

Learning Maximum Lag for Grouped Graphical Granger Models

Amit Dhurandhar

IBM T.J. Watson, Yorktown Heights, NY, USA, 10598

adhuran@us.ibm.com

Abstract—Temporal causal modeling has been a highly active research area in the last few decades. Temporal or time series data arises in a wide array of application domains ranging from medicine to finance. Deciphering the causal relationships between the various time series can be critical in understanding and consequently, enhancing the efficacy of the underlying processes in these domains. Grouped graphical modeling methods such as Granger methods provide an efficient alternative for finding out such dependencies. A key parameter which affects the performance of these methods is the maximum lag. The maximum lag specifies the extent to which one has to look into the past to predict the future. A smaller than required value of the lag will result in missing important dependencies while an excessively large value of the lag will increase the computational complexity alongwith the addition of noisy dependencies. In this paper, we propose a novel approach for estimating this key parameter efficiently. One of the primary advantages of this approach is that it can, in a principled manner, incorporate prior knowledge of dependencies that are known to exist between certain pairs of time series out of the entire set and use this information to estimate the lag for the entire set. This ability to extrapolate the lag from a known subset to the entire set, in order to get better estimates of the overall lag efficiently, makes such an approach attractive in practice.

Keywords—lag; granger; modeling

I. INTRODUCTION

Weather forecasting, business intelligence, supply chain management are a few but important application areas in which we observe (multiple) time series data. Given these multiple time series, the goal in most of these domains is to use the available time series data of the past to make accurate predictions of future events and trends. In addition to this primary goal, an important task is also to identify causal relationships between time series wherein, data from one time series significantly helps in making predictions about another time series. This problem of identifying causal relationships between various time series is a highly active research area [1], [2], [3].

Graphical modeling techniques which use Bayesian networks and other causal networks [4], [5], [6], [7] have been considered as a viable option for modeling causality in the past. Statistical tests [8] such as specific hypothesis tests have also been designed to identify causality between the various temporal variables. Both the graphical methods as well as the statistical tests mentioned above are focused on deciphering causal relationships between temporal variables and not between time series. In other words, these techniques

are unable to answer questions of the form, "does time series A cause time series B ?". Granger causality [1], a widely accepted notion of causality in econometrics, tries to answer exactly this question. In particular, it says that time series A causes B , if the current observations in A and B together, predict the future observations in B significantly more accurately, than the predictions obtained by using just the current observations in B . Recently, there has been a surge of methods that combine this notion of causality with regression algorithms [9], [2] namely; Group Lasso, Group boosting to name a important few. These methods fall under a class of methods commonly referred to as Grouped Graphical Granger Methods (GGGM). The primary advantage of using these methods as against performing the Granger test for every pair of potentially causally related time series is their scalability. If the number of time series is large, which is likely to be the case in most real life problems, then performing the Granger test for each pair individually can be tedious.

The GGGM have a parameter called the maximum lag, which is a key parameter that needs to be specified in order to decipher the appropriate causal relationships. The maximum lag for a set of time series signifies the number of time units one must look into the past to make accurate predictions of current and future events. Accurately estimating this parameter is critical to the overall performance of the GGGM since, a smaller than required value of the lag will result in missing important causal relationships while an excessively large value of the lag will increase the computational complexity alongwith the addition of noisy dependencies. A standard approach to estimate this parameter is to try different values for the lag and then to choose a value where the error in regression is significantly lower than a smaller value of the lag but not significantly worse than that for higher values of the lag. In other words, choosing the smallest value of the lag after which the learning rate more or less flattens out. One of the main issues with this approach is that, it cannot incorporate additional prior information that the user might have. Moreover, if we get large lags then we cannot be sure if the lag we have is the true value or a noisy value obtained because of overfitting (since, there are many variables).

In many real circumstances depending on the domain, we might have information about the shape of the distribution that the lags might follow as well as knowledge about certain

pairs of time series (with their lags) that we know for sure are causally related and we want to somehow leverage this information to get a better estimate of the maximum lag for the entire set. A striking example of such a scenario is in business intelligence where we may have different key performance indicators (KPI) (as time series) for a store such as; sales amount, sales quantity, promotion quantity, inventory age and sales age. We know here for sure that more the promotion quantity (i.e. quantity of goods that is on sale or has some rewards associated with it) more will be the sales quantity (i.e. amount sold) which will lead to higher sales amount. Hence, we know that promotion quantity has a causal relationship to sales quantity and both have a causal relationship with sales amount. In many cases we may even know the lag for these causally related pairs. In other cases, we may know with reasonable confidence that the lag is most likely to be small (say around or below 2 weeks) since, the specific promotions were run for approximately that period, in which case we may assume the lags are drawn from a distribution resembling an exponential. We do not know however, the relationships between the other pairs of time series and we want to find out other existing relationships using the information we have. Another example is supply chain where KPI's such as goods received by a retailer from a manufacturer and goods ordered by a retailer would be causally related but the lags here are more likely to be normally distributed with a moderately high mean than exponentially, since there can be a reasonable time delay from when the order is made to the time that the shipment is received. Other problems such as gene regulatory network discovery, weather forecasting also may contain this kind of additional information, which if modeled appropriately can aid in better estimating the maximum lag leading to performance improvements in causal discovery for GGGM.

