Learning Temporal Causal Graphs for Relational Time-Series Analysis

Yan Liu, Alexandru Niculescu-Mizil, Aurélie Lozano {LIUYA, ANICULE, ACLOZANO}@US.IBM.COM
IBM T.J. Watson Research Center, 1101 Kitchawan Rd, Yorktown Heights, NY 10598, USA

Yong Lu YONG_LU@HMS.HARVARD.EDU
Harvard Medical School, Harvard University, Boston, MA 02115, USA

Abstract

Learning temporal causal graph structures from multivariate time-series data reveals important dependency relationships between current observations and histories, and provides a better understanding of complex systems. In this paper, we examine learning tasks where one is presented with *multiple* multivariate time-series, as well as a relational graph among the different time-series. We propose an L_1 regularized hidden Markov random field regression framework to leverage the information provided by the relational graph and jointly infer more accurate temporal causal structures for all time-series. We test the proposed model on climate modeling and cross-species microarray data analysis applications.

1. Introduction

Identifying causality in multivariate time-series data is a topic of significant interest due to its many applications in fields as diverse as neuroscience (Song et al., 2009), economics (Arnold et al., 2007), climate science (Lozano et al., 2009b), and microbiology (Lozano et al., 2009a).

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.[1]

Given such relational time-series data, one faces the question of how to infer the causal structure for each time-series in a 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 that share the same causal structures, 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 model will have the same causal graph. The particular notion of causality we use in this paper is that of "Granger Causality" (Granger, 1980), 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,

[1] It is important not to confuse the relational graph, which represents relationships among the different time series, with the causal graph, which represents causal relationships among the individual variables within a multivariate time series.

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

then the past values of the former should be helpful in predicting the future values of the latter. Following (Arnold et al., 2007) 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. As we will see in Section 4, this leads to a form of transfer learning when inferring the causal graphs associated with different states, which makes our model applicable even in situations where partitioning the time-series might not seem appropriate.

We test our model on two synthetic datasets, and apply it to climate measurement data and immune response microarray data from multiple species. The experiment results show that our model not only performs better than other alternatives, but also has the capability to provide useful scientific insights.

2. Notations and Assumptions

We are given M multivariate time-series $X^{(1)}, ..., X^{(M)}$. Each time-series consists of a set of observations at N_i consecutive time points: $X^{(i)} = \{\vec{x}_1^{(i)}, \ldots, \vec{x}_{N_i}^{(i)}\}$, and each observation measures P variables: $\vec{x}_t^{(i)} = \{x_{1,t}^{(i)}, \ldots, x_{P,t}^{(i)}\}$. We will use $\vec{x^{(i)}}_{t..t'} = \{\vec{x^{(i)}}_t^T, \vec{x^{(i)}}_{t-1}^T, \ldots, \vec{x^{(i)}}_{t'}^T\}$ to denote the vector composed of the concatenation of all observations between times t' and t. We assume that all time-series measure the same P variables, and that the observations for all time-series are made at regular time intervals. When the context is clear, we will omit the superscript (i). We will also abuse the notation and use x_p to refer to the name of a particular variable in a time-series. Given a relational graph G_r over the M time-series, we write $(i, j) \in G_r$ to denote that time-series $X^{(i)}$ and $X^{(j)}$ are connected by an edge in G_r.

3. Temporal Causal Graphs and Granger Causality

For time series analysis, it is important to reveal the causal dependencies between current and past observations and represent them by *temporal causal graphs*. In this paper we will use Granger Causality, which states that a variable is the cause of another if past values of the former are helpful in predicting the future values of the later.[2] In the case of multivariate time series, a popular approach is to apply regression algorithms with variable selection, where current values of each variable are regressed on past values of all variables up to a maximum "lag" L:

$$x_{p,t} = \sum_{r=1}^{P} \sum_{l=1}^{L} \beta_{p,r,l} \cdot x_{r,t-l} + \varepsilon_p \quad (1)$$

with ε_p distributed $N(0, 1)$ for all $p \in \{1, .., P\}$ and all $t \in \{L + 1, ...N\}$.

