5. Granger causality and Bayesian network inference applied on insufficient number of data points for non-linear model. The grey edges in the inferred network structures indicate undetected causalities in our defined toy model. For each sample size n, we simulated a data set of 100 realizations of n time points. The Bayesian network structure represents a model average from these 100 realizations. High-confidence arcs, appearing in at least 95% of the networks are shown. The Granger causality inferred the structure according to the 95% confidence interval constructed by using the bootstrap method. (Ⅰ) The sample size is 300. (Ⅱ) The sample size is 150. (Ⅲ) The sample size is 50.

Fig. 6. Granger causality and Bayesian network inference applied on a stochastic coefficients non-linear model: the parameters in polynomial equation are randomly generated in the interval [-2,2]. (Ⅰ) We applied both approaches on different sample size (from 300 to 900). For each sample size, we generated 100 different coefficient vectors, so the total number of directed interactions for each sample size is 500. (a) The percentage of detected true positive causalities for both approaches. (b) Time cost for both approaches. (Ⅱ) For sample size 900, the derived causality (1 represents positive causality and 0 represents negative) is plotted with the absolute value of corresponding coefficients. For visualization purpose, the figure for Granger causality is shifted upward.

Fig. 7. Granger causality approaches and Bayesian network inference approaches applied on experimental data (small sample size). The experiment measures the intensity of 7 genes in two cases of Arabidopsis Leaf: mock (normal) and infected. (Ⅰ)The time traces of 7 genes are plotted. There are 4 realizations of 24 time points. The time interval is 2 hours. (Ⅱ) The network structures are derived by using dynamic Bayesian network inference. All the genes are numbered as shown. Interestingly, after infection, the total network structure is changed. (a) The network structure for mock case. (b) the network structure for infected case. (Ⅲ) The network structures are derived by using Granger causality. (a) The network structure for mock case. (b) the network structure for infected case. (c) Using bootstrapping method to construct a 95% confidence intervals. For visualization purpose, all the directed edges are numbered and enumerate them into the table.



Num	Edge	Num	Edge
1	1←2	11	3←4
2	1←3	12	3←5
3	1←4	13	4←1
4	1←5	14	4←2
5	2←1	15	4←3
6	2←3	16	4←5
7	2←4	17	5←1
8	2←5	18	5←2
9	3←1	19	5←3
10	3←2	20	5←4

Num	Node	Num	Node	Num	Node	Num	Node
1	$X_1(n-3)$	6	$X_2(n-1)$	11	$X_4(n-2)$	16	$X_1(n)$
2	$X_1(n-2)$	7	$X_3(n-3)$	12	$X_4(n-1)$	17	$X_2(n)$
3	$X_1(n-1)$	8	$X_3(n-2)$	13	$X_5(n-3)$	18	$X_3(n)$
4	$X_2(n-3)$	9	$X_3(n-1)$	14	$X_5(n-2)$	19	$X_4(n)$
5	$X_2(n-2)$	10	$X_4(n-3)$	15	$X_5(n-1)$	20	$X_5(n)$