In this paper, we propose an approach based on order statistics that can use this additional information in a principled manner and provide accurate estimates for the maximum lag. One of the primary advantages of this approach is that it is highly efficient and can be used as a quick diagnostic to get an estimate of the maximum lag. Our method as we will see in the experimental section tends to give accurate estimates of the maximum lag even when the lags are correlated.

The rest of the paper is organized as follows. In Section II, we take a closer look at Granger causality and GGGM. In Section III, we formally describe the basic setting and provide a list of desirable properties that a reasonable approach should be able to successfully incorporate. In Section IV, we first state and justify the modeling decisions we make. We then derive the estimator for the maximum lag. If the lags for the known causally related pairs are not given, we give a simple and efficient algorithm for finding these lags. In Section V, we perform synthetic and real data experiments to test the performance of the estimator. We discuss limitations

of the current approach and promising future directions in section VI.

II. PRELIMINARIES

In this section we discuss Granger causality and a class of efficient methods namely, GGGM, that can be used to decipher causal relationships when the number of time series is large.

A. Granger Causality

Clive Granger, a Nobel prize winning economist, gave an operational definition of causality around three decades back [1]. This notion of causality is commonly referred to as Granger causality. Loosely speaking, he stated that time series X "causes" time series Y , if future values of Y predicted using past values of X and Y is significantly more accurate, than just using past values of Y . In other words, X plays a substantial role in predicting Y .

Formally, let $X = \{x_t\}_{t=1}^L$ and $Y = \{y_t\}_{t=1}^L$ be the variables of the time series X and Y (of length L) respectively. Let d denote the maximum lag for this pair of time series¹. The Granger test then performs two linear regressions which are given by,

$$y_t \approx \sum_{i=1}^d e_i y_{t-i} + \sum_{i=1}^d f_i x_{t-i} \quad (1)$$

$$y_t \approx \sum_{i=1}^d e_i y_{t-i} \quad (2)$$

and checks to see if the estimates obtained from equation 1 are more accurate than those obtained from equation 2 in a statistically significant way. If so, then X is said to Granger cause Y .

B. Grouped Graphical Granger Methods

GGGM are a class of methods, that perform non-linear regression with group variable selection, in order to figure out existing causal relationships between different time series. These methods efficiently decipher causations when given a large set of time series and are shown to be more desirable than their pairwise counterparts [3]. Some examples of the regression algorithms that these methods use are the Group Lasso [9], the recently introduced Group Boosting and Adaptive Group Boosting [2]. Given a set of N time series $\{X^1, X^2, \dots, X^N\}$ these methods report for each $i \in \{1, 2, \dots, N\}$ the time series $\{X^j : j \neq i, j \in \{1, 2, \dots, N\}\}$ that "Granger cause" time series X_i . We will now provide details about the Group Lasso since, we deploy this method for causal discovery in the experiments. Notice that the focus of the paper is finding out the maximum lag for these models and not to compare these models and hence we choose a

¹Notice that the value of d greatly affects output of the Granger test

single method that is commonly used in practice for our experiments.

In a particular run of the GGM using Group Lasso, assume that X^i is the time series we want to regress in terms of $X^1, \dots, X^N$. Given a lag of d and L being the length of the time series (or the maximum length considered), let $Y = (x_L^i, x_{L-1}^i, \dots, x_{1+d}^i)^T$ denote the vector of variables to be regressed belonging to time series X^i . Let $X = (Z_1, \dots, Z_{L-d})^T$, where $Z_i = (x_{L-i}^1, \dots, x_{L-d-i+1}^1, \dots, x_{L-i}^N, \dots, x_{L-d-i+1}^N)^T$ be a $(L-d) \times dN$ predictor matrix. With this the Group Lasso solves the following objective,

$$\alpha(\lambda) = \underset{\alpha}{\operatorname{argmin}} \|Y - X\alpha\|^2 + \lambda \sum_{i=1}^N \|\alpha_i\|$$

where λ is a regularization parameter, α is a $dN \times 1$ vector of coefficients and α_i is a $d \times 1$ vector containing coefficients from α with the following indices $\{(i-1)d+1, \dots, id\}$. Minimizing the l_2 norm for groups of coefficients forces entire time series to be either relevant or irrelevant.