The parameters $\beta_{p,r,t}$ are estimated under the sparsifying Laplacian prior. Let β be a $P \times P \cdot L$ matrix whose p^{th} column is $\{\beta_{p,1,1}, .., \beta_{p,P,L}\}$. Then

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \log(\Phi_\lambda(X|\beta))$$

where

$$\Phi_\lambda(X|\beta) = (\lambda/2)^{(N-L)} \exp(-\lambda\|\beta\|_1) \cdot \prod_{t=L+1}^{N} \frac{1}{(2\pi)^{P/2}}$$
$$\times \exp(-\frac{1}{2}(\vec{x}_t - \vec{x}_{(t-1)..(t-L)} \cdot \beta)^T \cdot (\vec{x}_t - \vec{x}_{(t-1)..(t-L)} \cdot \beta)) \quad (2)$$

Standard lasso can be used to obtain $\hat{\beta}$. If any of $\hat{\beta}_{p,r,\cdot}$ is non-zero it means that the past values of variable x_r have a significant effect on the prediction of the variable x_p, hence an arc from x_r to x_p is added in the temporal causal graph.

Recent advances in regularization theory have led to a series of extensions of the original lasso algorithm, such as elastic net (Zou & Hastie, 2005) and group lasso (Yuan & Lin, 2006). These methods have been adapted to temporal graphical modeling with successful applications to biology and climate modeling domains (Lozano et al., 2009b;a), and they can also be readily used in our framework.

4. L_1 Regularized Hidden Markov Random Field Regression

In many applications one is given M multivariate time-series $X^{(1)}, ..., X^{(M)}$ rather than a single one. Often the time-series are not independent, but related as represented by a relational graph G_r. The challenge is how to leverage this rich information and infer more accurately the temporal causal structures for all M time-series.

The basic idea of our approach is to assign to each time-series a hidden state, which determines the parameters β (and thus the temporal causal graph), with

[2] Granger Causality is not meant to be equivalent to true causality, but is merely intended to provide useful information regarding causation.

the hidden states assignments guided by the prior information in the relational graph G_r. To this end, we model the data generating process by a hidden Markov random field over the relational graph G_r which defines a joint probability over time-series $X^{(1)}, ..., X^{(M)}$ and hidden states $s^{(1)}, ..., s^{(M)}$ as follows:

$$P_\lambda(X^{(1)}, \ldots, X^{(M)}, s^{(1)}, \ldots, s^{(M)}|\boldsymbol{\beta}, w) = \frac{1}{Z}\prod_{i=1}^{M}\Phi_\lambda(X^{(i)}|\beta^{(s^{(i)})})\prod_{(i,j)\in G_r}\Phi(s^{(i)}, s^{(j)}|w)$$

The node potentials, $\Phi_\lambda(X^{(i)}|\beta^{(s^{(i)})})$, are the same as those in Section 3 (eq. 2), but now there is a different set of parameters $\beta^{(s)}$ for each state s. The edge potentials $\Phi(s^{(i)}, s^{(j)}|w)$ are defined as $\Phi(s^{(i)}, s^{(j)}|w) = \exp(\sum_{s,s'} w_{ss'}\delta_{ss'}(s^{(i)}, s^{(j)}))$, where δ is the indicator function, i.e. $\delta_{ss'}(s^{(i)}, s^{(j)}) = 1$ if $s^{(i)} = s$ and $s^{(j)} = s'$, and 0 otherwise; $w_{ss'}$ is the parameter to capture the similarity between state s and s', which is similar to the transition probability in HMM. Z is the normalization constant. By our definition of node potentials, the value of Z will only be affected by the edge potentials, i.e. $Z = \sum_{s^{(1)},\ldots,s^{(M)}}\exp(\sum_{(i,j)\in G_r} w_{s^{(i)}s^{(j)}})$.

4.1. Parameter Estimation

There are two sets of parameters in the model, namely $\boldsymbol{\beta}$ and w. Since the values of the state variables $s^{(1)}, \ldots, s^{(M)}$ are not known, EM algorithm (Bilmes, 1998) is applied to estimate the parameters. Note that the expected value of the log likelihood function Q_λ is:

$$\begin{aligned}Q_\lambda &= \sum_{s^{(1)},\ldots,s^{(M)}} P(\{s^{(i)}\}|\{X^{(i)}\}, \widetilde{\boldsymbol{\beta}}, \widetilde{w})\cdot\log P_\lambda(\{X^{(i)}\}, \{s^{(i)}\}|\boldsymbol{\beta}, w)\\ &= \sum_{i=1}^{M}\sum_{s=1}^{S} P(s^{(i)} = s|\{X^{(i)}\}, \widetilde{\boldsymbol{\beta}}, \widetilde{w})\cdot\log(\Phi_\lambda(X^{(i)}|\beta^{(s)}))\\ &+ \sum_{(i,j)\in G_r}\sum_{s,s'=1}^{S} P(s^{(i)} = s, s^{(j)} = s'|\{X^{(i)}\}, \widetilde{\boldsymbol{\beta}}, \widetilde{w})\cdot\log(\Phi(s, s'|w))\\ &- \log(Z)\end{aligned}$$

