Causality Quantification and Its Applications: Structuring and Modeling of Multivariate Time Series

Takashi Shibuya, Tatsuya Harada, Yasuo Kuniyoshi
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo 113-8656, Japan
{takashi, harada, kuniyosh}@isi.imi.i.u-tokyo.ac.jp

ABSTRACT

Time series prediction is an important issue in a wide range of areas. There are various real world processes whose states vary continuously, and those processes may have influences on each other. If the past information of one process X improves the predictability of another process Y, X is said to have a causal influence on Y. In order to make good predictions, it is necessary to identify the appropriate causal relationships. In addition, the processes to be modeled may include symbolic data as well as numerical data. Therefore, it is important to deal with symbolic and numerical time series seamlessly when attempting to detect causality.

In this paper, we propose a new method for quantifying the strength of the causal influence from one time series to another. The proposed method can represent the strength of causality as the number of bits, whether each of two time series is symbolic or numerical. The proposed method can quantify causality even from a small number of samples. In addition, we propose structuring and modeling methods for multivariate time series using causal relationships of two time series. Our structuring and modeling methods can also deal with data sets which include both types of time series. Experimental results demonstrate that our methods can perform well even if the number of samples is small.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics—*time series analysis*

General Terms

Algorithms, Theory

Keywords

Information Theory, Entropy, Autoregressive Model, Markov Chain, Time Series Prediction

KDD'09, June 28–July 1, 2009, Paris, France.

1. INTRODUCTION

Time series prediction is an important issue in a wide range of areas such as finance [4], weather forecasting [1], and transportation planning [7]. For example, economists predict economic trends from indicators such as exchange rates, stock prices, and GDPs. The conclusions drawn by the economists then affect traders' decisions. The actions of the traders are then reflected back into the economic indicators, creating a cycle. Thus, there are various real world processes whose states vary continuously, and those processes may have influences on each other. When past values of one time series provide significant information about future values of another, the relationship of the two time series is called "causality", in distinction from correlation. In order to make good predictions, it is necessary to identify the appropriate causal relationships. In addition, the processes to be modeled may include symbolic data (weather, text, etc.) as well as numerical data (temperatures, stock prices, etc.). Therefore, it is important to be able to handle seamlessly mixed symbolic and numerical time series when attempting to detect causality. If causality can be properly evaluated, it is possible to predict multivariate time series by estimating their structure or modeling their dynamics.

Time series prediction for multivariate data is a well studied problem, especially in economics. Common methods are vector autoregressive models (VAR models) and dynamic factor models (DFMs) [11, 4]. VAR models can detect totally idealized relationships. The drawback of VAR models is that a large number of samples are required to obtain good parameter estimates. DFMs, on the other hand, avoid the need for a large number of samples by compressing the original multivariate variables into a smaller set of unobserved factors. However, DFMs cannot describe causal relationships among observed variables. Also, in some cases the compression does not work, if the original variables cannot be represented well by a smaller number of factors.

We focus on *causality measures* in order to solve these problems. Causality measures quantify the strength of a causal influence from one time series to another. Many causality measures have been proposed [9, 6], because detection of causality is a significant challenge in economics [5, 3] and biology [2]. Since causal relationships of just two time series can be estimated from a small number of samples, we expect that it is possible to reduce the number of samples needed for modeling multivariate time series by combining bivariate models. In contrast to the implicit modeling of relationships performed by the compression used in DFMs, using causality measures makes it possible to describe ex-

plicit casual relationships between observed variables. However, there is a problem remaining. Conventional causality measures cannot seamlessly handle time series with mixed symbolic and numerical data. The causality measures which can deal with such series are needed.

In this paper, we propose a set of new causality measures, which can represent the strength of a causal influence in a "common currency", regardless of whether each of two variables is numerical or symbolic. This common currency is a bit count. The proposed measures can quantify causality even if the number of samples is small. In addition, we propose structuring and modeling methods for multivariate time series based on causal relationships of two time series. Our methods can also deal with data sets that include both symbolic and numerical time series. Experimental results demonstrate that our methods can perform well even if the data is of limited size and includes mixed numerical/symbolic components.

The remainder of this paper is organized as follows. Section 2 discusses existing causality measures, and proposes the new method. Our structuring and modeling methods are given in Section 3. Section 4 reviews the results of our experiments. Section 5 is a brief conclusion.

2. CAUSALITY QUANTIFICATION

We introduce existing causality measures and point out their problems. Then, we propose a set of new causality measures.

2.1 Existing methods

2.1.1 *Transfer entropy*

The *transfer entropy*, proposed by Schreiber [10], is an information theoretic causality measure that evaluates causality by calculating the information one variable contains about another.

Assume that X, Y are the time series variables, which indicate x_t and y_t at time t respectively, and that the two variables may be approximated by stationary Markov processes of order k, l. Then, the dynamics of Y can be expressed by a transition probability $p(y_t|\boldsymbol{y}_{t-1}^{(l)})$, where $\boldsymbol{y}_{t-1}^{(l)}$ denotes $(y_{t-1}, \ldots, y_{t-l})^T$. Given that the past states $\boldsymbol{y}_{t-1}^{(l)}$ are known, the entropy of the subsequent state y_t is given by the following formula:

$$\sum_{y_t, \boldsymbol{y}_{t-1}^{(l)}} p(y_t, \boldsymbol{y}_{t-1}^{(l)}) \log_2 \frac{1}{p(y_t|\boldsymbol{y}_{t-1}^{(l)})}.$$

Generally, it is necessary to estimate the transition probability because the true transition probability $p(y_t|\boldsymbol{y}_{t-1}^{(l)})$ is unknown. When an estimate transition probability $q(y_t|\boldsymbol{y}_{t-1}^{(l)})$ is used instead of $p(y_t|\boldsymbol{y}_{t-1}^{(l)})$, the *Kullback-Leibler distance* expresses the difference between the two probability distributions. This distance can be interpreted as the code length penalty paid for using the model q when the real probability is p. Kullback-Leibler distance is given by the following formula:

$$\sum_{y_t, \boldsymbol{y}_{t-1}^{(l)}} p(y_t, \boldsymbol{y}_{t-1}^{(l)}) \log_2 \frac{p(y_t|\boldsymbol{y}_{t-1}^{(l)})}{q(y_t|\boldsymbol{y}_{t-1}^{(l)})}.$$