Fig. 1

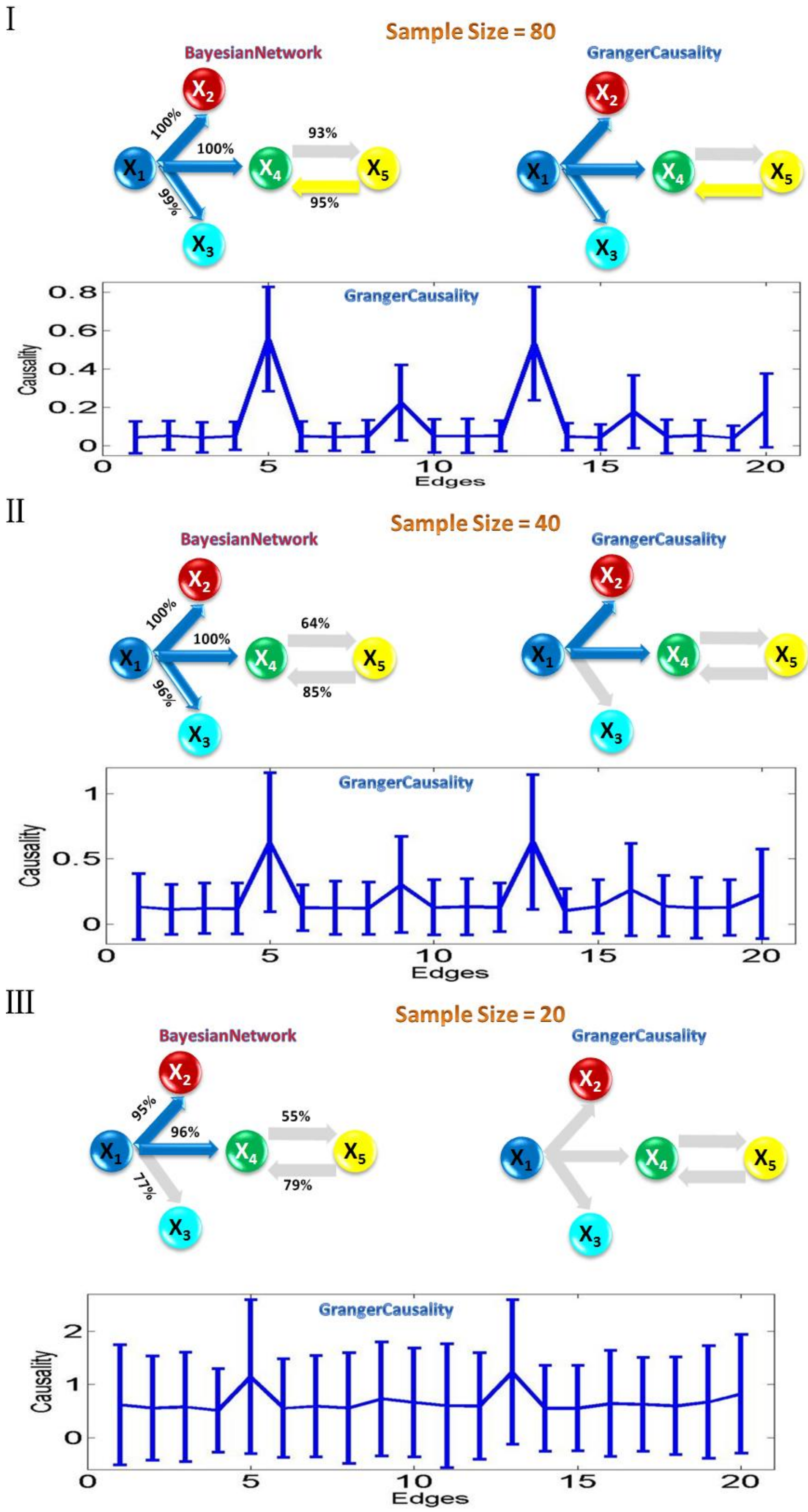
