ITGH: Information-theoretic Granger Causal Inference on Heterogeneous Data

Sahar Behzadi[1], Benjamin Schelling[1], and Claudia Plant[1,2]

[1] Faculty of Computer Science, Data Mining, University of Vienna, Vienna, Austria
[2] ds:UniVie, University of Vienna, Vienna, Austria
{sahar.behzadi,benjamin.schelling,claudia.plant}@univie.ac.at

Abstract. Granger causality for time series states that a cause improves the predictability of its effect. That is, given two time series x and y, we are interested in detecting the causal relations among them considering the previous observations of both time series. Although, most of the algorithms are designed for causal inference among homogeneous processes where only time series from a specific distribution (mostly Gaussian) are given, many applications generate a mixture of various time series from different distributions. We utilize Generalized Linear Models (GLM) to propose a general information-theoretic framework for causal inference on heterogeneous data sets. We regard the challenge of causality detection as a data compression problem employing the Minimum Description Length (MDL) principle. By balancing the goodness-of-fit and the model complexity we automatically find the causal relations. Extensive experiments on synthetic and real-world data sets confirm the advantages of our algorithm ITGH (for **I**nformation-**T**heoretic **G**ranger causal inference on **H**eterogeneous data) compared to other algorithms.

1 Introduction

Discovery of causal networks from observational data, where no certain information about their distribution is provided, is a fundamental problem with many applications in science. Among several notions of causality, Granger causality [7] is a popular method for causal inference in time series due to its computational simplicity. It states that a cause improves the predictability of its effect in the future. That is, given two time series x and y, considering the previous observations of y together with x improves the predictability of x if y causes x. There are various algorithms in this area depending on how we measure the predictability. Usually, any improvement in the predictability is measured in terms of variance of the prediction errors (known as Granger test, shortly GT).

In this paper we establish our method based on an information-theoretic measurement of the predictability. That is, we regard the challenge of causal inference as a data compression problem. In other words, employing the *Minimum Description Length* (MDL) principle, y causes x if considering the past of y together with x decreases the number of bits required to encode x. Unlike other information-theoretic approaches (e.g. entropy-based algorithms [16]), we incorporate complexity of the models in the MDL-principle. Thus, it leads to a natural trade-off among model complexity and goodness-of-fit while avoiding

over–fitting. Although Granger causality is well-studied, most of the algorithms are designed for homogeneous data sets where time series from a specific distribution are provided. Recently, Budhathoki *et al.* proposed a MDL-based algorithm designed for causal inference on binary time series [6]. Additive Noise Models (ANMs) have been proposed for either continuous [13] or discrete [12] time series. Graphical Granger approaches, which are popular due to their efficiency, mostly consider additive causal relations with a certain Gaussian assumption, e.g. TCML [1] or [2]. Despite the efficiency of homogeneous algorithms, many applications generate heterogeneous data, i.e. a mixture of various time series from different distributions. Moreover, transforming a time series to another time series with a specific distribution leads to inaccuracy. Therefore, applying an algorithm designed for homogeneous data sets on heterogeneous data does not guarantee a high performance.

Thus, integrating processes of various distributions without any transformation or certain assumptions sounds crucial. In this paper, we utilize *Generalized Linear Models* (GLMs) to extend the notion of Granger causality and introduce an integrative information-theoretic framework for causal inference on heterogeneous data regardless of time series distributions. Moreover, unlike many other algorithms, we aim at detecting causal networks. To the best of our knowledge, almost all of the existing algorithms are designed based on a pairwise testing approach which is inefficient in causal network discovery for large causal networks. To avoid this issue, we propose our MDL-based greedy algorithm (ITGH) to detect heterogeneous Granger causal relations in a GLM framework. Our approach consists of the following contributions:

Effectiveness: We introduce a MDL-based indicator for detecting Granger causal relations when ensuring the effectiveness by balancing goodness-of-fit and model complexity;

Heterogeneity: Applying the GLM methodology, we propose our heterogeneous MDL-based algorithm to discover the causal interactions among a wide variety of time series from the exponential family;

Scalability: Due to the proposed greedy approach, we might not find the overall optimal solution, but it makes ITGH scalable and convenient to be used in practice. Moreover, our extensive experiments confirm its efficiency;

Comprehensiveness: Our approach is comprehensive in the sense that we avoid any assumption about the distribution of data by applying an information-theoretic approach.

In the following, first, we present the related work in Section 2. In Section 3, we elaborate the theoretical aspects of ITGH providing the required background. In Section 4, we introduce our greedy algorithm ITGH. Extensive experiments on synthetic and real-world data sets are demonstrated in Section 5.