Here, if the current state y_t of Y is independent of the past states $\boldsymbol{x}_{t-1}^{(k)}$ of X, then *the generalized Markov property*,

$$p(y_t|\boldsymbol{y}_{t-1}^{(l)}, \boldsymbol{x}_{t-1}^{(k)}) = p(y_t|\boldsymbol{y}_{t-1}^{(l)}) \quad (1)$$

holds. Since the two probability distributions deviate from the generalized Markov property (Equation (1)) if the state of X has some kind of influence on Y, the causality from X to Y can be quantified by the Kullback-Leibler distance. Therefore, the average information about Y contained by X is given by the following formula for the transfer entropy:

$$T_{X\to Y} = \sum_{y_t, \boldsymbol{y}_{t-1}^{(l)}, \boldsymbol{x}_{t-1}^{(k)}} p(y_t, \boldsymbol{y}_{t-1}^{(l)}, \boldsymbol{x}_{t-1}^{(k)}) \times \log_2 \frac{p(y_t|\boldsymbol{y}_{t-1}^{(l)}, \boldsymbol{x}_{t-1}^{(k)})}{p(y_t|\boldsymbol{y}_{t-1}^{(l)})}. \quad (2)$$

This causality measure can be applied to symbolic time series without preprocessings. However, when this measure is applied to numerical time series, some kind of preprocessings or assumptions are needed. In the comparison experiments of this paper, we used the histograms of the embedding vectors (naïve histogram technique). In this case, the number of bins r is needed as an additional parameter.

2.1.2 *Continuous transfer entropy*

Continuous transfer entropy proposed by Kaiser et al. [8] is a causality measure derived from transfer entropy by assuming that numerical time series X, Y are generated by Gaussian processes.

Transfer entropy can be expressed as a sum of Shannon entropies:

$$T_{X\to Y} = H\left(\boldsymbol{Y}_{t-1}^{(l)} \otimes \boldsymbol{X}_{t-1}^{(k)}\right) - H\left(Y_t \otimes \boldsymbol{Y}_{t-1}^{(l)} \otimes \boldsymbol{X}_{t-1}^{(k)}\right) + H\left(\boldsymbol{Y}_{t-1}^{(l)}\right) - H\left(Y_t \otimes \boldsymbol{Y}_{t-1}^{(l)}\right) \quad (3)$$

$$\text{where} \quad H\left(\boldsymbol{Y}_{t-1}^{(l)}\right) = \int p(\boldsymbol{y}_{t-1}^{(l)}) \log_2 \frac{1}{p(\boldsymbol{y}_{t-1}^{(l)})} d\boldsymbol{y}_{t-1}^{(l)}.$$

For processes with Gaussian distributions, Shannon entropy can be expressed as following formula:

$$H\left(\boldsymbol{Y}_{t-1}^{(l)}\right) = \frac{1}{2} l \log_2 (2\pi e) + \frac{1}{2} \log_2 \left|C_{\boldsymbol{y}_{t-1}^{(l)}}\right| \quad (4)$$

where $C_{\boldsymbol{y}_{t-1}^{(l)}}$ is the covariance matrix of $\boldsymbol{y}_{t-1}^{(l)}$, and $\left|C_{\boldsymbol{y}_{t-1}^{(l)}}\right|$ denotes the determinant of $C_{\boldsymbol{y}_{t-1}^{(l)}}$.

From Equation (3) and Equation (4), the following formula is obtained as continuous transfer entropy.

$$T_{X\to Y} = \frac{1}{2} \log_2 \frac{\left|C_{\boldsymbol{y}_{t-1}^{(l)} \otimes \boldsymbol{x}_{t-1}^{(k)}}\right| \left|C_{y_t \otimes \boldsymbol{y}_{t-1}^{(l)}}\right|}{\left|C_{y_t \otimes \boldsymbol{y}_{t-1}^{(l)} \otimes \boldsymbol{x}_{t-1}^{(k)}}\right| \left|C_{\boldsymbol{y}_{t-1}^{(l)}}\right|}. \quad (5)$$

2.1.3 *Granger causality*

Granger causality is the method proposed by Granger [5]. The approach of this method is to examine if the prediction of one variable can be improved by incorporating the past information of another variable.

First, consider the autoregressive model (AR model):

$$y_t = a_0 + a^T \boldsymbol{y}_{t-1}^{(l)} + \epsilon_t^{(y)} \quad (6)$$

where a_0 is a constant term and $a \in R^l$ is a regression coefficient vector, determined by least squares. $\epsilon_t^{(y)}$ is white noise with zero mean and σ_y^2 variance, $\epsilon_t^{(y)} \sim N\left(0, \sigma_y^2\right)$.

Next, consider the following regression model:

$$y_t = b_0 + b^T y_{t-1}^{(l)} + c^T x_{t-1}^{(k)} + \epsilon_t^{(y|x)}. \quad (7)$$

As in the AR model, b_0 is a constant term and $b \in R^l$, and $c \in R^k$ are regression coefficient vectors determined by least squares. $\epsilon_t^{(y|x)}$ is white noise with zero mean and $\sigma_{y|x}^2$ variance, $\epsilon_t^{(y|x)} \sim N\left(0, \sigma_{y|x}^2\right)$.

Here, the modeling performance of the two models can be evaluated by the comparison of σ_y^2 and $\sigma_{y|x}^2$. If X has a causal influence on Y, then $\sigma_y^2 > \sigma_{y|x}^2$. However, it is not obvious how to compare the two variances, and multiple comparison methods have been proposed such as $\sigma_y^2 - \sigma_{y|x}^2$ and $\sigma_y^2/\sigma_{y|x}^2$ [9. 6]. One issue with Granger causality is that the magnitude relation of the measured values can be reversed by using different comparison methods. In the experiments of this paper, we used $\sigma_y^2 - \sigma_{y|x}^2$ as Granger causality.

Granger causality cannot be applied to symbolic time series because it is based on regression models.

2.2 Proposed method

We derive a set of new causality measures that deal with symbolic and numerical time series seamlessly by combining the regression model and the information theory.