III. SETUP

We have seen in the previous sections that the maximum lag is a key parameter in GGM. The causations reported can vary significantly for different values of the lag. In this section, we first describe the basic setup. Based on this setup and the problem we are trying to solve, we state some intuitions that a desirable approach should incorporate.

Let there be N time series $X^1, X^2, \dots, X^N$ of length L . Given this there are $T = \frac{N(N-1)}{2}$ pairs which can potentially be causally related. If there are certain pairs that you know for sure are not causally related you can consider T to be the total number of pairs after their removal. Note that there might be a group of time series (3 or more) that may be causally related in which case we consider every pair in that group to be potentially causally related and part of the T pairs. Thus, T is the effective number of pairs that may be causally related. Out of these T pairs assume that for a subset M of these we know for sure that they are causally related. We may in some scenarios even know the lag for each of these M pairs or at least a range in which the lags may lie. Moreover, depending on the domain we would have a fairly good idea of the shape distribution that these lags are likely to follow. Given all this information the goal is to estimate the maximum lag d_T for the T pairs.

Considering the goal and the basic setup it seems reasonable that any good estimator of d_T should capture at least the following intuitions:

- 1) As M approaches T the estimator should increasingly trust the available data in predicting maximum lag, as against the prior distributions specified over the lags.
- 2) Ideally, we should get the exact value of the maximum lag. However, if this is not possible, it is better to

get a reasonably accurate but conservative estimate of the maximum lag than an optimistic one since, in the latter case we will definitely miss important causations, which is not the case with the former. Moreover, a not too conservative estimate will in all likelihood not add many, if any, false positives.

- 3) The distributions allowed for modeling prior information and the ones learned using the data should be able to represent a variety of shapes viz. exponential decreasing, uniform, bell-shaped etc. depending on the application. In addition, the domain should be bounded since, GGM regress a finite number of variables, that is time series of length L .
- 4) The estimated maximum lag for the T pairs of time series should at least be d_M i.e. $d_T \geq d_M$, which is the maximum lag for the M causally related pairs.

IV. MODELING

In this section, we derive the estimator for the maximum lag for the T pairs of time series. Initially, we state and justify certain modeling decisions that we make. We then derive the estimator assuming that we know the lags for the M pairs, in a way that is consistent with the intuitions mentioned in the previous section. Finally, we give an algorithm to figure out the lags for the M pairs if they are not known apriori.

A. Assumptions

We assume that the ratio of the lags to the length of the time series or maximum length considered i.e. $0 \leq \frac{d}{L} \leq 1$, are drawn independently from a beta distribution for all T pairs. Formally, we assume that $\frac{d_i}{L} \sim \beta(a, b) \forall i \in \{1, \dots, T\}$, where d_i is the lag for the i^{th} pair and a, b are the shape parameters of the beta distribution.

We choose the beta distribution since, it captures intuitions mentioned in item 3 in the previous section. The distribution has a bounded domain $([0,1])$ and by varying the parameters (a, b) we can get different shapes as per the application. For example, if we want the lags to be drawn from a flat uniform distribution we can set $a = 1, b = 1$. If we want an exponential kind of behavior we can set a to be small and b to be large depending on the rate of decay we desire. If we want a bell-shaped (normal like) behavior we can set a and b to both have the same large value depending on the variance we desire. Hence, a beta distribution is a good choice if one wants the flexibility to model different shapes.

Though the beta may be able to model different shapes, the assumption that lags are drawn independently may not match reality. It is very much likely that the lags are correlated. For example, more fresh goods stocked-in at a store causes more sales which leads to higher revenue. The lag between stocked-in fresh goods and revenue is in all likelihood positively correlated with the lag between stocked-in and sales and also with the lag between sales

and revenue. Another example is in weather forecasting where a change in temperature may cause a change pressure which results in increased wind speeds. Here the lag between temperature and wind speeds is positively correlated with the lag between temperature and pressure as well as the lag between pressure and wind speed. We hence can see that the lags can be correlated in very realistic situations. However, all of these lags are positively correlated. It is hard to come up with a realistic scenario where a larger lag between a pair of time series leads to smaller lag between some other pair. It is important to note here that we are arguing for the sign of correlation between lags and not between time series. There are many real life scenarios where two time series may be negatively correlated such as increased inventory at a store implies less sales at that store, however when it comes to lags, a realistic situation which leads to negative correlation does not seem easy to find. The reason we mention this is that, if the lags are in fact positively correlated, assuming them to be independent leads to a conservative estimate of d_T [10], which is consistent with intuition 2 mentioned before. In Section VI, we will discuss some preliminary ideas about how one might proceed to model such correlations, but this is part of future work and beyond the scope of this paper.

As we have seen our assumptions are consistent with our intuitions and given these we will derive our estimator of d_T .