2 Related Work

Granger causality states that a cause (y) efficiently improves the predictability of its effect (x). There are various approaches to infer the causality depending on how to measure the predictability. Typically, any improvement in the predictability is measured in terms of variance of the error by a hypothesis testing

approach [10,14]. Moreover, graphical Granger methods are designed based on a penalized estimation of *vector autoregressive* (VAR) models [1,18]. The intention in this approach is that, if y causes x it has non-zero coefficients in the VAR model corresponding to x. First, Arnold *et al.* [1] proposed a Lasso penalized estimation for VAR models (TCML). As an extension, Bahadori and Liu [3] proposed a semi-parametric algorithm for non-Gaussian time series. Recently, authors in [5] employed adaptive Lasso to generalize this approach to the heterogeneous cases (HGGM). As another category, probabilistic approaches interpret the predictability as the improvement in the likelihood. Among them, Kim and Brown [8] introduced a probabilistic framework (SFGC) for Granger causal inference on mixed data sets by a pairwise testing of the maximum likelihood ratio. The approach is FDR-based where the statistical power of this methods rapidly decreases with increasing the number of hypotheses. As another approach, information-theoretic methods detect the causal direction by introducing a causal indicator. Among them, transfer entropy, shortly TEN, is designed based on Shannon's entropy [16] to infer linear and non-linear causal relations. In this approach, it is more likely that the causal direction with the lower entropy corresponds to the true causal relation. However, due to pairwise testing and its dependency on the lag variable, the computational complexity of TEN is exponential in the lag parameter. On the other hand, compression-based algorithms apply the Kolmogorov complexity and define a causal indicator based on the MDL-principle. Unlike the entropy-based approach, we incorporate the complexity of the models in the MDL-principle leading to more efficiency. Recently, Budhathoki *et al.* [6] proposed a MDL-based algorithm (CUTE) to infer the Granger causality among event sequences in a pairwise testing manner. This algorithm is designed only for binary time series. To the best of our knowledge, ITGH is the only algorithm in this approach which deals with discrete and continuous time series and supports the heterogeneity of data sets.

3 Theory

How to detect the Granger causal direction among any two time series? How to extend this concept to a general heterogeneous case? Could an information-theoretic approach lead to causal inference? These are fundamental questions we address in this section while providing the required background, simultaneously.

3.1 Granger Causality

Granger causality, introduced in the area of economics [7], is a well-known notion for causal inference among time series. Granger causality captures the temporal causal relations among time series although it is not meant to be always equivalent to the true causality since the question of "true causality" is deeply philosophical. Let $x = \{x^t | t = 1, \ldots, n\}$ and $y = \{y^t | t = 1, \ldots, n\}$ denote two stationary time series x and y up to time n, respectively. Moreover, let $\mathcal{I}(t)$ be all the information accumulated since time t and $\mathcal{I}_{\neg y}(t)$ denote all the information apart from the specified time series y up to time t.

Definition 1. ***Granger Causality:*** *Given two time series x and y, y Granger-causes x if including previous values of y along with x improves the predictability of x, i.e. $\mathcal{P}(x^t | \mathcal{I}_{\neg y}(t-1)) < \mathcal{P}(x^t | \mathcal{I}(t-1))$ where $\mathcal{P}$ denotes the predictability.*

More precisely, let Model 1 denote the *autoregressive* (AR) model of order d (the lag) corresponding to time series x and Model 2 denote the *vector autoregressive* (VAR) model w.r.t. x including the lagged observations of x and y.

$$x^t = \gamma_{t-d} \cdot x^{t-d} + ... + \gamma_{t-1} \cdot x^{t-1} + \epsilon^t \quad \text{(Model 1)}$$

$$x^t = \alpha_{t-d} \cdot x^{t-d} + ... + \alpha_{t-1} \cdot x^{t-1} + \beta_{t-d} \cdot y^{t-d} + ... + \beta_{t-1} \cdot y^{t-1} + \epsilon^t \quad \text{(Model 2)}$$

Thus, y causes x if the second model improves the predictability of x.

Here, the processes are assumed to be Gaussian in Model 1 and 2 and hence a linear model is considered overall. Moreover, in a linear model the error term (ϵ^t) is an additive Gaussian white noise with mean 0 and variance 1. However, these assumptions are not necessarily true in most of the applications. Thus, it is crucial to generalize the linear models to the non-linear cases in the sense that we include time series from various distributions and avoid any information loss resulted by a simple conversion.