where S is the number of hidden states. For the M-step, we estimate the values of the parameters $\boldsymbol{\beta}$ and w that maximize Q_λ. Due to the form of Q_λ we have:

$$\begin{aligned}\hat{\beta}^{(s)} &= \operatorname*{argmax}_{\beta^{(s)}}\sum_{i=1}^{M}\widetilde{P}^{(i)}(s)\cdot\log(\Phi(X^{(i)}|\beta^{(s)}))\\ &= \operatorname*{argmax}_{\beta^{(s)}}\sum_{i=1}^{M}\sum_{t=L+1}^{N_i}(\sqrt{\widetilde{P}^{(i)}(s)}\cdot\vec{x}_t^{(i)} - \sqrt{\widetilde{P}^{(i)}(s)}\cdot\vec{x}_{(t-1)..(t-L)}^{(i)}\cdot\beta^{(s)})^T\\ &\times(\sqrt{\widetilde{P}^{(i)}(s)}\cdot\vec{x}_t^{(i)} - \sqrt{\widetilde{P}^{(i)}(s)}\cdot\vec{x}_{(t-1)..(t-L)}^{(i)}\cdot\beta^{(s)})\\ &+\lambda\cdot\|\beta^{(s)}\|_1\cdot\sum_{i=1}^{M}\widetilde{P}^{(i)}(s)\end{aligned}$$

where $\widetilde{P}^{(i)}(s) = P(s^{(i)} = s|\{X^{(i)}\}, \widetilde{\boldsymbol{\beta}}, \widetilde{w})$ is the marginal probability that time-series i belongs to state s under the parameters estimated in the last iteration. The intuition is that the parameters $\hat{\beta}^{(s)}$ for each state s are learned using all the data available, with observations from time-series that are more likely to belong to state s receiving higher weights. The solutions $\hat{\beta}^{(s)}$ are obtained by applying standard lasso (or any other L_1 based regression technique) to the reweighted data.

The fact that the parameters $\hat{\beta}^{(s)}$ for each state are estimated using the same data, but with different weighting, leads to an implicit form of transfer learning, where the parameters corresponding to different states are encouraged to be similar to each other. The transfer of information between states s and s' is regulated by the angle between the vectors $\{\widetilde{P}^{(1)}(s), \ldots, \widetilde{P}^{(M)}(s)\}$ and $\{\widetilde{P}^{(1)}(s'), \ldots, \widetilde{P}^{(M)}(s')\}$. If these two vectors are collinear, then $\hat{\beta}^{(s)}$ and $\hat{\beta}^{(s')}$ will necessarily be identical, while if the two vectors are orthogonal, there will be no transfer between states s and s'. Between these two extremes, the lower the angle, the closer $\hat{\beta}^{(s)}$ and $\hat{\beta}^{(s')}$ will be. While beyond the scope of this paper, exploring the connections between our formalism and other transfer learning approaches is a very interesting direction for future work.

The parameters w are estimated using gradient descent, with the derivative Q_λ with respect to $w_{s,s'}$ being:

$$\frac{\partial Q}{\partial w_{ss'}} = \sum_{(i,j)\in G_r}(P(s^{(i)} = s, s^{(j)} = s'|\{X^{(i)}\}, \widetilde{\beta}, \widetilde{w})\delta(s, s')) - \sum_{(i,j)\in G_r} E[\delta_{ss'}(s^{(i)}, s^{(j)})|\{X^{(i)}\}, \boldsymbol{\beta}, w]$$

For E-step, we use loopy belief propagation (Murphy et al., 1999) to compute the marginal of individual nodes $P(s^{(i)} = s|\{X^{(i)}\}, \widetilde{\boldsymbol{\beta}}, \widetilde{w})$ and edges $P(s^{(i)} = s, s^{(j)} = s'|\{X^{(i)}\}, \widetilde{\beta}, \widetilde{w})$.

Although having a different motivation, the L_1 regularized hMRF model proposed in this section can be viewed as a generalization of the mixture of regression model (Bishop, 2007). Indeed, if G_r is taken to be the empty graph, and we eliminate the L_1 penalty on the parameters $\boldsymbol{\beta}$, then our model reduces to the mixture of regression model.

