Evaluating Temporal Observation-Based Causal Discovery Techniques Applied to Road Driver Behaviour

Rhys Howard RHYSHOWARD@LIVE.COM and **Lars Kunze** LARS@ROBOTS.OX.AC.UK
Oxford Robotics Institute, Dept. of Engineering Science, University of Oxford, 17 Parks Road, Oxford, OX1 3PJ, United Kingdom

Editors: Mihaela van der Schaar, Dominik Janzing and Cheng Zhang

Abstract

Autonomous robots are required to reason about the behaviour of dynamic agents in their environment. The creation of models to describe these relationships is typically accomplished through the application of causal discovery techniques. However, as it stands observational causal discovery techniques struggle to adequately cope with conditions such as causal sparsity and non-stationarity typically seen during online usage in autonomous agent domains. Meanwhile, interventional techniques are not always feasible due to domain restrictions. In order to better explore the issues facing observational techniques and promote further discussion of these topics we carry out a benchmark across 10 contemporary observational temporal causal discovery methods in the domain of autonomous driving. By evaluating these methods upon causal scenes drawn from real world datasets in addition to those generated synthetically we highlight where improvements need to be made in order to facilitate the application of causal discovery techniques to the aforementioned use-cases. Finally, we discuss potential directions for future work that could help better tackle the difficulties currently experienced by state of the art techniques.

Keywords: Causal Discovery, Time Series Data Analysis, Autonomous Driving

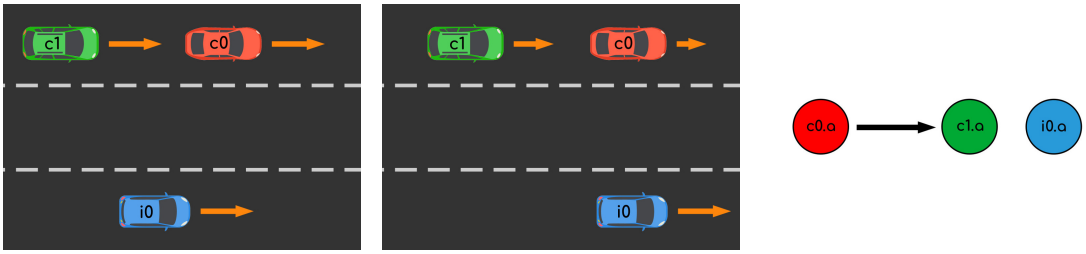
1. Introduction

With an increase in robots operating among humans, there is a growing need for these autonomous agents to understand how the actions of one agent - human or robotic - may affect the behaviour of other agents. In this paper we present a benchmark evaluation of contemporary temporal causal discovery methods in attempting to find causal relationships between agents.

The incentive for building a causal model of agent behaviours is to allow consideration of humans or other robots when planning, and then being able to retrospectively explain plans and outcomes (Hellström, 2021). Accounting for causality has also been identified as a critical component in developing trustworthy AI systems Ganguly et al. (2023).This type of work has become increasingly relevant due to the rise in robots operating in close proximity to humans, particularly in high risk domains such as driving. Human drivers must constantly be aware of other vehicles and act in a way as to not only achieve their objectives but simultaneously be aware of the consequences of their actions, both direct and indirect. By furthering discussion in this area we aim to encourage work towards developing a similar reasoning model for AI and robotics.

In order to explore the best paths forward for developing a system of causal discovery for agent behaviour the state of the art must first be evaluated. The methods covered in this work consist of those presented by Assaad et al. (2022), the NAVAR method presented by Bussmann et al. (2021) and the CD-NOD method presented by Zhang et al. (2017). In contrast to this work, the work of

© 2023 R. Howard & L. Kunze.



(*a*) Initial frame. (*b*) Post-**c0** braking. (*c*) Causal graph.

Figure 1: Illustration of the basic causal interaction we consider: The two-agent convoy scenario. Here **c0** and **c1** represent the head and tail of a two vehicle convoy respectively, while **i0** acts as an entirely independent third agent. (Freepik, 2022)

Assaad et al. only evaluates the methods on artificial data generated from common causal graph structures and fMRI simulation (Smith et al., 2011).

Working in an online fashion to establish causal links between agent behaviour presents a few challenges for the aforementioned methods. Most applications of causal discovery are upon problems with large amounts of available data where processing can be carried out in an offline setting (e.g. medicine, economics, sociology). Due to the quantity and nature of the data, there are likely to be a greater number of causal interactions resulting from causal relationships. However if data is processed in an online fashion and causal interactions indicative of causal relationships are infrequent, this leads to a condition we term causal sparsity. Under this condition, causal discovery methods must contend with being able to carry out discovery with substantially fewer observations of causal relationships being exhibited than would usually be the case. Most causal discovery tasks also assume causal stationarity in the ground truth causal relationships, either due to the synthetically generated data or due to the very nature of the problem being considered. Causal stationarity refers to causal relationships that remain constant throughout the observed time window. Because we consider human drivers, which vary in their reaction times and in the manner of their reactions, non-stationarity in their causal relationships is inevitable.

With the motivation, content and challenges in relation to this work laid out, we identify the following as the key contributions of this work:

- A benchmark evaluation of contemporary temporal causal discovery techniques applied to real world scenes which follow a predetermined Autonomous Driving (AD) scenario (See Sec. 2).
- A comparison of real world scene performance with performance on scenes synthetically generated with ideal conditions.
- A discussion of the difficulties associated with causal discovery on agent behavioural interactions on a scene-by-scene basis, and the limitations presented by some methods as a result.

We note that while this paper does focus upon an AD scenario, that the challenges encountered while tackling this problem are likely shared by other areas of autonomous robotics, particularly those where an interest in reasoning about agent behavioural interaction is present.

2. Problem Definition

To the author's knowledge the methods under consideration have not had their performance tested in the domain of AD before. Therefore we intend to carry out our investigation using a simple scenario that aims to establish the viability of the techniques in question. The types of causal relationships we are considering here are behavioural and not physical interactions between agents as the intended application of the discovered causal models is for decision explanation and introspection. In order to consider two agents as causally linked the behaviour of the causing agent must either be sufficient or necessary to produce some of the behaviour exhibited by the effected agent.

The scenario in question describes two primary agents **c0** and **c1** which form the head and tail of a two vehicle convoy. Throughout the course of the scenario, **c0** will at one point noticeably accelerate or decelerate and in response, **c1** will accelerate or decelerate accordingly, forming the basis of a causal relationship between the two agents. With the aim of demonstrating the precision of the approaches, we introduce a third independent agent **i0**. As the name suggests, the behaviour of **i0** is entirely independent of **c0** and **c1**, and as a result there should be no causal relationships detected between the convoy agents and **i0**. An illustration of this scenario is shown in Fig. 1.

We can prove that either an acceleration or deceleration case shows a valid causal link between agents by considering the alternate scenarios in which the lead vehicle did not accelerate/decelerate. In the deceleration case, as it is clear that that **c0** braking is sufficient to cause **c1** to brake, as **c1** would collide with **c0** otherwise. We can also consider the acceleration of **c0** necessary in causing **c1** to accelerate, as **c1** could not accelerate to a speed greater than **c0** without risking collision.

In terms of the numeric variables used in the causal discovery process, all three agents will rely purely either on their linear acceleration magnitudes denoted as **c0.a**, **c1.a**, and **i0.a** respectively or on their linear velocity magnitudes again denoted **c0.v**, **c1.v**, and **i0.v** respectively.

