Proceedings of Machine Learning Research vol TBD:1–16, 2023

Learning Conditional Granger Causal Temporal Networks

Ananth Balashankar[a], Srikanth Jagabathula[b], and Lakshminarayanan Subramanian[a]

[a]Department of Computer Science, New York University
[b]NYU Stern School of Business

Abstract

Granger-causality derived from observational time series data is used in many real-world applications where timely interventions are infeasible. However, discovering Granger-causal links in large temporal networks with a large number of nodes and time-lags can lead to *millions* of time-lagged model parameters, which requires us to make sparsity and overlap assumptions. In this paper, we propose to learn time-lagged model parameters with the objective of improving recall of links, while learning to defer predictions when the overlap assumption is violated over observed time series. By learning such conditional time-lagged models, we demonstrate a 25% increase in the area under the precision-recall curve for discovering Granger-causal links combined with a 18-25% improvement in forecasting accuracy across three popular and diverse datasets from different disciplines (DREAM3 gene expression, MoCAP human motion recognition and New York Times news-based stock price prediction) with correspondingly large temporal networks, over several baseline models including Multivariate Autoregression, Neural Granger Causality, Graph Neural Networks and Graph Attention models. The observed improvement in Granger-causal link discovery is significant and can potentially further improve prediction accuracy and modeling efficiency in downstream real-world applications leveraging these popular datasets.

1. Introduction

Granger causality Granger (1969) in time series data is important in many real world applications in economics Arnold et al. (2007a), climate science Lozano et al. (2009b), earth sciences Runge et al. (2019), and biology Lozano et al. (2009a). The knowledge of the Granger-causal structure allows us to build prediction models to make fine-grained time series forecasts conditioned on specific covariate values. For instance, the knowledge that a specific set of genes interfere with the expression of another gene allows us to build accurate gene regulatory networks, which assist in generating hypotheses for drug discovery. In practice, interventions are often infeasible because of the large dimensionality of data and making inference from real-time observations becomes inevitable as placing controls are either impractical, or even unethical. As a result, the inferences and forecasts must be done using only observational data. In this work, we assume that the underlying Granger-causal structure is specified to the extent that we know which covariates affect the outcome variable of interest. However, we lack information on how long the effect lasts and if it holds under all covariate distributions. For example, in the DREAM3 gene expression network task Prill et al. (2010), where multiple genes can express and influence other genes, experiments are carried out where as part of a treatment, certain catalyzing agents are introduced over a time period and the corresponding gene expression time series are observed. Here, although we know that there are a specific set of 100 genes that potentially Granger-cause each other, we

© 2023 .

do not know how the effect of "one gene regulating another" varies temporally. The number of time-lagged parameters in such a time series model can quickly grow with the maximum allowed time-lag (Table 1). In other words, because each of the gene expression varies over time, we need to know how the expression of one gene at one point in time regulates another gene's expression at a future time in a parameter efficient manner. This issue is unique to Granger-Causality over time series data in scenarios with limited data, relative to the high dimensionality of permissible time-lagged covariate distributions.

A common way to address this issue has been to be conservative and train prediction models over large time windows to capture any long-term effects of covariates Yang et al. (2018). This approach often runs into data sparsity issues and poor granger-causal link discovery accuracy Spirtes (2001). To deal with this issue, prior graphical granger methods Arnold et al. (2007b) have artificially imposed sparsity constraints on the model parameters forcing co-efficients to collapse to zero, thereby reducing time-lagged parameters to estimate (from $600-700$ to ≤ 100). However, Granger-Causality was not designed for large temporal networks with millions of time-lagged parameters; in fact, economists have often warned against blindly applying it over a large number of variables Peters et al. (2013). One of the issues in applying Granger-causality directly to large temporal networks, is that large window sizes also result in the increase in chances of a violation of the positivity assumption Rosenbaum and Rubin (1983); Vapnik (1999) - a condition necessary for consistency of the Graphical Granger methods, i.e certain time-lagged covariate distributions have been rarely or never observed previously and hence extrapolating the granger-causal links to such covariate distributions may be erroneous. In such scenarios, it might be best to defer prediction rather than predicting by extrapolating incorrectly.

To deal with the above challenges, in this paper, we propose a methodology to improve the recall of Granger-causal links by conditionally allowing for prediction deferrals. Specifically, given an outcome variable, a treatment variable, and a collection of covariates which are known to have passed the bivariate Granger-causal test (all of which are time series), our method parametrizes each of the Granger-causal links with two parameters: (a) the maximum window size, δ, and (b) the variance threshold, ρ. The maximum window size specifies a bound on how long the treatment effect lasts and the variance threshold specifies when predictions are deferred. In particular, if the variance of our estimator for a given covariate value is above ρ, then we defer the prediction, suggesting that we do not have sufficient confidence in whether the treatment takes effect under the given covariate value based on the observations. Our method chooses the values of δ and ρ to optimize a Granger-causal link recall metric, which is computed using only those covariate distributions whose variance is below the threshold ρ. A small value of δ will result in fewer deferrals during training and thus a more reliable estimate of the recall metric, but might miss several links with longer time lags yielding lower recall, and the model becomes too sensitive to perturbations over smaller time windows. Very large values of δ also result in lower recall because they result in a large number of deferrals during training and consequently less evidence from data to support link discovery. Thus our formulation trades-off the model's temporal sensitivity and overlap-based robustness, to learn predictive models that have high accuracy and are consistent with the known Granger-causal links.

