8-1-2014

Causal Graph Discovery For Hydrological Time Series Knowledge Discovery

Piraporn Jangyodsuk

Dong-Jun Seo

Jean Gao

CAUSAL GRAPH DISCOVERY FOR HYDROLOGICAL TIME SERIES KNOWLEDGE DISCOVERY

PIRAPORN JANGYODSUK (1), DONG-JUN SEO (2), GAO JEAN (1)
(1): Computer Science and Engineering Department, University of Texas at Arlington, 500 UTA Boulevard, Arlington, TX 76019
(2): Department of Civil Engineering, 416 Yates Street, Arlington, TX 76019

Causal relationship delivers important information in hydrological study to explore the causes of abnormal hydrology phenomena such as drought and flood, which will help improving our prediction and response ability to natural disasters. In this paper, we propose a new approach, mutual information causal (MI-Causal), for causal relationship discovery in time series data, which embodies the advantages of existing approaches and overcomes the limitations to satisfy the need from hydrological domain. Every time series data contain information from its causes and this information can be transferred to its effect. From this idea, we can create a causal graph in the same conditions based approaches but do not require high number of independency tests and causal relation calculation. Furthermore, the lead time is reported in the discovery of causal relationship, which is missing current causality research. The experimental results from both synthetic and real time hydrological data show that our proposed method outperforms regression approaches and Bayesian based approaches.

CAUSAL DISCOVERY ALGORITHM

Definition of causality

Causal inference or causal relationship discovery is an important task in hydrological study to explore the causes of abnormal hydrology phenomena such as drought and flood, which will help improving our prediction and response ability to natural disasters. Different from generic causality study where causal relation discovery is sufficient, for extreme hydrological situation prediction and modeling, we need not only to construct a causal graph to reveal the contributing factors, but also to provide the lead time of each cause to its effect. Lead time is the time difference between the occurrence of lead and effect.

There are two widely used causality definitions, one is from Granger [1] and the other is from Pearl [2]. Granger's causality has been widely used in hydrology, economics and finance. Granger utilizes linear auto-regressive model to identify causal relationships between time series. The major disadvantage of Granger's causality is its limitation to linear model. Research has been carried out using either Structure Equation Modeling (SEM) approach such as Shimizu *et al.* [3], Zhang *et al.* [4], Lacerda *et al.* [5], and Mooij *et al.* [6] or regression approach such as Haufe *et al.* [7], Hoyer *et al.* [8], Liu *et al.* [9], and Lazano *et al.* [10].

In Pearl's causality, causal relationships are represented by a directed acyclic graph (DAG) and conditional dependencies between variables. This definition gives more flexibility and is not limited to linearity. The pioneer works are the SGS algorithm from Verma *et al.* [11] and the PC algorithm from Spirtes *et al.* [12]. Many works, following Pearl's idea, contributes improvement to the PC algorithm, such as Wang *et al.* [13], Claassen *et al.* [14], VanderWeele *et al.* [15], and Ramsahai [16]. Our proposed, Mutual Information Causal (MI-Causal) algorithm will be based on Pearl's definition because of its ability to identify non-linear causal model.

Since the construction of Pearl's causal graphs needs a tool for conditional dependency measurement, mutual information is chosen for this task. We also find that mutual information's chain rule can reduce the number of dependency tests needed the graph construction. As a result, this algorithm is much faster than other methods.

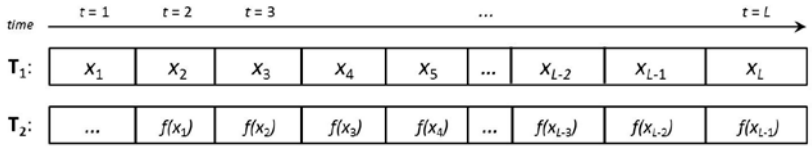

Figure 1. A causal relationship between two time series, T_1 and T_2, where T_1 is the cause of T_2 and leading time equals one step.