Finally, while we will not refrain from working with egocentric data, the scenario described above is not focused upon an ego agent/vehicle. Here the use of the term ego refers to data gathered or problems considered from a single agent/vehicle's perspective. In contrast to previous works (Kim and Canny, 2017; Li et al., 2020a) which have looked at causal discovery from a single point of view, we instead aim to investigate the described scenario in a perspective-agnostic manner.

3. Related Work

3.1. Causal Reasoning

Causal reasoning as a well defined area of study spawned out of work by Pearl (2009), expanding upon Bayesian networks (Pearl, 1985). Causal models provide information regarding the generative processes involved in the creation of data, rather than just the associations between variables/events. Causal discovery is the process of inferring causal models, and in the context of AD we hope to discover causal links between the behaviour of road agents.

Causal discovery can be carried out at one of three layers (Pearl and Mackenzie, 2019): Observational, Interventional, or Counterfactual. Interventional methods which inherently require experimentation are unsuitable for the AD domain we consider (Eberhardt and Scheines, 2007; Kocaoglu et al., 2017; Addanki et al., 2020). Meanwhile, counterfactual methods while interesting theoretically, have had little attention applied to them thus far outside of a meteorological simulation based approach relating to climate change (Hannart et al., 2016). As such, in this paper we focus on the application of contemporary observational approaches to the AD domain.

3.2. Causal Discovery on Time Series Data

Glymour et al. (2019) provides a description of contemporary observational causal discovery methods as well as discussing several difficulties typically encountered by these methods, including the extension of said methods to time series data. Expanding upon this Assaad et al. (2022) conducts a survey of observational approaches applied to time series data, and provides a quantitative evaluation of the approaches on synthetic and real data. The evaluated methods in question are the Granger causality based PWGC (Granger, 1969), MVGC (Geweke, 1982), and TCDF (Nauta et al., 2019) approaches, the constraint based PCMCI (Runge et al., 2019), oCSE (Sun et al., 2015), and tsFCI (Entner and Hoyer, 2010) approaches, the noise based VarLiNGAM (Hyvärinen et al., 2010) and TiMINo (Peters et al., 2013) approaches, and the score based DYNOTEARS (Pamfil et al., 2020) approach. Of these approaches oCSE is not relevant to this paper, as it assumes causal relationships only occur over a time lag of 1, which is far too little for the task we consider. In general, the methods show limited performance when self-causal relationships are excluded, which contributes towards our selection of a simple scenario to evaluate the approaches upon. In addition to the methods reviewed by Assaad et al., we also consider the Granger causality based NAVAR method presented by Bussmann et al. (2021) and the non-stationarity-based method CD-NOD presented by Zhang et al. (2017). NAVAR takes a similar approach to TCDF (Nauta et al., 2019) in that it applies neural learning techniques to the causal discovery problem, meanwhile CD-NOD bears some resemblance to traditional constraint-based methods, but additionally exploits non-stationarity to discover causal direction.

3.3. Causal Discovery in the Autonomous Driving Domain

While works combining causal discovery in the domain of AD are rare, there have been a few recent works which focus on this area. One such work is that carried out by McDuff et al. (2021). They present a detailed simulation environment to facilitate future causal research, and contrast this environment with a drastically simplified environment to demonstrate the increased challenge and realism that comes with the detail of the approach they describe. In their work they test their simulated scenarios against three causal discovery techniques.

Of the three methods evaluated, the NRI method (Kipf et al., 2018) operates on 2D particle systems and motion capture data, the NS-DR method (Yi et al., 2020) presents itself by testing on 3D particle systems, and the V-CDN method (Li et al., 2020b) consists of an entire pipeline dedicated to detecting key points on clothing and determining the physical relationships between said points with a causal model. All three of these methods can be considered visual causal discovery in that they are restricted to working upon visual input. Furthermore, due to all three of them utilising neural networks it increases the difficulty of performance verification and increases the risk of performance degradation from domain shift.

In terms of other works tackling causal discovery in the domain of AD, Kim and Canny (2017) and Li et al. (2020a) both attempt to discover which regions of a camera's view are responsible for invoking certain ego agent behaviours. The main limitation of such an approach is that it is limited to ego vehicles utilising cameras, cannot reason about the actions of other agents, and once again relies upon convolutional neural network techniques. Meanwhile de Haan et al. (2019) and Samsami et al. (2021) both utilise causal discovery as part of an imitation learning process for AD control, as opposed to our goal of evaluating behavioural interactions between agents.

4. Background

A causal directed graph is defined as $G = (V, E)$ where V corresponds to a set of variables and $E \subseteq [(v^x, v^y) \mid v^x \neq v^y, (v^x, v^y) \in V \times V]$ corresponds to a set of edges that describe causal relationships between the aforementioned variables. An edge $(v^x, v^y) \in E$ describes a causal relationship whereby variable v^x has a causal effect upon variable v^y. Causal Directed Acyclical Graphs (DAGs) comprise a subset of causal directed graphs that are without any cycles, a feature which is exploited by some causal discovery approaches when discovering DAGs. If the variables being considered consist of time series, a subscript is added to denote the variable at the specified time. For example, the variable $v^x \in V$ is denoted at time t as v^x_t.

The goal of causal discovery is to derive an approximation of the causal directed graph $\hat{G} = (\hat{V}, \hat{E})$. Here $\hat{E}$ is simply the hypothesised causal links, meanwhile $\hat{V}$ consists of the approximated set of variables, as there can exist the possibility of hidden confounders, which some causal discovery approaches make efforts to detect. It should be noted that while an assumption of causal sufficiency ensures that $\hat{V} = V$, we cannot rule out the presence of hidden confounders. The potential implications of this will be discussed in more detail in Sec. 6, but for the purposes of the discussing methodology only the three variables associated with each respective agent will ever be considered as part of our evaluation.

In terms of the problem defined in Sec. 2, we are looking to discover a causal temporal summary graph, in which there is an edge directed from the time series relating to agent **c0** to the time series relating to the agent **c1** and none besides that (See Fig. 1). A summary graph just means that provided there exists a causal relationship between two time series within a given time window given by τ an edge will be present in the summary graph to represent said relationship.

5. Observation-Based Temporal Causal Discovery

Here the we will briefly introduce the methods under consideration based upon a selection of those presented by Assaad et al. (2022) in their survey paper, in addition to the NAVAR method (Bussmann et al., 2021). We refer the reader to their papers for more substantial explanations of the methods paraphrased here.

5.1. Granger Causality Based Approaches

Granger causality is built upon the concept that causes should provide unique information that enables the prediction of their effects (Granger, 1969). The first two Granger causality based approaches assume a linear relationship between variables, which should indeed be true of the variables we consider, as the tail convoy vehicle should roughly mirror the speed and acceleration of the head convoy vehicle, albeit with a time lag. The Pair-Wise Granger Causality (PWGC) (Granger, 1969) approach determines the likelihood of a causal link between variables by considering two autoregressive models:

$$v^y_t = \sum_{t'=1}^{\tau} (a^y_{t'} v^y_{t-t'}) + \varepsilon^y_t \tag{PW-res.}$$

$$v^y_t = \sum_{t'=1}^{\tau} (a^y_{t'} v^y_{t-t'} + a^x_{t'} v^x_{t-t'}) + \varepsilon^y_t \tag{PW-full}$$

where $a^x_{t'}, a^y_{t'} \in \mathbb{R}$ represent coefficients specific to a time lag t' for variables v^x and v^y respectively and τ is the maximum tag lag considered. Meanwhile ε^y and ε^y are white-noise time series which represent the influence of exogenous non-confounding factors. If v^x does indeed have a causal effect upon v^y we would expect v^x to possess unique information allowing us to better predict v^y with (PW-full) than with (PW-res.). An F-test can be applied to an accuracy metric (e.g. Residual Sum of Squares) for each of the models to determine whether the difference is significant enough to consider a causal link to have been discovered.