Prior work does not consider link recovery but instead focuses on optimizing prediction accuracy using general purpose sparsity inducing techniques. In particular, multivariate

Dataset	n	δ	# Parameters	# Targets
MoCAP	54	2,024	113,344	12
DREAM3	46	10,500	504,000	5
Stocks	100	29,200	2,978,400	10

Table 1: Problem of Over-parameterization in Time-Series Granger-Causal Models

Auto-Regression linear models (VAR) are trained to optimize prediction accuracy while inducing sparsity in the time lag parameters through Group Lasso penalty Lozano et al. (2009a) regularization; links are included by comparing the prediction accuracy of the model with and without the treatment variable Conover and Iman (1981) to test for significance. The non-linear version of the above VAR Granger Causality models have also been proposed Tank et al. (2018) which could model additive effects of the past of each series in a decoupled manner. Sequence prediction models You et al. (2018) and graph attention models Veličković et al. (2018) which model the neighborhood of nodes to learn Granger-causal links have also been studied. In the structured prediction task, identifying how much covariate overlap exists between variables has been studied Jesson et al. (2020). We build off of these constraints and demonstrate that augmenting the condition of overlap violation ensures that prediction models which learn to defer when specific covariate distributions have not been previously observed are better at discovering Granger-causal links.

Learning such robust time-lagged Granger-causal models can be of immense importance in various real world scenarios of causal discovery. For example, in our running example of gene expression networks, using time series data from multiple such experiments carried out in laboratory settings along with our Granger-causal parametrization framework, we can extract optimal lag parameters between different genes to understand when (time-lag) and how (covariate overlap) they influence each other's expression. Similarly, in the human motion capture MoCAP task CMU (2009), we are able to improve the area under the precision-recall curve (AUC-PR) of detecting Granger-causal links in the human activity recognition dataset than baseline Granger-causality Granger (2001); Tank et al. (2018) methods. Finally, given time series of Granger-causal news events, we improve monthly forecasting accuracy of stock prices. Each of these three tasks have a large set of time-lagged parameters (113K - 2.9M as per Table 1), but aim to predict a very small number of target variables. The conditional temporal prediction models we have developed has

- Reduced the number of parameters to learn by 2-3 orders, and achieved 25% better AUC-PR in discovering Granger-causal links than comparative baselines
- Improved prediction accuracy over held-out time series by 18-25% across three datasets in MoCap activity recognition, DREAM3 gene regulatory networks detection and the New York Times news-based stock price prediction tasks.
- Formalized the trade-off between time-lag sensitivity and overlap-based robustness and showed that artificially increasing the maximum time-lag leads to an over-specified model with sub-optimal link recovery and prediction accuracy.

2. Granger Causal Link Recovery

Problem Setup: We consider a time-series forecasting problem where the goal is to predict the value $y(t+1)$ at time $t+1$ of an outcome times series using the past observations $\mathbf{X}(t,\delta) = \{\mathbf{x_1}(t,\delta), \mathbf{x_2}(t,\delta), ..\mathbf{x_n}(t,\delta)\}$, the treatment time series $\mathbf{v}(t,\delta)$, the historical outcomes $\mathbf{y}(t,\delta)$, each of them a vector of values evaluated up to δ discrete time steps back in time from timestamp t. We assume we are given a predictive model m, which outputs the future values of the outcome time series from the past observations:

$$\hat{y}(t+1) = m\left(\mathbf{X}(t,\delta), \mathbf{v}(t,\delta), \mathbf{y}(t,\delta)\right).$$

Each of the above variables $\{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}, \mathbf{v}, \mathbf{y}\}(t,\delta)$ are time-lagged multivariate variables with δ time-lagged values going back from time instant t. This means the predictive model has a total of $(n+2)\cdot\delta$ variables as input to predict the value of the outcome at time $t+1$. We also assume we are given non-parametric links of interest of the form $\mathbf{v} \rightarrow \mathbf{y}$. We now train this predictive model on historical observational data and use it to make fine-grained predictions for each covariate value $(\mathbf{x}, \mathbf{v})$. Implementing this approach requires us to answer two questions:

1. How long does the treatment effect last?
2. Under what covariate values does the treatment take effect?

Answering the first question allows to set the correct value of δ. The second question is important when dealing with observational data because the positivity assumption in covariate overlap $P(\mathbf{v}(t,\delta) = v|\mathbf{X}(t,\delta) = \mathbf{x}) > 0$ is often not met and blindly assuming it holds can lead to incorrect extrapolation.

Prior work has focused on answering the first question using sparsity inducing methods Arnold et al. (2007b), while we argue that the second question is of equal importance to learn a robust time-lagged Granger-causal model and has implications on answering the first question. We formulate this as a joint learning problem with the *goal* to learn a compact Granger-causality aware time series predictive model which also learns to deferfrom making over-confident predictions for time periods with very few conditional treatment and covariate observations in the data Jesson et al. (2020).