3.2 General Causal Framework

We extend the Granger causality to a general GLM framework where a wide variety of distributions are included and no transformation is required. GLM, introduced by Nelder and Baker in [11], is a natural extension of the linear regression to the case where time series can have any distribution from the exponential family. Therefore, the response variable is not a simple linear combination of covariates but its mean value is related to the covariates by a *link function*. Corresponding to every distribution, there is an appropriate canonical link function [11]. Thus, we generalize the models introduced in Section 3.1 as follows (Model 1 $\rightarrow$ Model 3 and Model 2 $\rightarrow$ Model 4):

$$E(x^t|x) = g(\gamma_{t-d} \cdot x^{t-d} + ... + \gamma_{t-1} \cdot x^{t-1}) + \epsilon^t \quad \text{(Model 3)}$$

$$E(x^t|x,y) = g(\alpha_{t-d} \cdot x^{t-d} + ... + \alpha_{t-1} \cdot x^{t-1} + \beta_{t-d} \cdot y^{t-d} + ... + \beta_{t-1} \cdot y^{t-1}) + \epsilon^t \quad \text{(Model 4)}$$

where g is the appropriate link function w.r.t. the distribution of time series x. GLM relaxes the Gaussianity assumptions about the involved time series and the error term. Therefore, ϵ^t does not necessarily follow a standard Gaussian distribution and it can have any distribution from the exponential family leading to more accurate models. In the following we denote Model 3 and Model 4 as M_x and M_{xy}, respectively. Thus, y causes x if M_{xy} results in an improvement in the predictability of x compared to M_x. Next, we propose an information-theoretic approach to measure the improvement in the predictability.

3.3 Information-theoretic measuring of Causal Dependencies

How to measure the predictability? In this paper, we regard measuring the predictability to a compression problem. That is, we employ the description length [4] of time series in the sense that the more predictable a time series is the less number of bits is required to compress and describe it.

MDL-Principle Essentially, MDL [4] is a well-known model selection approach to evaluate various models and find the most accurate one considering the minimum description length criteria. MDL-principle regards the model selection challenge to a data compression problem in the sense that more accurate models lead to less compression cost. Let $\mathcal{M}$ denote a set of various candidate models representing your data. Following the two-part MDL [4], the best fitting model $M \in \mathcal{M}$ is the one which minimizes $DL(D, M) = DL(D|M) + DL(M)$ where $DL(D|M)$ concerns the description length of the data set D encoded by means of the model M and $DL(M)$ represents the model complexity, i.e. cost of encoding the model itself.

We consider $DL(D, M)$ as a model selection indicator. That is, employing a coding scheme, the number of bits required to encode the data indicates the accuracy of the model used in the coding process. According to the Shannon coding theorem [17], the ideal code length is related to the likelihood and is bounded by the entropy. More precisely, for an outcome a the number of bits required for coding is defined by $\log_2 \frac{1}{PDF(a)}$, where $PDF(.)$ shows the *probability density function* (a relative likelihood of a) with the assumption that $\lim_{PDF(a)\to 0^+} PDF(a)\log_2(PDF(a)) = 0$. This coding scheme is also known as *log loss*. As a consequence, we assign shorter bit strings to the outcomes with higher probability and longer bit strings to outcomes with lower probability. Thus, the better the model fits the data, the more likely the observations are and hence the less the compression cost is.

Causal Inference by MDL Back to Section 3.2, let $P(x^t|x^{t-d}, ..., x^{t-1})$ denote the predictive model w.r.t. Model 3 showing the probability of an outcome $x^t, t = 1, ..., n$ w.r.t. the lagged observations of x up to time $t-1$. We assume that P belongs to a class of prediction strategies, i.e. $P \in \mathcal{P}$. Thus, following MDL-principle, the coding cost of time series x assuming Model 3 is defined as:

$$DL(x|M_x) = \sum_{t=d}^{n} -\log P(x^t|x^{t-d}, ..., x^{t-1}) \tag{5}$$

Moreover, let $P(x^t|x^{t-d}, ..., x^{t-1}, y^{t-d}, ..., y^{t-1})$ denote the predictive model w.r.t. Model 4 assuming the past observations of x and y. Analogously, the coding cost of time series x assuming Model 4 is defined as:

$$DL(x|M_{xy}) = \sum_{t=d}^{n} -\log P(x^t|x^{t-d}, ..., x^{t-1}, y^{t-d}, ..., y^{t-1}) \tag{6}$$

Referring to the generalized definition of Granger causality (Section 3.2), time series y causes x when using M_{xy} instead of M_x improves the predictability of x. That is, if y causes x, including y leads to higher probability for the observations in x, i.e. $P(x^t|x^{t-d}, ..., x^{t-1}) < P(x^t|x^{t-d}, ..., x^{t-1}, y^{t-d}, ..., y^{t-1})$. Since higher probabilities (more accurate models) result the smaller number of required bits for encoding the data (Section 3.3), therefore $DL(x|M_{xy}) < DL(x|M_x)$.