The PWGC approach is limited in that it only ever considers pairs of variables and can therefore struggle with mediators relationships present in the underlying causal graph. Multi-Variate Granger Causality (MVGC) (Geweke, 1982) builds upon PWGC by proposing the following autoregressive models at the cost of higher computational overhead:

$$v^y_t = \sum_{v^i \in V^{\neg x}} (\sum_{t'=1}^{\tau} (a^i_{t'} v^i_{t-t'})) + \varepsilon^y_t \tag{MV-res.}$$

$$v^y_t = \sum_{v^i \in V} (\sum_{t'=1}^{\tau} (a^i_{t'} v^i_{t-t'})) + \varepsilon^y_t \tag{MV-full}$$

where $V^{\neg x} = V \setminus \{v^x\}$. Unlike PWGC, which only considers self causation in the restricted case and the additional information provided by a single variable in the full case, MVGC considers all bar one variable in (MV-res.) and all variables in (MV-full). Because this approach considers all information available while only excluding the information of one variable that is being examined as a cause, it is not only able to tackle mediator causal relationships, but is also robust against against confounding variables as their influence is captured in both models with equal measure.

The Temporal Causal Discovery Framework (TCDF) (Nauta et al., 2019) and Neural Additive Vector Auto-Regression (NAVAR) (Bussmann et al., 2021) approach both offer neural network based approaches to model non-linear causal relationships. TCDF utilises a form of attention-based convolutional neural network to predict for a given variable provided the historic values of all variables, the attention values attributed during training are then used to determine potential causal parents for the given variable. Meanwhile NAVAR trains a network - a Mutli-Layer Perceptron (MLP) in our experiments - for each variable to predict for all other variables, in this case relying upon the variance a variable provides in its prediction contributions to determine the presence of causal relationships.

5.2. Noise-Based Approaches

Noise-based approaches share a theoretical similarity with Granger causality based approaches in that these approaches also rely upon examining the flow of information between variables. However, in contrast to Granger causality, noise-based methods do not operate upon identifying information possessed by variables useful for predicting other variables. Instead they attempt to identify the direction of causal links by identifying information variables hold about the noise of other variables.

To explain the premise of the first noise-based method, VarLiNGAM (Hyvärinen et al., 2010), consider the following bivariate example of LiNGAM (Shimizu et al., 2006) its non-temporal predecessor. Let v^x have a causal effect upon v^y and that their distribution be generated as follows:

$$v^x = \varepsilon^x \tag{1}$$

$$v^y = a^{x,y} v^x + \varepsilon^y \tag{2}$$

where ε^x and ε^y represent noise from exogenous non-confounding factors that affect the variables. Under this distribution, v^y captures information on the noise provided by ε^x because its value is derived from v^x. However, v^x does not capture any information on ε^y, establishing an asymmetry that LiNGAM aims to exploit. It is important to note that ε^x and ε^y must not be jointly Gaussian, as under these conditions it is impossible to determine the causal direction.

In order to solve the directions of causal links the model reframes the underlying model responsible for the time series as $V = AV + \varepsilon$ where V is a vector of variables, ε is a vector of noise/error values, and A is a strictly lower triangular matrix which describes the direction of causal links between the variables of V. If the model is further refined to $V = B\varepsilon$ where $B = (I - A)^{-1}$ we can aim to try and solve for A. The initial paper on LiNGAM (Shimizu et al., 2006) applied Independent Component Analysis (ICA) (Comon, 1994) in order to achieve this. However the method explored by this paper utilises an extension named DirectLiNGAM (Shimizu et al., 2011). This method constructs an auto-regressive model and recursively checks independence between each variable acting as a predictor and the residuals given by applying that predictor to other variables. The most independent predictor is placed highest in the causal hierarchy. In each subsequent step the remaining variables are substituted for their residuals obtained during the previous step to remove the influence of variables with established positions. Since the above process only establishes the direction of causation, the strength of the causal effect can be determined by conventional covariance-based regression, before pruning is carried out by applying the Adaptive Lasso method (Zou, 2006). From here the extension to VarLiNGAM (Hyvärinen et al., 2010) is achieved by considering the variables in question over a time window defined by the maximum time lag τ:

$$V_t = \sum_{t'=1}^{\tau} (A_{t'} V_{t-t'}) + \varepsilon_t \tag{3}$$

Provided one can approximate $(A_{t'})_{1 \geq t' \geq \tau}$ the steps of DirectLiNGAM can be followed in similar fashion albeit working with time series rather than regular non-temporal variables. The use of the Adaptive Lasso remains important as prior to this step the number of causal links will be of the order $\mathcal{O}(|V|^2 \tau)$, however the risk for our problem at least more concerns the potential for the inclusion of false positives in the summary graph.

The other noise-based approach, TiMINo (Peters et al., 2013) consists not so much of a single model as much as it describes a class of models. The models described by TiMINo are assumed to adhere to the form:

$$v_t^x = f^x(pa(v_t^x, t), pa(v_t^x, t-1), ..., pa(v_t^x, t-\tau), \varepsilon_t^x) \tag{4}$$

where $pa(v_t^x, t')$ represents the causal parents of v_t^x from the time t' up to maximum time lag given by τ. The function f^x and additive noise ε_t^x are partially dependent upon one another, as if f^x is non-linear, ε_t^x should be Gaussian, and if f^x is linear, ε_t^x should be non-Gaussian. Alternatively these requirements can be relaxed if the data follows a time structure - as opposed to independent and identically distributed time indexed variables - that the joint data distribution is faithful to and there is a lack of cycles. Due to the nature of the data being worked with, this latter case captures our task more closely. In terms of how TiMINo operates, it utilises a supplied regression method and independence test. The implementation utilised by this paper (Assaad et al., 2022) utilises a linear regression model and a cross covariance based independence test with Bonferroni correction. TiMINo proceeds by learning a predictor for each time series and then determining how independent each time series involved in the predictor is from the residuals produced from applying the predictor.

The time series with the predictor that produces the greatest level of independence is deemed to be at the bottom of the causal hierarchy as little or no information regarding its own noise is present in other time series. This process is repeated until a full causal hierarchy is established. At this point the causal parents of each time series are refined by removing those time series unnecessary to produce independent residuals.

5.3. Constraint-Based Approaches

Constraint-based approaches make the assumption that the conditional independences seen in the probability distributions exhibited by observed data are reflective of the underlying causal graph structure, a condition referred to as causal faithfulness. Constraint-based approaches typically work upon the time series considered in the form of a window causal graph, where each relative time lag - time series combination represents a node within the graph. Both of the constraint-based methods we consider share a similar initial process of first identifying which which nodes are unconditionally independent before progressively checking independence upon adjacent node pairs conditional upon their neighbours. Once adjacencies remain stable one can look for unshielded triples and convert these into collider structures as these are the only structures that can be directly identified under the assumptions of the methods considered.

One of the oldest constraint-based approaches is the Peter-Clark (PC) algorithm (Spirtes et al., 2001). Following the steps outlined above the PC algorithm proceeds to iteratively apply 3 rules for refining the causal direction of adjacent nodes. Depending upon the data available it will not always be possible to orient every edge. The resulting graph describes a Markov Equivalence Class (MEC) which contains all the possible full directed causal graphs based upon the remaining undirected edges, should any be present. The PC with Momentary Conditional Independence (PCMCI) (Runge et al., 2019) extends the PC algorithm to better work with time series. While the initial steps are similar to the PC algorithm, the Momentary Conditional Independence test is designed to avoid the influence of auto-correlations by evaluating the level of dependence between nodes while conditioning upon the parents of both nodes. It is important to note that PCMCI can be used with any conditional independence test, though for this paper only the partial correlation (Baba et al., 2004) approach is considered.