B. Derivation

Assume we know the lags for the M pairs of time series and d_M is the maximum lag for these pairs. Let $\beta(a_p, b_p)$ be the distribution over the lags (divided by L , the length of the time series or maximum length considered) based on the users prior knowledge. Given the lags for the M pairs we can learn the parameters for the beta distribution from the data using maximum likelihood estimation (mle). There are no simple closed form formulae to estimate these parameters, however, as for many complicated distributions the values of the parameters have been tabulated as functions of the available sample [11]. Using these tables the mle estimates for the parameters can be found extremely efficiently, that is in $O(M)$ time.

Let $\beta(a_l, b_l)$ be the distribution over the lags (divided by L) learned from the data. Notice that, we are trying to estimate the maximum lag for the T pairs and hence we need to find the distribution of the maximum order statistic given the individual distributions over the lags.

In general, if we have N independently and identically distributed random variables $X_1, \dots, X_N$ with cumulative distribution function (cdf) $F(x)$, the cdf of the k^{th} order statistic $F_k(x)$ [12] is given by,

$$F_k(x) = \sum_{i=k}^N F^i(x)(1 - F(x))^{N-i}$$

Hence in our case, since we are looking for the distribution of the max order statistic, this distribution would be given by $F_N(x) = F^N(x)$. Based on the prior distribution over the lags, the distribution over the maximum lag would be $\beta^T(a_p, b_p)$, while based on the data the distribution would be $\beta^T(a_l, b_l)$. Considering intuition 1 in the previous section, we combine these two sources of information to get a mixture distribution over the maximum lag,

$$\frac{\bar{d}_T}{L} \sim \frac{M}{T} \beta^T(a_l, b_l) + \frac{T-M}{T} \beta^T(a_p, b_p)$$

where $\bar{d}_T$ is our estimator of d_T . Thus a point estimate for the maximum lag could be the mean of this distribution i.e. $\bar{d}_T = L(\frac{M}{T} \mu_l + \frac{T-M}{T} \mu_p)$, where $\mu_i = \text{mean}(\beta^T(a_i, b_i))$. Note however, that intuition 4 is still not captured, since this estimated value may not necessarily be greater than or equal to d_M . Hence, taking this into consideration our final estimator is given by,

$$\bar{d}_T = \max([L(\frac{M}{T} \mu_l + \frac{T-M}{T} \mu_p)], d_M) \quad (3)$$

where $[.]$ indicates rounding to the closest integer. Note that there is no closed form formula for μ_l or μ_p for arbitrary choice of parameters of the beta distribution. However, these means can be estimated efficiently in the following manner. We first create an array containing the cdf values of the corresponding distribution over the max order statistic, for a particular quantization of the domain of the beta distribution. For example, in case the quantization is 100, the array would have 100 cdf values for the following values of the domain, $\{0.01, 0.02, \dots, 1\}$. We then sample from a Uniform(0,1) and using inverse probability integral transform [10] we would map to the closest (or interpolated) value in the domain to get samples from the respective distribution over the max order statistic. Averaging these samples would give us an estimate of the corresponding mean.

Special case: If the lags (divided by L) turn out to be distributed according to an uniform distribution which is given by $\beta(1, 1)$, then the mean of the distribution of the max order statistic has a nice closed form solution. Given T Uniform(0,1) random variables, the distribution of the max order statistic is a $\beta(T, 1)$, whose mean is $\frac{T}{T+1}$. Hence, in this special case we do not need to sample to approximate the mean as we have a simple closed form formula.

C. Estimating lags for the M time series

In the previous subsection we derived an estimator for the maximum lag assuming that we know the lags for the M causally related time series. However, this might always not be true. In fact, in many cases we may know the range in which the lag is likely to lie rather than the exact value. Sometimes we may have no such information, in which case

the range would be the entire time series i.e. $(1, \dots, L)$. In such scenarios we can use the procedure given in Algorithm 1 to find the lag for a pair of causally related time series.

Algorithm 1 Estimating lags for a pair of causally related time series.

Input: Time series X and Y of length L where X causes Y , a range for the lag $(l, \dots, u)$ where $1 \leq l \leq u \leq L$ and error threshold $\epsilon > 0$.

Output: Lag given by d .