Time-Lagged Granger-Causal Model Assumptions: Learning the time-lag parameters based on temporal predictions have been studied in the domain of Granger causality Granger (2001), where the links are established based on the lag parameter in a time series of the causal variable that provide the highest reduction in regression error of predicting the effect variable in a multivariate setting. These models make assumptions of sparsity, i.e for a given $(\mathbf{v}, \mathbf{y})$ only a small number of time-lagged variables of $\mathbf{X}, \mathbf{y}, \mathbf{v}$ are predictive of future values of $y(t+1)$. Many sparsity enforcing methods have been proposed like the Lasso regularization Arnold et al. (2007b) which minimize the number of non-zero weights in a linear model Tank et al. (2018); Balashankar et al. (2019), a minimal generative model graph to learn directed information graphs Quinn et al. (2015), or propose an auxiliary task based regularization of jointly predicting the causal graph and optimal predictors Kyono et al. (2020) or propose a recall-based regularization method to model autocorrelated time series with latent confounders Gamella and Heinze-Deml (2020). One *overlooked assumption* in the above approaches in overcoming the overparameterization issue is that of the

positivity assumption in covariate overlap ($P(\mathbf{v}(t,\delta) = v|\mathbf{X}(t,\delta) = \mathbf{x}) > 0$) Imbens (2004), between treatment and outcome, given observational data. We overcome the limitation of prior methods by addressing violations of the positivity assumption explicitly through covariate conditional variance estimation.

3. Methodology

Under the ignorability (no unobserved confounders) assumption, we can either estimate the importance of the treatment on the outcome variable from observed data by either slicing data based on treatment Shalit et al. (2017) or incorporating the treatment variable as another covariate Gelman and Hill (2006); Lawes et al. (2012); Jesson et al. (2020). In Algorithm 1, we adopt the latter approach and learn a time-lagged prediction model that optimizes the recall of Granger-causal links with the target variable y, using the covariates $\mathbf{X}$ and the treatment variable $\mathbf{v}$. Now we explain the Granger-causal link significance test used, and how we use a conditioned version of it to incorporate the covariate overlap assumption. Finally, we tie all of these components into a Bayesian optimization algorithm that maximizes the recall of the Granger-causal links.

To overcome the intractability of ensuring the overlap assumption for large number of covariates, we use the approach used by Jesson et al. (2020). We estimate the lack of overlap using the conditional variance - $\hat{V}(\mathbf{x},\delta)$ of the predictive models $\hat{m}, \tilde{m}$, which predict $\hat{y}(t+1)$ (Eqn 5). We use a non-parametric Conover Squared Ranks (SR) test-statistic used for testing for equality of variance in prediction errors Conover and Iman (1981), to approximate the *new information* that helps in improving the prediction accuracy of models. Once the prediction models are trained on certain splits of the data for a given outcome, we then estimate the variance by the bootstrapping method and evaluating the covariate variance on numerous held-out development splits. For a given threshold ρ, we adopt a trimming policy Hernán and Robins (2010) with the rejection policy, conditioning on covariates $\mathbf{x}$, where the covariate variance $Var[\hat{V}(\mathbf{x},\delta)]$ is above a threshold ρ Jesson et al. (2020), and then compute the **Time-Lagged Conditional Treatment Importance -** $TCTI(\delta,\rho)$.

$$TCTI(\delta,\rho) = \mathbb{E}_{\mathbf{x}:Var[\hat{V}(\mathbf{x},\delta)]\leq\rho}\hat{V}(\mathbf{x},\delta) \tag{1}$$

This formulation outlines that when the predictive causality test statistic has high variance, we should not rely on those slices as they violate the overlap assumption. If the estimated value of TCTI passes the SR test with α significance ($\alpha = 0.05$) under a $\mathrm{F}(1,(n+2)\delta+1)$ distribution, we then consider the treatment variable v to have an α-significant Granger-causal link with y. Our choice to place a fixed threshold on the variance is to understand the effect of this variance on the recall of the Granger causal links, which cannot be done without imposing a strict discretization (either to choose a link or not), by using the variance directly in the predictive modeling. This process is then repeated for all treatment variables to give the recall $w(\delta,\rho)$, which we will maximize:

$$w(\delta,\rho) = \frac{\#\text{ links }\alpha\text{-significant with }(\delta,\rho)\text{ as per Eqn 1}}{\#\text{ Granger-causal links}} \tag{2}$$

Computing TCTI: Since the values of importance weights can vary depending on the choice of models, for example in the case of linear regression - they are coefficients, whereas in non-linear network models, there are attention weights, activation vector alignment - similar to Granger methods Granger et al. (2000), we use a non-parametric Conover Squared Ranks (SR) test for equality of variance Conover and Iman (1981), to test if the treatment variable provides any *new information* that helps with the prediction accuracy of models: $\hat{m}$ with and $\tilde{m}$ without the treatment variable as input. The test is run on the prediction errors $\epsilon(\hat{y}_{(\mathbf{x},\delta)}(t+1)), \epsilon(\tilde{y}_{(\mathbf{x},\delta)}(t+1))$ produced by prediction models $\hat{m}, \tilde{m}$ respectively on held-out temporally disjoint test data.

$$\hat{y}_{(\mathbf{x},\delta)}(t+1) = \hat{m}(\mathbf{X}(t,\delta) = \mathbf{x}, \mathbf{v}(t,\delta), \mathbf{y}(t,\delta)) \tag{3}$$

$$\tilde{y}_{(\mathbf{x},\delta)}(t+1) = \tilde{m}(\mathbf{X}(t,\delta) = \mathbf{x}, \mathbf{y}(t,\delta)) \tag{4}$$

$$\hat{V}(\mathbf{x},\delta) = SR(\epsilon(\hat{y}_{(\mathbf{x},\delta)}(t+1)), \epsilon(\tilde{y}_{(\mathbf{x},\delta)}(t+1))) \tag{5}$$