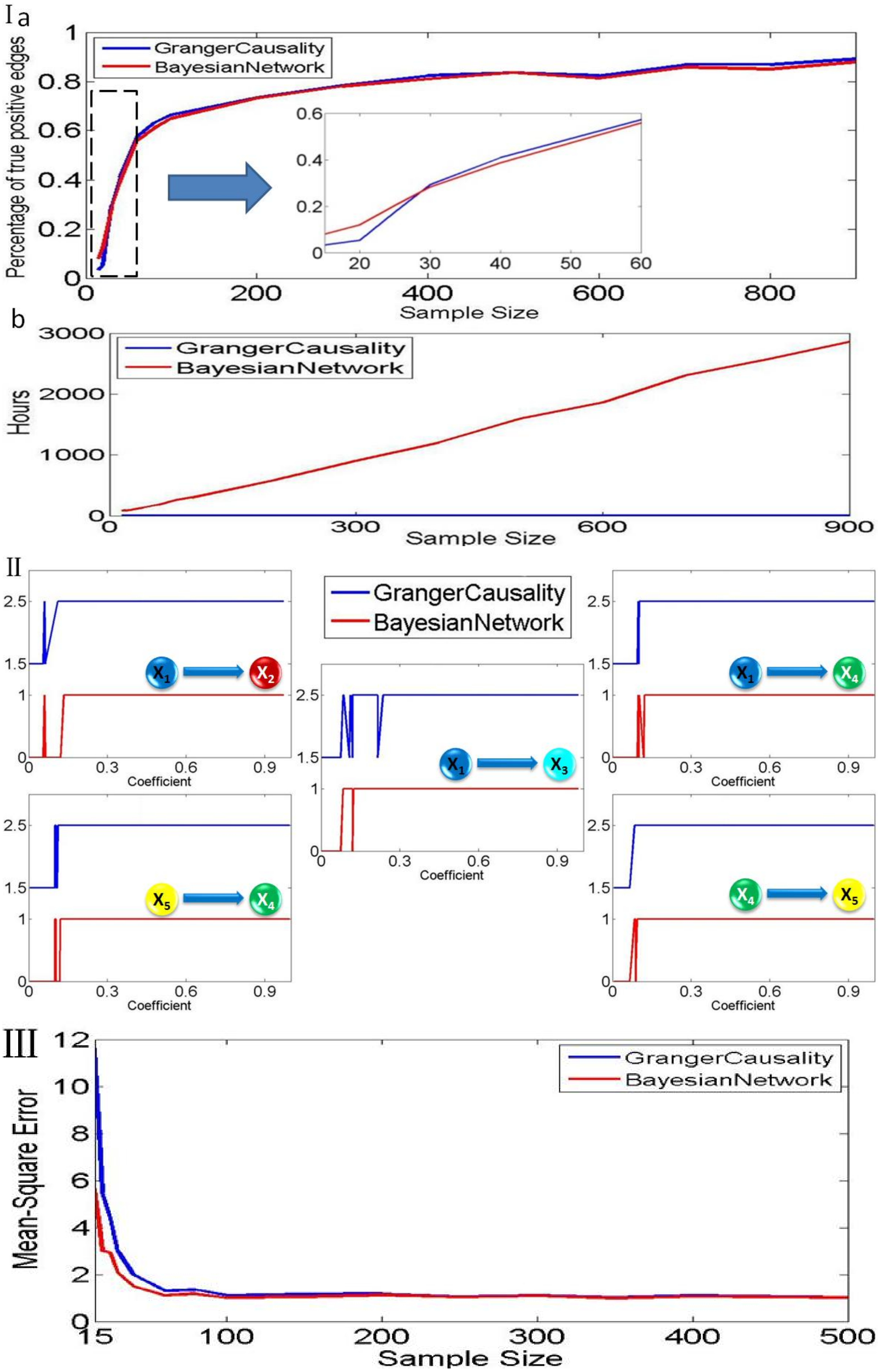
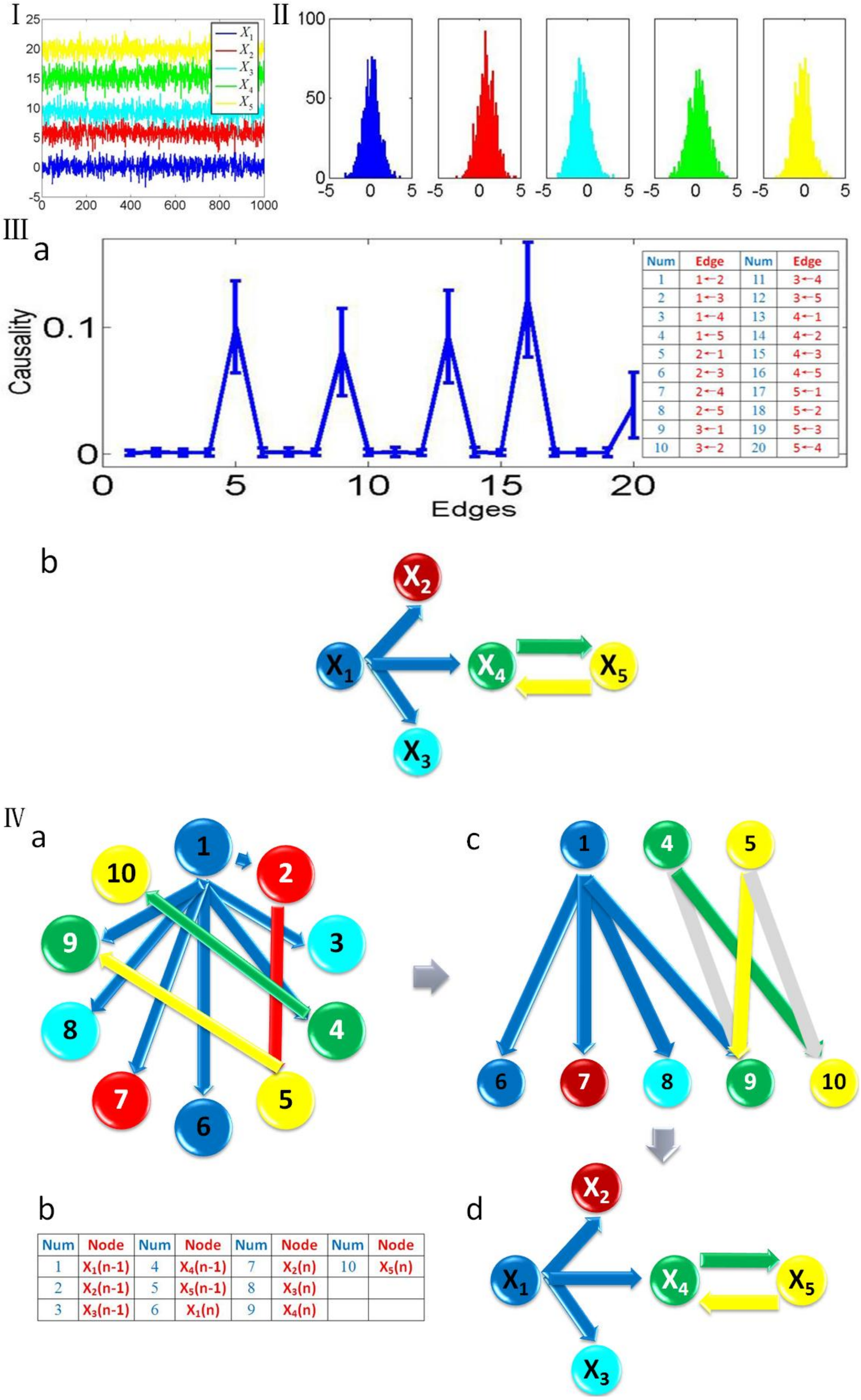

