Causal Temporal Regime Structure Learning

Abdellah Rahmani
LTS4, EPFL
Lausanne, Switzerland

Pascal Frossard
LTS4, EPFL
Lausanne, Switzerland

Abstract

Understanding causal relationships in multivariate time series is essential for predicting and controlling dynamic systems in fields like economics, neuroscience, and climate science. However, existing causal discovery methods often assume stationarity, limiting their effectiveness when time series consist of sequential regimes, consecutive temporal segments with unknown boundaries and changing causal structures. In this work, we firstly introduce a framework to describe and model such time series. Then, we present CASTOR, a novel method that concurrently learns the Directed Acyclic Graph (DAG) for each regime while determining the number of regimes and their sequential arrangement. CASTOR optimizes the data log-likelihood using an expectation-maximization algorithm, alternating between assigning regime indices (expectation step) and inferring causal relationships in each regime (maximization step). We establish the identifiability of the regimes and DAGs within our framework. Extensive experiments show that CASTOR consistently outperforms existing causal discovery models in detecting different regimes and learning their DAGs across various settings, including linear and nonlinear causal relationships, on both synthetic and real world datasets.

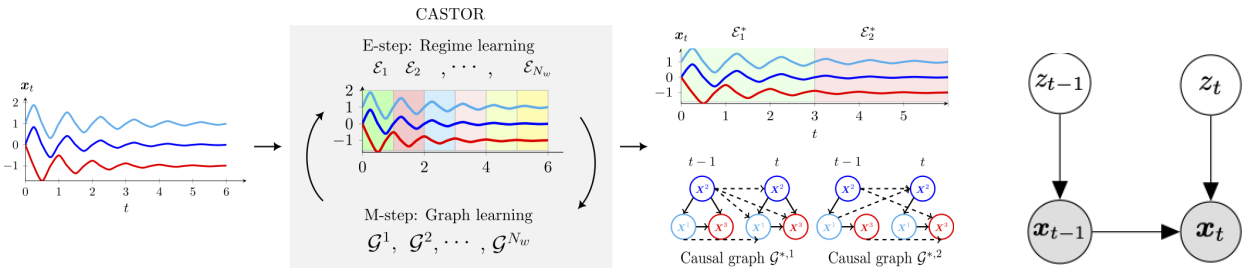
1 INTRODUCTION

Causal structure learning from multivariate time series (MTS) variables is essential in many fields like disease evolution (Shen et al., 2020) or climate science (Runge et al., 2019), as causal structures reveal complex real-world mechanisms. Recent approaches have focused on extracting causality from observational data, handling both linear and nonlinear relationships, and accommodating instantaneous and time-lagged connections (Pamfil et al., 2020; Löwe et al., 2022; Runge, 2018; Gong et al., 2022). However, these methods often assume that time series data come from a single regime governed by one causal graph, which is inadequate for real-world scenarios. For example, causal dependencies vary across different climatic regimes (Karmouche et al., 2023) and financial settings (Huang et al., 2020), time series are characterized by multiple unknown regimes in neurological studies regarding epilepsy (Rahmani et al., 2023). In practice, successive regimes may originate from different causal models over the same variables, each described by a distinct temporal causal graph.

Recent research has addressed causal discovery from multivariate time series (MTS) with various regimes. For example, CD-NOD (Huang et al., 2020) detects change points and produces a summary causal graph where each variable's parents are the union of its parents across all existing regimes. However, CD-NOD cannot infer individual causal graphs for each regime, does not highlight edges that appear or disappear between regimes, and is incapable of identifying recurring regimes, which is a crucial limitation in many scenarios. Another work addressing MTS composed of multiple regimes is RPCMCI (Saggioro et al., 2020), which learns a temporal graph for each regime. However, it only infers time-lagged relationships and requires prior knowledge of the number of regimes and transitions between them. Overall, these methods make restrictive assumptions that limit their applicability in practical settings and cannot simultaneously identify the number of regimes and their corresponding indices, i.e., their start and end points.

To address these limitations, we first present a new framework that formulates Structural Equation Models (SEMs) and Causal Graphical Models (CGMs) for MTS composed of multiple regimes. We introduce CASTOR, which, to the best of our knowledge, is the first method

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.



(a) (b)

Figure 1: (a) An illustration of CASTOR processing an input MTS using an EM procedure to infer two regimes determining their partitions ($\mathcal{E}_1^*$ and $\mathcal{E}_2^*$) and learning the temporal causal graphs. Dashed edges represent time-lagged links; solid arrows indicate instantaneous links. (b) CASTOR's graphical model for lag $L = 1$: observed variables are depicted in grey, and the latent variables are uncolored.

designed to learn causal relationships, determine the number of regimes, and uncover their arrangement from MTS with multiple regimes, each corresponding to an MTS block. The method is presented conceptually in Figure 1a. Unlike other methods, CASTOR does not require prior knowledge of the number of regimes or their indices. It optimizes the data log-likelihood using an expectation-maximization algorithm (EM), alternating between assigning regime indices (expectation step) and inferring causal relationships in each regime (maximization step). We prove that in our framework, comprising multivariate time series with multiple regimes modeled by Gaussian structural equation models with equal error variances, both the regimes and their corresponding graphs are identifiable up to a permutation of the regime labels.

Our extensive comparative analysis with causal discovery models tailored for MTS with multiple regimes shows that CASTOR consistently outperforms them across various scenarios in both structure learning and regime detection. Furthermore, we compare CASTOR with models that assume stationary MTS by providing them the ground truth regime partition information. Even with this advantage, our approach demonstrates similar or superior performance in inferring DAGs on both synthetic and real-world datasets. We finally apply CASTOR to two real-world datasets, IT monitoring data and Biosphere-Atmosphere data where the results show its ability to detect regimes and also generate explanatory DAGs. The main contributions of this paper can be summarized as follows:

- We formulate a new SEMs and CGMs for MTS composed of multiple regimes.
- We present, CASTOR, the first method designed to learn the number of regimes, their indices and their corresponding DAG from MTS with multiple regimes.
- We show that the exact maximization of the score function identifies the ground truth regimes and graphs up to a permutation in the case of Gaussian noise with equal variance.
- We show that CASTOR outperforms state-of-the-art methods in a wide variety of conditions, including linear and non-linear causal relationships and different number of nodes and regimes on both synthetic and real-world datasets.

Related works. Assaad et al. (2022) offer an extensive survey of methods for learning temporal causal relationships. Granger causality is the primary approach used for causal discovery from MTS (Löwe et al., 2022; Bussmann et al., 2021; Xu et al., 2019). However, it is unable to accommodate instantaneous effects. DYNOTEARS (Pamfil et al., 2020), on the other hand, leverages the acyclicity constraint established by Zheng et al. (2018) to continuously relax the DAG and differentiably learn instantaneous and time lagged structures. However, DYNOTEARS is still limited to linear functional forms. TiMINo (Peters et al., 2013) provides a general theoretical framework for temporal causal discovery with functional causal models. However, the aforementioned methods assume that MTS are composed of a single regime.