5. Discussion of Alternative Approaches

Another approach to learning temporal causal models from relational multivariate time-series is to extend the weighted linear regression with L_1 penalty. Specifically, one can define a similarity function, $f(i, j)$, between time-series i and j based on the relational graph

G_r. For example $f(i,j)$ could be defined as an exponentially decaying function of the distance between time-series i and j in G_r. One then infers the temporal causal graphs for time-series i by using the data from all time-series weighted by f:

$$\hat{\beta}^{(i)} = \underset{\beta^{(i)}}{\text{argmin}} \quad \sum_{j=1}^{M} f(i,j)\Big(\sum_{t=L+1}^{N_j} (||\vec{x}_t^{(j)} - \vec{x}_{t-1..t-L}^{(j)}\beta^{(i)}||_2) + \lambda \cdot ||\beta^{(i)}||_1 \cdot \sum_{j=1}^{M} f(i,j)$$

A formulation similar in spirit to the one above has been used by (Song et al., 2009) to learn time-varying temporal graphs. A major disadvantage of this type of models is that they ignore the long-range dependencies. For example, in climate application, London and Seattle are distant in locations, but the two cities might have similar weather patterns and therefore share the same temporal graphs. Another disadvantage is that one has to define the similarity function, which is a difficult task in real applications.

A similar approach to the one above is to state the problem in a regularized transfer learning formulation. Using a similarity function $f(i,j)$ defined as above, one can define a joint loss function $\mathcal{L}$ by imposing a regularizer based on assumption that similar time-series should have similar coefficients. Therefore we have

$$\mathcal{L} = \sum_{i=1}^{M}\sum_{t=L+1}^{N^{(i)}} (\vec{x}_t^{(i)} - \vec{x}_{t-1..t-L}^{(i)}\beta^{(i)})^T(\vec{x}_t^{(i)} - \vec{x}_{t-1..t-L}^{(i)}\beta^{(i)}) + \lambda_1 \sum_{i,j} w_{i,j}(\beta^{(i)} - \beta^{(j)})^T(\beta^{(i)} - \beta^{(j)}) + \lambda_2 \cdot \sum_{i=1}^{M} ||\beta^{(i)}||_1 \quad (3)$$

This approach would suffer from the same disadvantages as the previous one, i.e. the lack of long-range dependency and the need to specify the similarity function $f(i,j)$.

6. Experimental Results

To examine the effectiveness of the proposed algorithm, we conduct experiments on two synthetic datasets, and two real applications: spatial-temporal climate data analysis and cross-species innate immune response analysis.

6.1. Synthetic Data

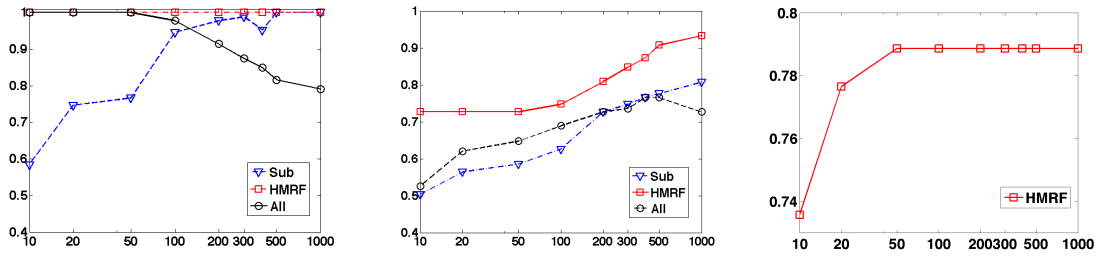
In this experiment, we generate synthetic datasets using an Ising model on a 10×10 grid with coefficients as follows: $w(s,s') = 1$ if $s = s'$ and 0.5 otherwise. We first obtain the state sequence from a Gibbs sampler and then sample the observations associated with each node from linear Gaussian model, whose temporal graphs of 5 variables (p = 5) are determined by its state. More specifically:

Simulation Data I assumes that state 1 corresponds to a sparse causal graph, i.e. the precision matrix is an AR(1) model in which $(\Sigma^{-1})_{ii} = 1$, $(\Sigma^{-1})_{i,i-1} = (\Sigma^{-1})_{i-1,i} = 0.5$, and state 2 corresponds to a dense causal graph with precision matrix as $(\sum^{-1})_{ii} = 2$, $(\sum^{-1})_{ii'} = 1$ where $i \neq i'$ (examples also used in (Yuan & Lin, 2007; Friedman et al., 2008)). The goal of conducting experiments on this dataset is to verify whether our algorithm is able to recover the sparse causal graph from the data with mixed observations from the dense causal graph.