The other constraint-based method, tsFCI (Entner and Hoyer, 2010) is based upon the earlier FCI algorithm (Zhang, 2008). The FCI follows the same initial process as the PC algorithm, however when it comes to the iterative application of rules there are 10 rules as opposed to 3. The purpose of these additional rules to allow FCI to accommodate the presence of exogenous hidden confounder ancestors and hidden descendants that have been inadvertently conditioned upon (e.g. selection bias). To achieve this FCI works upon the concept of a Maximal Ancestral Graph (MAG) rather than a MEC, that allows for bi-directional edges representing hidden confounder ancestors, and non-directional edges representing conditioned upon hidden descendants.

5.4. Non-Stationarity-Based Approaches

While most approaches struggle with non-stationarity, approaches such as CD-NOD proposed by Zhang et al. (2017) actively aim to exploit non-stationarity. For the most part CD-NOD resembles a constraint-based approach except with the addition of a surrogate variable designed to capture either domain-based or time-based distribution shifts. The discovery of the causal graph skeleton is carried out in a similar manner to a method such as PC, relying upon conditional independence

tests and applying orientation rules. However, in determining causal direction, CD-NOD bears some similarities to a noise-based approach, except rather than examining the propagation of noise CD-NOD examines the influence of variable specific parameters calculated as a function of the surrogate variable. Since the actual parameters are unknown CD-NOD tackles the problem through the application of Kernel Density Estimation (KDE). One limitation of this approach is that if the surrogate variable has control over a confounding variable, CD-NOD is reliant upon the influence of the confounding variable being weak compared to the surrogate controlled variable parameters.

In the aforementioned work, CD-NOD was purely intended for application on instantaneous causal relationships, however it was discussed that it was natural to extend the method to include time lagged relationships, this was subsequently accomplished by Huang et al. (2020). However a shortcoming of this approach as a whole in the context of this work, is while CD-NOD can capture shifts in causal model parameterisation, it cannot capture changes to the causal skeleton that occur with time, nor can it capture shifts in causal direction with time. This is also true of the time lagged extension, meaning that a causal relationship that varies in time lag over time but not with causal effect cannot be adequately captured. Nevertheless of the approaches evaluated in this paper, CD-NOD offers the best framework to provide robustness against non-stationarity.

5.5. Score-Based Approaches

Score-based approaches view the causal discovery process as the task of finding a causal graph which maximises a scoring metric of how well the data fits the supplied graph. Since an exhaustive search would be computationally intractable this is typically tackled as an optimisation problem. In terms of the score metrics used for this type of approach, the Bayesian Information Criterion (BIC) (Schwarz, 1978), Bayesian Dirichlet equivalence (BDe) (Heckerman et al., 1995) score, and Cross-Validation (CV) (Peña et al., 2005) technique have all been proposed as options.

However, the approach we consider for evaluation, DYNOTEARS (Pamfil et al., 2020) defines a new metric for this purpose. The following consists of the metric they define without any of the components relating to instantaneous causal relationships and considering a single independent realization of each underlying causal model, as this better reflects the problem we consider:

$$f(A) = \ell(A) + \lambda_A \max_{v^j \in V} \sum_{t'=1}^{\tau} \sum_{v^i \in V} (a_{t'}^{i,j}) \tag{5}$$

$$\ell(A) = \frac{1}{2n} \sum_{t=\tau}^{t_{max}} \sum_{v^i \in V} (v_t^i - \sum_{t'=1}^{\tau} \sum_{v^j \in V} (a_{t'}^{i,j} v_{t-t'}^j)) \tag{6}$$

where $n = t_{max} + 1 - \tau$, t_{max} is the largest time index present in the time series, λ_A is a regularisation constant, and $a_{t'}^{i,j}$ corresponds to the i-th row and j-th column of A at the time lag of t'. Here (6) describes a loss term derived from the difference between the time series values predicted by A and the observed time series values, meanwhile (5) reflects an overall score derived from combining (6) with a regularisation term. After optimising A we can then construct a causal graph by thresholding each element $a_{t'}^{i,j}$ within A, and adding a graph edge should the value prove great enough. As a result A reflects both a weight matrix updated while learning a linear predictive model with DYNOTEARS, and an adjacency matrix which is later converted to a causal graph by applying a threshold to its elements.

6. Experiments

6.1. Datasets & Code

The experiments were carried out on three datasets independently, the Lyft Level 5 Prediction dataset, the High-D dataset and a synthetic dataset. All the scenes in this dataset exhibit the causal relations/interactions described in Sec. 2. We make the code used to carry out these experiments as well as the data associated with them that can be publicly shared available in a Git repository[1].

LYFT LEVEL 5 PREDICTION

This dataset is comprised of over a thousand hours of driving data collected over 20 autonomous vehicles operating in Palo Alto, California (Houston et al., 2020). The data itself is comprised of egocentric data (e.g. position/orientation of capturing autonomous vehicle), labeling of vehicles and pedestrians complete with position, orientation and bounding boxes, and the status of any perceived traffic lights. This time series data is structured into a series of scenes, each $\sim 30\ s$ in length and captured at $10\ Hz$. In addition to these scenes, there is a static semantic map describing the road network the autonomous vehicles were operating on.

In order to transform the data available into two-agent convoy scenes upon which causal discovery could be carried out, a manual inspection was carried out upon each scene in turn. This was used to determine which scenes exhibited the scenario laid out in Sec. 2, and within these scenes which agents - besides the ego vehicle - would form part of the scene. In total 50 scenes were extracted from the dataset in this manner, all of which were at least $10\ s$ in length following any trimming of the scene that needed to be carried out. Finally, the velocities and accelerations of agents were estimated based upon the change in agent positions between frames, and in order to account for positional jitter, the velocity and acceleration were smoothed via a 15 frame moving average.

HIGH-D

In contrast to the variety of roads within Palo Alto covered by the Lyft dataset, the High-D dataset focuses upon stretches of highway at six locations across Germany (Krajewski et al., 2018). It offers greater coverage of the roads it considers by capturing overhead footage using a drone. This is automatically annotated, resulting in a positional error of labelled vehicles typically under $10\ cm$. The recordings that make up the dataset were captured at $25\ Hz$ over the course of several minutes. As opposed to a detailed semantic map, the dataset simply offers the y position of the lane markings in pixel space, as all data is aligned to have the lanes run parallel to the x axis. Unlike the Lyft dataset, the High-D dataset also provides stable velocity and acceleration values out of the box, and as such these did not need to be calculated. The data was however resampled to $10\ Hz$ as to maintain a similar maximum time lag in seconds.

Due to additional details in terms of lane occupancy and inter-vehicular adjacency provided by the dataset, it was possible to automate the process of creating scenes by simply searching the dataset for instances where this convoy-like causal relationship occurs and then selecting a third vehicle from another lane to act as **i0**. Once again, all produced scenes were of at least $10\ s$ in length, though most were longer due to the wide area captured by the drone camera. This led to a total of 3395 scenes being extracted.