Several studies have sought to tackle the challenge of causal discovery in Non-stationary time series data (Huang et al., 2020; Günther et al., 2023; Saggioro et al., 2020). Remarkably, Huang et al. (2020) address the setting of time series composed of different regimes by modulating causal relationships through a regime index. While it provides a summary graph highlighting behavioral changes across regimes, they cannot infer individual causal graphs neither the exact number of regime. Saggioro et al. (2020) assume knowledge of the number of regimes and propose the inference of only time-lagged links. Furthermore, they evaluate their algorithm on graphs with a limited number of nodes. Finally, Balsells-Rodas et al. (2024) addresses first-order regime-dependent causal discovery from MTS with multiple regimes. They proved that first-order Markov switching models with non-linear Gaussian transitions are identifiable up to permuta-

tions. Their work offers also a practical algorithms for regime-dependent causal discovery in time series data. However, its primary limitation is the assumption of solely time-lagged relationships, with the theory being restricted to a single time lag. (Detailed related work in Appendix A).

2 FRAMEWORK

In this section, we first introduce our notations. Then we define the temporal causal graph, describe the setting of MTS with multiple regimes, and present a new SEMs for their representation.

Notation. Matrices, vectors, and scalars are denoted by uppercase bold $\boldsymbol{G}_\tau$, lowercase bold $\boldsymbol{x_t}$ and lowercase normal letters $x^i_{t-\tau}$, respectively. Ground-truth variables are indicated with an asterisk, such as $\mathcal{G}^*$. We assume all distributions have densities $p(\boldsymbol{x}_t)$ w.r.t. the Lebesgue measure. The notation $[|0:L|]$ represents the set of integers $\{0, ..., L\}$ and $|\cdot|$ denotes set cardinality. We denote a temporal causal graph (Definition 1) as $\mathcal{G} = (\mathbf{V}, E)$, represented by a collection of adjacency matrices $\boldsymbol{G}_{\tau\in[|0:L|]} = \{\boldsymbol{G}_0, \dots, \boldsymbol{G}_L\}$. Following Gong et al. (2022), $\mathbf{Pa}^i_{\mathcal{G}}(< t)$ refers to the lagged parents of node i in $\mathcal{G}$ at previous time $t-\tau$ with $1 \leq \tau \leq L$, while $\mathbf{Pa}^i_{\mathcal{G}}(t)$ denotes the instantaneous parents at the current time t (i.e., $\tau = 0$). $(\boldsymbol{x}_t)_{t\in\mathcal{T}} = (x^i_t)_{i\in\mathbf{V},t\in\mathcal{T}}$ represent a MTS of $|\mathbf{V}| = d$ components and length $|\mathcal{T}|$. $\mathcal{T}$ is the time index set.

Definition 1 (Temporal Causal Graph)**.** *The temporal causal graph, associated with the MTS* $(\boldsymbol{x_t})_{t\in\mathcal{T}}$*, is defined by a DAG* $\mathcal{G} = (\mathbf{V}, E)$ *and a fixed maximum lag* L*. Its vertices* $\mathbf{V}$ *consists of the set of components* $x^1_{t'}, \dots, x^d_{t'}$ *for each* $t' \in [|t-L:t|]$ *. The edges* E *of the graph are defined as follows:* $\forall\tau \in [|1:L|]$ *variables* $x^i_{t-\tau}$ *and* x^j_t *are connected by a lag-specific directed link* $x^i_{t-\tau} \to x^j_t$ *in* $\mathcal{G}$ *pointing forward in time if and only if* x^i *at time* $t-\tau$ *causes* x^j *at time* t*. Then the coefficient* $[G_\tau]_{ij}$ *associated with the adjacency matrix* $\boldsymbol{G}_\tau \in \mathcal{M}_d(\mathbb{R})$ *will be non-null and* $x^i \in \mathbf{Pa}^j_{\mathcal{G}}(< t)$ *if* $\tau \neq 0$*. For instantaneous links* ($\tau = 0$)*, we can not have self loops i.e.* $i \neq j$*. If* $\tau = 0$*, we have an edge* $x^i_t \to x^j_t$ *and* $x^i \in \mathbf{Pa}^j_{\mathcal{G}}(t)$ *if and only if* x^i *at time* t *causes* x^j *at time* t*.*

MTS with multiple regimes assumption. A MTS can exhibit either a single regime as assumed in prior works like Rhino (Gong et al., 2022) and DYNOTEARS (Pamfil et al., 2020) or K different non-overlapping regimes, as in our approach. For a MTS $(\boldsymbol{x}_t)_{t\in\mathcal{T}}$ composed of K disjoint regimes, each regime u is a stationary MTS block with a minimum duration ζ and has its own temporal causal graph $\mathcal{G}^u$, as defined in definition 1. We denote the set of these temporal causal graphs as $\mathcal{G} = (\mathcal{G}^u)_{u\in[|1:K|]}$. Regimes occur sequentially, with the constraint that a subsequent regime v (where $v = u+1$) cannot commence until at least ζ time units have passed since the start of the preceding one u, and also persists for a minimum of ζ samples. Additionally, if regime u reoccurs, its duration in the second appearance is also no less than ζ samples (Minimum regime duration). The indices corresponding to all occurrences of regime u are stored in a set denoted by $\mathcal{E}_u$. The collection $\mathcal{E} = (\mathcal{E}_u)_{u\in[|1:K|]}$ represents the unique time partition of the MTS $(\boldsymbol{x}_t)_{t\in\mathcal{T}}$ composed of K different regimes. In many application areas (Karmouche et al., 2023; Rahmani et al., 2023), non-stationarity can be modeled not through abrupt or continuous changes but rather as piecewise constant regimes. Importantly, the graphs $\mathcal{G}^u$ are regime-dependent, meaning that they vary across different regimes (i.e., $\mathcal{G}^u \neq \mathcal{G}^v$).

SEMs for MTS with multiple regimes. We now propose a novel functional form for SEMs that incorporates linear or non-linear relations, instantaneous links and multiple regimes. We have $\forall u \in [|1:K|], \forall t \in \mathcal{E}_u$:

$$x^i_t = g^u_i\left(\mathbf{Pa}^i_{\mathcal{G}^u}(< t), \mathbf{Pa}^i_{\mathcal{G}^u}(t)\right) + \epsilon^i_t, \tag{1}$$

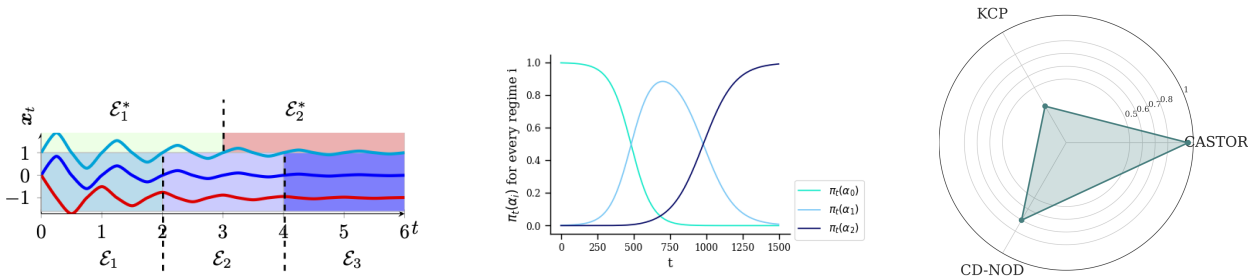
where g^u_i is a general differentiable linear or non-linear function and $\epsilon^i_t \sim \mathcal{N}(0,1)$, follows to a normal distribution. By assuming Causal Markov property, we can define the associated Causal Graphical Model (CGM), with $\boldsymbol{x}_{<t}$ refers $\{\boldsymbol{x}_{t-L}, ..., \boldsymbol{x}_{t-1}\}$, $\forall u \in [|1:K|], \forall t \in \mathcal{E}_u$:

$$p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{<t}, \mathcal{G}^u\right) = \prod_{i=1}^{d} p\left(x^i_t \mid \mathbf{Pa}^i_{\mathcal{G}^u}(< t), \mathbf{Pa}^i_{\mathcal{G}^u}(t)\right). \tag{2}$$