Simulation Data II data are generated from causal graphs with similar graph structures: state 1 shares the same distribution as state 1 in Simulation Data I, and state 2 has precision matrix of $(\sum^{-1})_{ii} = 1$, $(\sum^{-1})_{i,i-1} = (\sum^{-1})_{i-1,i} = 0.5$, $(\sum^{-1})_{i,i-2} = (\sum^{-1})_{i-2,i} = 0.25$ (see Figure 2 (1a) and (2a) for graph structure). Our goal in this experiment is to examine whether the algorithm can recover the true graphs when the underlying two causal graphs are similar. This setting mimics the real applications and better showcases the advantage of our model. Therefore we will focus more on this dataset.

In the experiment, we generate $N = 500$ samples for each node. The penalty term λ is selected by BIC. We compare the performance of our model with two other baselines: one is aggregating all the data from different nodes to learn a single graph (referred to as "ALL"), and the other is learning a graph using data from individual node only (referred to as "SUB"). We evaluate their performance on structure learning using the F1-measure, i.e. viewing the causal modeling problem as that of predicting the inclusion of the edges in the true graph (Silva et al., 2006). The results are presented in Table 1 and show that hMRF achieves statistically significant better performance than competing methods on both Simulation Data I and II.

To analyze how the sample size influence performance, we vary N from 10, 20, 50, 100, 200, 300, 400, 500 to 1000. In addition to the structure prediction performance, for hMRF model we also examine the performance of predicting the true state associated with each time-series with F1-measure. We repeat each experiment 30 times and report the average in Figure 1. Figure 1 shows that: (1) as expected, the performance of both structure learning and state prediction by hMRF model increases when we increase the number of sample size per node. Notice that the performance is very



(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 II: (a) performance (in F_1) of structure learning by competing methods for component graph associated with state 1. x axis: number of sample size per node, y axis: F_1 score. (b) performance (in F_1) of structure learning by competing methods for component graph associated with state 2. (c) performance of state prediction. x axis: number of sample size per task, y axis: F_1 scores.

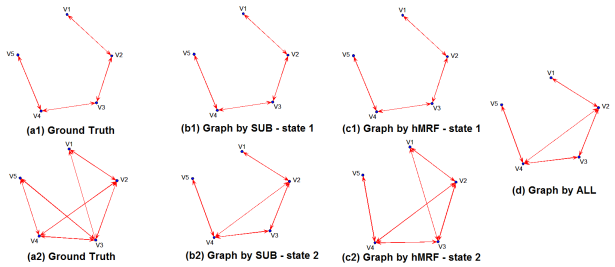


Figure 2. Example of learned graphs by different methods: (a1, b1) true component graphs of state 1 and state 2; (a2, b2) learned graphs by baseline method SUB; (a3, b3) learned graphs by hMRF; (4) learned graph by baseline method ALL

reasonable (around 80.0 in F1) even when there are only 200-300 examples per node, which demonstrates the power of hMRF to leverage useful information from other nodes. (2) When the sample size is small, ALL predicts perfectly for the causal graph of state 1 because it represents the strongest causal connections in both graphs. As the sample size increases, additional dependencies present in the causal graph (i.e. those corresponding to state 2) become more apparent, and ALL is forced to add additional arcs to account for these dependencies. This will lower the precision and hence the F_1 score for state 1, explaining the degradation of performance in Figure 1(a). The hMRF on the other hand has the flexibility to learn different graphs for the two states, so its performance does not degrade. (3) The performance by baseline method SUB is the least desirable given its low F1 scores and instability. Figure 2 shows an example of learned graph by different methods when the sample size N is 200. As we can see, hMRF produces graphs closest to the ground truth.