2.2.1 *From numerical data to numerical data*

Let X, Y be numerical time series. When it is assumed that Y is ordered by AR model (Equation (6)), the conditional probability distribution $p(y_t|y_{t-1}^{(l)})$ is given by the following formula:

$$p(y_t|y_{t-1}^{(l)}) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\left(y_t - a_0 - a^T y_{t-1}^{(l)}\right)^2}{2\sigma_y^2}\right).$$

In AR model, the probability distribution of $y_{t-1}^{(l)}$ is a multivariate normal distribution. Let $\mu_{y_{t-1}^{(l)}}$ denote the mean of $y_{t-1}^{(l)}$, and $C_{y_{t-1}^{(l)}}$ denote the covariance matrix of $y_{t-1}^{(l)}$. Then, the joint probability distribution $p(y_t, y_{t-1}^{(l)})$ is the multivariate normal distribution given by the following formula:

$$p(y_t, y_{t-1}^{(l)}) = \frac{1}{\sqrt{(2\pi)^{l+1}\sigma_y^2\left|C_{y_{t-1}^{(l)}}\right|}} \exp\left(-\frac{1}{2}\psi_t^{(l+1)T} X \psi_t^{(l+1)}\right)$$

$$\text{where} \quad \psi_t^{(l+1)} = \begin{bmatrix} y_t - a_0 - a^T \mu_{y_{t-1}^{(l)}} \\ y_{t-1}^{(l)} - \mu_{y_{t-1}^{(l)}} \end{bmatrix}$$

$$X = \begin{bmatrix} \frac{1}{\sigma_y^2} & -\frac{a^T}{\sigma_y^2} \\ -\frac{a}{\sigma_y^2} & \frac{aa^T}{\sigma_y^2} + C_{y_{t-1}^{(l)}}^{-1} \end{bmatrix}.$$

The determinant of the covariance matrix of this distribution is $\sigma_y^2\left|C_{y_{t-1}^{(l)}}\right|$. Therefore, the following formula is obtained:

$$\left|C_{y_t \otimes y_{t-1}^{(l)}}\right| = \sigma_y^2\left|C_{y_{t-1}^{(l)}}\right|. \quad (8)$$

Similarly, as to the joint probability distribution of $y_t \otimes y_{t-1}^{(l)} \otimes x_{t-1}^{(k)}$, the following formula is obtained if Equation (7) is assumed.

$$\left|C_{y_t \otimes y_{t-1}^{(l)} \otimes x_{t-1}^{(k)}}\right| = \sigma_{y|x}^2\left|C_{y_{t-1}^{(l)} \otimes x_{t-1}^{(k)}}\right|. \quad (9)$$

Since time series data ordered by a regression model is assumed to be a multivariate normal distribution, the following formula is obtained by substituting Equation (8) and Equation (9) for Equation (5).

$$T_{X \to Y} = \frac{1}{2}\log_2 \frac{\sigma_y^2}{\sigma_{y|x}^2}. \quad (10)$$

This causality measure represents the average number of bits of information about Y that is contained in X, within the framework of Granger causality.

2.2.2 *From symbolic data to numerical data*

Let X be symbolic time series, and Y be numerical time series. When it is assumed that Y is ordered by a normal distribution, the transfer entropy is given by the the following formula:

$$\sum_{x_{t-1}^{(k)}} \iint p(y_t, y_{t-1}^{(l)}, x_{t-1}^{(k)}) \times \log_2 \frac{p(y_t|y_{t-1}^{(l)}, x_{t-1}^{(k)})}{p(y_t|y_{t-1}^{(l)})} dy_t dy_{t-1}^{(l)}$$

$$= \sum_{x_{t-1}^{(k)}} p(x_{t-1}^{(k)}) \iint p(y_t, y_{t-1}^{(l)}|x_{t-1}^{(k)}) \times \log_2 \frac{p(y_t|y_{t-1}^{(l)}, x_{t-1}^{(k)})}{p(y_t|y_{t-1}^{(l)})} dy_t dy_{t-1}^{(l)}$$

$$= \frac{1}{2}\sum_{x_{t-1}^{(k)}} p(x_{t-1}^{(k)}) \log_2 \frac{\left|C_{y_{t-1}^{(l)}}\left(x_{t-1}^{(k)}\right)\right|\left|C_{y_t \otimes y_{t-1}^{(l)}}\right|}{\left|C_{y_t \otimes y_{t-1}^{(l)}}\left(x_{t-1}^{(k)}\right)\right|\left|C_{y_{t-1}^{(l)}}\right|}$$

where $C_{y_{t-1}^{(l)}}\left(x_{t-1}^{(k)}\right)$ is the covariance matrix of $y_{t-1}^{(l)}$ corresponding to $x_{t-1}^{(k)}$, the states of $X_{t-1}^{(k)}$.

Here, consider the Switching AR model:

$$y_t = b_0\left(x_{t-1}^{(k)}\right) + b\left(x_{t-1}^{(k)}\right)^T y_{t-1}^{(l)} + \epsilon_t^{(y|x)}$$

where $b_0\left(x_{t-1}^{(k)}\right)$ and $b\left(x_{t-1}^{(k)}\right)$ denote the constant term and the coefficient which switch corresponding to $x_{t-1}^{(k)}$. Let $\sigma_{y|x}^2\left(x_{t-1}^{(k)}\right)$ be the variance of $\epsilon_t^{(y|x)}$ corresponding to $x_{t-1}^{(k)}$. Then, the following formula is obtained.

$$T_{X \to Y} = \frac{1}{2}\sum_{x_{t-1}^{(k)}} p(x_{t-1}^{(k)}) \log_2 \frac{\sigma_y^2}{\sigma_{y|x}^2\left(x_{t-1}^{(k)}\right)}. \quad (11)$$

This causality measure represents the average number of bits of information about Y contained in X. This measure denotes how much predictability of y_t is improved by preparing a different model for each state of $X_{t-1}^{(k)}$.

2.2.3 *From numerical data to symbolic data*

Let X be numerical time series, and Y be symbolic time series. When it is assumed that X is distributed according

Table 1: Overview of proposed causality measures

Types of data	Measure
Symbolic data → Symbolic data	Equation (2)
Numerical data → Numerical data	Equation (10)
Symbolic data → Numerical data	Equation (11)
Numerical data → Symbolic data	Equation (12)

to a normal distribution, the transfer entropy is given by the the following formula:

$$\begin{aligned} T_{X\to Y} &= \sum_{y_t, \boldsymbol{y}_{t-1}^{(l)}} \int p(y_t, \boldsymbol{y}_{t-1}^{(l)}, \boldsymbol{x}_{t-1}^{(k)}) \times \\ &\qquad \log_2 \frac{p(y_t|\boldsymbol{y}_{t-1}^{(l)}, \boldsymbol{x}_{t-1}^{(k)})}{p(y_t|\boldsymbol{y}_{t-1}^{(l)})} d\boldsymbol{x}_{t-1}^{(k)} \\ &= \sum_{y_t, \boldsymbol{y}_{t-1}^{(l)}} p(y_t, \boldsymbol{y}_{t-1}^{(l)}) \int p(\boldsymbol{x}_{t-1}^{(k)}|y_t, \boldsymbol{y}_{t-1}^{(l)}) \times \\ &\qquad \log_2 \frac{p(\boldsymbol{x}_{t-1}^{(k)}|y_t, \boldsymbol{y}_{t-1}^{(l)})}{p(\boldsymbol{x}_{t-1}^{(k)}|\boldsymbol{y}_{t-1}^{(l)})} d\boldsymbol{x}_{t-1}^{(k)} \\ &= \frac{1}{2} \sum_{y_t, \boldsymbol{y}_{t-1}^{(l)}} p(y_t, \boldsymbol{y}_{t-1}^{(l)}) \log_2 \frac{\left| C_{\boldsymbol{x}_{t-1}^{(k)}} \left(\boldsymbol{y}_{t-1}^{(l)} \right) \right|}{\left| C_{\boldsymbol{x}_{t-1}^{(k)}} \left(y_t, \boldsymbol{y}_{t-1}^{(l)} \right) \right|}. \end{aligned} \tag{12}$$