When the MTS consists of K unknown regimes, it cannot be represented by a single DAG. A new formulation describing the CGM in such scenarios is as follows:

$$p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{<t}\right) = \sum_{u=1}^{K} p(z_{t,u}) \cdot p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{<t}, \mathcal{G}^u\right), \tag{3}$$

where $p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{<t}, \mathcal{G}^u\right)$ is specified in Eq (2), while $p(z_{t,u})$ models the probability of $\boldsymbol{x}_t$ belonging to regime u ($\boldsymbol{x}_t$ belongs to regime u if $z_{t,u} = 1$ Figure 1b). As we explained above, the regimes are non-overlapping hence the $p(z_{t,u}) = \mathbb{1}_{\mathcal{E}_u}(t)$ is an indicator function, defined as $\mathbb{1}_{\mathcal{E}_u}(t) = 1$ if $t \in \mathcal{E}_u$ and 0 otherwise. Previous works assume prior knowledge of time partition $\mathcal{E}$ or report a summary causal graph (Huang et al., 2020), falling short of elucidating the full temporal causal graph. In the next section we present CASTOR, a causal discovery method tailored for MTS with multiple regimes.



(a) (b) (c)

Figure 2: (a) Initialization with $N_w = 3$ windows: $\mathcal{E}_1$ and $\mathcal{E}_3$ are pure regimes; $\mathcal{E}_2$ is impure, containing samples from ground-truth regimes $\mathcal{E}_1^*$ and $\mathcal{E}_2^*$ ($K = 2$). (b) Illustration of $\pi(\boldsymbol{\alpha}^u, t)$ after CASTOR's first iteration with equal windows of 500 samples for an MTS of 1500 samples with two ground-truth regimes: $\mathcal{E}_1^* = [|0 : 799|]$ and $\mathcal{E}_2^* = [|800 : 1500|]$. (c) Comparison between CASTOR, CD-NOD and KCP on regime detection for a MTS of 10 nodes and 4 regimes using accuracy metric.

3 CASTOR: CAUSAL TEMPORAL REGIME STRUCTURE LEARNING

We propose a method to jointly learn the number of regimes K, their indices $\mathcal{E} = (\mathcal{E}_u)_{u\in[|1:K|]}$, and the corresponding DAGs $\mathcal{G} = (\mathcal{G}^u)_{u\in[|1:K|]}$ from a MTS $(\boldsymbol{x}_t)_{t\in\mathcal{T}}$ with unknown regimes by maximizing the log-likelihood:

$$\log p\left(\boldsymbol{x}_{0:|\mathcal{T}|}\right) = \sum_{t=0}^{|\mathcal{T}|} \log\left(\sum_{u=1}^{K} \mathbb{1}_{\mathcal{E}_u}(t)\, p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{<t}, \mathcal{G}^u\right)\right). \tag{4}$$

Learning the DAGs $\mathcal{G} = (\mathcal{G}^u)_{u\in[|1:K|]}$, concurrently entails the estimation of the regime distribution $p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{<t}, \mathcal{G}^u\right)$. We model CASTOR's estimation of the joint density of the u^{th} regime by:

$$f^u\left(\boldsymbol{x}_t\right) := \prod_{i=1}^{d} f_i^u\left(\mathbf{Pa}^i_{\mathcal{G}^u}(<t), \mathbf{Pa}^i_{\mathcal{G}^u}(t)\right), \tag{5}$$

where f^u is a distribution family. It is important to highlight that while f^u can in theory be any distribution, in this particular study, we assume normal noise, used by many works (Pamfil et al. (2020); Huang et al. (2020)) and for which they showed the identifiability of causal graphs for one regime (Peters and Bühlmann, 2014). As a result from SEM Eq (1), our distribution f^u will be a Gaussian distribution.
Section 3.1 presents the challenges of the learning problem, while Sections 3.2 and 3.3 detail the procedures for regime learning and graph learning.

3.1 Challenges of the learning problem and EM choice justification

Since the regime indices are unknown, the learning problem is challenging. The sum inside the logarithm in Eq (4) renders the log-likelihood intractable. To address this, we employ the EM algorithm (Dempster et al., 1977), which introduces variables $\gamma_{t,u}$, that the posterior probability $p\left(\boldsymbol{z}_t|\boldsymbol{x}_t, \boldsymbol{x}_{<t}\right)$, to model regime participation and alternates between regime learning (E-step) and graph learning (M-step). Hence we have the expected log likelihood $\mathbb{E}[\log p] = \mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}}\left[\log p\left(\boldsymbol{x}_{0:T}, \boldsymbol{z}_{0:T}\right)\right]$:

$$\mathbb{E}[\log p] = \sum_{t=0}^{|\mathcal{T}|}\sum_{u=1}^{K} \gamma_{t,u} \log\Big(\mathbb{1}_{\mathcal{E}_u}(t)\cdot f^u\left(\boldsymbol{x}_t\right)\Big), \tag{6}$$

However, the logarithm term may be zero, causing the log-likelihood to diverge. Additionally, applying EM requires prior knowledge of the number of regimes, which we assume is unavailable.

We aim to learn continuous functions that solve the divergence problem, but also robust against random regime switching for a given sample $\boldsymbol{x}_t$. We want a sample $\boldsymbol{x}_t$, if belonging to regime u in the current iteration, to be assigned to the neighboring regimes $(u-1, u+1)$ or the same regime u in the next iteration. We employ the soft-max function $\pi(\boldsymbol{\alpha}^u, t) = \frac{\exp(\alpha_1^u t + \alpha_0^u)}{\sum_{j=1}^{N_w}\exp(\alpha_1^j t + \alpha_0^j)}$ (function of $\boldsymbol{\alpha}^u \in \mathbb{R}^2$ and time index t).

Regarding the second challenge, the unavailability of the number of regimes, CASTOR initially divides the MTS into $N_w > K$ equal time windows in the initialisation step (the length of the initialized windows is greater than ζ minimum regime duration), where each window represents one initial regime estimate. Our initialization scheme builds some initial pure regimes (regimes composed of samples from the same ground truth regime) and other impure ones (regimes composed of samples from two neighboring ground truth regimes), idea illustrated in Figure 4a. The expected log-likelihood of Eq (6) becomes:

$$\mathbb{E}[\log p] = \sum_{t=0}^{|\mathcal{T}|}\sum_{u=1}^{N_w} \gamma_{t,u}\ \log\Big(\pi(\boldsymbol{\alpha}^u, t)\, f^u(\boldsymbol{x}_t)\Big).$$

After our initialization scheme, we start with some pure regimes and impure regimes due to the fact that our ini-

tial windows are sufficiently small. CASTOR estimates the graphs (by learning $f^u(\boldsymbol{x}_t)$) for the different initial regimes and also learns the parameters $\boldsymbol{\alpha}^u$ that maximizes the alignment between $\gamma_{t,u}$ (rectangular function) and $\pi(\boldsymbol{\alpha}^u, t)$ (example on Figure 4b) using the M-step (subsection 3.3). This alignment is desirable to ensure stability in the assignment of the samples. Then, our method alternates between updating the regime indices $\mathcal{E} = (\mathcal{E}_u)_{u\in[|1:N_w|]}$ during the E-step (subsection 3.2) and learning the temporal causal graphs $(\mathcal{G}^u)_{u\in[|1:N_w|]}$ along with new $\pi(\boldsymbol{\alpha}^u, t)$ during the M-step. This process repeats until a predefined maximum number of iterations is reached.