Although mutual information cannot identify the direction of causality, the sequence of events can be exploited. For time series data, without loss of generality, we can exploit the sequence of events by assuming cause must occur before its effect. As such, our proposed algorithm, MI-Causal, can find causal relationships and causal directions by using mutual information and time information. In the following sections, the proposed algorithm is explained in detail. Then, the experimental result on hydrological data will be discussed.

MI-Causal algorithm

Due to the exploitation of time information, the MI-Causal has a parameter, called leading time. The leading time of a cause to its effect is the time difference from the occurrence of that cause to the occurrence of its effect. Leading time and causal relationship in MI-Causal are based on the following three assumptions:

1. **Cause must occur before its effects.** The leading time of a cause to any of its effects must be greater than zero.
2. **The leading time of a cause-effect pair is consistent.** Leading time does not change or drift as time goes. As shown in Figure 1, the value of T_2 in every time step is correlated to the value of T_1 from the previous time step with a causal function, $f(\bullet)$.
3. **Causal relationship is not limited to linearity.** The causal function, $f(\bullet)$, in Figure 1 can be linear or non-linear function.

The MI-Causal returns a causal graph in which the leading time of each cause-effect pair is presented on its causal edge. An example of causal graph, shown in Figure 2, represents a system with seven time series, X_1, X_2, …, X_7, and four causal relationships as follows,

1. X_2 from the previous step is the cause of X_1.
2. X_5 from the previous three steps and X_7 from the previous two steps are causes of X_2.

3. X_6 from the previous step is a cause of X_3.

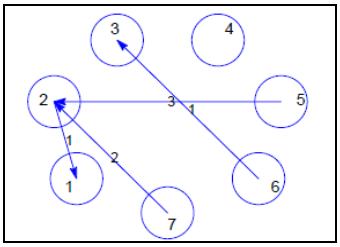

Figure 2. An example of a causal graph returned from the MI-Causal on a data set with seven time series and four causal relationships.

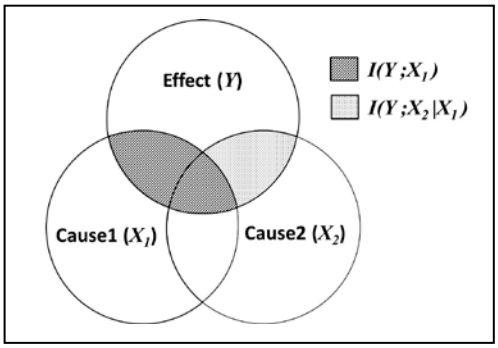

Figure 3. Mutual information between an effect, Y, and its two causes, X_1 and X_2. ($N = 2$).

$$I(X_{i:N};Y) = \sum_{i=i}^{N} I(X_i;Y \mid X_{1:i-1}) . \quad (1)$$

According to mutual information's chain rule in Eq. (1), mutual information between an effect, Y, and its N causes, $X_{1:N}$, can be incrementally constructed from mutual information between Y and a cause, X_i, conditioned on all previously found causes, $X_{1:i-1}$. Thus, the MI-Causal focuses on one effect and searches for one cause at a time using the maximum conditional mutual information, conditioned on all discovered causes of that effect, as the selection criteria. The MI-Causal continues searching for causes until all the time series left do not provide more information about the effect. Then, it starts searching for causes of another effect.

For a data set with M time series, X_i where $1 \le i \le M$, and the length of time series is L, data points in each time series are ordered from the farthest past, $t = 1$, to present, $t = L$, so the time series X_i can be written as

$$X_i = \left(x_i^{\,1}, x_i^{\,2}, \ldots, x_i^{\,L}\right). \quad (2)$$

Because this algorithm also finds the leading time from a cause to its effect, the original M time series are pre-processed to create $M \times (P+1)$ time series representing M time series at leading time from 0 to P. The MI-Causal creates a set, V_G, containing the pre-processed time series. Each of the pre-processed time series is cut from the original time series as follows

$$X_i^{\,p} = \left(x_i^{\,P-p+1}, x_i^{\,P-p}, ..., x_i^{\,L-p}\right), \quad (3)$$

where $1 \le i \le M$ and $0 \le p \le P$. The time series, X^p_i, represents the time series X_i whose leading time is p. Time series with zero leading time, X^0_i where $1 \le i \le M$, are separated from V_G and put into the set of effects, V_E.