The SR test we use, is the non-parametric alternative of the Levene's test Brown and Forsythe (1974), which itself is the robust alternative for non-normal distributions to the 1-way between-groups analysis of variance (ANOVA) Fisher (1992) test to detect equality of population means. We use this test over the parametric ones as we do not make any assumption of the distribution of variance (normal), as our test of overlap violation cannot work if we already assume that there is an underlying normal distribution. The input for the prediction model are δ lagged time series of covariates, treatments and outcomes, and we will control the length of this time series as part of our methodology. We vary the δ and train jointly - warm starting hidden model parameters as the time lag δ increases, instead of training a separate model from random initialization per value of δ. Thus, we are able to compute the temporal lag that maximizes the prediction accuracy of the target variable that is Granger-causally linked to covariates. We characterize that a Granger-causal link to be supported by the observed data (or to be recalled), if adding a Granger-causal variable's temporal lag causes an increase in the prediction accuracy of the outcome conditioned on the covariate $\mathbf{X}(t,\delta) = \mathbf{x}$, as show by a test statistic with a p-value below the statistical significance threshold ($\alpha = 0.05$) under a $\mathrm{F}(1, n+1)$ distribution, as compared to not incorporating the Granger-causal variable at all. Otherwise, we characterize that Granger-causal link as not yet observed in the data. To convert the time-sensitivity into a variance based estimate, we compare the prediction errors ϵ and characterize it by the inequality of variance $\hat{V}(\mathbf{x},\delta)$ given by the test statistic of the Squared Ranks (SR) test. Models $\hat{m}, \tilde{m}$ are trained and tested (Eqn 5) on temporally disjoint time series data.

Overlap-based Conditioning: We now have to overcome the intractability of conditioning on the large number of covariates. Here too, we use the previously trained models: $\hat{m}, \tilde{m}$ to predict the target variable, but use the variance of the treatment importance estimate Jesson et al. (2020). Once the prediction models $\hat{m}, \tilde{m}$ are trained on certain splits of the data for a given outcome, we then estimate the variance by a bootstrapping method and evaluating $\hat{V}(\mathbf{x},\delta)$ on numerous held-out development splits.

$$Var[\hat{V}(\mathbf{x},\delta)] = Var[SR(\epsilon(\hat{y}_{(\mathbf{x},\delta)}(t+1)), \epsilon(\tilde{y}_{(\mathbf{x},\delta)}(t+1)))] \tag{6}$$

For a given threshold ρ, we adopt a trimming policy Hernán and Robins (2010) with the rejection policy, conditioning on covariates where the variance of $Var[\hat{V}(\mathbf{x}, \delta)]$ is above a threshold ρ Jesson et al. (2020), and then compute the *Time-Lagged Conditional Treatment Importance* - $TCTI(\delta, \rho)$. While the models trained are dependent on δ, the computation of $TCTI(\delta, \rho)$ is done after the training is completed, rather than at training time.

Fine-tuning hyperparameters: Our formulation outlines that when the predictive causality test statistic has high variance, we should not rely on those slices as they violate the overlap assumption. For a given time-sensitivity lag parameter, this trade-off with overlap assumption can be understood by varying the hyper-parameters together. We see that for thresholds: ρ, such that the overlap condition $\mathbf{x} : Var[\hat{V}(\mathbf{x}, \delta)] \leq \rho$ is satisfied for all covariates $\mathbf{x}$, then there would be no deferral, and such a $TCTI(\delta, \rho)$ would directly evaluate the Granger-causality of the links, and hence $w(\delta, \rho)$ would be equal to 1. We now outline 3 approaches we used to fine-tune the hyper-parameters δ, ρ.

Grid Search: By using a grid search for values of $\delta \in \{1, 2, \ldots T\}$, $\rho \in \{\eta, 2\eta, \ldots k\eta\}$, for each Granger-causal link in the dataset, we search for the value δ^*, ρ^* that maximizes the $w(\delta, \rho)$. This way, we search among all Granger-causal links, the sensitivity and robustness parameters that is best supported by the observed data. This can be time-consuming to be done for each model and we can upper-bound the time lag parameter T to train the model.

Random Search: Instead of an exhaustive grid search, in this approach, we sample values of δ, ρ from a uniform distribution and choose the parameters that maximizes the $w(\delta, \rho)$. Here, we were able to heuristically choose a bound larger than $T, k\eta$ respectively and can search among values not explored by the grid search. Although we can control the number of hyper-parameters to train and evaluate the models against, the computational cost in training the models remain.

Bayesian Optimization: To learn which hyper-parameters δ, ρ result in high recall of Granger-causal links as per the significance level $\alpha = 0.05$, we used Bayesian optimization with the probability prior f parameterized by θ to be drawn from Gaussian processes. Specifically, we maximize the fraction of links w validated with a significance level for a given value of (δ, ρ). The covariance kernel chosen is the ARD Matern 5/2 Kernel Snoek et al. (2012), which has been demonstrated to capture realistic hyper-parameter distributions in neural networks, while resulting in sampling functions that are twice differentiable. We choose hyper-parameters in parallel using the Bayesian roll out method, where the acquisition function is optimized the utility of expected improvement per trial.

Inference: Once we obtain the optimal δ^*, ρ^*, for each outcome variable, we use the trained models to make time-series predictions over unseen data. If the variance for a given covariate as pre-computed by the test statistic in Eqn 5 is above the threshold ρ^* at training time, we continue to defer to make predictions on those covariates at inference time. The resulting time-lagged prediction model then is used directly to infer the Granger-causal links based on the non-zero model parameters in the models (Sec 4).