Justification of EM choice. We argue that inferring the regimes and learning the associated DAGs are intertwined problems, which makes the EM algorithm a suitable choice for this learning task. We investigate the possibility of solving the problem of causal structure learning from MTS with multiple regimes by first using change point detection method to identify regime indices, then applying existing causal discovery methods for each regime. However, as shown in Figure 2c, standard change point detection methods failed to detect regimes resulting from changes in causal mechanisms. Specifically, when comparing CASTOR and CD-NOD (Huang et al., 2020) with KCP (Arlot et al., 2019), a state-of-the-art change point detection method, we observed that KCP was unable to detect the regimes. This is likely because changes in causal mechanisms involve shifts in conditional distributions, which are harder for KCP to detect. Further details are provided in the Appendix E.10.

3.2 Expectation step: Regime learning

During the E-step, CASTOR updates $\gamma_{t,u}$ (as shown in Eq (7), derivation details in Appendix C), the probability of $\boldsymbol{x}_t$ belonging to regime u depends on two factors, the position of $\boldsymbol{x}_t$ within the current regime and whether the current regime is pure or impure.

$$\begin{aligned}\gamma_{t,u} &= \frac{p(z_{t,u}=1)\, p(\boldsymbol{x}_t \mid \boldsymbol{x}_{<t}, z_{t,u}=1, \mathcal{G}^u)}{\sum_{j=1}^{N_w} p(z_{t,j}=1)\, p(\boldsymbol{x}_t \mid \boldsymbol{x}_{<t}, z_{t,j}=1, \mathcal{G}^j)} \\ &= \frac{\pi(\boldsymbol{\alpha}^u, t)\, f^u(\boldsymbol{x}_t)}{\sum_{j=1}^{N_w} \pi(\boldsymbol{\alpha}^j, t)\, f^j(\boldsymbol{x}_t)} \propto \pi(\boldsymbol{\alpha}^u, t)\, f^u(\boldsymbol{x}_t)\end{aligned} \tag{7}$$

Case 1: When $\boldsymbol{x}_t$ belongs to a pure regime u and is far from the border in the current iteration, $\pi(\boldsymbol{\alpha}^u, t)$ is high (e.g., $\pi(\boldsymbol{\alpha}^0, t \in [0, 300])$ in Figure 4b). The graph for u is meaningful because it was learned on pure data, so $f^u(\boldsymbol{x}_t)$ is also high. Consequently, $\gamma_{t,u} \propto \pi(\boldsymbol{\alpha}^u, t)\, f^u(\boldsymbol{x}_t)$ (Eq (7)) remains maximal, keeping $\boldsymbol{x}_t$ in regime u in the next iteration.

Case 2: $\boldsymbol{x}_t$ belongs to pure regime u and is near the border in the current iteration. In this situation, $\pi(\boldsymbol{\alpha}^u, t)$ and $\pi(\boldsymbol{\alpha}^{u+1}, t)$ are approximately equal (e.g., $\pi(\boldsymbol{\alpha}^0, t \in [350, 500])$ and $\pi(\boldsymbol{\alpha}^1, t \in [350, 500])$ in Figure 4b). Since the graph for regime u was learned from pure data, $f^u(\boldsymbol{x}_t)$ remains high. Therefore, $\gamma_{t,u}$ stays maximal, keeping $\boldsymbol{x}_t$ in regime u in the next iteration.

Case 3: $\boldsymbol{x}_t$ belongs to impure regime $u+1$ and is near the border in the current iteration. Here, $\pi(\boldsymbol{\alpha}^u, t)$ and $\pi(\boldsymbol{\alpha}^{u+1}, t)$ are roughly equal (e.g., $\pi(\boldsymbol{\alpha}^0, t \in [501, 650])$ and $\pi(\boldsymbol{\alpha}^1, t \in [501, 650])$ in Figure 4b). However, because the graph of regime u is more meaningful (learned from pure data), $f^u(\boldsymbol{x}_t) > f^{u+1}(\boldsymbol{x}_t)$. Thus, $\gamma_{t,u} > \gamma_{t,u+1}$, causing $\boldsymbol{x}_t$ to switch from regime $u+1$ to u in the next iteration.

Case 4: $\boldsymbol{x}_t$ belongs to impure regime $u+1$ and is far from the border (e.g., $t \in [650, 850]$ in Figure 4b). In this case, it's uncertain whether $\boldsymbol{x}_t$ will switch regimes in the next iteration. However, as the pure regime u expands with each iteration, $\boldsymbol{x}_t$ will eventually be near the border of regime $u+1$, bringing us back to Case 3.

For simplicity reasons, we explicit these cases from one border but the same thing happens in the other border which accelerates convergence. If a sample $\boldsymbol{x}_t$ belongs to regime u, it will never be allocated to a non neighboring regime v due to the fact that $\pi(\boldsymbol{\alpha}^u, t) >> \pi(\boldsymbol{\alpha}^v, t)$ and $f^u(\boldsymbol{x}_t) >> f^v(\boldsymbol{x}_t)$.

Example. In Figure 4b, after learning the graphs and parameters $\boldsymbol{\alpha}^u$ for the first iteration where the regimes are equal windows of 500 data points. The samples $\boldsymbol{x}_t$, where $t \in [|350, 500|]$ belonging to regime 0 in the previous iteration, are more likely to stay in regime 0 or transition to the neighboring regime 1 ($\pi(\boldsymbol{\alpha}^1, t \in [350, 500])$ same range as $\pi(\boldsymbol{\alpha}^0, t \in [350, 500])$) than the non neighboring regime 2 ($\pi(\boldsymbol{\alpha}^2, t \in [350, 500])$ almost 0).

After updating $\gamma_{t,u}$, for each sample $\boldsymbol{x}_t$, CASTOR assigns a value of 1 to the most probable regime u (with the highest $\gamma_{t,u}$), and 0 to others. Additionally, CASTOR filters out regimes with insufficient samples (fewer than ζ, the minimum regime duration, defined as a hyper-parameter). Discarded regime samples are then reassigned to the nearest regime in terms of probability $\gamma_{t,u}$ in the subsequent iteration which is in general a neighboring regime ensured by the way we set up the probability $\gamma_{t,u} \propto \pi(\boldsymbol{\alpha}^u, t)\, f^u(\boldsymbol{x}_t)$.

