Learning Temporal Causal Graphs for Relational Time-Series Analysis *

Yan Liu
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
liuya@us.ibm.com

Alexandru Niculescu-Mizil
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
anicule@us.ibm.com

Aurelie Lozano
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598
aclozano@us.ibm.com

Identifying causality in multivariate time-series data is a topic or significant interest due to its many applications in fields as diverse as neuroscience, economics, climate science, and microbiology to name a few.

In many applications, one is presented with multiple multivariate time-series rather than a single one. For instance, climate and meteorological data are collected at a variety of different location on the globe, with different instruments and measurement protocols; gene expression microarray data are collected for different species, under different conditions, and by different labs. Moreover, one can usually identify relationships between these different time-series, such as time-series being collected at neighboring locations in the case of climate data, or microarray experiments being conducted on the same species, or under the same conditions. These relationships define a "relational graph" among the different time-series where related time-series are connected by an edge.

Given such relational time-series data, one faces the question of how to infer the causal structure for each time-series in manner that is more flexible than requiring a common causal graph for all time-series, while, at the same time, avoiding the brittleness due to data scarcity if one were to independently learn a different causal structure for each time-series. At a first approximation, the solution we propose in this paper can be viewed as finding a middle ground between these two extremes by partitioning the time-series into subsets of that share the same causal structure, and pooling the observations from all the time-series in a subset to learn more robust causal graphs.

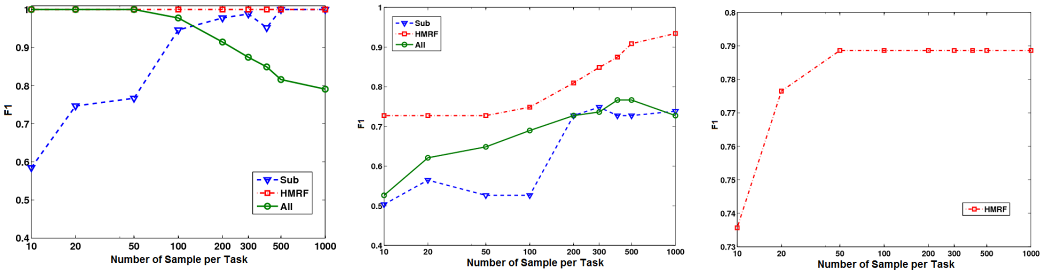
Specifically, we define a hidden Markov Random Field (hMRF) on the relational graph, and assign a hidden state to each node (time-series). Nodes that share the same state in the hMRF will have the same causal graph. The particular notion of causality we use in this paper is that of "Granger Causality" [Gra80], which has proven useful as an *operational* notion of causality in time series analysis in the area of econometrics, and has become popular in many other fields. Granger causality is based on the intuition that a cause should necessarily precede its effect, and in particular that, if a variable causally affects another, then the past values of the former should be helpful in predicting the future values of the latter. Following [ALA07] we use an L1 regularized regression approach to efficiently detect Granger causality in multivariate time-series.

While we described the model in terms of hard partitioning of the time-series to ease understanding, in reality the model maintains a soft partitioning throughout learning. This leads to a form of transfer learning when inferring the causal graphs associated with different states. This makes our model applicable even in situations where partitioning the time-series might not seem appropriate.

We test our model on a synthetic dataset where the relational graph is a 10x10 grid, and data is generated from two hidden states. Results are presented in Figure 1 and show higher performance when compared to learning a single graph for all time-series (ALL), and learning independent graphs for each time-series (SUB).

We also apply the model to a climate modeling problem. The learned causal graphs for three

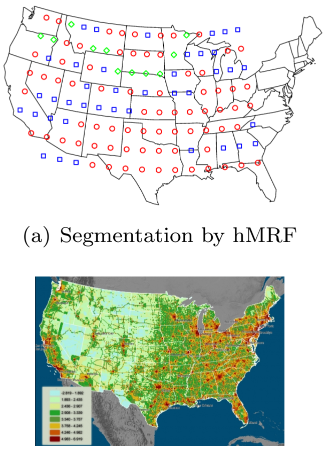
[*] Topic: Learning Algorithms; Preference: Oral



(a) F_1 scores of State 1 Graph (b) F_1 scores of State 2 Graph (c) State prediction

Figure 1: Comparison results on Simulation Data.

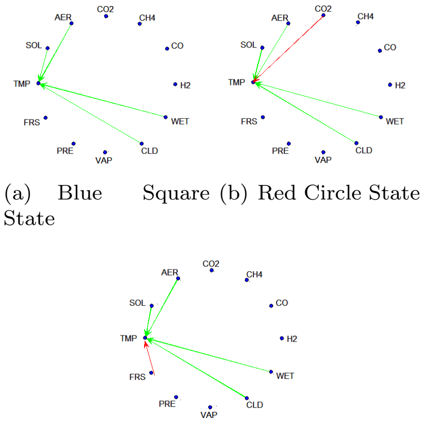
hidden states are shown in Figure 3 (we only show temperature for better visualization). Figure 2(a) shows the assignment of locations to the different states. The results seem reasonable (compared with the US CO2 concentration map in Figure 2(b)) in that the green (diamonds) state corresponds to the mid-north part of the country, where the region is cold and temperature is affected by the number of frost days, the red (circles) state represents the developed regions in the south, west and east of the US, where the CO2 concentration is high enough to influence temperature (i.e. the greenhouse effect), while the blue (squares) state is dominant in central less populated area with less CO2 concentration.



(a) Segmentation by hMRF

(b) Map of US CO2 Concentration(`http://www.purdue.edu/eas/carbon/vulcan/GEarth`)

Figure 2: Predicted states for each location.



(a) Blue Square State (b) Red Circle State

(c) Green Diamond State

Figure 3: Causal graphs learned for each state.

References

[ALA07] A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical granger methods. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (SIGKDD-07)*, 2007.

[Gra80] C. Granger. Testing for causlity: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.