4. Evaluation

Datasets: DREAM3: The DREAM3 gene expression network inference challenge Prill et al. (2010) consists of 5 datasets, 2 for E.Coli and 3 for Yeast, with each dataset containing 100 time series. Each of the time series has 46 variables, each of them gene expression

Algorithm 1 BayesOpt for Exploring the Time-Overlap Trade-offs

$(\delta, \rho) \sim$ U: uniform random distribution, $\hat{\delta} \leftarrow 0$
$M = \{\}, \hat{w} \leftarrow 0, \psi \leftarrow$stopping criterion
while $\hat{w} < 1 - \psi$ **do**
  Update BayesOpt acquisition function
  Acquire $D = \{k$ values of $\delta\}$ from BayesOpt
  Update $max_\delta \leftarrow max_{\delta \in D}(\delta, max_\delta)$
  Train $\hat{m}_{max_\delta}, \tilde{m}_{max_\delta}$ and update $\hat{\delta} \leftarrow max_\delta$ if $max_\delta < \hat{\delta}$
  Acquire $R = \{k$ values of $\rho\}$ from BayesOpt
  Update f, δ^*, ρ^* over R to maximize $w(\delta, \rho)$: Eqn (2)
end while
Return δ^*, ρ^*

replicates observed at 21 time instants. For each of these 46 variables, we consider in a round-robin version that one of those variables is the outcome, one is the treatment, and the rest as covariates. This allows us to estimate $TCTI(\delta, \rho)$ for a total of 2070 combinations of treatment and outcomes, given the covariates. The ground truth contains directed Granger-causal links between 46 replicates, and through our prediction models, we parametrize each of the Granger-causal links by sweeping over values of δ, ρ with the maximum significant value of $TCTI(\delta, \rho)$. In parallel, we also report the AUROC (Area under the Receiver-Operator Curve) and AUC-PR (Area under the precision recall curve) for the classification task of detecting the binary gene expression.

MoCAP: The CMU MoCAP dataset CMU (2009) consists of motion sensor data, for 54 joints, collected from two subjects for a total of 2024 time points. Here, we are given the Granger-causal links of the human skeleton and we learn the time lag and overlap parameters for the classification of each of the activities - jumping jacks, side twists, arm circles, etc. Here, we report the AUC-PR (Area under the Precision-Recall curve) for detecting the human activity based on the movements in the human joints, along with the recall of the Granger-causal links in the human skeleton.

Stock Price Prediction: In the stock prediction task, the outcomes are each of the 10 stock prices from 2010-13 (with 2013 as the test split), and for the treatment variable, we are given the top 10 financial news based Granger-causal factors and a further 100 co-variates extracted from NY Times by Balashankar et al. (2019). Here too, we measure the Root Mean-Squared Errors (RMSE) of the predicted values and the fraction of time instants where we had to defer the prediction.

Baselines: We present a few comparable baseline prediction models built to incorporate Granger-causality for time series and discuss how compact time-lagged versions of these models have been learnt. These models are the baselines we evaluate in Section 5.

Multivariate linear auto-regressive models: Multivariate Auto-Regression linear models (VAR) take the time series of the treatment, covariates and the lagged values of the target variable as input to predict the target variable for future time instants. Here, we compare the prediction accuracy of the model with and without the treatment variable using the non-parametric Squared Rank test Conover and Iman (1981) to test the significance of a

non-zero coefficient in matrix C. Additionally, we also add the Group Lasso penalty Lozano et al. (2009a) that has been shown to overcome the need of precisely estimating the time lag by applying the Lasso regularization.

$$y(t+1) = A \cdot y(t,\delta) + B \cdot \mathbf{X}(t,\delta) + C \cdot \mathbf{v}(t,\delta) + D \tag{7}$$

Neural Granger causality: The non-linear version of the above VAR Granger Causality models have also been proposed Tank et al. (2018) which could model additive effects of the past of each series in a decoupled manner. Here, by modeling the task of Granger causality using componentwise multilayer perceptrons and recurrent neural networks, all time series are captured in an input layer of the neural network having a total of $\delta \cdot n^2 \cdot W$ parameters, where W is the number of hidden units in the input layer. In order to model long time lags in Granger causality, they use component-wise recurrent neural networks (RNN) for each time series. Similar to the linear model, to enforce sparsity, Lasso penalty and the hierarchical group Lasso penalty have been proposed, which chooses a suitable lag for each of the time series - but ignores the covariate overlap violation.

Generative Graph Neural Networks: Another approach proposed in You et al. (2018), is to model this as a sequence prediction by reducing the graph to Breadth-First-Search (BFS) based deterministic sequence. They use a hierarchical graph RNN structure to first model the node prediction problem. In our case, although we know all the nodes of the network ahead of time, we can use the edge prediction model and predict edges in a BFS sequence. While this is comparable to the Neural Granger Causality model, the number of parameters to be learnt is lesser: $\delta \cdot n \cdot W$.

Graph Attention Model: Also recently, with the success of attention models in natural language tasks like machine translation, attending over the neighborhood of nodes, instead of recurrent architectures has been shown to be specifically relevant for graphical causal modeling Veličković et al. (2018). This approach requires only the neighborhood of nodes and scales better than spectral representations of the graph, which need to be aware of the entire graph structure.