1. https://github.com/cognitive-robots/temporal-cd-evaluation-paper-resources

SYNTHETIC

In addition to running experiments upon the two real world datasets we also generated and ran experiments upon a synthetic dataset. Doing so facilitates differentiating between performance loss due to the nature of the scenario from performance loss due to real world complications (e.g. sensor noise, variations in human behaviour, etc.). In order to generate this dataset, a series of velocity goal objectives are assigned for the lead convoy agent and independent agent. A proportional error controller is then used to actuate the acceleration of the aforementioned agents while adhering to a set of linear kinematic constraints. Meanwhile the tail convoy agent is actuated by a proportional error controller that aims to maintain a convoy distance over relative velocity of $2.24\ s$, with a $0.5\ s$ time lag to mimic a reasonable human reaction time. In both of these cases the proportional error controller in question calculates the vehicle acceleration by multiplying the velocity/distance error by a predefined gain parameter. For our experiments we found a proportional gain of 1.0 sufficed to produce satisfactory causal scenes. The scenes were made to last for $50 - 70\ s$ and were generated at $10\ Hz$. Scenes were generated in such a way as to have 12 causal interactions within the convoy and 12 actions carried out by the independent agent.

6.2. Parameters

The key parameters universal to all the methods were the significance alpha and maximum time lag, in order to avoid assessing the efficacy of the methods with a poor parameter selection these values were individually varied. The values used for the significance alpha were 0.001, 0.005, 0.01, 0.03, 0.05 and 0.1, while using a maximum time lag of $2.5\ s$ based upon a conservative estimate of reaction and actuation times of drivers. Meanwhile the values used for the maximum time lag were $2.5\ s$, $3.6\ s$ and $4.9\ s$, while using a significance alpha of 0.05 based upon the previous work carried out by Assaad et al. (2022).

6.3. Evaluation Metric

To determine the efficacy of each evaluated method we calculate a True Positive (TP), False Positive (FP), and False Negative (FN) count for each causal scene. For a link within a discovered causal graph is a TP if present in the ground truth causal graph, and a false negative otherwise. Meanwhile, if a link present in the ground truth causal graph is not present in the discovered causal graph, this counts as a FN. The ground truth causal graph for all causal scenes we consider corresponds to the scenario described in Sec. 2 and illustrated in Fig. 1. From the previously calculated values we can derive the F_1 score for each causal scene and take the mean across scenes for a given method. For additional details on the calculation of the TP, FP and FN counts please see the appendix.

6.4. Results

The results are displayed in Fig. 2. Overall the best mean F_1 scores are provided by DYNOTEARS, MVGC and TiMINo. PCMCI and NAVAR also provide competitive results, but are outperformed by at least one other method in every case. PWGC, tsFCI, and CD-NOD generally under-perform compared with other methods, while TCDF and VarLiNGAM completely fail in almost every case.

While it is possible that these failures are due to an issue of implementation, TCDF directly calls the same code utilised by the original paper (Nauta et al., 2019) and VarLiNGAM directly calls code from the Python *lingam* package (Shimizu et al., 2011; Hyvärinen and Smith, 2013),

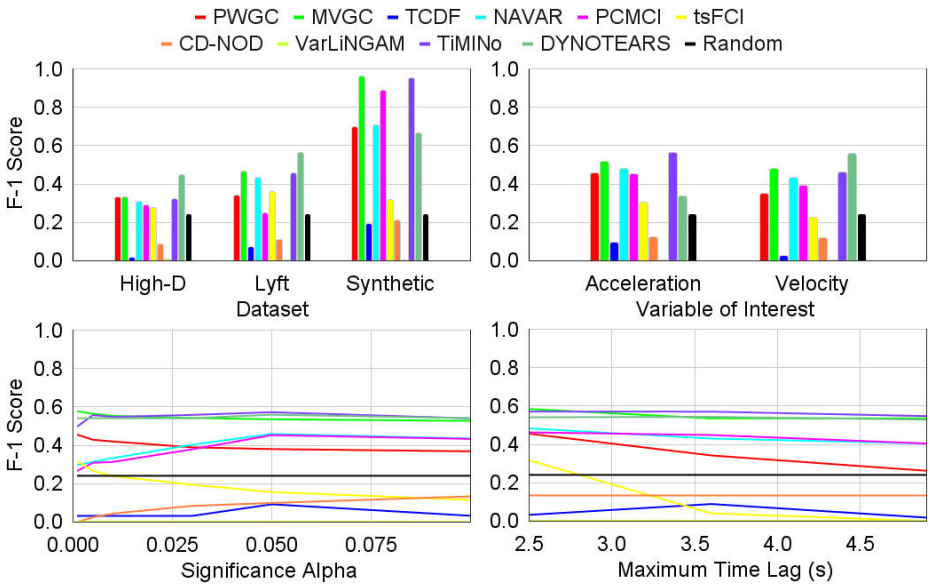


Figure 2: F_1 Score from applying each method to the Lyft, High-D and Synthetic datasets, for acceleration and velocity variables, while varying significance alpha and maximum time lag independently. For the dataset graph, the F_1 Score given for each method is based upon the optimal combination of parameters for the specific dataset-method combination. This process is similarly repeated in the other graphs, however the maximum F_1 Score is not taken across datasets in order to avoid the synthetic dataset dominating the figures. Instead the mean is taken across datasets, while still taking the optimal combination of parameters for parameters besides the parameter being varied.

making this unlikely. Theoretically speaking it is possible that TCDF is under-performing due to lack of training data, but this is made less likely by the fact NAVAR is competitive also training upon the same amount of data. Likewise with VarLiNGAM, it is possible that there is too weak of a direct coupling between variables for the noise of one variable to affect another, but considering TiMINo is competitive, this is again unlikely. Another possibility is that all variable noise is Gaussian, and TiMINo is still able to perform causal discovery despite this fact, while VarLiNGAM cannot.

In terms of parameter selection, significance alpha has the most influence although also the least consistent pattern of the two key parameters, making it an important parameter to tune for effective application of many methods. Interestingly DYNOTEARS appears to barely respond to changes in significance alpha, which could make it a better method to apply out of the box. Maximum time lag has a minimal impact across many methods, but in general over-estimating the maximum time lag leads to degradation in performance, thus a good approximation may suffice.

A clear issue illustrated by these results is the lack of readiness for these methods to be applied to real world data in these types of scenarios. While some methods such as MVGC and TiMINo are able to get close to a F_1 score of 1.0 on synthetic data, the greatest performance we see on real world data is from DYNOTEARS at ~ 0.565 F_1 score. This is clearly an unacceptable level of performance in a domain such as AD, demonstrating that either significant improvements need to be made to existing methodologies or alternate directions of research considered. This work does

however highlight which methods might be candidates for further work, namely the aforementioned MVGC, TiMINo and DYNOTEARS methods.

Between real world data and synthetic data the main differences that could explain the degradation of performance are: increased levels of noise, causal non-stationarity and causal sparsity. Of these the latter two are of greater interest to us, as issues resulting from noise do not directly relate to causality research and can potentially be resolved at a lower level of abstraction (e.g. more accurate localisation). In terms of causal non-stationarity, the synthetic dataset effectively has stationarity as a result of the agent reaction time always being $0.5\ s$, meanwhile real world drivers feature inconsistent reaction times. As for causal sparsity, real world scenes we consider are typically $30\ s$ or under and contain a few causal interactions within the convoy at most. Meanwhile the synthetic scenes are $50 - 70\ s$ in length and contain 12 causal interactions within the convoy. This highlights a key issue with attempting causal discovery in an online system where causal interactions may be infrequent and the observation window brief. The combination of these issues also likely explains the relatively poor performance of a method leveraging non-stationarity like CD-NOD. Here synthetic scenes provide ample yet mostly stationary data, while real scenes provide little in the way of data, that may or may not be non-stationary dependent upon whether more than a single causal interaction occurs.