The next part of MDL incorporates the model complexity. Thus, we say, M_{xy} fits the characteristics of the data more appropriately only if it is beneficial in terms of the model cost too, i.e. $DL(x|M_{xy}) + DL(M_{xy}) < DL(x|M_x) + DL(M_x)$. In the next section we introduce the model complexity in more detail.

3.4 Heterogeneous MDL-based Granger Causal Framework

Given p time series $x_1, ..., x_p$, the generalized VAR model of order d w.r.t. x_i is ${x_i}^t = g_i(\mathcal{X}^t \cdot \beta_i)$ where $\mathcal{X}^t$ denotes a concatenated vector of lagged observations $X^{t-d}, ..., X^{t-1}$ corresponding to all p time series and d is the lag. β_i is the regression coefficient vector consisting of $p \times d$ coefficients. Now we extend the MDL-based definition of Granger causality to a general form.

Definition 2. ***Multivariate MDL-based Granger Causality:*** *Let $\mathcal{C}_i$ denote the set of all causal time series corresponding to x_i together with x_i itself where $|\mathcal{C}_i| \leq p$ for $i = 1, ..., n$. Then, $M_{\mathcal{C}_i}$ is a generalized VAR model w.r.t. x_i including the lagged observations of time series in $\mathcal{C}_i$. Moreover, Let $M_{\mathcal{C}_i \cup x_j}$ represent the generalized VAR model w.r.t. x_i including all causal time series together with x_j. Then, x_j Granger-causes x_i if $DL(x_i, M_{\mathcal{C}_i \cup x_j}) < DL(x_i, M_{\mathcal{C}_i})$.*

In the following we clarify how to encode a time series and compute the corresponding description length ($DL(.)$).

Predictive Coding Scheme: One of the well-known approaches to encode time series is the predictive coding scheme where the prediction error w.r.t. a time series together with the parameters of the corresponding predictive model are encoded and transmitted. This scheme comprise three major components, i.e. a prediction model, the error term and an encoder. As a prediction model for a time series $x_i, i = 1, ..., p$ we consider the generalized VAR model as introduced in Definition 2. Let $\hat{x_i}^t$ be the predicted value of x_i at time t. Then, the prediction error ${e_i}^t$ is the difference between the observed value ${x_i}^t$ and the estimated value $\hat{x_i}^t$, i.e. ${e_i}^t = {x_i}^t - \hat{x_i}^t$. Finally the prediction error needs to be encoded by a an encoder and transmitted to the receiver along with parameters of the prediction model.

Fit the Distribution: Most of the time only observational data is provided in practice where the true distributions for the time series are not known. In this paper, we follow the MDL-principle discussed in Section 3.3 to find the most fitting predictive model for the data. That is, we assume a set of candidate prediction strategies from the exponential family. Considering every candidate, we estimate the parameters for the generalized AR model (M_x) employing an estimator (e.g. maximum likelihood). As discussed in Section 3.3, the more a model fits the data, the smaller the description length is. More precisely, let $\mathcal{P} = \{P_1, ..., P_m\}$ denote the set of the candidate prediction strategies (probability distributions) from the exponential family e.g. Gaussian, Poisson or Gamma. Thus, the optimal predictive model $P \in \mathcal{P}$ w.r.t. x is defined as $P = \min_{\forall P_i \in \mathcal{P}} DL_i(x, M_x)$

Objective Function: Considering the predictive coding scheme, the prediction error needs to be encoded. In order to correctly decode the data, the model as well is required to be coded and transferred. We first focus on the error coding costs then on the model complexity and finally we introduce our integrative objective function for heterogeneous time series.