3.3 Maximization step: Graph learning

CASTOR utilizes the binary regime indices $\gamma_{t,u}$ learned in the E-step to estimate the DAGs for each regime and learn the parameters $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}^u, u \in [|1, N_w|]\}$ that align $\pi(\boldsymbol{\alpha}^u, t)$ with $\gamma_{t,u}$ by maximizing the following equation:

$$\sup_{\mathcal{G},\boldsymbol{\alpha}} \frac{1}{|\mathcal{T}|} \sum_{u=1}^{N_w} \sum_{t=0}^{|\mathcal{T}|} \gamma_{t,u} \log \pi\left(\boldsymbol{\alpha}^u, t\right) f^u(\boldsymbol{x}_t) - \lambda |\mathcal{G}^u|, \quad \text{s.t.} \quad \boldsymbol{G}_0^u \text{ is a DAG.}$$

The maximization of the aforementioned equation can be decomposed into two distinct maximization problems. The first problem, regime alignment, focuses on aligning $\pi\left(\boldsymbol{\alpha}^u, t\right)$ with $\gamma_{t,u}$:

$$\sup_{\boldsymbol{\alpha}} \frac{1}{|\mathcal{T}|} \sum_{u=1}^{N_w} \sum_{t=1}^{|\mathcal{T}|} \gamma_{t,u} \log \pi\left(\boldsymbol{\alpha}^u, t\right), \tag{8}$$

while the second one, graph learning, involves estimating DAGs for every regime:

$$\mathcal{S}(\mathcal{G}, \mathcal{E}) := \sup_{\mathcal{G}} \frac{1}{|\mathcal{T}|} \sum_{u=1}^{N_w} \sum_{t \in \mathcal{E}_u} \log f^u(\boldsymbol{x}_t) - \lambda |\mathcal{G}^u|, \quad \text{s.t., } \boldsymbol{G}_0^u \text{ is a DAG.} \tag{9}$$

where $\mathcal{G}$ stands for $\mathcal{G} = (\mathcal{G}^u)_{u \in [|1:N_w|]}$, $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}^u, \forall u \in [|1, N_w|]\}$, $|\mathcal{G}^u|$ is the number of edges in the temporal causal graph of regime u and we note $\mathcal{S}(\mathcal{G}, \mathcal{E})$ the score function of CASTOR. The first term in CASTOR's score function is the averaged log-likelihood over data while the second term is a penalty term with positive small coefficient λ that controls the sparsity constraint. We further impose an acyclicity constraint on the adjacency matrix $\boldsymbol{G}_0^u$ of instantaneous links. The other adjacency matrices $\boldsymbol{G}_{\tau \in [|1:L|]}^u$ are inherently acyclic by definition, because these matrices establish links between variables at time t and their time-lagged parents at time $t - \tau$. It is worth noting that the optimization for regime alignment remains the same for both linear and nonlinear causal relationships. However, the graph learning problem differs between these two settings.

3.4 Linear case

For linear causal relationships, the SEM (Eq (1)) is:

$$\forall u \in [|1:, K|], \forall t \in \mathcal{E}_u : \boldsymbol{x}_t = \boldsymbol{x}_t \boldsymbol{G}_0^u + \sum_{\tau=1}^{L} \boldsymbol{x}_{t-\tau} \boldsymbol{G}_\tau^u + \boldsymbol{\epsilon}_t,$$

where $\epsilon_t \sim \mathcal{N}(0, I)$. Thus, $\forall u \in [|1:, K|], \forall t \in \mathcal{E}_u$:

$$\boldsymbol{x}_t \mid \boldsymbol{x}_{<t} \sim \mathcal{N}(\boldsymbol{x}_t \boldsymbol{G}_0^u + \sum_{\tau=1}^{L} \boldsymbol{x}_{t-\tau} \boldsymbol{G}_\tau^u, I).$$

Using CASTOR's score function results in the following minimization problem (details in Appendix C):

$$\min_{\mathcal{G}} \frac{1}{|\mathcal{T}|} \sum_{u=1}^{N_w} \sum_{t=1}^{|\mathcal{T}|} \gamma_{t,u} \left\| \boldsymbol{x}_t - \left(\boldsymbol{x}_t \boldsymbol{G}_0^u + \sum_{\tau=1}^{L} \boldsymbol{x}_{t-\tau} \boldsymbol{G}_\tau^u \right) \right\|_F^2 + \lambda |\mathcal{G}^u| + \frac{\rho}{2} h\left(\boldsymbol{G}_0^u\right)^2 + \alpha h\left(\boldsymbol{G}_0^u\right), \tag{10}$$

where $\mathcal{G}$ stands for $\mathcal{G} = (\mathcal{G}^u)_{u \in [|1:N_w|]}$ and α, ρ characterize the strength of the DAG penalty. The function $h(\boldsymbol{G}) = \operatorname{tr}\left(e^{\boldsymbol{G} \odot \boldsymbol{G}}\right) - d$ corresponds to the acyclicity constraints proposed in Zheng et al. (2018) ($\odot$ is the Hadamard product). For example, let $\boldsymbol{G}_0$ be the adjacency graph for instantaneous relation. The constrain condition requires $h(\boldsymbol{G}_0) = 0$. We employ an augmented Lagrangian method (Zheng et al., 2018; Pamfil et al., 2020; Brouillard et al., 2020; Liu and Kuang, 2023) to address the optimization challenge that incorporates the acyclicity constraints.

Algorithm 1 CASTOR algorithm

Input: MTS $\boldsymbol{X}$, window size W, lag L, max iteration N_{iter}, min regime duration ζ
for $i = 1$ **to** N_{iter} **do**

$\gamma_{t,u} \leftarrow \frac{\pi_t(\alpha_u) f^u(\boldsymbol{x}_t)}{\sum_{j=1}^{N_w} \pi_t(\alpha_j) f^j(\boldsymbol{x}_t)}$ `Expectation step`

$\alpha \leftarrow \operatorname{argmin}_\alpha \sum_{u=1}^{N_w} \sum_{t=1}^{|\mathcal{T}|} \gamma_{t,u} \log\left(\pi_t(\alpha_u)\right)$ `Regime alignment`

$\mathcal{G} \leftarrow$ argmin of Eq (10) or (11) `Graph learning`

if $\sum_t^{|\mathcal{T}|} \gamma_{t,u} \leq \zeta$ **then**
$\forall t : \gamma_{t,u} \leftarrow 0$
end if
end for
Output: γ, $\mathcal{G}$

3.5 Non-linear case

For non-linear causal relationships in the MTS $(\boldsymbol{x_t})_{t \in \mathcal{T}}$, we estimate the distribution parameters f^u (Eq (9)) by modeling the non-linear SEM (Eq (1)). According to Eq (1), each component x_t^i follows a Gaussian distribution:

$$x_t^i \mid \boldsymbol{x}_{<t} \sim \mathcal{N}(g_i^u(\mathbf{Pa}^i_{\mathcal{G}^u}(<t), \mathbf{Pa}^i_{\mathcal{G}^u}(t)), 1).$$

We employ Neural Networks (NN) to capture the non-linearity, as in Zheng et al. (2020); Brouillard et al. (2020); Liu and Kuang (2023). For each regime u and component i, a separate and small NN_i^u models the distribution parameters. We aggregate the lagged variables into $\boldsymbol{x}_t^{\text{lag}} = [\boldsymbol{x}_{t-1} | \cdots | \boldsymbol{x}_{t-L}]$ and input both $\boldsymbol{x}_t$ and $\boldsymbol{x}_t^{\text{lag}}$ into NN_i^u to predict $\hat{x}_t^i$, thereby estimating f_i^u. Our neural networks are defined as,

$$\forall i \in [|1:d|] : \text{NN}_i^u(\boldsymbol{x}_t, \boldsymbol{x}_t^{\text{lag}}) = \psi_i^u\left(\phi_i^u(\boldsymbol{x}_t), \phi_i^{u,\text{lag}}(\boldsymbol{x}_t^{\text{lag}})\right),$$

where ψ_i^u consists of locally connected layers (Zheng et al., 2020) and activation functions, and ϕ_i^u, $\phi_i^{u,\text{lag}}$ are composed of linear layers and sigmoid functions. The locally connected layers help to capture variable dependencies in the initial layer.