Table 1. Comparison results of structure learning on simulation data (sample size per node = 500)

Algorithm	Simulation I (F_1)		Simulation II (F_1)	
	State 1	State 2	State 1	State 2
hMRF	**0.9251**	**0.7577**	**1.000**	**0.9085**
ALL	0.8191	0.6388	0.8160	0.7664
SUB	0.7429	0.7273	1.000	0.7753

6.2. Applications to Climate Modeling

Climate change is one of the most critical socio-technological issues mankind faces in the new century (IPCC, 2007). An important challenge in understanding climate change is to uncover the causal relationships between the various climate observations and forcing factors, which can be of either natural or anthropogenic (human) origin. We use monthly measurements of climate and climate forcing variables, including temperature (TMP), precipitation (PRE), vapor (VAP), cloud cover (CLD), wet days (WET), and frost days (FRS), green house gases (Methane (CH4), Carbon Dioxide (CO2), Hydrogen (H2) and carbon monoxide (CO), solar radiation (SOL) and aerosols (AER) from CRU (`http://www.cru.uea.ac.uk/cru/data`), NOAA (`http://www.esrl.noaa.gov/gmd/dv/ftpdata.html`), and NCDC (`http://rredc.nrel.gov/solar/old_data/nsrdb/`). The relational multivariate time series data span 13 years (from 1990 to 2002) on a 2.5×2.5 degree grid that covers most of the United States. [3]

In the experiment, we use a maximal lag of 4 and vary the number of hidden states from 2 to 4. Here we show the causal graph learned from the climate data with 3 hidden states in Figure 4 (since more states will result in almost redundant graphs). For better

[3]The climate data are already interpolated on a grid.

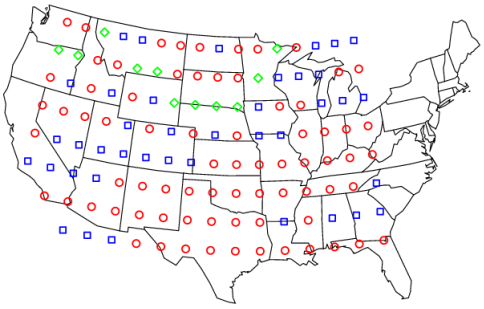
visualization, we only show the edges with in-links to temperature (TMP) since it is the most concerning factor in global warming. We are comparing three different approaches: ALL (results shown in Figure 4(a)), SUB (Figure 4(b) and 4(c)), and hMRF (Figure 4(d)- 4(f)). In terms of overall graph structure, we can see that the two edges unique to particular locations (e.g. Figure 4(c)) are missing in the graph by ALL (Figure 4(a)) while the graphs by SUB are not stable, i.e. the causal graphs change significantly from location to location, probably due to the insufficient number of observations at each location. We can also see that the causal graphs of different states learned by hMRF share great similarity and the common part (green edges) seems very reasonable, i.e. the temperature is mostly decided by solar radiance (SOL), but is also affected by cloud, wet days and aerosol. To understand the difference in the causal graphs between different states, we show the US locations associated with the causal graphs (determined by the state assignment for each location by our model) in Figure 3(a). The results seem reasonable (compared with the US CO2 concentration map in Figure 3(b)) in that the green (diamonds) state corresponds to the mid-north part of the country, where the region is cold and the 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.

Notice that in this experiment we assume the climate temporal graph remains the same over time but varies across locations. The time-invariant assumption may not be true over a long time period, such as millions of years, but is reasonable for a short period, e.g. 20 years or so, as is in our experiment.

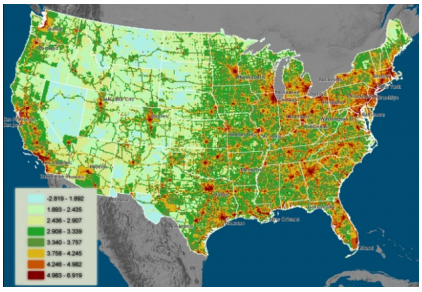
6.3. Applications to Cross-Species Gene Regulatory Network Discovery

Most multicellular organisms rely on their immune system to defend against the infection from a multitude of pathogens. To understand the roles and possible interplays between different types of immune cells, it is important to identify both the common responses of different immune cells, as well as responses unique to a certain cell type or species.