Following the properties of a GLM framework, the prediction errors can have any distribution from the exponential family [11]. Since the true distribution for the error term is also unknown, we employ our proposed fitting procedure, discussed in the previous section, to find the most accurate distribution w.r.t. the error term. Thus, the coding cost of the error e_i w.r.t. x_i is defined as:

$$DL(x_i|M_{\mathcal{C}_i}) = DL(e_i) = \sum_{t=1}^{n} -\log PDF_e({e_i}^t|{e_i}^{t-1}, ..., {e_i}^{t-d}) \quad (7)$$

where $PDF_e(.)$ is the most accurate model w.r.t. e_i and n is the length of time series x_i. Moreover, assuming the prediction model $M_{\mathcal{C}_i}$ w.r.t. x_i, the parameters in this model are the regression coefficients or β_i (a vector of length $p \times d$) plus g_i, the appropriate link function. Following a central result from the theory of MDL [15], the parameter costs to model n observations of x_i w.r.t. the prediction model $M_{\mathcal{C}_i}$ is approximated by $DL(M_{\mathcal{C}_i}) = \frac{m_i}{2} \log n$ where m_i denote the number of parameters in $M_{\mathcal{C}_i}$, i.e. $m_i = p \times d + 1$. The model cost depends logarithmically on the length of time series x_i. The intention behind this formulation is that for shorter time series the parameters do not need to be coded with very high precision. However, we consider time series with the same length in this paper. Altogether, for a data set D consisting of time series $x_1, ..., x_p$ our MDL-based objective function is defined as $DL(D, M) = \sum_{i=1}^{p} DL(x_i|M_{\mathcal{C}_i}) + DL(M_{\mathcal{C}_i})$ where $M = \{M_{\mathcal{C}_i}|i = 1, ..., p\}$.

Algorithm 1 Granger Causal Network Detection by ITGH

```
1: ITGH (X = [x_1, ..., x_p])
2: adj = [0] // Output, a p × p adjacency matrix
3: fitDistribution(X);
4: for all x_i in X do
5:    S_i:= Sorted time series according to their dependencies w.r.t. x_i
6:    C_i = {x_i} // The set of all causal time series w.r.t. x_i
7:    DL_I = 0 // The cost including the candidate time series
8:    DL_E = 0 // The cost excluding the candidate
9:    while DL_I ≤ DL_E do
10:      x_j := The candidate, the first time series in S_i
11:      DL_I = DL(x_i, M_{C_i ∪ x_j})
12:      DL_E = DL(x_i, M_{C_i})
13:      if DL_I ≤ DL_E then
14:         adj(i, j) = 1 // x_j causes x_i
15:         remove x_j from S_i
16:         C_i = C_i ∪ x_j
17:      end if
18:   end while
19: end for
20: return (adj)
```

4 ITGH Algorithm

To cope with the inefficiency resulted by a pairwise testing, we propose our greedy-based ITGH algorithm consisting of two main building blocks: (1) fitting a distribution to the time series and (2) detecting the Granger causal network in a greedy way. Considering *fitDistribution(.)* in Algorithm 1, once we find the most accurate fitted distribution w.r.t. every time series as explained already. Then, we use this information as an assumption in our greedy algorithm. To be fair,

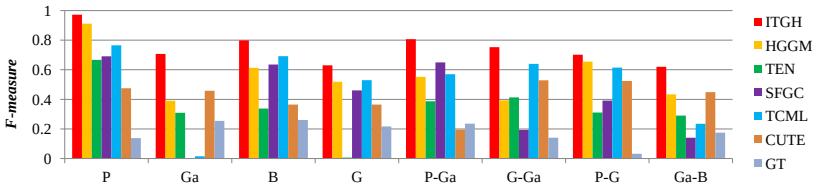

Fig. 1. Investigating the accuracy. P: Poisson, Ga: Gamma, G: Gaussian, B: Bernoulli

we also input the fitted distributions to other comparison methods. Moreover, for every x_i, we sort $x_1, ..., x_p$ based on their dependencies in the corresponding regression model. In fact (also inspired by [1]), the time series with the higher dependency w.r.t. x_i has the higher coefficients in the regression model. Thus, we iteratively include the time series with the higher dependency w.r.t. x_i in the regression model as far as this procedure improves the compression cost of x_i. Essentially, for a candidate x_j we compute the description length of x_i (see Definition 2) considering two models $M_{\mathcal{C}_i}$ and $M_{\mathcal{C}_i \cup x_j}$. If including x_j pays off in terms of the compression cost, we keep including the next time series. Otherwise, the procedure terminates when no further causes exist for x_i. The output of this algorithm is an adjacency matrix for the Granger causal network. ITGH is deterministic in the sense that investigating the causal relations for p time series in any random order leads to the same causal graph. The runtime complexity of ITGH in the best case is $\mathcal{O}(p^2 \log(p)) + \mathcal{O}(pc^2 n)$ and in the worst case is $\mathcal{O}(p^2 \log(p)) + \mathcal{O}(p^2 c^2 n)$ where c is $d \times |\mathcal{C}_i \cup x_j|$. However, mostly in reality $p \ll n$ which means the runtime complexity of ITGH is highly depending on n leading to a complexity of order $\mathcal{O}(c^2 n)$. For a detailed analysis of the computational complexity please check the appendix.