7. Discussion & Conclusions

The experiments establish that the evaluated approaches struggle with tackling even a simple causal scenario in the AD domain when working with real world data. However, these methods constitute the contemporary work of the temporal causal discovery field, indicating that further development of existing methodologies or new lines of thought entirely may be necessary to overcoming the new challenges presented within the AD domain.

The chief qualities of these new types of problems are causal non-stationarity and causal sparsity. The first of these has been identified in some recent literature as a matter of concern (Glymour et al., 2019). There has been some progress in this direction with one work applying a non-stationary causal discovery approach to medical data (Strobl, 2017) and another very recent work attempting causal discovery on conditionally stationary data (Rodas et al., 2021). This latter example is interesting because it explores a physical causal relationship between particles through a spring, which more closely mirrors the types of relationships present in the AD domain.

Causal sparsity as a quality is harder to overcome, as all observational approaches inherently rely upon evidence that is likely to be lacking when working in an online scenario-by-scenario fashion. It is here where a theoretical counterfactual approach to causal discovery might succeed by utilising approximations of individual agent behaviour to discover inter-agent causal relationships. While such an approach has been applied in the domain of climate science (Hannart et al., 2016), to the author's knowledge no such research has been conducted in relation to autonomous agents.

Thus we suggest for future work that both of these avenues should undoubtedly be explored further, in doing so open up new avenues of causal reasoning between agents in autonomous robotics.

Acknowledgments

This work was supported by the EPSRC project RAILS (grant reference: EP/W011344/1) and the Oxford Robotics Institute research project RobotCycle.

References

Raghavendra Addanki, Shiva Kasiviswanathan, Andrew Mcgregor, and Cameron Musco. Efficient intervention design for causal discovery with latents. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 63–73. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/addanki20a.html.

H AKAIKE. Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd International Symposium on Information Theory, 1973*, pages 267–281. Akademiai Kiado, 1973.

Charles K Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.

Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004. doi: https://doi.org/10.1111/j.1467-842X.2004.00360.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-842X.2004.00360.x.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246. URL http://www.jstor.org/stable/2346101.

Bart Bussmann, Jannes Nys, and Steven Latré. Neural additive vector autoregression models for causal discovery in time series. In Carlos Soares and Luis Torgo, editors, *Discovery Science*, pages 446–460, Cham, 2021. Springer International Publishing. ISBN 978-3-030-88942-5.

William S. Cleveland and Susan J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988. ISSN 01621459. URL http://www.jstor.org/stable/2289282.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994. ISSN 0165-1684. doi: https://doi.org/10.1016/0165-1684(94)90029-9. URL https://www.sciencedirect.com/science/article/pii/0165168494900299. Higher Order Statistics.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/947018640bf36a2bb609d3557a285329-Paper.pdf.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007. doi: 10.1086/525638. URL https://doi.org/10.1086/525638.

Doris Entner and Patrik O Hoyer. On causal discovery from time series data using fci. In *5th European Workshop on Probabilistic Graphical Models*, pages 121–128. Helsinki Institute for Information Technology HIIT, 2010.

Freepik. Flat car collection in top view. https://www.freepik.com/free-vector/flat-car-collection-top-view_1349616.htm, 2022. Accessed: 2022-02-09.

Niloy Ganguly, Dren Fazlija, Maryam Badar, Marco Fisichella, Sandipan Sikdar, Johanna Schrader, Jonas Wallat, Koustav Rudra, Manolis Koubarakis, Gourab K. Patro, Wadhah Zai El Amri, and Wolfgang Nejdl. A review of the role of causality in developing trustworthy ai systems, 2023. URL https://arxiv.org/abs/2302.06975.

John Geweke. Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77(378):304–313, 1982. doi: 10.1080/01621459.1982.10477803. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1982.10477803.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00524. URL https://www.frontiersin.org/article/10.3389/fgene.2019.00524.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912791.

A. Hannart, J. Pearl, F. E. L. Otto, P. Naveau, and M. Ghil. Causal counterfactual theory for the attribution of weather and climate-related events. *Bulletin of the American Meteorological Society*, 97(1):99 – 110, 2016. doi: 10.1175/BAMS-D-14-00034.1. URL https://journals.ametsoc.org/view/journals/bams/97/1/bams-d-14-00034.1.xml.

David Heckerman, Dan Geiger, and David M Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20(3):197–243, 1995.

Thomas Hellström. The relevance of causation in robotics: A review, categorization, and analysis. *Paladyn, Journal of Behavioral Robotics*, 12(1):238–255, 2021. doi: doi:10.1515/pjbr-2021-0017. URL https://doi.org/10.1515/pjbr-2021-0017.

J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska. One thousand and one hours: Self-driving motion prediction dataset. https://level-5.global/level5/data/, 2020.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020. URL http://jmlr.org/papers/v21/19-232.html.

Aapo Hyvärinen and Stephen M Smith. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14:111–152, 2013.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11 (56):1709–1731, 2010. URL http://jmlr.org/papers/v11/hyvarinen10a.html.

Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2688–2697. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/kipf18a.html.

Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental design for learning causal graphs with latent variables. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/291d43c696d8c3704cdbe0a72ade5f6c-Abstract.html.

Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2118–2125, 2018. doi: 10.1109/ITSC.2018.8569552.

Chengxi Li, Stanley H. Chan, and Yi-Ting Chen. Who make drivers stop? towards driver-centric risk assessment: Risk object identification via causal inference. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10711–10718, 2020a. doi: 10.1109/IROS45743.2020.9341072.

Yunzhu Li, Antonio Torralba, Anima Anandkumar, Dieter Fox, and Animesh Garg. Causal discovery in physical systems from videos. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9180–9192. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/6822951732be44edf818dc5a97d32ca6-Abstract.html.

Daniel McDuff, Yale Song, Jiyoung Lee, Vibhav Vineet, Sai Vemprala, Nicholas Gyde, Hadi Salman, Shuang Ma, Kwanghoon Sohn, and Ashish Kapoor. Causalcity: Complex simulations with agency for causal discovery and reasoning, 2021.

Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019. ISSN 2504-4990. doi: 10.3390/make1010019. URL https://www.mdpi.com/2504-4990/1/1/19.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1595–1605. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/pamfil20a.html.

Judea Pearl. Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th conference of the Cognitive Science Society, University of California, Irvine, CA, USA*, pages 15–17, 1985.

Judea Pearl. *Causality*. Cambridge University Press, 2nd edition, 2009.

Judea Pearl and Dana Mackenzie. *The Book of Why*. Penguin Books, 2019.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/hash/47d1e990583c9c67424d369f3414728e-Abstract.html.

Jose M. Peña, Johan Björkegren, and Jesper Tegnér. Learning dynamic bayesian network models via cross-validation. *Pattern Recognition Letters*, 26(14):2295–2308, 2005. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2005.04.005. URL https://www.sciencedirect.com/science/article/pii/S0167865505001212.

Joseph Ramsey, Peter Spirtes, and Jiji Zhang. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, page 401–408, Arlington, Virginia, USA, 2006. AUAI Press. ISBN 0974903922.

Carles Balsells Rodas, Ruibo Tu, and Hedvig Kjellstrom. Causal discovery from conditionally stationary time-series, 2021.

Jose Rojas. Autonomous driving dataset visualization with python and vizviewer. https://towardsdatascience.com/24ce3d3d11a0, 2020. Accessed: 2022-08-18.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5 (11), 2019. doi: 10.1126/sciadv.aau4996. URL https://advances.sciencemag.org/content/5/11/eaau4996.