This causality measure represents the average number of bits of information about Y contained in X. This measure denotes how much the distribution of $\boldsymbol{x}_{t-1}^{(k)}$ is changed by taking into consideration not only $\boldsymbol{y}_{t-1}^{(l)}$ but also y_t.

2.2.4 *Overview of proposed method*

By using three proposed causality measures and the ordinary transfer entropy, causal influences of all combinations of symbolic and numerical time series can be represented by the average numbers of bits (Table 1). Moreover, the proposed method can be utilized directly for the modeling of multivariate time series because this method is based on regression models that represent the dynamics of the variables.

Compared to existing methods, the proposed method has some other advantages. The proposed method requires a smaller number of parameters. Though the proposed measures use the two embedding dimension parameters k, l, these are common to all causality measures [9]. In addition, the proposed measures are invariant to linear transformations or normalization of numerical time series.

3. STRUCTURING AND MODELING

3.1 Structuring method based on causality

Estimation of the causal structure hidden in data is an important and a well studied subject. There are two main kinds of estimation methods. One is to select the best structure by searching and evaluating entire structures, the other is to build the structure by examining each causal influence from one variable to another. The former works well at estimating the true structure if there are enough samples. However, the number of possible structures increases exponentially depending on the number of variables: when the number of variables is N, the number of structures is

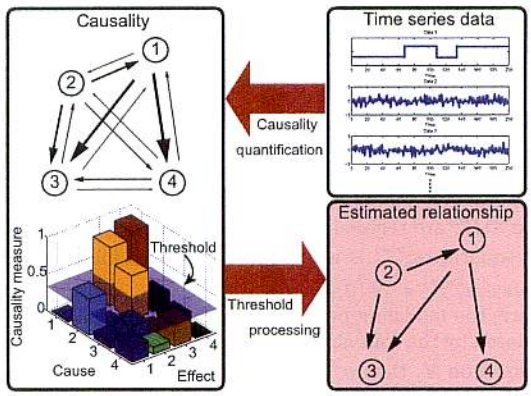


Figure 1: Overview of proposed structuring method.

$O\left(e^{N^2}\right)$. It is difficult to search through the large parameter space created by a data set with many variables. On the other hand, the latter needs fewer samples in order to estimate relationships. We focus on this advantage. Even when sufficient samples can be observed, the latter method can sooner model observed time series and make predictions. We propose a new structuring method, which structures multivariate time series using causality measures. Our causality measures proposed in Section 2 can represent the strength of a causal influence as a bit count, regardless of whether each of two time series is symbolic or numerical. As a result, the proposed structuring method can be applied to data sets which include both symbolic and numerical time series. The detailed method is as follows.

When a data set of N time series are given, $N(N-1)$ causality measures can be calculated, representing each pair of variables but excluding self-interactions. Structuring can be achieved by filtering the $N(N-1)$ values with a threshold, leaving behind only the significant causal links (Figure 1).

3.2 Modeling method based on causality

Modeling of multivariate time series can be realized by combining the bivariate models which are used in calculations of causality measures. In the combination process, the values of the causality measures are used as weights. By using weights based on the values, models which have strong causal influences are stressed, while models with weak causal influences are ignored. The detailed method is as follows.

Let Y denote the objective numerical variable, and X_1, ..., X_{N-1} denote the other variables. Then, the modeling of Y can be realized by combining the regression models $y_t = f(\boldsymbol{y}_{t-1}^{(l)}, \boldsymbol{x}_{i,t-1}^{(k)})$ which is used in the calculation of the causality measure $T_{X_i \to Y}$:

$$y_t = \frac{\sum_{i=1}^{N-1} T_{X_i \to Y} \times f(\boldsymbol{y}_{t-1}^{(l)}, \boldsymbol{x}_{i,t-1}^{(k)})}{\sum_{i=1}^{N-1} T_{X_i \to Y}}$$

where, if X_i is a numerical time series,

$$f(\boldsymbol{y}_{t-1}^{(l)}, \boldsymbol{x}_{i,t-1}^{(k)}) = b_{i,0} + \boldsymbol{b}_i^T \boldsymbol{y}_{t-1}^{(l)} + \boldsymbol{c}_i^T \boldsymbol{x}_{i,t-1}^{(k)}$$

and, if X_i is a symbolic time series,

$$f(\boldsymbol{y}_{t-1}^{(l)}, \boldsymbol{x}_{i,t-1}^{(k)}) = b_{i,0}\left(\boldsymbol{x}_{i,t-1}^{(k)}\right) + \boldsymbol{b}_i \left(\boldsymbol{x}_{i,t-1}^{(k)}\right)^T \boldsymbol{y}_{t-1}^{(l)}.$$

In the above formulation, all causal relationships are taken into account. However, the model can be simplified by filtering the values of the causality measures with a threshold T_{thre}. The filtering can be realized by the following processing:

$$T_{X_i \to Y} = 0 \quad \text{if } T_{X_i \to Y} \leq T_{thre}.$$

Unlike *normal* VAR models, our proposed method can model data sets that include not only numerical but also symbolic time series. In addition, this method has some other advantages. First, the model can be estimated even from a small number of samples because the model is made by combining bivariate models. If all variables are numerical, the required number of samples is independent of the number of variables, while *normal* VAR models need many more samples. Second, when new variables are added to the already learned model or unnecessary variables are deleted from it, the bivariate models which are used in it are available for the estimation of a new model. In the case of VAR models, the model must be estimated from the beginning.

4. EXPERIMENT

In this section, we first examine the performance of the proposed and existing causality measures. The data sets used for evaluation have arbitrary combinations of numerical and symbolic data (with the exception of symbolic-symbolic data, which just gives the transfer entropy). Next, we conduct experiments on structuring and modeling methods by examining whether relationships among multivariate time series can be estimated by combining causal relationships of two time series.