Initialize $lower = l$, $upper = u$, $d = 0$

repeat

 Initialize $mid = \lfloor \frac{lower+upper}{2} \rfloor$

 Regress Y in terms of X for a lag $upper$ and let the error of regression be e_u

 Regress Y in terms of X for a lag mid and let the error be e_c

 Regress Y in terms of X for a lag $lower$ and let the error be e_l

if $e_u - e_c \geq \epsilon$ **then**

$lower = mid$

else if $e_c - e_l \geq \epsilon$ **then**

$upper = mid$

else

$d = lower$

end if

until $(d == lower \text{ or } lower == upper)$

$d = lower$

The procedure essentially does a binary search in the given range and outputs a lag d for which the error in regression is significantly lower than that for a smaller lag but almost the same for higher lags. The time complexity of this procedure is $O(\log(u-l)T_R)$, where $\{l, \dots, u\}$ is the range and $O(T_R)$ is the time it takes to perform regression which depends on the method used. The algorithm is thus simple and efficient. It is easy to see that the error of regression cannot increase as the lag increases since, we have a superset of the variables for larger lags of what we have for smaller lags, leading to a better fit to the available data. *The lag that we get from this procedure is very likely to be the true lag (or at least extremely close to it) and not a consequence of overfitting in the case that it is large (i.e. close to L), since we know apriori that the pair under consideration is causally related.* If we did not have that information we could not have made this claim with much conviction. Moreover, using Algorithm 1 to estimate the lags for the entire set would take $O(T \log(u-l)T_R)$ which is more expensive than our proposed solution of using the derived estimator after the lags for the M causally related time series have been deciphered.

Table I

THE TABLE SHOWS THE ERROR I.E. MEAN (ROUNDED TO 3 DECIMALS) $\pm$ 95% CONFIDENCE INTERVAL (ROUNDED TO 2 DECIMALS) OF THE ESTIMATOR UNDER DIFFERENT SETTINGS WHEN THE LAGS FOR THE M CAUSALLY RELATED PAIRS OF TIME SERIES ARE KNOWN.

a, b	M	ρ	Error
1,1	100	0.1	0.003+-0.01
1,1	100	0.5	0.053+-0.07
1,1	100	0.8	0.131+-0.12
1,1	5000	0.1	0
1,1	5000	0.5	0.002+-0.01
1,1	5000	0.8	0.012+-0.05
1,5	100	0.1	0.006+-0.01
1,5	100	0.5	0.061+-0.04
1,5	100	0.8	0.07+-0.06
1,5	5000	0.1	0.004+- 0.01
1,5	5000	0.5	0.006+-0.02
1,5	5000	0.8	0.008+-0.03
5,5	100	0.1	0.004+-0.04
5,5	100	0.5	0.02+-0.06
5,5	100	0.8	0.155+-0.08
5,5	5000	0.1	0.002+-0.03
5,5	5000	0.5	0.01+-0.04
5,5	5000	0.8	0.018+-0.06

D. Time Complexity

From the previous subsections we know that there are two scenarios, the first where the lags for the M causally related time series are known and second where these lags are not known. It is interesting to see what the time complexities are for these two cases and how they compare with standard approaches.

With this, if we know the lags for the M causally related time series then the lag for the entire set can be estimated in $O(M)$ time. This is the case since, parameter estimation for the beta and obtaining the final estimate for the maximum lag using our estimator takes $O(M)$ time. If we do not know these lags then the maximum lag can be estimated in $O(M \log(u-l)T_R)$ time, since Algorithm 1 needs to be run M times. The standard model selection approaches such as cross-validation or those based on regularization, which essentially try out different lags and which are unable to incorporate the available domain knowledge take, $\Omega(T(u-l)T_R)$ time. Hence, our approach is significantly more efficient than these other approaches.

V. EXPERIMENTS

In this section, we perform synthetic and real data experiments in order to evaluate our estimator. As we will see our estimator is robust and gives accurate estimates of the maximum lag even when assumptions such as the lags being independently distributed are violated. In addition, we compare the accuracy of our approach with a standard approach used for these applications namely; 10-fold cross-validation, which as we mentioned earlier is much more computationally intensive.

Table II

THE TABLE SHOWS THE ERROR I.E. MEAN (ROUNDED TO 3 DECIMALS) $\pm$ 95% CONFIDENCE INTERVAL (ROUNDED TO 2 DECIMALS) OF OUR ESTIMATOR AS WELL AS THAT OF 10-FOLD CROSS-VALIDATION (*CV*) UNDER DIFFERENT SETTINGS WHEN THE LAGS FOR THE M CAUSALLY RELATED PAIRS OF TIME SERIES ARE UNKNOWN.