Mohammad Reza Samsami, Mohammadhossein Bahari, Saber Salehkaleybar, and Alexandre Alahi. Causal imitative model for autonomous driving, 2021. URL https://arxiv.org/abs/2112.03908.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. ISSN 00905364. URL http://www.jstor.org/stable/2958889.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006. URL http://jmlr.org/papers/v7/shimizu06a.html.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12 (33):1225–1248, 2011. URL http://jmlr.org/papers/v12/shimizu11a.html.

Stephen M. Smith, Karla L. Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F. Beckmann, Thomas E. Nichols, Joseph D. Ramsey, and Mark W. Woolrich. Network modelling methods for fmri. *NeuroImage*, 54(2):875–891, 2011. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2010.08.063. URL https://www.sciencedirect.com/science/article/pii/S1053811910011602.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, 2nd edition, December 2001. ISBN ARRAY(0x39e112d8). URL https://ideas.repec.org/b/mtp/titles/0262194406.html.

Eric V. Strobl. *Causal discovery under non-stationary feedback*. PhD thesis, Department of Biomedical Informatics, University of Pittsburgh, 2017. URL https://www.proquest.com/dissertations-theses/causal-discovery-under-non-stationary-feedback/docview/2013196949/se-2?accountid=13042.

Jie Sun, Dane Taylor, and Erik M. Bollt. Causal network inference by optimal causation entropy. *SIAM Journal on Applied Dynamical Systems*, 14(1):73–106, 2015. doi: 10.1137/140956166. URL https://doi.org/10.1137/140956166.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, volume 8, 2020.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896, 2008. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2008.08.001. URL https://www.sciencedirect.com/science/article/pii/S0004370208001008.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 804–813, Arlington, Virginia, USA, 2011. AUAI Press. ISBN 9780974903972.

Kun Zhang, Biwei Huang, Jiji Zhang, Clark Glymour, and Bernhard Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *IJCAI: Proceedings of the Conference*, volume 2017, page 1347. NIH Public Access, 2017.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. doi: 10.1198/016214506000000735. URL https://doi.org/10.1198/016214506000000735.

Appendix A. Data Preparation

Because the real world datasets used in this paper are not open for public download we cannot share the data supplied as input for the experiments, only the data resulting from the experiments. However, in order to ensure the experiments are reproducible, we document the pre-processing steps taken on the datasets.

A.1. Lyft Level 5 Prediction

The first step in extracting scenes from the Lyft dataset is to extract the map information stored in the Protobuf format semantic map file using the "extract_map.py" script. Once this has been done, the agent data can be extracted via the "extract_agents.py" script.

At this point there are two options available, if the scenes can be visualised in some fashion[2] one can select scenes that encapsulate the two convoy scenario. From these scenes it is necessary to identify the ID number of the second convoy vehicle as well as choosing an independent vehicle. It is important to select an independent vehicle that is present in the scene for a majority of the time, as the scene will be trimmed to the time window where all agents are present. Furthermore it is also greatly preferable to select an independent agent that is not stationary for the entirety of the scene as to present a fair challenge to the methods under consideration. Once these details have been identified, "convert_to_two_agent_followed_scene.py" and "convert_to_two_agent_follower_scene.py" can be used to convert the extracted agent data to CSV time series files.

Alternatively, under the "conversion" directory the "agent_json_data_to_two_agent_convoy.sh" script will convert the selection of scenes and associated agent IDs used in the main paper to CSV time series files. Using this script requires setting the three environment variables, the input directory, output directory and the base Lyft dataset script directory (i.e. lyft_prediction_dataset_tools).

A.2. High-D

With the additional information provided by the High-D dataset the process of scene extraction can be automated. Calling the "extract_two_agent_convoy_scenes.py" script while passing the path to the High-D data directory will complete the whole extraction process. It is necessary to set a minimum scene length value and a minimum proportion velocity change for potential convoy agents, we selected $10\ s$ and 0.2 respectively for these values.

A.3. Synthetic

For documentation of the steps to generate synthetic data, we refer the reader to the Git repository linked in Sec. 6.1. However, to summarise one can either call "create_two_agent_convoy_scene.py" or "create_two_agent_convoy_scenes.py" to generate a single or multiple causal scenes to be output to the specified file path or directory path respectively. Below the parameters used to generate the set of synthetic causal scenes used in this paper are documented.

2. The tool utilised by the authors is part of another project and is not currently available to the public. However, the L5Kit provided by Lyft Level 5 on GitHub is quite useful for visualising scenes, and there is documentation of better visualisation methods available (Rojas, 2020).

- Variable: {"Acceleration", "Velocity"}
  - Each of these parameter settings was used once independently to generate 100 acceleration-based causal scenes and 100 velocity-based causal scenes.
- Frequency: $10.0\ Hz$
- Duration: $50.0 - 70.0\ s$
- Convoy Actions: 12
- Independent Actions: 12
- Minimum Convoy Distance: $10.0\ m$
- Maximum Convoy Distance: $100.0\ m$
- Proportional Coefficient: 1.0
- Integral Coefficient: 0.0
- Differential Coefficient: 0.0
- Minimum Action Interval: $1.0\ s$
- Minimum Velocity: $0.0\ m/s$
- Maximum Velocity: $44.7\ m/s$
- Minimum Start Velocity: $4.47\ m/s$
- Maximum Start Velocity: $26.8\ m/s$
- Minimum Acceleration: $-6.56\ m/s^2$
- Maximum Acceleration: $3.5\ m/s^2$
- Safe Distance Over Velocity: $2.24\ s$
- Reaction Time $0.5\ s$
- Fixed Actuary Noise: $0.1 - 1.6\ m/s^2$
- Proportional Actuary Noise: $0.1 - 1.6$
- Fixed Sensory Noise: $0.01 - 0.16\ m$
- Proportional Sensory Noise: $0.005 - 0.08$

Appendix B. Experiment Parameters

Here we present the parameter values used by each method, and offer explanations for certain parameter choices. If an explanation is not offered it either is a default value based upon those utilised by previous works Assaad et al. (2022); Bussmann et al. (2021); Zhang et al. (2017), or one of the two parameters justified in the main paper (i.e. maximum time lag and significance alpha). We do not claim to assign optimal parameters to methods as optimising multiple parameters across 10 temporal causal discovery methods was deemed infeasible within the scope of the work. That being said, we would like to explore carrying out experiments for multiple significance alphas as part of future work.

- Pairwise Granger
  - Significance Alpha: 0.05
  - Statistical Test: Regression Sum of Squares (SSR) F-Test
  - Maximum Time Lag: 25 / 2.5 s
- Multivariate Granger
  - Significance Alpha: 0.05
  - Statistical Test: Chi-Squared
  - Multiple Hypothesis Test Correction: Benjamini-Hochberg False Discovery Rate (Benjamini and Hochberg, 1995)
  - VAR Model Estimation Regression Mode: Ordinary Least Squares
  - Information Criteria Regression Mode: Locally Weighted Regression (Cleveland and Devlin, 1988)
  - Model Order: Akaike Information Criterion (AKAIKE, 1973)
  - Maximum Time Lag: 25 / 2.5 s
  - Maximum Autocovariance Lags: 1000
  - Random Seed: Undefined
- TCDF
  - Significance Alpha: 0.05
  - Epochs: 1000
  - Learning Rate: 0.01
  - Hidden Layers: 1
  - Kernel Size: 5
    - While utilising a single hidden layer and identical values for kernel size and dilation coefficient, the maximum time lag consists of the kernel size/dilation coefficient value squared. Therefore in order to match the maximum time lag of 25 assigned to other methods we assign 5 to the kernel size/dilation coefficient here.
  - Dilation Coefficient: 5

    - * See above.
  - Random Seed: 1111
  - Optimizer: Adam (Kingma and Ba, 2014)
  - CUDA: False
    - * While CUDA usage might theoretically increase the speed of TCDF, we felt it best to avoid executing methods on the GPU where possible. This way the performance reflected in the main paper should act as a lower bound on the performance that can be improved with CUDA usage, rather than the reverse.