5 Experiments

To assess the performance of ITGH we conduct several experiments on synthetic and real-world data sets in terms of *F-measure*. We compare ITGH to SFGC [8], TEN [16] and HGGM [5] which are designed to deal with heterogeneous data sets. Moreover, we compare our algorithm to TCML [1], CUTE [6] and the basic Granger test (GT) [7] to investigate the effect of assuming a specific (mostly Gaussian) distribution for non-Gaussian processes or transforming time series. ITGH is implemented in MATLAB and for the other comparison methods we used their publicly available implementations and recommended parameter settings. The source code and data sets are publicly available at: `https://tinyurl.com/yar5yuoq`.

5.1 Synthetic Experiments

In any synthetic experiment, we report the average performance of 50 iterations performed on different data sets with the given characteristics. The length of generated time series is always 1,000 except it is explicitly mentioned. Unless otherwise stated, we assume a random dependency level (strength of causal relations) among time series. In all the synthetic experiments we input the lag parameter as well as the true distributions to all the algorithms.

Accuracy: In this experiment we generated various data sets from different distributions. Two discrete (Poisson and Bernoulli) and two continuous

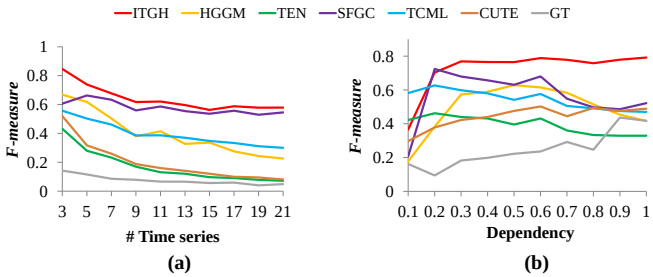

Fig. 2. Various experiments on synthetic data sets concerning the effectiveness.

(Gamma and Gaussian) distributions were selected to cover some of possible combinations of distributions. Every data set consists of four time series with three causal relations where in mixed data sets the heterogeneity factor is 70%-30% (e.g. 3 Poisson and 1 Gaussian). As it is observable in Figure 1, regardless of the homogeneity or heterogeneity of the data or even the distribution of the time series, ITGH outperforms other algorithms by a wide margin. Interestingly, confirming the advantages of an MDL approach applied in a GLM framework we outperform TCML on Gaussian data set although it is designed specifically for Gaussian time series and performs better than other algorithms on such data sets. On the other side, we outperform CUTE on the Bernoulli data set due to the inefficiency of pairwise testing compared to our proposed greedy approach. In the following we focus on a mixture of time series having Poisson and Gamma distribution as a representative for heterogeneous data sets.

Effectiveness: This experiment specifically investigates the effectiveness of the greedy approach in ITGH in terms of *F-measure* when the number of time series is increasing. Here we generate heterogeneous data sets where in any case 70% of the time series are Poisson and 30% are Gamma distributed and the number of causal relations is equal to 0.67% of the number of time series. It is already expected that the performance of an exhaustive pairwise testing approach is decreasing when dealing with larger graphs. Figure 2a confirms our expectation and illustrates the constantly descending performance of HGGM, TEN and CUTE. As excepted, GT and SFGC are quite stable. However, GT is the worst algorithm in this experiment resulting in a maximum *F-measure* of 0.14. Moreover, this experiment shows the advantages of ITGH and SFGC compared to other algorithms regardless of the number of time series, although in the beginning their performance is affected by growing the causal graph.

Dependency: We refer to the coefficients of VAR models as the dependency which essentially show the strength of causal relations. In this experiment we investigate the performance of the algorithms concerning various dependencies ranging from 0.1 to 1. Analogously, we focus on data sets where a mixture of 3 Poisson and 1 Gamma time series are generated. In Figure 2b any ascending or descending trend shows the inefficiency while a constant trend confirms the ability of an algorithm to deal with strong and weak causal relations. ITGH generally outperforms other competitors in terms of *F-measure* and unlike other algorithms, varying the dependency does not influence the performance of our algorithm significantly. Ignoring the starting point, the stable trend of ITGH confirms the efficiency of our algorithm even for lower dependency levels. Unexpectedly, the performance of TCML, SFGC and TEN is slightly descending in this experiment.