a, b	M	ρ	Error	CV Error
1,1	100	0.1	0.004 $\pm$ 0.01	0.15 $\pm$ 0.01
1,1	100	0.5	0.05 $\pm$ 0.08	0.143 $\pm$ 0.02
1,1	100	0.8	0.135 $\pm$ 0.13	0.121 $\pm$ 0.01
1,1	5000	0.1	0.001	0.15 $\pm$ 0.03
1,1	5000	0.5	0.004 $\pm$ 0.01	0.143 $\pm$ 0.01
1,1	5000	0.8	0.011 $\pm$ 0.03	0.121 $\pm$ 0.02
1,5	100	0.1	0.007 $\pm$ 0.02	0.132 $\pm$ 0.05
1,5	100	0.5	0.066 $\pm$ 0.03	0.113 $\pm$ 0.04
1,5	100	0.8	0.071 $\pm$ 0.06	0.09 $\pm$ 0.04
1,5	5000	0.1	0.005 $\pm$ 0.01	0.132 $\pm$ 0.05
1,5	5000	0.5	0.007 $\pm$ 0.01	0.113 $\pm$ 0.04
1,5	5000	0.8	0.009 $\pm$ 0.04	0.09 $\pm$ 0.04
5,5	100	0.1	0.002 $\pm$ 0.05	0.091 $\pm$ 0.03
5,5	100	0.5	0.02 $\pm$ 0.07	0.087 $\pm$ 0.02
5,5	100	0.8	0.161 $\pm$ 0.06	0.082 $\pm$ 0.01
5,5	5000	0.1	0.001 $\pm$ 0.05	0.091 $\pm$ 0.03
5,5	5000	0.5	0.01 $\pm$ 0.04	0.087 $\pm$ 0.02
5,5	5000	0.8	0.02 $\pm$ 0.05	0.082 $\pm$ 0.01

A. Synthetic Experiments

Through our experiments on synthetic data, we want to find out how our method behaves for different shapes of the lag distribution, for different amounts of pairwise correlation between the lags and for different values of M . In the first set of experiments, we assume that the M lags are given, while in the second set of experiments we assume that only the M pairs of causally related time series are given but the lags are not for the same set of time series. In this case, we generate the time series using the standard Vector Auto-Regression (VAR) model [13] and the regression method used in Algorithm 1 is Group Lasso. An overview of the procedure used to conduct experiments is as follows:

- 1) Let the total number of time series be Q . Sample $T = \frac{Q(Q-1)}{2}$ lags that are correlated from a multivariate beta distribution. The procedure for doing so is as follows [14]: Specify a, b, ρ where a, b are parameters of a univariate beta distribution and ρ is the pairwise correlation between lags. Generate a sample p from $\beta(a, b)$, followed by a sample k from $\text{Binomial}(N, p)$ where $N = \lceil \frac{(a+b)\rho}{1-\rho} \rceil$. Finally, generate T samples from $\beta(a+k, N+b-1)$ and after multiplying by L round the values to the closest integer. Note that these sequentially generated T samples are a single sample from a T variate beta distribution with marginals $\beta(a, b)$ and pairwise correlation ρ .
- 2) Assign each lag to a directed edge on a complete graph

(with no cycles)² uniformly at random. The vertices of this graph denote individual time series.

- 3) Choose a value for $M \leq T$.
- 4) Randomly sample M lags without replacement from the set of T lags. In the first set of experiments, just note the M lags since we assume that they are readily available. In the second set of experiments where we assume the M lags are not known, generate Q time series using the VAR model in accordance with the T lags. The way to accomplish this, is to have d_T (max lag in the T lags) coefficient matrices. An entry (i, j) in the coefficient matrix A_r where $i, j \in \{1, \dots, Q\}$, $r \in \{1, \dots, d_T\}$ is 0, if time series j does not "cause" time series i or if the lag between i and j is less than r , otherwise the entry is a sample from a $N(0, 1)$. For our method, identify the M pairs of time series that correspond to the randomly sampled M lags and find out the lags for these M pairs using Algorithm 1. For 10-fold cross-validation, randomly partition the T pairs of time series into 10 parts and compute the average error in regression (average L_1 loss) for different choices of lags.
- 5) In the first set of experiments where the M lags are known, estimate the maximum lag using our method. In the second set of experiments, estimate the maximum lag using our method and then using cross-validation where the smallest lag which gives an average error in regression ≤ 0.1 is chosen as an estimate of the maximum lag. In both these sets of experiments compute the absolute error $\frac{|d_T - \bar{d}_T|}{L}$ for the methods involved where $\bar{d}_T$ is the corresponding estimate of the true lag d_T .
- 6) Repeat steps 4 and 5 multiple times (100 times) and find the average absolute error with a 95% confidence interval from all runs.
- 7) Repeat steps 3, 4, 5, 6 for different values of M .
- 8) Repeat all the previous steps for different levels of correlation and different choices of a, b which will produce different shape distributions.

In all the experiments we set $Q = 100$ and the length of the time series is also set (L) to 100. The values of a, b that we set for the beta are; $a = 1, b = 5$ (exponential shaped distribution), $a = 1, b = 1$ (uniform distribution) and $a = 5, b = 5$ (bell (normal) shaped distribution). We run the experiments for three values of ρ ; $\rho = 0.1$ (low correlation), $\rho = 0.5$ (medium correlation) and $\rho = 0.8$ (high correlation). The values of M that we used in the experiments are; $M = 100$ (low M) and $M = 5000$ (high M). The prior distributions we assume over the lags in each of the cases are betas with the corresponding a and

²Such a graph can be created by starting at a random node and adding outgoing edges to all other vertices. Then choosing another node and adding outgoing edges to all but the nodes that point to it. Repeat the previous step $Q - 1$ times.