For each node i, the instantaneous and time-lagged interactions with node j are captured by the norms

of the corresponding columns in the first layer weight matrices:

$$[\boldsymbol{G}_0^u]_{ij} = \|\Theta_i^u(\text{column } j)\|_2\,,$$
$$[\boldsymbol{G}_\tau^u]_{ij} = \left\|\Theta_i^{u,\text{lag}}(\text{column } ((\tau-1)\cdot d + j))\right\|_2\,,$$

where Θ_i^u and $\Theta_i^{u,\text{lag}}$ are parameters of the first layers of the NNs ϕ_i^u and $\phi_i^{u,\text{lag}}$, respectively. The matrix $\Theta_i^{\text{lag}} \in \mathcal{M}_{d,dL}(\mathbb{R})$ has dL columns, with L as the maximum lag. Incorporating NNs into the maximization step (Eq. 9) results in the following minimization problem (details in Appendix C):

$$\min_{\theta,\mathcal{G}} \quad \frac{1}{|\mathcal{T}|}\sum_{u=1}^{N_w}\sum_{t=1}^{|\mathcal{T}|}\sum_{i=1}^{d}\gamma_{t,u}\mathcal{L}\left(x_t^i, \psi_i^u\left(\phi_i^u(\boldsymbol{x}_t), \phi_i^{u,\text{lag}}(\boldsymbol{x}_t^{\text{lag}})\right)\right) + \lambda|\mathcal{G}^u| + \frac{\rho}{2}h(\boldsymbol{G}_0^u)^2 + \alpha h(\boldsymbol{G}_0^u), \tag{11}$$

where $\mathcal{L}$ is a least squares loss, θ includes all network parameters, and $\mathcal{G} = (\mathcal{G}^u)_{u\in[|1:N_w|]}$. We enforce the acyclicity constraint using an augmented Lagrangian term. Algorithm 1 summarizes our CASTOR model for both linear and nonlinear causal relationships.

4 IDENTIFIABILITY RESULTS

In this section, we present the identifiability results of our framework, highlighting that the causal structure can be recovered from observational data only. Following Brouillard et al. (2020); Gong et al. (2022); Liu and Kuang (2023), we outline our assumptions below:

Definition 2 (Causal Stationarity, Runge (2018))**.** *A stationary time series process $(\boldsymbol{x}_t)_{t\in\mathcal{T}}$ with graph $\mathcal{G}$ is called causally stationary over a time index set $\mathcal{T}$ if and only if for all links $\boldsymbol{x}_{t-\tau}^i \to \boldsymbol{x}_t^j$ in the graph*

$$x_{t-\tau}^i \not\perp\!\!\!\perp x_t^j \mid \boldsymbol{x}_{<t} \setminus \{x_{t-\tau}^i\}.$$

Assumption 1 (Causal Stationarity for MTS with multiple regime)**.** *A MTS $(\boldsymbol{x}_t)_{t\in\mathcal{T}}$ with K regimes, graph set $(\mathcal{G}^u)_{u\in[|1:K|]}$, and regime partition $\mathcal{E} = (\mathcal{E}_u)_{u\in[|1:K|]}$ is **causally stationary** over the time index set $\mathcal{T}$ if, for each regime $u \in [|1:K|]$, the sub-series $(\boldsymbol{x}_t)_{t\in\mathcal{E}_u}$ is causally stationary with graph $\mathcal{G}^u$ as defined in definition 2.*

Assumption 2 (Causal Markov Property (CPM))**.** *A set of joint distributions $(p(\cdot|\mathcal{G}^u))_{u\in[|1:K|]}$ satisfies the **CPM** with respect to the DAGs $(\mathcal{G}^u)_{u\in[|1:K|]}$ if, for each $u \in [|1:K|]$, the distribution $p(\cdot|\mathcal{G}^u)$ satisfies the CPM relative to the DAG $\mathcal{G}^u$. Specifically, in every regime u, each variable is independent of its non-descendants given its parents.*

Assumption 3 (Causal Minimality)**.** *Given a set of DAGs $(\mathcal{G}^u)_{u\in[|1:K|]}$ and a set of joint distribution $(p(\cdot|\mathcal{G}^u))_{u\in[|1:K|]}$, we say that this set of distributions satisfies causal minimality w.r.t. the set of DAGs $(\mathcal{G}^u)_{u\in[|1:K|]}$ if for every u: $p(\cdot|\mathcal{G}^u)$ is Markovian w.r.t the DAG $\mathcal{G}^u$ but not to any proper subgraph of $\mathcal{G}^u$.*

Assumption 4 (Causal Sufficiency)**.** *A set of observed variables $\boldsymbol{V}$ is causally sufficient for a process $\boldsymbol{x}_t$ if and only if in the process every common cause of any two or more variables in $\boldsymbol{V}$ is in $\boldsymbol{V}$ or has the same value for all units in the population.*

Using the above assumptions and operating in the settings outlined in Brouillard et al. (2020); Liu and Kuang (2023), Theorem 1 states that the ground truth solution $(\mathcal{G}^*, \mathcal{E}^*)$ uniquely maximizes the score defined in Eq (9), up to a permutation.

Theorem 1. *Assume SEMs with Gaussian noise, presented in Eq(1), that satisfy the causal Markov property, stationarity, minimality and sufficiency. If each regime has enough data and the penalty coefficients in Eq (10-11) are sufficiently small, it holds asymptotically that for any estimation*

$$\mathcal{S}\left(\mathcal{G}^*, \mathcal{E}^*\right) > \mathcal{S}(\hat{\mathcal{G}}, \hat{\mathcal{E}})$$

If any of the estimated graphs $\hat{\mathcal{G}}^u$ represents an edge disagreement with all the ground truth graphs $\mathcal{G}^*$ or any of the estimated regimes in $\hat{\mathcal{E}}$ is close to none of the ground truth regimes in the sense of Kullback–Leibler.

Full proof is provided in Appendix F. When this score is maximized CASTOR can identify true regimes and causal graphs up to a permutation. However, the convergence is not always guaranteed due to EM instability and the non-convexity of acyclicity constraints. Moreover, Theorem 1 does not provide additional information on the ranking of various solutions. Given two sub optimums, one closer to the ground truth solution w.r.t. KL divergence, Theorem 2 shows the closest one has the higher score.