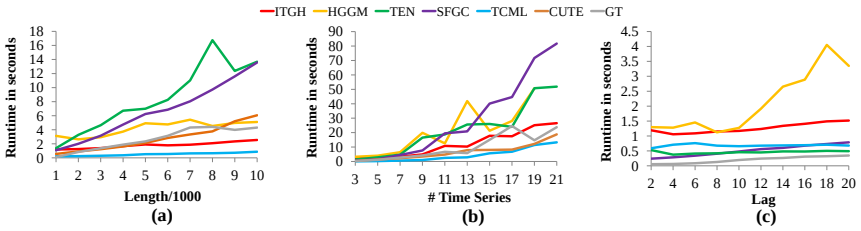

Fig. 3. Investigating the scalability in various experiments.

Scalability: While investigating the Scalability we generate data sets with the same setting as previous experiment concerning the effectiveness. During the first experiment we vary the length of time series ranging from 1,000 to 10,000 when the number of time series is set to five. As Figure 3a depicts, ITGH is the second fastest algorithm in this experiment and outperforms HGGM, TEN and SFGC. Together with TCML, our algorithm shows a perfect stable trend when increasing the length of time series. In the other experiment we iteratively increase the number of time series. As expected, all the algorithms have an increasing trend (Figure 3b). However, we outperform other heterogeneous algorithms in this experiment as well. Finally, algorithms are investigated when the lag is increasing. Except HGGM, all other algorithms are almost stable in this experiment (Figure 3c). Although ITGH seems to be relatively time-consuming compared to others in this experiment, its runtime is less than 1.5 seconds and still reasonable.

5.2 Real Applications with Ground Truth

We conduct various experiments on publicly available real-world data sets where a valid ground truth is provided. Table 1 summarizes the characteristics of the data sets while we input the same fitted distribution to every algorithm resulted by *fitDistribution(.)* procedure. To be fair, we report the best result for any algorithm in Table 1 in terms of *F-measure* when considering various lags ranging from 1 to 20. Moreover, we conducted various experiments on the lag variable in appendix which is specially interesting in real-world experiments. For the data sets marked with *, the ground truth is given partially and the information about some interactions is missing. Therefore, corresponding to any data we report the average *F-measure* w.r.t. the causal pairs where the true information is given. As it is clear from Table 1, ITGH outperforms other algorithms on almost all the data sets (except *Spike Train*). However, because of the space limitation a detailed analysis of the results as well as data sets is not possible here, please check the appendix.

5.3 Application to Climatology

What causes the climate changes? In this experiment, we investigate causal relations between the climate observations and various natural and artificial forcing factors when no ground truth is provided. The data set, provided in [9], is publicly available. We consider the monthly measurements of 11 factors over 13 years (from 1990 to 2002) in two states in the US, i.e. Montana and Louisiana: temperature (TMP), precipitation (PRE), vapor (VAP), cloud cover (CLD), wet

Data set	*Distribution*	*Length*	*ITGH*	*SFGC*	*HGGM*	*TEN*	*TCML*	*CUTE*	*GT*
Traffic	1 P, 1 B	254	**1.00**	0.67	**1.00**	0.00	0.00	**1.00**	0.00
Ozone	2 G	365	**1.00**	0.50	0.67	0.00	0.40	0.00	0.67
Speed	2 Ga	202	**1.00**	0.00	**1.00**	0.00	0.00	**1.00**	0.00
Temperature	2 G	168	**1.00**	0.00	0.00	**1.00**	0.40	**1.00**	0.67
Mooij	2 G	16382	**1.00**	0.67	0.67	0.00	0.00	**1.00**	0.67
* Moffat	2 Ga, 1 G	721	**1.00**	0.33	0.67	0.50	0.45	0.00	0.67
* Abalone	1 P, 3 G	4177	**1.00**	0.56	0.00	**1.00**	0.00	**1.00**	0.67
* Energy	1 P, 2 G	9504	**0.89**	0.00	0.55	0.30	0.56	0.67	0.67
Spike Train	4 B	1000	0.62	0.47	0.00	0.57	0.20	**0.76**	0.50

Table 1. Comparison on real data sets including a ground truth in terms of *F-measure*.