I
Sample Size = 80
BayesianNetwork
GrangerCausality
X1
X2
X3
X4
X5
100%
100%
99%
93%
95%
GrangerCausality
Causality
0.8
0.6
0.4
0.2
0
0
5
10
15
20
Edges
II
Sample Size = 40
BayesianNetwork
GrangerCausality
100%
100%
96%
64%
85%
GrangerCausality
Causality
1
0.5
0
Edges
III
Sample Size = 20
BayesianNetwork
GrangerCausality
95%
96%
77%
55%
79%
GrangerCausality
Causality
2
1
0
Edges

Fig. 2



I a
Percentage of true positive edges
GrangerCausality
BayesianNetwork
0
0.2
0.4
0.6
0.8
1
0
200
400
600
800
Sample Size
0
0.2
0.4
0.6
20
30
40
50
60
b
Hours
3000
2000
1000
0
0
300
600
900
Sample Size
GrangerCausality
BayesianNetwork
II
GrangerCausality
BayesianNetwork
X_1 → X_2
X_1 → X_3
X_1 → X_4
X_5 → X_4
X_4 → X_5
2.5
1.5
1
0
0.3
0.6
0.9
Coefficient
III
Mean-Square Error
12
10
8
6
4
2
0
15
100
200
300
400
500
Sample Size
GrangerCausality
BayesianNetwork

Fig. 3



Num	Edge	Num	Edge
1	1←2	11	3←4
2	1←3	12	3←5
3	1←4	13	4←1
4	1←5	14	4←2
5	2←1	15	4←3
6	2←3	16	4←5
7	2←4	17	5←1
8	2←5	18	5←2
9	3←1	19	5←3
10	3←2	20	5←4

Num	Node	Num	Node	Num	Node	Num	Node
1	$X_1(n-1)$	4	$X_4(n-1)$	7	$X_2(n)$	10	$X_5(n)$
2	$X_2(n-1)$	5	$X_5(n-1)$	8	$X_3(n)$		
3	$X_3(n-1)$	6	$X_1(n)$	9	$X_4(n)$		