In this experiment, we applied our algorithm to recovering the causal graphs between genes in immune response system across different cell types. Specifically, we use the time-series microarray datasets on



(a) Segmentation by hMRF

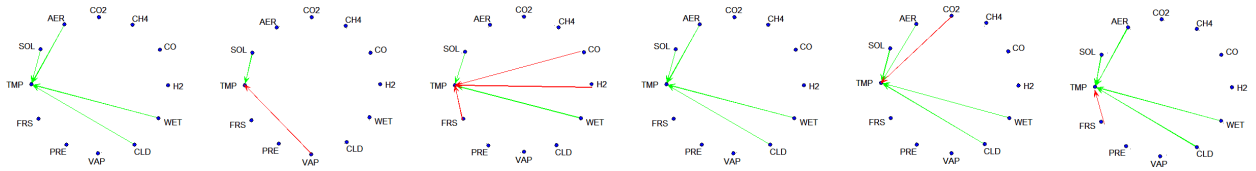


(b) Map of US CO2 concentration(http://www.purdue.edu/eas/carbon/vulcan/GEarth)

Figure 3. Predicted labels of underlying hidden states for each location. Green diamonds: state 1; Red circle: state 2; Blue square: state 3

innate immune response of human and mouse in (Lu et al., 2010). The gene expression experiments were done on macrophages (M) and dendritic cells (DC) in humans and mice, under the infection of two types of bacteria, Gram-positive (P) and Gram-negative (N). The 39 microarray experiments are grouped into seven datasets based on cell types, and referred to as "human.DC.N", "human.DC.P", "human.M.N", "human.M.P", "mouse.DC.N", "mouse.M.N" and "mouse.M.P" respectively. In order to explore information sharing across species/cell types, we select the common regulatory genes that either themselves or their orthologs can be found in all the datasets, resulting in a set of 789 common genes across species.

We construct the relational graph as follows: there is an edge between the experiments on the same species since many genes should exhibit similar regulatory relations across experiments; there is also an edge between the same cell type across different species because some of the genes may share similar regulatory functions as their orthologs. We varied the number of hidden states from 2 to 7 and set to 4 based on Bayesian information criterion (BIC) score. We ran experiments for a maximum lag of 2.



(a) ALL (b) SUB: Loc-1 (c) SUB: Loc-2 (d) Blue State (e) Red State (f) Green State

Figure 4. Causal graph learned by ALL (A), SUB (B-C) and hMRF (D-F). The location in (B) is [30.475,-114.75] and the location in (C) is [42.975, -99.75]. Green edge: common edges shared by (D-F); red edge: additional edges unique to each state.

Table 2. Percentage of overlap between bootstrap graphs and original graphs

Cell Type	% of Overlap	Cell Type	% of Overlap
human.DC.N	0.7572	mouse.DC.N	0.7713
human.DC.P	0.7569	mouse.M.N	0.7510
human.M.N	0.7541	mouse.M.P	0.7527
human.M.P	0.7575		

Table 3. Top 10 genes by out-degress in the learned graphs by different methods

hMRF		ALL		SUB	
EntrezID	Edge #	EntrezID	Edges #	EntrezID	Edges #
FTH1	182	PTGS2	170	ACVR2A	224
IL1R2	110	ACVR2A	157	VPS45	179
B2M	104	CXCL10	154	PTGS2	175
VIM	75	DUSP2	145	NFE2	172
CXCL10	74	PPIB	140	FTH1	168
RPL37	71	FMO1	136	FOS	167
LSP1	70	PECAM1	135	PECAM1	162
DRA	68	NR4A1	132	FPR1	160
MSN	66	MCM4	131	CDC6	157
CD14	60	IL7R	128	LSP1	140

First, we evaluate the performance of our method by applying the Bootstrap procedure (Davison & Hinkley, 1997). More precisely, given the original lagged data matrix, we randomly draw B datasets by sampling with replacement the rows of the original data matrix, so that each dataset has the same number of rows as the original lagged data matrix. We then apply our method to each of the B bootstrap datasets. Comparing the "original graph" (i.e. the graph obtained by using the original dataset) with the "bootstrap graphs" (i.e. those obtained using the bootstrap datasets) allows us to get a measure of confidence in the causal relationships identified in the "original graph". In particular, for each causal relationship identified in the "original graph", we can get confidence in that relationship by counting the number of times it appears in the "bootstrap graphs". As shown in Table 2, the causal relationships identified by our method in the "original graph" appear on the average 75.2% of the time in the "bootstrap graphs", which demonstrates that hMRF produces stable graphs.