Other Related Work: We aim to understand the assumptions required to identify the optimal time sensitivity parameters of Granger-causal links in time series data, once the direction and presence of the Granger-causal links have already been defined. Prior work in this space has focused on methodologies to increase recall of the causal links in auto-correlated time series Gerhardus and Runge (2020) or regularize over unseen parts of the causal graph Kyono et al. (2020). However, such methods fail to quantify when it might be even possible to recover the optimal time lag parameters in an observed data distribution. The covariate relationship in time series have been explored in Granger causality with group boosting methods Lozano et al. (2009a) or Markov random field regression Liu et al. (2010) which capture the non-linear group information between variables in a time series. Prior methods Riva and Bellazzi (1996) that maintain a set of probabilistic causal models and perform model selection can also benefit from the quantification of the trade-off between overlap and temporal parameters in longitudinal data.

Experiment setup: Each time the prediction models $\hat{m}, \tilde{m}$ are updated in Step 7 of Algorithm 1, the model parameters are updated inline with the optimization objective of the baseline model with the modified max_δ. In the case of the VAR model, the training is

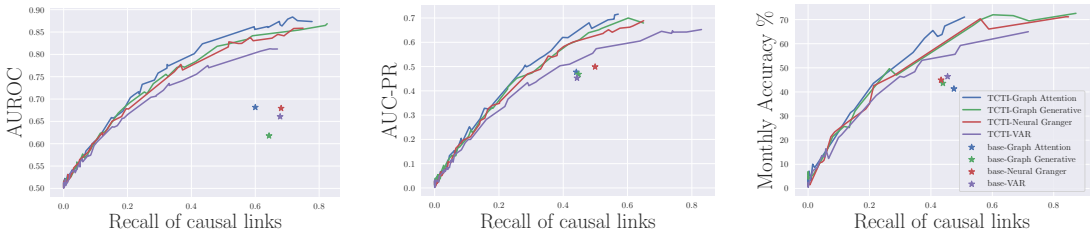

Figure 1: **Recall of Granger-Causal Links vs Prediction Accuracy (a) DREAM3:** Across each of the 5 outcomes in the DREAM3 dataset, we see that as the recall of the Granger-causal links in the gene expression network increases, the AUROC of the time series of the gene expression level also increases. **(b) MoCAP:** As the recall of the Granger-causal links of the human skeleton network increases, the AUC-PR in the human activity recognition task increases. **(c) Stock:** As the recall of the Granger-causal links of the financial news factors increases, the prediction accuracy of 10 stock prices increases.

using the same Group Lasso penalty as that of the baseline model. In the Neural Granger Model, the weight terms for the hierarchical and component-wise Lasso penalty are the same as that used for the baseline model without TCTI. For the Generative Graph model, the model parameters are updated as per the node and edge prediction loss, but other hyperparameters remain the same as that of the baseline model - such as learning rate, architecture of the GraphRNN (4 layers of GRU cells, 128 hidden state dimensions for node prediction; and 4 layers of GRU cells and 16 hidden state dimensions for edge prediction). For the Graph Attention Network baseline, we update the model parameters of the model like attention weights; but architectural hyperparameters remain the same in our TCTI version - we use the transductive learning set up with 8 attention heads each computing 8 features, followed by a single attention head for the prediction task.

5. Results

We now demonstrate the efficiency of our approach in Granger-Causal link discovery, while showing that our prediction deferrals do not reduce prediction accuracy. Also, we show choosing the right set of temporal and covariate hyperparameters are critical for this improvement, as compared to generic sparsity inducing baselines where the choice is adhoc.

Granger-Causal Link Discovery: We demonstrate that our proposed method by optimizing $TCTI(\delta, \rho)$ and hence the Granger-causal link recall, we improve both the recall of Granger-causal links and the prediction accuracy across 3 datasets and 4 baseline Granger causal models (see baselines Sec 4 in detailed methods). In Figure 1, we see that the prediction accuracy (y-axis) improves by 18-21% and the recall of Granger-causal links (x-axis) improves by 25% across 3 datasets for each of the four prediction models' TCTI-{VAR, Neural Granger, Graph Generative, Graph Attention}, when the time lag and overlap parameters are optimized with 25 random restarts. The baseline models base-{VAR, Neural Granger, Graph Generative, Graph Attention} indicated as dots in the plot, have the sparsity constraint enforced through Lasso regularization loss, and end up with a lower recall of Granger-causal links i.e more Granger-causal links are not used for prediction, with low

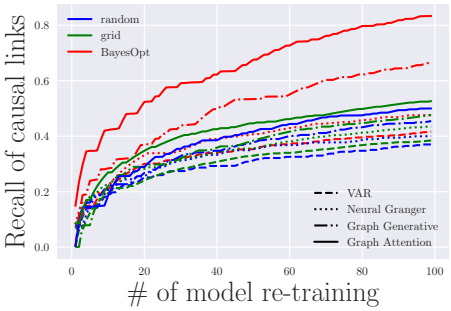

Figure 2: **Sampling efficiency among hyperparameter search methods** The number of re-trainings required to fine-tune the time sensitivity and robustness overlap parameters to improve the recall of Granger-causal links in the DREAM gene expression dataset

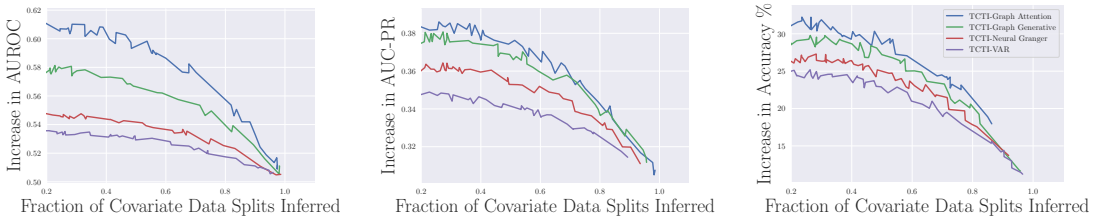

Figure 3: **Prediction Deferrals effect on Accuracy** Choosing δ based on an overlap based constraint, the prediction accuracy increases on the remaining test samples on (a) DREAM3 (b) MoCAP (c) Stock datasets for 4 Granger causal models. As the prediction models choose to defer on larger fractions of the covariate data splits, the increase in accuracy is higher.