4.1 Causality quantification

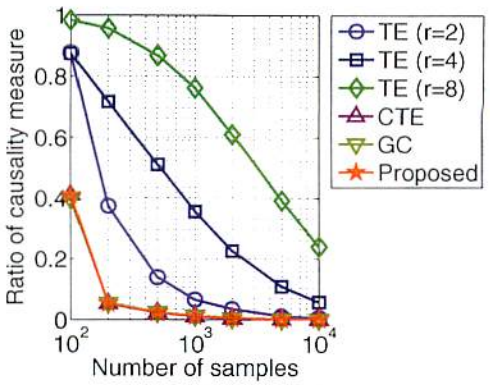
4.1.1 *From numerical data to numerical data*

In the first experiment, we examine the performance of causality quantification methods for numerical time series. We compared the ability of the different methods to distinguish the presence or absence of a causal influence on bivariate data sets. Each data set consists of two numerical time series C_1 and C_2. C_1 and C_2 are constructed using by the following model:

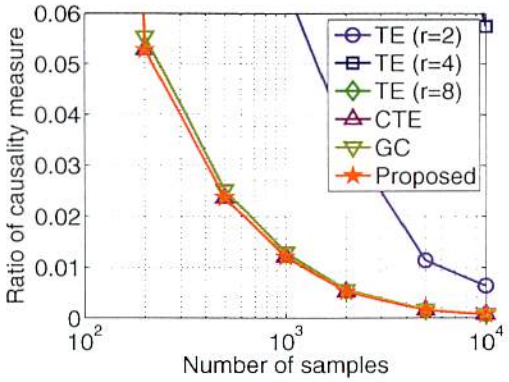
$$c_{1,t} = 0.3c_{1,t-1} + 0.3c_{2,t-1} + \epsilon_{1,t}$$
$$c_{2,t} = 0.3c_{2,t-1} + \epsilon_{2,t}$$

where $\epsilon_{1,t}$ and $\epsilon_{2,t}$ are white noises with zero mean and unit variance, $\epsilon_{1,t} \sim N(0,1), \epsilon_{2,t} \sim N(0,1)$. In this model, C_2 has a causal influence on C_1 whereas C_1 does not on C_2. We generate 100 independent data sets from random initial values, and compare the proposed measure with transfer entropy (TE), continuous transfer entropy (CTE), and Granger causality (GC). We set two embedding dimension parameters of causality measures to $k = 1$ and $l = 1$ according to the model of data sets. As to TE, we compare three different parameters (the numbers of bins), $r = 2$, $r = 4$, and $r = 8$. In this experiment, we change the number of samples $N = 100$, 200, 500, 1000, 2000, 5000, and 10000.

Figure 2 depicts the result. The horizontal axis denotes the number of samples used in calculation of causality measures. The vertical axis denotes ratios of causality measures $T_{C_1 \to C_2}/T_{C_2 \to C_1}$. The shown values are means over 100 iterations. The smaller $T_{C_1 \to C_2}/T_{C_2 \to C_1}$ is, the more clearly



(a) Overall view.



(b) Closeup of (a).

Figure 2: The performance of causality measures when varying the number of samples (from numerical data to numerical data).

the presence or absence of causality is distinguished. We observe that the proposed measure, CTE, and GC successfully indicate the presence or absence of causality because these three methods do not quantize numerical time series. Note that the graphs of these methods overlap. Though the three methods do not differ in performance on the pure numerical data sets, the proposed measure has the advantages that it describes the dynamics of time series, and represents the strength of causal influences as the number of bits. As to TE, the score of $r = 2$ is better than $r = 8$. This is because, as the number of bins increases, more samples is needed for the exact estimation of probability distributions. In fact, $T_{C_2 \to C_1}$ of $r = 2$ converges near $N = 1000$ whereas that of $r = 8$ does not even at $N = 10000$. TE is very sensitive to the parameter r.

4.1.2 *From symbolic data to numerical data*

In the second experiment, we examine performance of causality quantification from symbolic to numerical time series, using trivariate data sets. Each data set consists of two symbolic time series S_1, S_2, and one numerical time series C. S_1, S_2, and C are ordered by the following model:

$$p(S_{1,t} = 0 | S_{1,t-1} = 0) = 0.9$$
$$p(S_{1,t} = 1 | S_{1,t-1} = 0) = 0.1$$

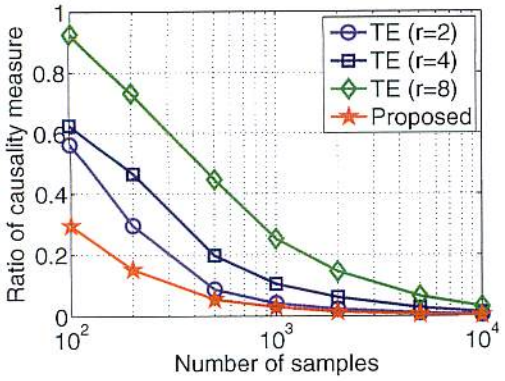


Figure 3: The performance of causality measures when varying the number of samples (from symbolic data to numerical data).

$$p(S_{1,t} = 1|S_{1,t-1} = 1) = 0.9$$
$$p(S_{1,t} = 0|S_{1,t-1} = 1) = 0.1$$

$$p(S_{2,t} = 0|S_{2,t-1} = 0) = 0.9$$
$$p(S_{2,t} = 1|S_{2,t-1} = 0) = 0.1$$
$$p(S_{2,t} = 1|S_{2,t-1} = 1) = 0.9$$
$$p(S_{2,t} = 0|S_{2,t-1} = 1) = 0.1$$

$$c_t = \begin{cases} 0.3c_{t-1} + \epsilon_t & \text{if } s_{1,t-1} = 0 \\ 0.9c_{t-1} + \epsilon_t & \text{if } s_{1,t-1} = 1. \end{cases}$$

where ϵ_t is white noise with zero mean and unit variance, $\epsilon_t \sim N(0,1)$. In this model, C is influenced by S_1, not by S_2. We generate 100 independent data sets from random initial values, and compare the proposed measure with transfer entropy (TE) of the parameter $r = 2$, $r = 4$, or $r = 8$. We set two embedding dimension parameters to $k = 1$ and $l = 1$ according to the model of data sets. We change the number of samples $N = 100$, 200, 500, 1000, 2000, 5000, and 10000.