Fig. 4

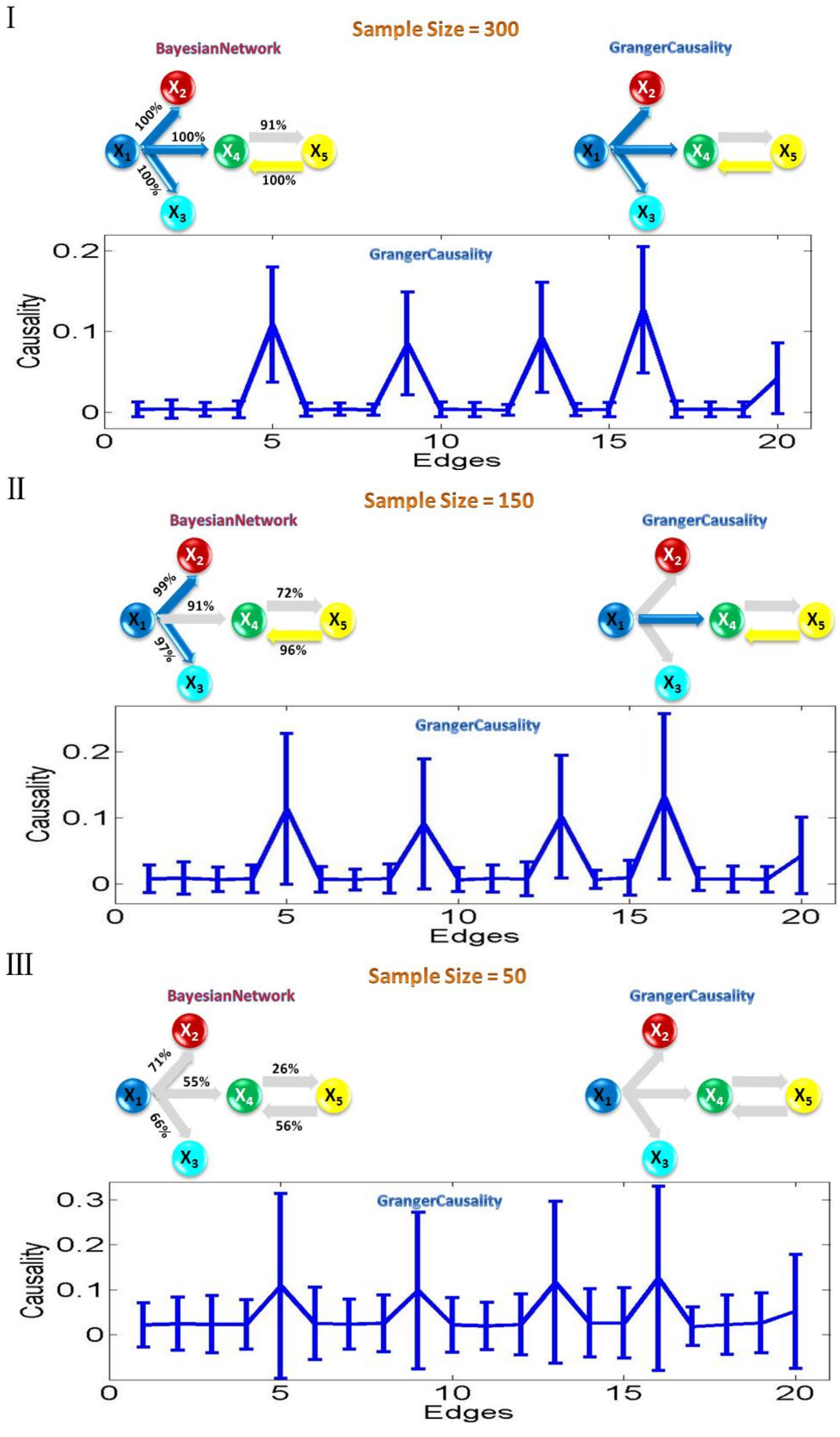

I
Sample Size = 300
BayesianNetwork
GrangerCausality
X_1
X_2
X_3
X_4
X_5
100%
100%
100%
91%
100%
GrangerCausality
Causality
0.2
0.1
0
Edges
0
5
10
15
20
II
Sample Size = 150
BayesianNetwork
GrangerCausality
99%
91%
97%
72%
96%
GrangerCausality
Causality
0.2
0.1
0
Edges
0
5
10
15
20
III
Sample Size = 50
BayesianNetwork
GrangerCausality
71%
55%
66%
26%
56%
GrangerCausality
Causality
0.3
0.2
0.1
0
Edges
0
5
10
15
20