We also compare the learned graphs generated by hMRF with the graphs learned by the other two baselines. Compared with SUB, our method has major advantages since some of the datasets, for example Human.DC.P and Mouse.M.N, have very limited number of time-series observations (1-2), and no reasonable graph can be generated by SUB. For fair comparison (in favor of the SUB method), we choose the dataset with the largest number of time-series observations, i.e. Human.M.N, to compare the results of different methods. One general observation is that the graphs by ALL (31,218 edges) and SUB (14,346 edges) are much denser than that by hMRF (7458 edges) while the three graphs share 4,071 edges in common. Sparse graphs do not necessarily suggest better performance, but around 54.6% commonality suggests that hMRF is able to provide a graph with much higher precisions. Figure 3 lists an example of 10 genes with the highest number of out-degrees in the learned graphs. From the results, we can see that hMRF not only shares some top-ranked genes with the other two algorithms, such as CXCL10, but also uniquely identifies important immune genes, such as IL1R2, HLA-DRA, and CD14, as well as B2M (Beta-2-microglobulin), which is a serum protein found in association with the major histocompatibility complex (MHC) class I heavy chain on the surface of nearly all nucleated cells; MSN (Moesin), which is localized to filopodia and other membranous protrusions that are important for cell-cell recognition and functions as cross-linkers between plasma membranes and actin-based cytoskeletons.

7. Conclusion

In this paper, we examine the task of learning temporal causal graphs from relational multivariate time-series data, which are available in an increasing number of applications. To narrow the gap between the rich information available in the data and existing solutions

in the literature, we proposed an L_1 penalized hidden Markov random field regression framework. It defines a Markov random field over the relational graph and jointly learn the causal graphs for all time-series so that it can leverage the information from all the data and learn more robust causal graphs. Experiment results on several datasets show that our model consistently outperforms alternative approaches. In addition, our algorithm provides interesting scientific insights which might lead to better understanding of causal relations in climate change and immune response.

For future work, we are interested in leveraging the insights of other work on causality, i.e. those with much stronger classes of constraints (e.g. (Spirtes et al., 1993)) for better temporal causal modeling. We are also interested in examining the relationships between our model and transfer learning, especially its connections to weighted linear regression and regularized transfer learning with graph Laplacian, as described in the paper.

Acknowledgement

We sincerely thank Naoki Abe, Hongfei Li, Jonathan Hosking, Rick Lawrence and Piotr Mirowski for discussing the ideas in the paper. We thank anonymous reviewers for their valuable suggestions.

References

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

Bilmes, Jeff. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, 1998.

Bishop, Christopher M. *Pattern Recognition and Machine Learning*. Springer, 2007.

Davison, A. C. and Hinkley, D. V. *Bootstrap Methods and Their Application (Cambridge Series in Statistical and Probabilistic Mathematics , No 1)*. Cambridge University Press, 1997.

Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. Sparse inverse covariance estimation with the graphical lasso. *Biostat*, 9(3):432–441, 2008.

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

IPCC. Climate change 2007 - the physical science basis working group i contribution to the fourth assessment report of the ipcc intergovernmental panel on climate change. *IPCC Fourth Assessment Report on scientific aspects of climate change for researchers, students, and policymakers*, 2007.

Lozano, A., Abe, N., Liu, Y., and Rosset, S. Grouped graphical granger modeling for gene expression regulatory networks discovery. In *Proceedings of International Conference on Intelligent Systems for Molecular Biology (ISMB-09)*, 2009a.

Lozano, A., Li, H., Niculescu-Mizil, A., Liu, Y., Perlich, C., Hosking, J., and Abe., N. Spatial-temporal causal modeling for climate change attribution. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (SIGKDD-09)*, 2009b.

Lu, Y., Mahony, S., Benos, P., Rosenfeld, R., Simon, I., Breeden, L., and Bar-Joseph, Z. Combined analysis reveals a core set of cycling genes. *Genome Biology*, 8(7):R146, 2010.

Murphy, Kevin P., Weiss, Yair, and Jordan, Michael I. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of Uncertainty in AI*, pp. 467–475, 1999.

Silva, Ricardo, Scheine, Richard, Glymour, Clark, and Spirtes, Peter. Learning the structure of linear latent variable models. *J. Mach. Learn. Res.*, 7:191–246, 2006. ISSN 1532-4435.

Song, Le, Kolar, Mladen, and Xing, Eric. Time-varying dynamic bayesian networks. In *Advances in Neural Information Processing Systems 22*, pp. 1732–1740. 2009.

Spirtes, P., Glymour, C., and Scheines, R. Causation, prediction, and search. *Lecture Notes in Statistics 81*, 1993.

Yuan, Ming and Lin, Yi. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

Yuan, Ming and Lin, Yi. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Zou, Hui and Hastie, Trevor. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.