Figure 3 shows the result. The horizontal axis denotes the number of samples used in calculation of causality measures. The vertical axis denotes ratios of causality measures $T_{S_2 \to C}/T_{S_1 \to C}$ (means over 100 repetitions). The smaller $T_{S_2 \to C}/T_{S_1 \to C}$ is, the more clearly the presence or absence of causality is distinguished. As evident from Figure 3, the proposed measure performs better than TE because the proposed method does not quantize numerical time series, on the other hand, TE is very sensitive to quantization.

4.1.3 *From numerical data to symbolic data*

In the third experiment, we examine the ability of the proposed method to calculate causality from numerical to symbolic time series. We use trivariate data sets, consisting of two numerical time series C_1, C_2, and one symbolic time series S. C_1, C_2, and S are created using the following model:

$$c_{1,t} = \epsilon_{1,t}$$
$$c_{2,t} = \epsilon_{2,t}$$

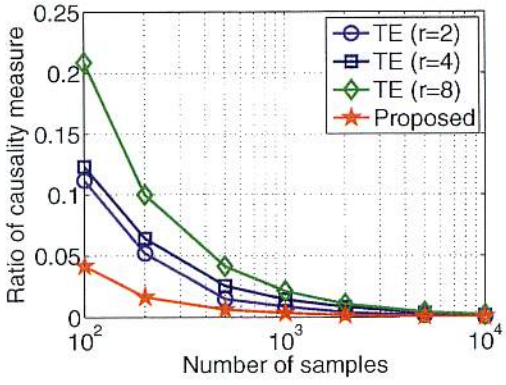


Figure 4: The performance of causality measures when varying the number of samples (from numerical data to symbolic data).

$$s_t = \begin{cases} s_{t-1} & \text{if } c_{1,t-1} > 1 \\ 1 & \text{if } c_{1,t-1} \leq 1 \text{ and } s_{t-1} = 0 \\ 0 & \text{if } c_{1,t-1} \leq 1 \text{ and } s_{t-1} = 1. \end{cases}$$

where $\epsilon_{1,t}$ and $\epsilon_{2,t}$ are white noises with zero mean and unit variance, $\epsilon_{1,t} \sim N(0,1), \epsilon_{2,t} \sim N(0,1)$. In this model, S is influenced by C_1, not by C_2. We generate 100 independent data sets from random initial values, and compare the proposed measure with the transfer entropy (TE) using parameter values $r = 2$, 4, or 8. We set the two embedding dimension parameters to $k = 1$ and $l = 1$. We change the number of samples $N = 100$, 200, 500, 1000, 2000, 5000, and 10000.

The result is displayed in Figure 4. The horizontal axis denotes the number of samples used in calculation of causality measures, and the vertical axis denotes ratios of causality measures $T_{C_2 \to S}/T_{C_1 \to S}$ (means over 100 iterations). The proposed method still performs better than TE.

4.2 Structuring and Modeling

4.2.1 *Mixed symbolic-numerical data*

In the fourth experiment, we examine the performance of our structuring and modeling methods using a model, consisting of four numerical variables (C_1, C_2, C_3, and C_4) and two symbolic symbolic variables (S_1 and S_2). Those six variables have the causal relationship depicted in Figure 5. All of these are generated by simple Markov processes. S_1 and S_2 are subject to the following transition probabilities:

$$p(S_{1,t} = 0|S_{1,t-1} = 0) = 0.9$$
$$p(S_{1,t} = 1|S_{1,t-1} = 0) = 0.1$$
$$p(S_{1,t} = 1|S_{1,t-1} = 1) = 0.9$$
$$p(S_{1,t} = 0|S_{1,t-1} = 1) = 0.1$$

$$p(S_{2,t} = 0|S_{2,t-1} = 0) = 0.9$$
$$p(S_{2,t} = 1|S_{2,t-1} = 0) = 0.1$$
$$p(S_{2,t} = 1|S_{2,t-1} = 1) = 0.9$$
$$p(S_{2,t} = 0|S_{2,t-1} = 1) = 0.1.$$

Variations of C_1, C_2, C_3, and C_4 are generated by the fol-

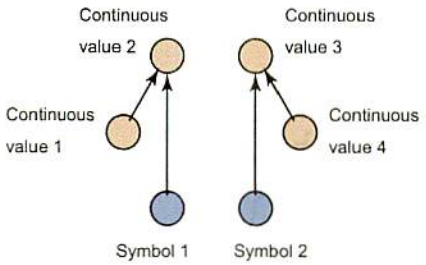


Figure 5: Causal relationship of the model which consists of both symbolic and numerical time series.

lowing formula:

$$c_{1,t} = 0.3c_{1,t-1} + \epsilon_{1,t}$$

$$c_{2,t} = \begin{cases} 0.3c_{2,t-1} + 0.3c_{1,t-1} + \epsilon_{2,t} & \text{if } s_{1,t-1} = 0 \\ 0.9c_{2,t-1} + 0.3c_{1,t-1} + \epsilon_{2,t} & \text{if } s_{1,t-1} = 1 \end{cases}$$

$$c_{3,t} = \begin{cases} 0.3c_{3,t-1} + 0.3c_{4,t-1} + \epsilon_{3,t} & \text{if } s_{2,t-1} = 0 \\ 0.9c_{3,t-1} + 0.3c_{4,t-1} + \epsilon_{3,t} & \text{if } s_{2,t-1} = 1 \end{cases}$$

$$c_{4,t} = 0.3c_{4,t-1} + \epsilon_{4,t}$$

where $\epsilon_{1,t}$, $\epsilon_{2,t}$, $\epsilon_{3,t}$, and $\epsilon_{4,t}$ are white noises with zero mean and 0.1 variance, $\epsilon_{1,t} \sim N(0, 0.1)$, $\epsilon_{2,t} \sim N(0, 0.1)$, $\epsilon_{3,t} \sim N(0, 0.1)$, $\epsilon_{4,t} \sim N(0, 0.1)$. We generate 100 independent data sets from random initial values. We set the two embedding dimension parameters to $k = 1$ and $l = 1$ as indicated by the model used to construct the data sets.

First, we examine the performance of our structuring method with 100 data sets. We structure each data set changing a threshold of calculated causality measures and compute mean of recall, precision, and f-measure by threshold, changing the number of samples $N = 100$, 200, 300, 500, 700, and 1000. We let P denote the number of causal links that can be correctly estimated, Q denote the number of causal links that is present in the original structure of Figure 5, and R denote the number of causal links that is present in the estimated structure. Then, the recall and the precision are defined as: Recall $= P/Q$, Precison $= P/R$. Because of tradeoff between these two scores, the total performance is evaluated by f-measure: F-measure $= \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$. Figure 6 shows relationships between recall and precision. Each line of Figure 6 ends halfway because, if all causality measures is weaker than the threshold used in the filtering phase, no links will be estimated, so $P = R = 0$. Figure 7 displays the relationship between maximum f-measure and the number of samples used. We observe that our method successfully discovered the structure of a comparatively simple data set, even though it includes both symbolic and numerical data.