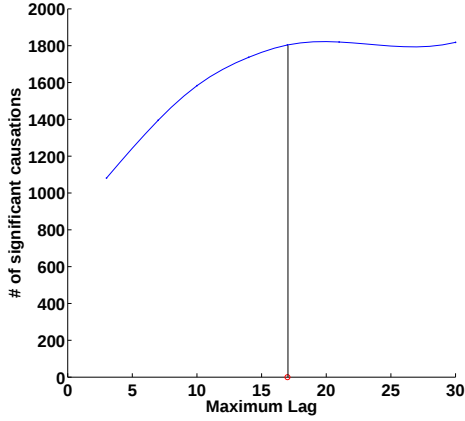


Figure 1. We observe the behavior of the number of significant causations deciphered against the max lag for the business intelligence dataset. The vertical line (at 17) on the plot denotes our estimate of the max lag and from its corresponding value on the curve we see that this estimate of the lag is sufficient to decipher important causations without being too conservative.

b parameters. The a and b learned from the lags of the M time series is done using mle. The ϵ parameter in algorithm 1 is set to 0.1.

From tables I and II we observe that the performance of the estimator is qualitatively the same, when lags for the M causally related pairs are known or have to be deciphered using algorithm 1. As expected, with increasing correlation between the lags (ρ) the estimates worsen, however, they are still pretty accurate for medium and high correlations, especially when M is large. The estimates improve with increasing M for a particular ρ , but are quite good for lower M when ρ is low or moderate. This implies that our estimator is robust to assumptions such as independence, being reasonably violated and hence, can prove to be useful in practice. Moreover, the estimates from our method compare favorably with cross-validation in most cases, except when the correlation is high and sample size is low. Given that cross-validation is also more computationally expensive, this is a strong case for our method.

B. Real Data Experiments

Business Intelligence: The first real dataset we test our method on is a business intelligence dataset. The dataset contains point of sales and age information of a product. In particular, for 107 stores we have the following 3 KPI's for each store namely; sales amount, sales quantity and sales age of products giving us a total of 321 time series. We know that sales quantity affects sales amount for the same store and sales quantity and amount between different stores also share causal relationships. However, we are interested in finding out if there are any causal relationships between sales age at different stores as well as sales age and the other sales KPI's within and between stores (termed freshness analytics). With

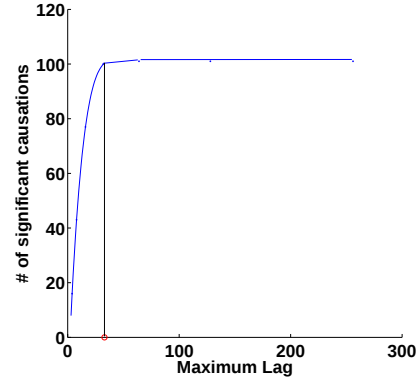


Figure 2. We observe the behavior of the number of significant causations deciphered against the max lag for the chip manufacturing dataset. The vertical line (at 33) on the plot denotes our estimate of the max lag and from its corresponding value on the curve we see that this estimate of the lag is sufficient to decipher important causations without being too conservative.

this we have $M = \frac{214(214-1)}{2} = 22791$ and $T = 51360$. We know the lags for the M pairs of time series to be 14 (days) for KPI's between stores and 1 for KPI's within stores. Hence, the maximum lag is 14 for the M pairs. The lengths of the time series are 145 (around 5 months worth of data) but we know that goods get replenished after a month which means $L = 30$. The prior belief is that the true lag can lie anywhere in this interval and hence, we assume a uniform prior over the (ratios of the) lags i.e. $\beta(1, 1)$.

With this information we estimate the maximum lag to be 17 (days). We evaluate this estimate by checking where on the curve in Figure 1 does a lag of 17 lie, i.e. what number of significant causations (i.e. causal strengths > 0.1) do we identify using Group Lasso for that lag. A good lag estimate would be one where we decipher most of the causations without the estimate being overly conservative. As we see in the figure, our estimate identifies most of the important causations without being excessively pessimistic and hence, seems to be a very good estimate of the true lag. The 10-fold cross-validation estimate (where average L_1 penalty ≤ 0.1) is 15, which seems to be a little too optimistic.

Manufacturing: The second real dataset we test our method on is a chip manufacturing dataset. The dataset has information about the measurements taken once a wafer (which contains around 80-100 chips) is manufactured. There are 30 such measurements which include 17 physical measurements such as average wafer speed rating etc. and 13 electrical measurements such as gate length etc. The length of each of the time series is 1161 (i.e. $L = 1161$). It is known that the physical measurements as well as the electrical measurements are causally related amongst themselves. However, the

causal relationships between the physical measurements and electrical measurements are not known. Given this we have $T = \frac{30(30-1)}{2} = 435$ and $M = \frac{17(17-1)}{2} + \frac{13(13-1)}{2} = 214$. The exact values of the lags for the M time series are not known and so we use algorithm 1 to find the lags ($\epsilon = 0.1$). It also believed that the lags tend to decay exponentially with mean around 35 and variance around 5. From this information we can compute a_p, b_p , which are the parameters of the prior.