Then, for each effect in V_E, the MI-Causal searches for one cause in each iteration and adds it to the set of causes C_i using conditional mutual information and mutual information's chain rule. The set of causes C_i contains all discovered causes of the effect X^0_i and $C_i \subseteq V_G$.

After discovering causes for all effects, the MI-Causal creates a causal graph consists of M nodes representing the original M time series. For each pair of causal relationships, the MI-Causal draws a directed edge from the cause to its effect with a number denoting the leading time on that edge.

Table 1. 32 Variables of the OHD-NOAA's hydrological data set

	Variable	Layer	Description
1	accmax	1	Maximum water equivalent since snow began to accumulate (mm)
2	adimpc	1	Additional impervious area water content (mm)
3	evap	1	Actual evapotranspiration (mm per dt)
4	liqw	1	Liquid water storage (mm)
5	lzfpc	1	Lower zone primary free water content (mm)
6	lzfsc	1	Lower zone supplemental water content (mm)
7	lztwc	1	Lower zone tension water content (mm)
8	pevap	1	Potential evapotranspiration (mm per dt)
9	rain	1	Rainfall forcing (mm per dt)
10	rmlt	1	Rain plus melt dept (mm)
11	runoff	1	Surface flow component (mm per dt)
12-15	smliq	1-4	Unfrozen volumetric soil moisture at Noah defined layers where layer 1 is the top layer
16	sndpt	1	Snow depth (mm)
17	snow	1	Snowfall forcing (mm per dt)
18	snowfrac	1	Snow cover fraction, dimensionless
19-22	soilm	1-4	Total volumetric soil moisture at Noah defined layers where layer 1 is the top layer
23-26	soilt	1-4	Soil temperature at Noah defined layers where layer 1 is the top layer
27	subflow	1	Subsurface flow component (mm per dt)
28	swe	1	Snow water equivalent (mm)
29	tem	1	Air temperature forcing (°C)
30	twe	1	Total water equivalent (mm)
31	uzfwc	1	Upper zone free water content (mm)
32	uztwc	1	Upper zone tension water content (mm)

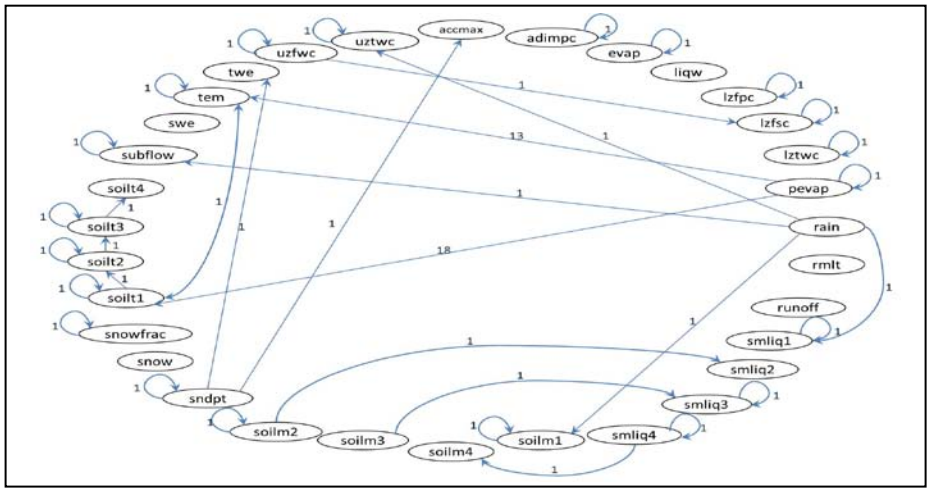

Figure 4. The causal graph from MI-Causal on OHD-NOAA's hydrological data set

Table 2. Causal relationships discovered by MI-Causal on OHD-NOAA's hydrological data set