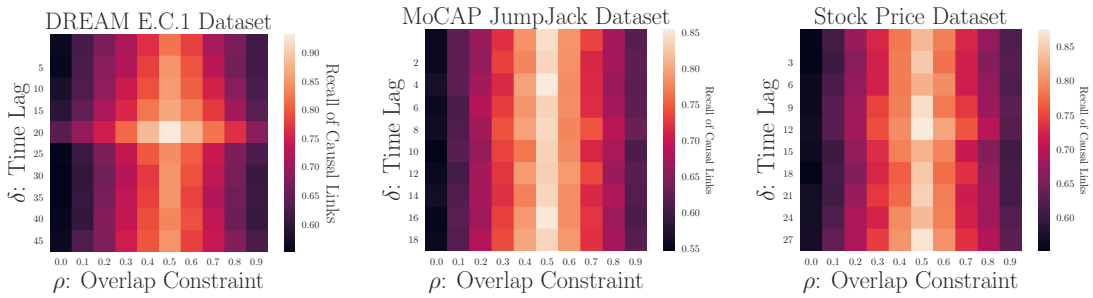

Figure 4: **Variation between time lag and overlap** parameters for the Graph Attention Model on 3 datasets shows the need to learn them jointly

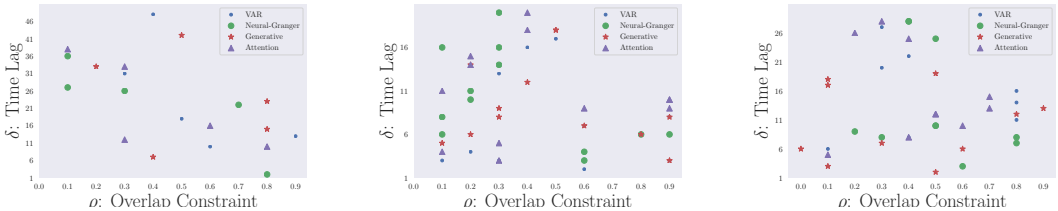

Figure 5: **Trade-off for a fixed prediction accuracy** Time sensitivity and robustness overlap among the fine-tuned Neural Granger Causal prediction models across Granger-causal links that provide the highest $TCTI(\delta, \rho)$ for (a) DREAM3, (b) MoCAP and (c) Stock Price prediction datasets

prediction accuracy. Further, to be comparable with baseline models which do not defer, we do not defer predictions at inference time, but only while learning the optimal time-lag and overlap parameters δ, ρ in Algorithm 1.

Importance of Prediction Deferrals: To understand how deferring predictions on covariates with high variance as per TCTI (δ, ρ) has helped learning better prediction models on the remaining covariates, we see that for each of the 3 tasks and 4 models in Figure 3, that there is an increase in prediction accuracy for the covariates we choose to predict as compared to the overall baseline model. Further, the increase is greater when the number of covariate slices deferred is greater. Thus by deferring predictions over covariates where overlap assumptions are violated, we improve the prediction accuracy and rely on robust Granger-causal links for prediction.

Variation in Time-Lag and Overlap Parameters: To understand how setting time-lag and overlap parameters is critical for the high performance of our approach, we plot the range of parameters required to achieve a fixed prediction accuracy across the 4 prediction models with our approach. In each of the dataset, we see that the distribution of parameters for time-lag, overlap-constraint for the links for a given Granger-causal model (the Graph Attention Model) has high variability as shown in Figure 4. Also, for a fixed prediction accuracy of outcomes, we see that the links for the 4 modeling choices are clustered into 3 groups - (low δ, high ρ), (low δ, low ρ), (high δ, low ρ). The lack of links with relatively high δ and high ρ further empirically affirms the trade-off between time-lag and overlap-constraints as shown in Figure 5.

Hyperparameter Optimization Method: We also see that across the 3 datasets, and the 4 modeling techniques - the choice of the hyperparameter fine-tuning methodology can impact the number of Granger-causal links we can recover with significant $TCTI(\delta, \rho)$ as shown in Figure 2. While grid search and random sampling provide good initial estimates of the maximum recall that can be achieved, we see that BayesOpt quickly outperforms these brute force search mechanisms for the same number of model re-trainings. This drastically reduces the time and cost required to identify the temporal and overlap characteristics of the Granger-causal links.

6. Conclusion

We have demonstrated the need to parameterize causal links with their associated temporal sensitivity with awareness of overlap assumption violations on 3 diverse datasets. In time series data, we show that there exists a trade-off between how temporally sensitive any prediction model incorporating the causal links can be while not compromising on the covariate overlap assumption. This allows us to further build better prediction models while not relying on data that lacks overlap while attempting to capture long term effects. Further, since the size of such temporal networks are quite large with millions of parameters of estimate, we have shown that a principled way of addressing covariate uncertainty through prediction deferrals can further improve the trust in practitioners.