Theorem 2. *Assume the same conditions as in Theorem 1, for any estimations $(\mathcal{G}, \mathcal{E})$ and $(\mathcal{G}', \mathcal{E}')$, such that $(\mathcal{G}, \mathcal{E})$ is closer to the optimal solution $(\mathcal{G}^*, \mathcal{E}^*)$ than $(\mathcal{G}', \mathcal{E}')$ in terms of Kullback–Leibler, it holds asymptotically: $\mathcal{S}(\mathcal{G}, \mathcal{E}) > \mathcal{S}(\mathcal{G}', \mathcal{E}')$.*

5 EXPERIMENTS

5.1 Synthetic data

Data generation. We conduct extensive experiments to evaluate CASTOR's performance on synthetic datasets (details in Appendix E.1). For ground truth graph generation, we use the Barabási-Albert model (degree 4) for instantaneous links and the Erdős–Rényi model (degree 1–2) for time-lagged relationships. For non-linear cases, the functions g_i^u in Eq (1) use random weights from a uniform distribution over]0,2] and activation functions randomly chosen from

{`Tanh, LeakyReLU, ReLU`}. We consider $L = 1$, while additional experiments with multiple lags are provided in the Appendix. Regime durations are randomly selected from {300, 400, 500, 600}. We test different numbers of nodes ({5, 10, 20, 40} for linear cases, {10, 20} for non-linear) and varying regime counts ($K \in \{2, 3, 4, 5\}$). Each combination of K and d nodes is repeated three times, resulting in over 60 distinct datasets, we provide `our code`.

Benchmarks. We benchmark our model against several baselines, including causal discovery methods for MTS with multiple regimes, such as CD-NOD (Huang et al., 2020) and RPCMCI (Saggioro et al., 2020). Since CD-NOD returns a summary graph (see Appendix E.10), we compute a comparable summary graph from CASTOR's output for fair evaluation. CASTOR is also compared with models for single-regime MTS, including Rhino, with and without historically dependent noise, (Gong et al., 2022), PCMCI+ , using Partial correlation for linear relationships and GPDC for non-linear ones, (Runge, 2020), DYNOTEARS (Pamfil et al., 2020), and VARLINGAM (Hyvärinen et al., 2010). *Given that these models cannot deal with multiple regimes, to make the evaluation fair, we put them in a more favorable position and provide these models with the true regime partition information. This is done by training the aforementioned models on each pure regime separately (regime governed by the same graph).*

Results and discussion. Table 1 presents the results for linear case, while Figure 3 summarizes the results for the non-linear setting. In the linear case, we notice that CASTOR and DYNOTEARS outperform all the baselines, either those designed for MTS with multiple regimes, such as RPCMCI and CD-NOD or the other methods that assumes stationarity (Table 1). We emphasize that CASTOR performs similarly to DYNOTEARS, even though the latter benefits from prior access to ground truth regime partitions by being trained separately on each pure regime.
In the non-linear case, CASTOR outperforms all the baseline on instantaneous link, Figure 3. PCMCI+ and Rhino w/o hist performs better than CASTOR in inferring time lagged links, however it is worth to note that these algorithms has access to the ground truth regime partition while CASTOR learns the number of regimes, their indices and the corresponding DAGs. In both scenarios (linear and non-linear) RPCMCI struggles to achieve convergence, particularly in settings with more than 3 different regimes due to its assumption of only inferring time-lagged relations. The comparison with CD-NOD on graph learning and also regime detection show that CASTOR outperforms CD-NOD, which is understandable because, CD-NOD learns one summary graph for the whole MTS and also expects only a few

Table 1: Average F1 scores for different models on linear SEMs with $d = 40$ nodes. K indicates the number of regimes, *Split* denotes whether regime separation is automatic (A) or manual (M), and *Type* classifies the graph as window (W) or summary (S). *Inst.* refers to instantaneous links, and *Lag* to time-lagged edges.

			$K = 3$		$K = 4$	
Model	Split	Type	Inst.	Lag	Inst.	Lag
VARLINGAM	M	W	9.83	1.13	10.9	1.43
Rhino	M	W	0.00	20.8	0.00	22.8
Rhino w/o hist	M	W	0.00	38.4	0.00	39.1
PCMCI+	M	W	54.1	84.6	53.7	86.1
DYNOTEARS	M	W	97.4	98.8	97.3	97.9
RPCMCI	A	W	-	18.4	-	-
CASTOR	A	W	**98.2**	**99.8**	**98.3**	**98.9**
CD-NOD	A	S	11.3		5.57	
CASTOR	A	S	**99.8**		**99.2**	

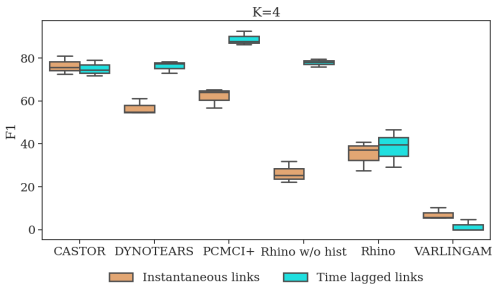
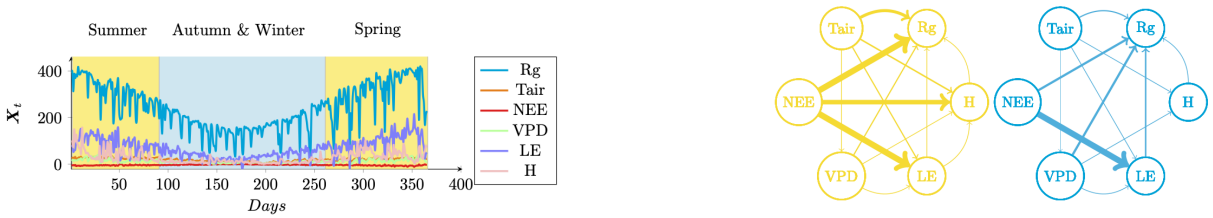


Figure 3: F1 scores by Models for 20 nodes and 4 regimes for non linear causal relationships. Orange indicates performance on instantaneous links, and sky-blue signifies performance on time-lagged relationships.

variables of the graph to be affected by the regime change (This assumption may not hold true in real scenarios such as epileptic seizures or climate science). However, CASTOR does not have this assumption and also learns one graph per regime (more details on Appendices E.9 and E.10.). Although our settings are identifiable, PCMCI+ infers a Markov equivalent class for the instantaneous links, which explains its performance deterioration in instantaneous relations and with a higher number of nodes. Rhino faces challenges in the absence of historical dependent noises (as confirmed by Figure 4 on page 24 of the Rhino paper). Moreover, Rhino utilizes ConvertibleGNN with Normalizing flows to learn the causal graphs. To train this model, a minimum of 50 time series of length 200 (10000 samples), all sharing the same causal graph is needed. Our ablation studies in Appendix E.4 highlight that neither the size of the window nor the minimum regime duration impact CASTOR performance on both regime detection and causal graph inferring. Additional results and evaluations using other metrics, such as SHD, that confirm the above findings are available in Appendix



(a) (b)

Figure 4: CASTOR's results on Biosphere-Atmosphere data.(a) CASTOR identifies two regimes, months with hot weather colored in yellow and other with cold weather colored in blue. (b) The instantaneous links are for the two regimes, with the blue graph corresponding to the blue regime, cold weather, and the yellow one to the yellow regime, hot weather.

Table 2: F1 Scores across IT data.

Model	Graph type	F1 Reg1	F1 Reg 2
PCMCI+	W	12.1	**29.6**
DYNOTEARS	W	18.2	**28.5**
Rhino	W	28.6	25.8
CASTOR	W	18.2	**28.5**
CASTOR non-lin	W	**40.0**	24.5
CD-NOD	S	23.5	
CASTOR non-lin	S	**36.8**	

E.7 and E.8. The comparison with CD-NOD (Huang et al., 2020) and KCP (Arlot et al., 2019) on the regime detection task is presented in Appendix E.10.