days (WET), frost days (FRS), green house gases including Methane (CH4), Carbon Dioxide (CO2), Hydrogen (H2) and carbon monoxide (CO) and solar radiation including global extraterrestrial (GLO). After fitting the distribution for any time series, we apply ITGH and other heterogeneous methods inputting the most appropriate distribution. The data providers suggested a maximum lag of 4 [9]. However, no exact information about the lag is given. Therefore, the lag is randomly set to 3 for Louisiana and 2 for Montana. Since the temperature is the most concerning factor in global warming and also for a better visualization, we focus on the factors which influence the temperature. Green house gases, specially CO2, as well as solar radiation are the most important factors in global warming. Moreover, depending on where a state is located, cold or warm region, various climate measurements influence the temperature. According to the annual average temperature of states in the US, Louisiana is located in the warm region where the CO2 concentration is also high. As Figure 4a shows, ITGH correctly detects CO2 and the solar radiation as causal factors for temperature (confirmed by [9]). Moreover, influencing the temperature by VAP is also plausible since Louisiana is located in the warm subtropical region. On the other side, the result of SFGC does not sound interpretable since it finds a causal relation among all the factors and the temperature, even the frost days per month. HGGM seems more efficient compared to SFGC, However, it does not find any effects caused by one of the most effective factors, i.e. CO2. Unlike Louisiana, Montana is located in the cold region. Therefore, the detected causal direction from the frost days and vapor to the temperature in Figure 4b is reasonable (also confirmed by [9]). However, HGGM is not able to find the relation among the frost days and the temperature. Moreover, the CO2 concentration in this state is not high. Therefore, CO2 does not influence the temperature in Montana dramatically. ITGH correctly does not consider a causal relation among CO2 and temperature while SFGC does. On the other side, HGGM is not able to find the effect of frost days, although it correctly recognizes the relation between CO2 and the temperature.

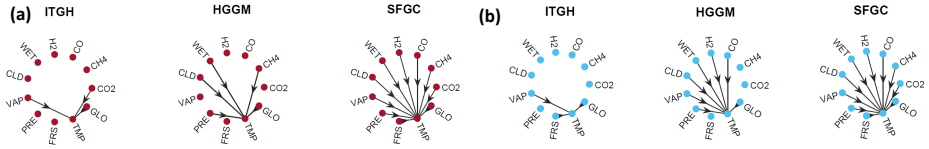

Fig. 4. Application to Climatology. a) causal graphs w.r.t. Louisiana, b) causal graphs w.r.t. Montana.

6 Conclusions and future work

In this paper we proposed ITGH, an information-theoretic algorithm for discovery of causal relations in a mixed data set while profiting of a GLM framework. Following the MDL-principle, we introduced an integrative objective function applicable for time series having distributions from the exponential family. Our greedy approach leads to an effective and efficient algorithm without any assumption about the distribution of the data. One of the avenues for future work is to employ our MDL-based approach to efficiently detect the anomalies in heterogeneous data sets.

References

1. Arnold, A., Liu, Y., Abe, N.: Temporal causal modelling with graphical Granger methods. In: KDD (2007)
2. Bahadori, M.T., Liu, Y.: Granger causality analysis in irregular time series. In: SDM (2012)
3. Bahadori, M.T., Liu, Y.: An examination of practical granger causality inference. In: SDM (2013)
4. Barron, A., Rissanen, J., Yu, B.: The minimum description length principle in coding and modeling. IEEE Transactions on Information Theory (1998)
5. Behzadi, S., Schindler, K., Plant, C.: Granger causality for heterogeneous processes. In: PAKDD (2019)
6. Budhathoki, K., Vreeken, J.: Causal inference on event sequences. In: SDM (2018)
7. Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica (1969)
8. Kim, S., Putrino, D., Ghosh, S., Brown, E.N.: A granger causality measure for point process models of ensemble neural spiking activity. PLOS Computational Biology (2011)
9. Liu, Y., Niculescu-Mizil, A., Lozano, A., Lu, Y.: Learning temporal causal graphs for relational time-series analysis. In: ICML (2010)
10. Lütkepohl, H.: New introduction to multiple time series analysis. Springer (2005)
11. Nelder, J.A., Baker, R.J.: Generalized linear models. Encyclopedia of statistical sciences (1972)
12. Peters, J., Janzing, D., Schölkopf, B.: Causal inference on discrete data using additive noise models. IEEE transactions on pattern analysis and machine intelligence (2011)
13. Peters, J., Mooij, J.M., Janzing, D., Schölkopf, B.: Causal discovery with continuous additive noise models. J. Mach. Learn. Res. **15**, 2009–2053 (2014)
14. Quinn, C.J., Coleman, T.P., Kiyavash, N., Hatsopoulos, N.G.: Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. Journal of Computational Neuroscience (2011)
15. Rissanen, J.: A universal prior for integers and estimation by minimum description length. The Annals of Statistics (1983)
16. Schreiber, T.: Measuring information transfer. Phys. Rev. Lett. (2000)
17. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27**(4), 623–56 (1948)
18. Shojaie, A., Michailidis, G.: Discovering graphical Granger causality using the truncating lasso penalty. Bioinformatics (2010)