Fig. 5

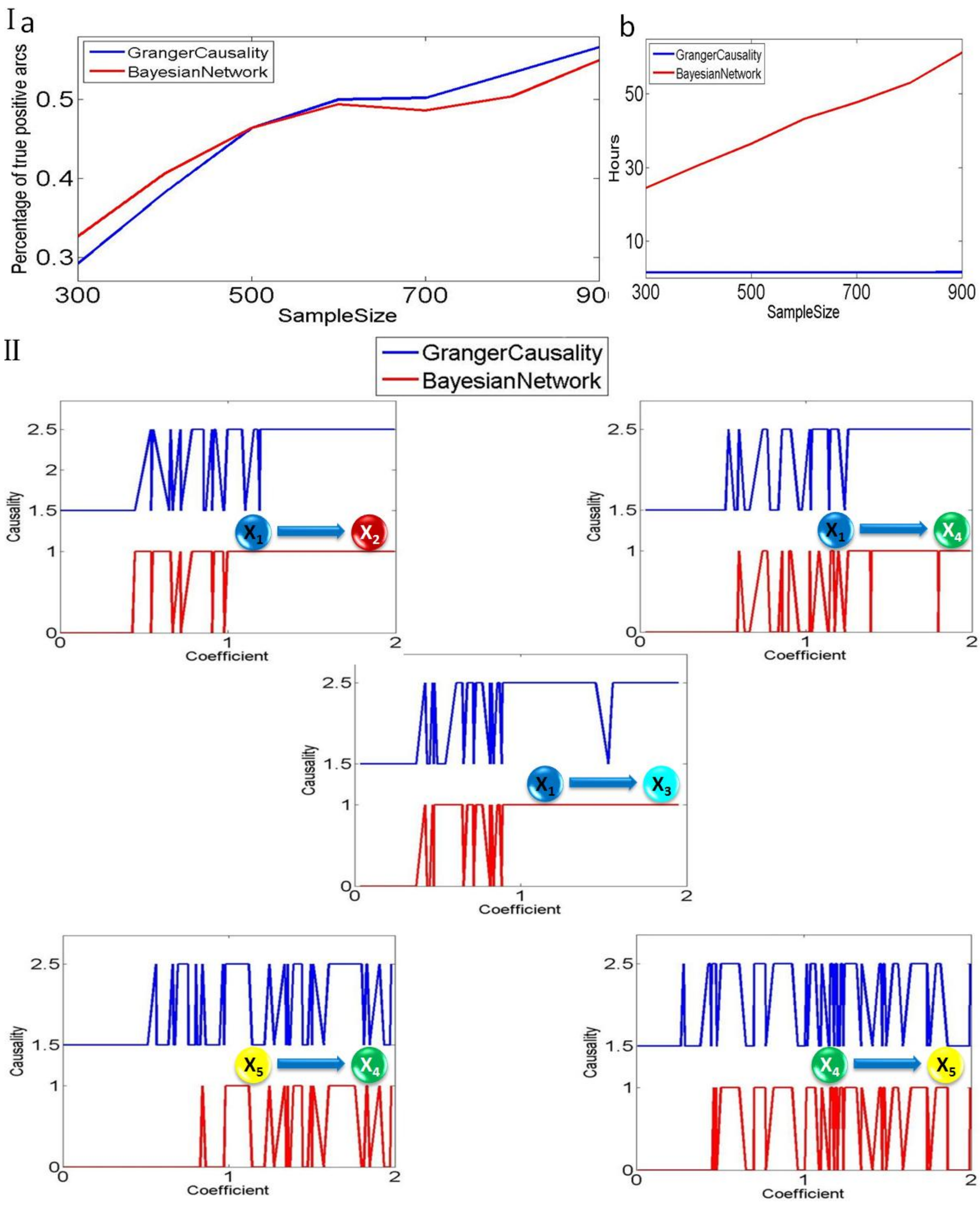
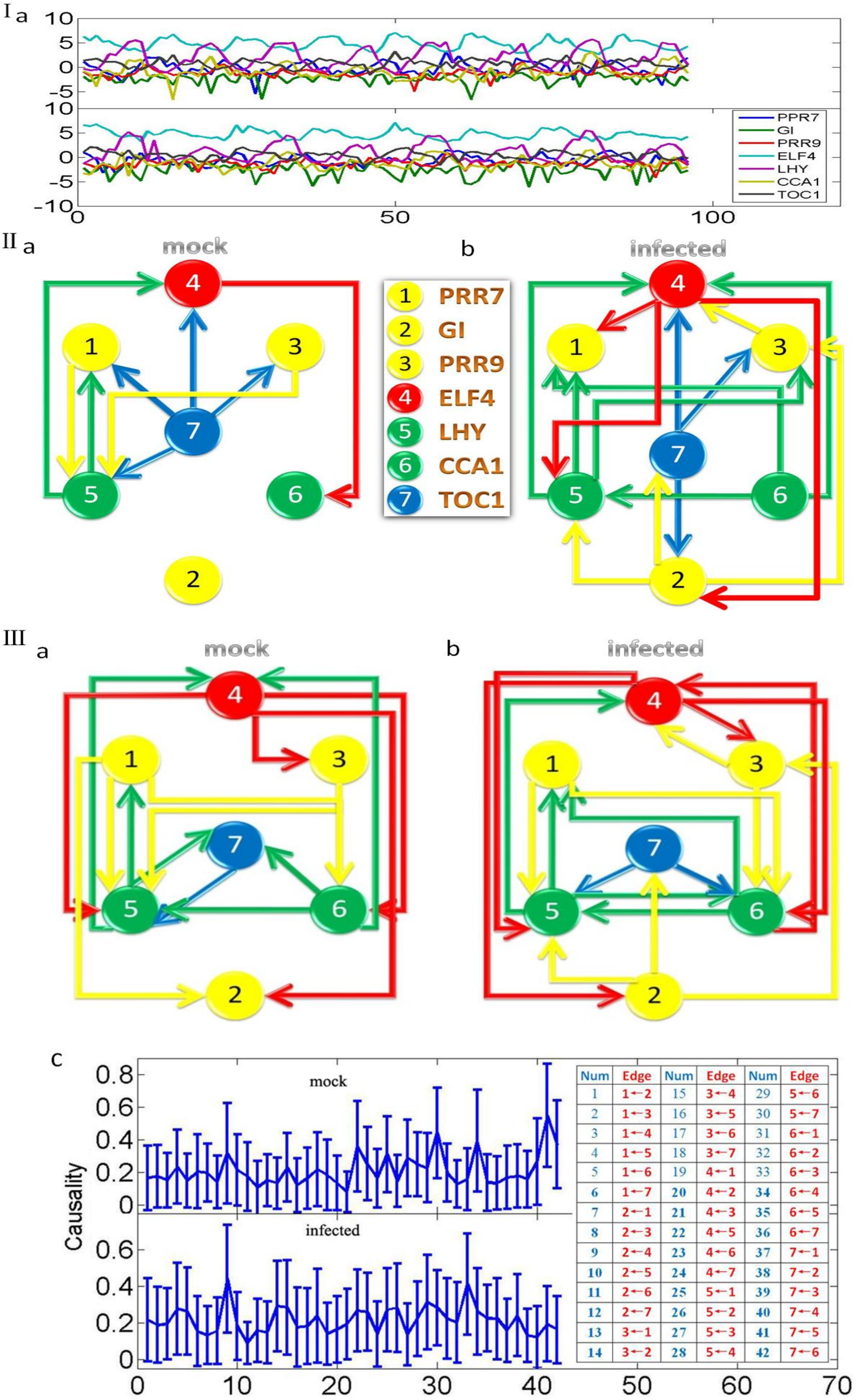

I a
GrangerCausality
BayesianNetwork
Percentage of true positive arcs
0.3
0.4
0.5
300
500
700
90
SampleSize
b
GrangerCausality
BayesianNetwork
Hours
10
30
50
300
500
700
900
SampleSize
II
GrangerCausality
BayesianNetwork
Causality
2.5
2
1.5
1
0
0
1
2
Coefficient
X_1
X_2
X_1
X_4
X_1
X_3
X_5
X_4
X_4
X_5

Fig. 6



Num	Edge	Num	Edge	Num	Edge
1	1←2	15	3←4	29	5←6
2	1←3	16	3←5	30	5←7
3	1←4	17	3←6	31	6←1
4	1←5	18	3←7	32	6←2
5	1←6	19	4←1	33	6←3
6	1←7	20	4←2	34	6←4
7	2←1	21	4←3	35	6←5
8	2←3	22	4←5	36	6←7
9	2←4	23	4←6	37	7←1
10	2←5	24	4←7	38	7←2
11	2←6	25	5←1	39	7←3
12	2←7	26	5←2	40	7←4
13	3←1	27	5←3	41	7←5
14	3←2	28	5←4	42	7←6

Fig. 7