With this we estimate the lag to be 33. Again, we evaluate this estimate by checking where on the curve in Figure 2 does a lag of 33 lie, i.e. what number of significant causations (i.e. causal strengths > 0.1) do we identify using Group Lasso for that lag. As mentioned before, a good lag estimate would be one where we decipher most of the causations without the estimate being overly conservative. As we see in the figure, our estimate identifies most of the important causations without being excessively pessimistic and hence, seems to be a very good estimate of the true lag. The 10-fold cross-validation estimate (where average L_1 penalty ≤ 0.1) is 47, which seems to be overtly pessimistic in this case.

VI. DISCUSSION

In the previous sections we listed some intuitions that a good estimator should capture and derived an estimator consistent with these intuitions. The manner in which we captured these intuitions in our derivation of the resultant estimator, may not be the only way of doing so. A bayesian approach to capture these intuitions rather than a mixture model approach, might also be plausible. However, in this case, it is not easy to see how one can derive an estimator that is efficiently computable and accurate at the same time.

One of the major limitations of our approach was that we assumed the lags to be independent, which not likely to be the case in practice. Though we showed that our estimator is robust to correlations that may exist between various lags, it would always be desirable if we could relax this assumption and derive an estimator which takes into account such correlations. A hint to achieving this might be to look closely at the procedure of generating data from a multidimensional beta distribution, described in the experimental section. This procedure generates correlated random variables (lags), which is what we desire. However, learning the parameters of this model or specifying priors does not seem to be a trivial task and needs further investigation.

To summarize, we proposed a novel approach to efficiently and accurately estimate the maximum lag for a set of causally related time series. This approach is able to capture prior knowledge that a user might posses in a principled manner and can be used as a quick diagnostic to estimate the maximum lag. Given these characteristics of our estimator along with it being robust to considerable

violations of the independence assumption, we believe, that it has the potential to be useful in practice.

ACKNOWLEDGEMENTS

Special thanks to Aurelie Lozano for providing the Group Lasso code and helpful discussions. We would also like to thank Naoki Abe, Alexandru Niculescu and Jonathan Hosking for providing useful comments and intruiging discussions. In addition, we would like to thank Jayant Kalagnanam, Stuart Seigel, Shubir Kapoor, Mary Helander and Tom Ervolina for providing the real dataset and motivating the problem addressed in this paper.

REFERENCES

- [1] C. Granger, "Testing for causality: a personal viewpoint," 2001.
- [2] Y. L. a. S. R. A. Lozano, N. Abe, "Grouped graphical granger modeling methods for temporal causal modeling," in *KDD*. ACM, 2009.
- [3] Y. L. a. N. A. A. Arnold, "Temporal causal modeling with graphical granger methods," in *KDD*. ACM, 2007.
- [4] I. N. a. D. P. N. Friedman, "Learning bayesian network structure from massive datasets: The "sparse candidate" algorithm," in *UAI*, 1999.
- [5] D. Heckerman, "A tutorial on learning with bayesian networks," MIT Press, Tech. Rep., 1996.
- [6] A. M. and P. Spirtes, "Graphical models for the identification of causal structures in multivariate time series models," in *5th Intl. Conf. on Computational Intelligence in Economics and Finance.*, 2006.
- [7] C. G. a. P. S. R. Silva, R. Scheine, "Learning the structure of linear latent variable models," *J. Mach. Learn. Res.*, vol. 7, pp. 191–246, 2006.
- [8] C. G. a. R. S. P. Spirtes, *Causation, Prediction and Search*, 1st ed. The MIT Press, 2001, vol. 1.
- [9] M. Y. and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B*, vol. 68, 2006.
- [10] Y. Tong, *Probabilistic Inequalities for Multivariate Distributions*, 1st ed. Academic Press, 1980.
- [11] R. P. a. L. H. R. Gnanadesikan, "Maximum likelihood estimation of the parameters of the beta distribution from smallest order statistics," *Technometrics*, vol. 9, pp. 607–620, 1967.
- [12] H. D. and H. Nagaraja, *Order Statistics*, 3rd ed. Wiley and Sons, 2003.
- [13] W. Enders, *Applied Econometric Time Series*, 2nd ed. Wiley and Sons, 2003.
- [14] I. H. a. W. S. A. Minhajuddin, "Simulating multivariate distributions with specific correlations," Southern Methodist University, Tech. Rep., 1996.