Next, we compare the prediction accuracy of the modeling methods by using cross validation with 100 data sets. We use the learned models to predict values of numerical variables at a given time step by looking at the values at the previous time step. We compare three methods: (1) AR model of each variable, which disregards causal influences, (2) switching VAR model, which switches VAR models according to the combination of states of the two symbolic variables, (3) our modeling method. In this experiment, we change the number of samples $N = 100$, 200, 300, 500, 700, and 1000. The result is displayed in Figure 8. The horizontal axis denotes the number of samples. The vertical axis denotes the mean squared errors of predicted values. As evident from Figure 8, our method makes better predictions than AR because our method takes into account causal relationships among variables. Compared to switching VAR, our method performs better when the number of samples is small. This result confirms that it is possible to reduce the number of samples needed for modeling multivariate time series by combining bivariate models.

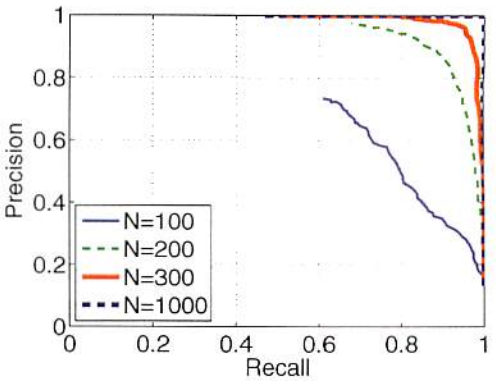


Figure 6: Recall-precision graph of the proposed structuring method on data which consists of both symbolic and numerical time series.

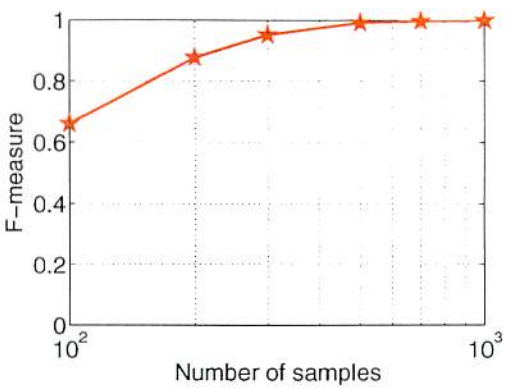


Figure 7: F-measure of the proposed structuring method on data which consists of both symbolic and numerical time series, when varying the number of samples.

4.2.2 *Genetic regulatory data*

In the fifth experiment, we conduct the same experiment using genetic regulatory model of [12], whose variables are all numerical. This causal links between genes in this model have a more complex network structure than the models used in the previous experiments. The values of gene expression levels are updated by two processes. First, the values at each step are produced by the following formula:

$$Y_t - Y_{t-1} = A(Y_{t-1} - T) + \epsilon.$$

where Y_t is a vector which represent the expression levels of all genes at time t. Second, the expression levels are

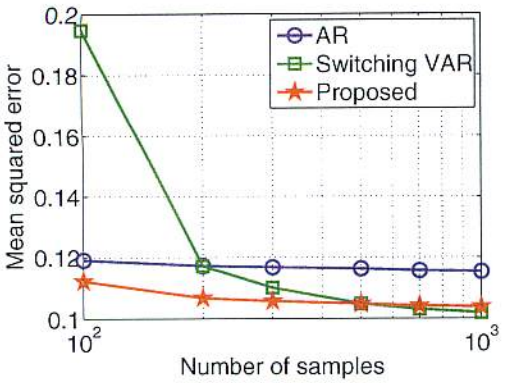


Figure 8: The performance of AR, VAR, and the proposed modeling method on data which consists of both symbolic and numerical time series.

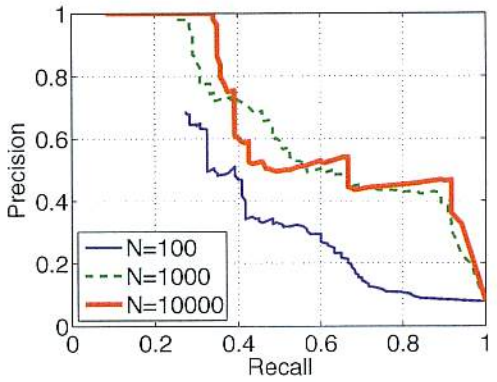


Figure 9: Recall-precision graph on the genetic regulatory data.

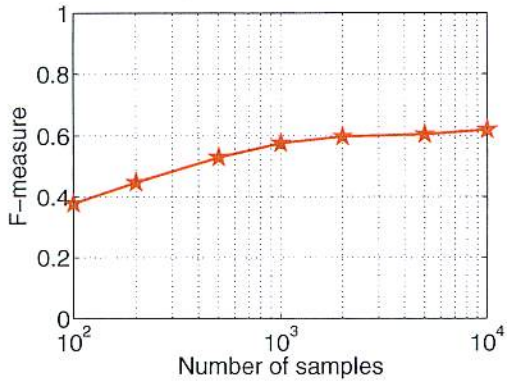


Figure 10: F-measure on the genetic regulatory data when varying the number of samples.

restricted by a floor and ceiling function to range from 0 to 100. The matrix A represents the causal relationships of all genes. If the (i, j)-th element of A is nonzero, the i-th gene has some influence on the j-th gene. The vector T is a constant vector. All elements of T are set to 50, the median value between the maximum and minimum. If a gene expression is at a level above 50, the regulatory effect on the genes occurs as specified in A; the higher the expression level, the stronger the specified effect. In contrast, if a gene expression is at a level below 50, its effect is in the opposite direction of that specified in A; the lower the expression level, the stronger the opposite effect. The ϵ term models biological noise and is drawn uniformly at random from the range -10 to 10. In this experiment, we use 10 different networks with 12 or 13 genes, and generate 10 independent data sets for each network using random initial values, for a total of 100 data sets. We set the two embedding dimension parameters to $k = 1$ and $l = 1$.