Effect	Cause	Leading day(s)
accmax	sndpt	1
adimpc	adimpc	1
evap	evap	1
lzfpc	lzfpc	1
lzfsc	lzfsc	1
lzfsc	uzfwc	1
lztwc	lztwc	1
pevap	pevap	1
smliq1	rain	1
smliq1	smliq1	1
smliq2	soilm2	1
smliq3	smliq3	1
smliq4	smliq4	1
sndpt	sndpt	1
snowfrac	snowfrac	1
soilm1	rain	1
soilm1	soilm1	1
soilm2	soilm2	1

Effect	Cause	Leading day(s)
soilm3	smliq3	1
soilm4	smliq4	1
soilt1	pevap	18
soilt1	tem	1
soilt2	soilt1	1
soilt2	soilt2	1
soilt3	soilt3	1
soilt4	soilt3	1
subflow	rain	1
subflow	subflow	1
tem	pevap	13
tem	soilt1	1
tem	tem	1
twe	sndpt	1
uzfwc	uzfwc	1
uztwc	rain	1
uztwc	uztwc	1

EXPERIMENTAL RESULTS

In collaboration with the Office of Hydrologic Development at the National Oceanic and Atmospheric Administration (OHD-NOAA), the 32-variable hydrological data set is available for experiment. The description of variables can be found in Table 1. Each variable consists of one time series per 4x4 km^2 area of the US except Alaska and Hawaii. The data was collected every 6 hours from January 2, 1979 to December 31, 2008. In this experiment, Arlington, TX

was chosen. The result is expected to correspond with the hydrologic cycle and the characteristics of the area, such as type of soil, and climate.

From the results in Figure 4 and Table 2, most of the result causal relationships seem reasonable comparing to the hydrologic cycle. The snow depth (sndpt) from the previous day affecting the maximum water since snow began to accumulate (accmax) is quite reasonable. Water in the top soil layer (smliq1 and soilm1) is affected by rain and water it had from the previous day. Rain is the cause of the sub-surface water (subflow) and the upper zone tension water content (uztwc). Small effect from snow comparing to rain is consistent with the climate in the chosen area because snow usually falls only one to two days per year. The causal relationships of soil temperature at four different layers are also according to nature. The temperature of the top soil layer is affected by the potential evapotranspiration (pevap), the air temperature (tem) and the top soil temperature (soilt1) itself. Other layers below that are affected by itself and the temperature from the soil layer right above it.

One interesting result is that the cycle between tem and soilt1. The air temperature (tem) from the previous day causes the current temperature of the top soil layer (soilt1) and vice versa. This cycle may be caused by a confounder of these two variables, for example the intensity of sun light, wind velocity, etc.

Some variables depend only on its value from the previous day, for example smliq3, smliq4, soilm2, soilm3, soilm4. These variables represent the amount of water contained in deeper soil layers. Intuitively, they should also be affected by water from the soil layer right above them just like the soil temperature. Actually, this result illustrates the ability of soil in the observed area to retain water.

CONCLUSION

In this paper, a new algorithm, called MI-Causal, for discovering quantitative, efficient causal relationship in time series hydrological data is proposed. This algorithm is based on Pearl's causality due to the fact that causal relationship in real world is not limited to linearity. Since this algorithm is designed for time series data, the information about the sequential order of events can be exploited to identify the direction of causality. Based on the mutual information's chain rule and information theory, this algorithm uses conditional mutual information to find causal relationships.