References

Andrew Arnold, Yan Liu, and Naoki Abe. Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 66–75, New York, NY, USA, 2007a. ACM. ISBN 978-1-59593-609-7. doi: 10.1145/1281192.1281203. URL http://doi.acm.org/10.1145/1281192.1281203.

Andrew Arnold, Yan Liu, and Naoki Abe. Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, page 66–75, New York, NY, USA, 2007b. Association for Computing Machinery. ISBN 9781595936097. doi: 10.1145/1281192.1281203. URL https://doi.org/10.1145/1281192.1281203.

Ananth Balashankar, Sunandan Chakraborty, Samuel Fraiberger, and Lakshminarayanan Subramanian. Identifying predictive causal factors from news streams. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2338–2348, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1238. URL https://www.aclweb.org/anthology/D19-1238.

Morton B Brown and Alan B Forsythe. Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346):364–367, 1974.

CMU. Carnegie mellon university motion capture database, 2009. URL http://mocap.cs.cmu.edu/.

William Conover and Ronald Iman. [rank transformations as a bridge between parametric and nonparametric statistics]: Rejoinder. *American Statistician - AMER STATIST*, 35: 124–129, 08 1981. doi: 10.1080/00031305.1981.10479327.

Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer, 1992.

Juan L. Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15464–15475. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/b197ffdef2ddc3308584dce7afa3661b-Paper.pdf.

Andrew Gelman and Jennifer Hill. *Causal inference using regression on the treatment variable*, page 167–198. Analytical Methods for Social Research. Cambridge University Press, 2006. doi: 10.1017/CBO9780511790942.012.

Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12615–12625. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/94e70705efae423efda1088614128d0b-Paper.pdf.

C. W. J. Granger. Essays in econometrics. chapter Investigating Causal Relations by Econometric Models and Cross-spectral Methods, pages 31–47. Harvard University Press, Cambridge, MA, USA, 2001. ISBN 0-521-79697-0. URL http://dl.acm.org/citation.cfm?id=781840.781842.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

Clive WJ Granger, Bwo-Nung Huangb, and Chin-Wei Yang. A bivariate causality between stock prices and exchange rates: evidence from recent asianflu. *The Quarterly Review of Economics and Finance*, 40(3):337–354, 2000.

Miguel A Hernán and James M Robins. Causal inference, 2010.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11637–11649. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/860b37e28ec7ba614f00f9246949561d-Paper.pdf.

Trent Kyono, Yao Zhang, and Mihaela van der Schaar. Castle: Regularization via auxiliary causal graph discovery. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1501–1512. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1068bceb19323fe72b2b344ccf85c254-Paper.pdf.

Carlene MM Lawes, Simon Thornley, Robert Young, Raewyn Hopkins, Roger Marshall, Wing Cheuk Chan, and Gary Jackson. Statin use in copd patients is associated with a

reduction in mortality: a national cohort study. *Primary Care Respiratory Journal*, 21 (1):35–40, 2012.

Yan Liu, Alexandru Niculescu-Mizil, Aurélie Lozano, and Yong Lu. Learning temporal causal graphs for relational time-series analysis. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 687–694, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Aurelie C. Lozano, Naoki Abe, Yan Liu, and Saharon Rosset. Grouped graphical granger modeling methods for temporal causal modeling. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, page 577–586, New York, NY, USA, 2009a. Association for Computing Machinery. ISBN 9781605584959. doi: 10.1145/1557019.1557085. URL https://doi.org/10.1145/1557019.1557085.

Aurelie C Lozano, Hongfei Li, Alexandru Niculescu-Mizil, Yan Liu, Claudia Perlich, Jonathan Hosking, and Naoki Abe. Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 587–596, 2009b.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, pages 154–162, 2013.

Robert J. Prill, Daniel Marbach, Julio Saez-Rodriguez, Peter K. Sorger, Leonidas G. Alexopoulos, Xiaowei Xue, Neil D. Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. Towards a rigorous assessment of systems biology models: The dream3 challenges. *PLOS ONE*, 5(2):1–18, 02 2010. doi: 10.1371/journal.pone.0009202. URL https://doi.org/10.1371/journal.pone.0009202.

Christopher J. Quinn, Negar Kiyavash, and Todd P. Coleman. Directed information graphs. *IEEE Transactions on Information Theory*, 61(12):6887–6909, 2015. doi: 10.1109/TIT.2015.2478440.

Alberto Riva and Riccardo Bellazzi. Learning temporal probabilistic causal models from longitudinal data. *Artificial Intelligence in Medicine*, 8(3):217–234, 1996. ISSN 0933-3657. doi: https://doi.org/10.1016/0933-3657(95)00034-8. URL https://www.sciencedirect.com/science/article/pii/0933365795000348. Temporal Reasoning in Medicine.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10 (1):1–13, 2019.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms, 2017.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

Peter Spirtes. An anytime algorithm for causal inference. In *AISTATS*, 2001.

A. Tank, I. Covert, N. Foti, A. Shojaie, and E. Fox. Neural Granger Causality for Nonlinear Time Series. *ArXiv e-prints*, February 2018.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, second edition, November 1999. ISBN 0387987800.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Albert C Yang, Chung-Kang Peng, and Norden E Huang. Causal decomposition in the mutual causation system. *Nature communications*, 9(1):1–10, 2018.

Jiaxuan You, Rex Ying, Xiang Ren, William L. Hamilton, and Jure Leskovec. Graphrnn: A deep generative model for graphs. *CoRR*, abs/1802.08773, 2018. URL http://arxiv.org/abs/1802.08773.