First, we examine the performance of our structuring method with 10 data sets from one network selected randomly. We structure each data set changing a threshold of calculated causality measures and compute mean of recall, precision, and f-measure by threshold, changing the number of samples $N = 100$, 200, 500, 1000, 2000, 5000, and 10000. Figure 9 displays relationships between recall and precision. Each line of Figure 9 ends halfway, for the same reason as mentioned above. Figure 10 shows the relationship between maximum f-measure and the number of samples. As evident from Figure 10, the f-measure does not change very much as the number of samples increases from 1000 to 10000. The reason is that our method extracts not only true causal relationships but also indirect relationships such as grandparent-grandchild ($X \rightarrow Y$ and $Y \rightarrow Z \implies X \rightarrow Z$) or brothers ($X \rightarrow Y$ and $X \rightarrow Z \implies Y \leftrightarrow Z$). Our structuring method cannot determine whether a relationship is direct or indirect because the causality measure consistently reports that one variable has a causal influence on the other. In order to apply our method to real world engineering problems, we need to determine how this disadvantage affects actual predictions.

Next, we conduct an experiment to compare the different modeling methods. We use the learned models to predict gene expression levels at a given time step using the information from the previous time step. We compare three methods: (1) AR model of each variable, (2) VAR model, (3) our modeling method. We conduct cross validation for 10 data sets generated from each structure, using a number of samples $N = 1000$, 2000, 3000, 5000, 7000, and 10000. The result is displayed in Figure 11. The horizontal axis denotes the number of samples, and the vertical axis denotes the mean squared errors of predicted values. Figure 11 shows a similar result for an experiment on data consisting of mixed symbolic and numerical time series.

We discuss the difference in effects of indirect relationships between structuring and modeling. In the structuring experiment on genetic regulatory data, indirect relationships were extracted because the causality measures between the corresponding variables were moderately large. The result shows that indirect relationships (grandparent-grandchild or sibling) prevent the estimation of the true structure. However, those indirect relationships may have a positive influence on the modeling results. The proposed causality measures to numerical time series (Equation (10) and Equation (12)) are based on the prediction accuracy of regression models. That is, indirect relationships that create high causality measures can help to make good predictions. This means that, if variables which connect observed variables are not observed, our modeling method can implicitly model effect of

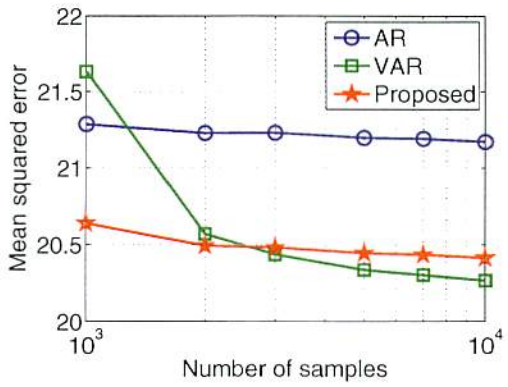


Figure 11: The performance of AR, VAR, and the proposed modeling method on the genetic regulatory data.

Table 2: Comparing mean of relative errors for AR and the proposed modeling method on the Japanese stock price data.

Method	$k,l=1$	$k,l=2$	$k,l=3$
Autoregression	2.40%	3.13%	3.71%
Proposed method	2.37%	2.84%	3.26%

the unobserved variables. Indirect relationships are bad for our structuring method but good for our modeling method. Note that our structuring method can estimate simple structures even if structures have both symbolic and numerical variables (Figure 7).

4.2.3 *Real stock price data*

Finally, we applied the proposed modeling method to the real-world data, which consists of the 225 closing stock prices used in the calculation of the Nikkei Stock Average from September 3, 2007 to March 25, 2008. We predict the closing prices by the day, using data of the past 20 business days (about 1 month). In this experiment, we change the embedding dimension k,l from 1 to 3. We cannot use VAR of 225 variables to estimate a model from only 20 time steps. We compare the prediction accuracy of the proposed method with AR.

Table 2 shows mean relative errors. The proposed method forecasts better than AR in three kinds of parameters. There are two possible reasons to explain this result. The first reason is that stock prices have direct causal influences on each other. If there are causal relationships, then naturally our method can make better predictions by exploiting those relationships. The second possible reason is that several stock prices have common causal sources. If there are indirect relationships between stock prices, our method can model those causal sources implicitly and make good predictions.

This result demonstrates that the proposed modeling method is effective even when it is applied to the real world data.

5. CONCLUSIONS

In this paper, we proposed a causality quantification method which can handle mixed symbolic and numerical time series data in a seamless fashion. The proposed method has better causality detection performance than existing methods, especially when the number of samples is small. In addition, we proposed structuring and modeling methods using our causality measures. Because our causality quantification method can deal with two types of time series, our modeling methods can also deal with data sets that include both. Experimental results demonstrated that our methods can perform well even if the number of samples is small. In modeling experiments, it is showed that our method can implicitly model unobserved variables which connect observed variables.

6. REFERENCES

[1] S. P. Charles, B. C. Bates, I. N. Smith, and J. P. Hughes. Statistical downscaling of daily precipitation from observed and modelled atmospheric fields. *Hydrological Processes*, 18(8):1373–1394, May 2004.

[2] Y. Chen, S. L. Bressler, and M. Ding. Frequency decomposition of conditional granger causality and application to multivariate neural field potential data. *Journal of Neuroscience Methods*, 150(2):228–237, January 2006.

[3] C. Dritsaki and M. Dritsaki-Bargiota. The causal relationship between stock, credit market and economic development: An empirical evidence for greece. *Economic Change and Restructuring*, 38(1):113–127, March 2005.

[4] G. Elliott, C. W. J. Granger, and A. Timmermann. *Handbook of Economic Forecasting, Volume1*. North Holland, Amsterdam, 2006.

[5] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, July 1969.

[6] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46, March 2007.

[7] W.-C. Hong, P.-F. Pai, S.-L. Yang, and R. Theng. Highway traffic forecasting by support vector regression model with tabu search algorithms. In *Proc. of the International Joint Conference on Neural Networks*, pages 1617–1621, October 2006.

[8] A. Kaiesr and T. Schreiber. Information transfer in continuous processes. *Physica D*, 166:43–62, June 2002.

[9] M. Lungarella, K. Ishiguro, Y. Kuniyoshi, and N. Otsu. Methods for quantifying the causal structure of bivariate time series. *International Journal of Bifurcation and Chaos*, 17(3):903–921, March 2007.

[10] T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, January 2000.

[11] J. H. Stock and M. W. Watson. New indexes of coincident and leading economic indicators. *NBER Working Paper*, pages 351–393, 1989.

[12] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis. Advances to bayesian network inference for generating causal networks from observational biological data. *Bioinformatics*, 20(18):3594–3603, December 2004.