- NAVAR
  - Significance Alpha: 0.05
  - Maximum Time Lag: 25 / 2.5 s
  - Hidden Nodes: 10
  - Hidden Layers: 1
  - Epochs: 2000
  - Batch Size: 32
  - Sparsity Penalty: 0.1
  - Weight Decay: 0.001
  - Dropout: 0.5
  - Learning Rate: 3.0×10^{-4}
  - Validation Proportion: 0.0
  - Network Type: Multi-Layer Perceptron
  - Normalize: True
  - Split Time Series: False

- PCMCI
  - Significance Alpha: 0.05 (Used for both "pc_alpha" and "alpha_level")
  - False Discovery Rate Method: Benjamini-Hochberg (Benjamini and Hochberg, 1995)
  - Minimum Time Lag: 0 / 0.0 s
  - Maximum Time Lag: 25 / 2.5 s
  - Maximum Number of Conditions to Test: Unrestricted
  - Maximum Number of Conditions of Y to Use: Unrestricted
  - Maximum Number of Conditions of X to Use: Unrestricted
  - Maximum Number of Combinations of Conditions: 1
  - Conditional Independence Test: Partial Correlation (Baba et al., 2004)

    * Partial correlation and conditional mutual information using K-nearest numbers were both used by Assaad et al. (2022). We tested both of these with the autonomous driving data, however the conditional mutual information approach took on average almost half an hour per run, making it intractable as an online causal discovery approach.

- tsFCI
  - Significance Alpha: 0.05
  - Maximum Time Lag: 25 / 2.5 s
  - Include Instant Effects: False
  - Data Type: Continuous
  - Algorithm: tsCFCI (A variant of tsFCI with less stringent faithfulness requirements) (Ramsey et al., 2006)
- CD-NOD
  - Significance Alpha: 0.05
  - Maximum Time Lag: ?
    * Although a maximum time lag parameter is described in the work discussing a time lagged relationship extension to CD-NOD (Huang et al., 2020) (i.e. P), there does not seem to be a way to specify this based upon the documentation available within the Git repository[3] nor from our attempts at interpreting the codebase.
  - Conditional Independence Test: Kernel-based Conditional Independence (KCI) Zhang et al. (2011)
  - Surrogate Variable: Time Index
  - Maximum Number of Conditioning Variables: 1
  - Type: 0 (Run all phases)
  - Pairwise: False
  - Bonferroni Correction: False
  - Conditional Independence Test GP Optimisation: True
  - Direction Determination GP Optimisation: True
  - Observational Variable Kernel Width: 0 (Automatically calculated via GP optimisation)
  - Time Index Kernel Width: 0.1
- VarLiNGAM
  - VAR Model Estimation Regression Mode: Ordinary Least Squares
  - Trend Assumption: Co-Constant, No Trend
  - Pruning: True

3. https://github.com/Biwei-Huang/Causal-Discovery-from-Nonstationary-Heterogeneous-Data

  - Significance Alpha: 0.05
  - Maximum Time Lag: 25 / 2.5 s
  - Regularisation Criterion: Bayesian Information Criterion (Schwarz, 1978)
  - Random Seed: Undefined
  - Algorithm: DirectLiNGAM (A variant of LiNGAM which typically executes faster) (Shimizu et al., 2011)

- TiMINo
  - Significance Alpha: 0.05
  - Maximum Time Lag: 25 / 2.5 s
  - Assumed Time Series Model: Linear
  - Independence Test: Cross Covariance
  - Include Instant Effects: False
  - Check for Confounders: False

- DYNOTEARS
  - Threshold for W: 0.01
  - Threshold for A: 0.01
  - Regularisation Constant for W: 0.05
  - Regularisation Constant for A: 0.05
  - Maximum Time Lag: 25 / 2.5 s
  - Maximum Number of Iterations: 100
  - Acyclicity Tolerance: 1.0×10^{-8}

- Random Causal Discovery
  - Edge Likelihood: 0.5
    * This was selected purely based upon it being the most naive approach to randomly constructing a graph (i.e. Flipping a coin for each potential edge effectively) as the intended use for this method was as a baseline.

Appendix C. Experiment Setup

C.1. Hardware

- CPU: AMD Ryzen 9 3950X
- GPUs: Nvidia GeForce RTX 2070 SUPER, Nvidia GeForce GTX 750 Ti
- Storage: Samsung Electronics 970 EVO Plus NVMe M.2 Internal SSD

C.2. Software

- Kernel: Linux version 5.4.0-137-generic
- OS: Ubuntu 20.04.5 LTS "Focal"
- GCC/G++: 10.3.0
- Python: 3.8.10
  - numpy: 1.23.3
  - pandas: 1.4.3
  - scikit-learn: 0.23.1
  - scipy: 1.4.1
  - statsmodels: 0.11.1
  - joblib: 0.15.1
  - graphviz: 0.8.4
  - networkx: 2.8.5
  - matplotlib: 3.1.2
  - torch: 1.12.1+cu113
  - l5kit: 1.1.0
- R: 4.2.2 "Innocent and Trusting"
- Matlab: R2022a Update 4 (9.12.0.2009381)

Appendix D. Experiment Process

Carrying out a set of experiments corresponds to setting the significance alpha and maximum time lag and carrying out a number of runs of "test_ad.py" while varying three parameters:

- Method: GrangerPW, GrangerMV, TCDF, NAVARMLP, PCMCIParCorr, tsFCI, CDNOD, VarLiNGAM, TiMINo, Dynotears
- Dataset: lyft, highd, synthetic
- Variable: acceleration, velocity

The first two of these correspond to the method to use and the dataset to apply it to. The third parameter specifies a sub-dataset in the sense that it determines whether or not to use acceleration or velocity as the variable of interest for agents.

Each run of "test_ad.py" applies the selected method to every scene in the sub-dataset, before checking the discovered causal graphs against the predetermined ground truth using the true positive, false positive, and false negative metrics defined in the main paper. The final output from the run is all of the graphs discovered for each scene, the individual F_1 Score, Precision, Recall and Runtime for each scene, and the mean and standard deviation of these metrics across all scenes.

In order to calculate F_1 Score, Precision and Recall, we calculated the True Positive (TP), False Positive (FP) and False Negative (FN) count for the i-th scenario as follows:

$$|TP_i| = \sum_{v^j \in V} |\hat{pa}_i(v^j) \cap pa(v^j)| \tag{7}$$

$$|FP_i| = \sum_{v^j \in V} |\hat{pa}_i(v^j) \setminus pa(v^j)| \tag{8}$$

$$|FN_i| = \sum_{v^j \in V} |pa(v^j) \setminus \hat{pa}_i(v^j)| \tag{9}$$

where $pa(v^j)$ represents the ground truth scenario independent parents for the time series v^j and $\hat{pa}_i(v^j)$ represents the parents discovered for the i-th scenario by applying a given method. Although some methods can detect self-causal relationships, these have no clear meaning in the context of agent behavioural interaction, and therefore these relationships are excluded from the evaluation.