The experimental results on OHD-NOAA's hydrological data set show that MI-Causal can discover causal relationships without conflict with the hydrologic cycle. In the future, this algorithm can be improved and extended to spatial-temporal causal relationship discovery and applied to important hydrological problems, such as discovering causes of drought and flood, and predicting future drought and flood.

REFERENCES

[1] Granger C., “Testing for Causality: A Personal Viewpoint”, Journal of Economic Dynamics and Control, Vol.2, No. 1 (1980), pp 329-352.

[2] Pearl J., “Causality: Models, Reasoning and Inference”, 2nd edition, Cambridge University Press, (2009).
[3] Shimizu S., Hoyer P.O., Hyvärinen A. and Kerminen A., “A Linear Non-Gaussian Acyclic Model for Causal Discovery”, Journal of Machine Learning Research, Vol.7, No. 12 (2006), pp 2003-2030.
[4] Zhang K. and Hoyer P.O., “Distinguishing Causes from Effects using Nonlinear Acyclic Causal Models”, Journal of Machine Learning Research Workshop and Conference Proceedings, Vol.6 (2010), pp 157-164.
[5] Lacerda G., Spirtes P., Ramsey J, and Hoyer P.O., “Discovering Cyclic Causal Models by Independent Components Analysis”, Proc. of the 24th Annual Conference on Uncertainty in Artificial Intelligence, Helsinki, Finland, (2008), pp 366-374.
[6] Mooij J.M., Janzing D., Heskes T. and Schölkopf B., “On Causal Discovery with Cyclic Additive Noise Models”, Advances in Neural Information Processing Systems, Vol. 24, (2011), pp 639-647.
[7] Haufe S., Müller K-R., Nolte G., and Krämer N., “Sparse Causal Discovery in Multivariate Time Series”, Journal of Machine Learning Research Workshop and Conference Proceedings, Vol.6 (2010), pp 97-106.
[8] Hoyer P.O., Janzing D., Mooij J.M., Peters J. and Schölkopf B., “Nonlinear Causal Discovery with Additive Noise Models”, Advances in Neural Information Processing Systems, Vol. 21, (2009), pp 689-696.
[9] Liu Y., Niculescu-Mizil A., Lozano A, and Lu Y., “Learning Temporal Causal Graphs for Relational Time-Series Analysis”, Proc. of the 27th International Conference on Machine Learning, Haifa, Israel, (2010), pp 687-694.
[10] Lozano A.C., Li H., Niculescu-Mizil A., Liu Y., Perlich C., Hosking J., and Abe N., “Spatial-temporal Causal Modeling for Climate Change Attribution”, Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, (2009), pp 587-596.
[11] Verma T. and Pearl J., “Equivalence and Synthesis of Causal Models”, Proc. of the 6th Annual Conference on Uncertainty in Artificial Intelligence, Cambridge, MA, (1990), pp 220-227.
[12] Spirtes P. and Glymour C., “An Algorithm for Fast Recovery of Sparse Causal Graphs”, Social Science Computer Review, Vol.9, No. 1 (1991), pp 62-72.
[13] Wang Z., and Chan L., “Using Bayesian Network Learning Algorithm to Discover Causal Relations in Multivariate Time Series”, IEEE 11th International Conference on Data Mining, Vancouver, Canada, (2011), pp 814-823.
[14] Claassen T. and Heskes T., “A Bayesian Approach to Constraint based Causal Inference”, Proc. of the 28th Annual Conference on Uncertainty in Artificial Intelligence, Catalina Island, California, (2012), pp 207-216.
[15] VanderWeele T.J., and Robins J.M., “Properties of Monotonic Effects on Directed Acyclic Graphs”, Journal Machine Learning Research, Vol.10, No. 3 (2009), pp 699-718.
[16] Ramsahai R.R., “Causal Bounds and Observable Constraints for Non-deterministic Models”, Journal Machine Learning Research, Vol.13, No. 3 (2012), pp 829-848.