5.2 Web activity dataset

We now evaluate CASTOR on two stacked IT monitoring time series datasets, firstly introduced by Bystrova et al. (2023) (details and analysis of the challenges of these datasets are in Appendix E.2), each comprising 1106 timestamps and 7 nodes, sourced from EasyVista. The web activity data is challenging, because it could present missing values, misaligned time series and partially sleeping time series due to inactivity of certain servers. IT experts are not sure that this data satisfies causal sufficiency assumption. Aït-Bachir et al. (2023) present a study on the performance of causal discovery method on this data and shows the same range of performance. We compared our method to a subset of the best models mentioned in the benchmark subsection. As it is evident from Table 2, CASTOR proficiently identifies the exact number of regimes and their indices. On regime 2, CASTOR (with linear relationships), PCMCI+ and DYNOTEARS outperform Rhino and CASTOR non-lin (uses NN for non linear relationships). However, in regime 1, CASTOR non-linear and Rhino outperform all other models. This superiority can be attributed to the fact that CASTOR non-lin and Rhino employs NNs to learn causal relationships, and the non-linearity in regime 1 complicates graph learning for the other models.

5.3 Biosphere–Atmosphere data

We demonstrate CASTOR's practical utility by applying it to biosphere-atmosphere data from the FLUXNET dataset (San Luis site, Argentina (Garcia et al., 2015)). Previously, Krich et al. (2021) analyzed these data using PCMCI+ to identify causal relationships among six variables: global radiation (R_g), air temperature (T_{air}), net ecosystem exchange (NEE), vapor pressure deficit (VPD), sensible heat (H), and latent heat flux (LE). Unlike this prior approach, which used PCMCI+ on overlapping three month windows, we automate both causal discovery and regime learning by providing CASTOR with one year of MTS data without predetermined intervals. Starting from four initial regimes based on non-overlapping three month windows and enforcing a minimum regime duration of two months, CASTOR identified two regimes after a few iterations: one spanning Autumn and Winter (April–September) and the other covering Spring and Summer. This partition achieved an accuracy of 85.4%. An initial observation is all variables appear as parents of global radiation, a counterintuitive result due to causal sufficiency assumption. CASTOR interpret global radiation as net radiation. Mathematically, net radiation is defined as $R_n = R_g - SW_\uparrow + LW_\downarrow - LW_\uparrow$, balancing global shortwave and longwave radiation components, and also satisfies the energy balance equation $R_n = H + LE + G$. Assuming causal sufficiency (all causal variables observed) leads CASTOR to incorrectly infer causal directions. For additional analysis, see Appendix E.12.

6 CONCLUSION

We present CASTOR, a method for learning causal relationships from MTS with multiple regimes. CASTOR learns the number of regimes, their indices, and infers causal graphs for each regime simultaneously. It outperforms existing causal discovery models in handling both linear and non-linear relationships across multiple regimes on synthetic and real datasets.

Acknowledgment

We thank Ali Mourtada, Nikolaos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jimenez, Anas Essounaini, Thibault Séjourné for helpful feedbacks and comments. This work was supported by the SNSF Sinergia project 'PEDESITE: Personalized Detection of Epileptic Seizure in the Internet of Things (IoT) Era'

References

Aït-Bachir, A., Assaad, C. K., de Bignicourt, C., Devijver, E., Ferreira, S., Gaussier, E., Mohanna, H., and Zan, L. (2023). Case studies of causal discovery from it monitoring time series. *arXiv Preprint arXiv:2307.15678*.

Arlot, S., Celisse, A., and Harchaoui, Z. (2019). A kernel multiple change-point algorithm via model selection. *Journal of machine learning research*.

Assaad, C. K., Devijver, E., and Gaussier, E. (2022). Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*.

Assaad, K., Devijver, E., Gaussier, E., and Ait-Bachir, A. (2021). A mixed noise and constraint-based approach to causal inference in time series. In *ECML PKDD*.

Balsells-Rodas, C., Wang, Y., and Li, Y. (2024). On the identifiability of switching dynamical systems. In *Forty-first International Conference on Machine Learning*.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*.

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. (2020). Differentiable causal discovery from interventional data. *In Advances in Neural Information Processing Systems*.

Bussmann, B., Nys, J., and Latré, S. (2021). Neural additive vector autoregression models for causal discovery in time series. In *Discovery Science*.

Bystrova, D., Assaad, C. K., Arbel, J., Devijver, E., Gaussier, E., and Thuiller, W. (2023). Causal discovery from time series with hybrids of constraint-based and noise-based algorithms. *arXiv e-prints*, pages arXiv–2306.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society*.

Entner, D. and Hoyer, P. O. (2010). On causal discovery from time series data using fci. *Probabilistic graphical models*.

Garcia, A., Bella, C., Houspanossian, J., Magliano, P., Jobbágy, E., Posse, G., Fernández, R., and Nosetto, M. (2015). Fluxnet2015 ar-slu san luis, dataset(2009-2011).

Gong, W., Jennings, J., Zhang, C., and Pawlowski, N. (2022). Rhino: Deep causal temporal relationship learning with history-dependent noise. *Preprint arXiv:2210.14706*.

Günther, W., Ninad, U., and Runge, J. (2023). Causal discovery for time series from multiple datasets with latent contexts. *Preprint arXiv:2306.12896*.

Hasan, U., Hossain, E., and Gani, M. O. (2023). A survey on causal discovery methods for iid and time series data. *Transactions on Machine Learning Research*.

Haufe, S., Müller, K.-R., Nolte, G., and Krämer, N. (2010). Sparse causal discovery in multivariate time series. In *causality: objectives and assessment*, pages 97–106. PMLR.

Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., and Schölkopf, B. (2020). Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(1).

Hyttinen, A., Eberhardt, F., and Järvisalo, M. (2014). Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pages 340–349.

Hyvärinen, A., Zhang, K., Shimizu, S., and Hoyer, P. O. (2010). Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5).

Karmouche, S., Galytska, E., Runge, J., Meehl, G. A., Phillips, A. S., Weigel, K., and Eyring, V. (2023). Regime-oriented causal model evaluation of atlantic–pacific teleconnections in cmip6. *Earth System Dynamics*.

Ke, N. R., Bilaniuk, O., Goyal, A., Bauer, S., Larochelle, H., Schölkopf, B., Mozer, M. C., Pal, C., and Bengio, Y. (2019). Learning neural causal models from unknown interventions. *Preprint arXiv:1910.01075*.

Krich, C., Migliavacca, M., Miralles, D. G., Kraemer, G., El-Madany, T. S., Reichstein, M., Runge, J., and Mahecha, M. D. (2021). Functional convergence of biosphere–atmosphere interactions in response to meteorological conditions. *Biogeosciences*.

Liu, C. and Kuang, K. (2023). Causal structure learning for latent intervened non-stationary data. In *International Conference on Machine Learning*.

Lorch, L., Rothfuss, J., Schölkopf, B., and Krause, A. (2021). Dibs: Differentiable bayesian structure learning. *In Advances in Neural Information Processing Systems*, 34:24111–24123.

Löwe, S., Madras, D., Zemel, R., and Welling, M. (2022). Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*.

Newman, M. (2018). *Networks*. Oxford university press.

Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., and Aragam, B. (2020). Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*.

Peters, J. and Bühlmann, P. (2014). Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228.

Peters, J., Janzing, D., and Schölkopf, B. (2013). Causal inference on time series using restricted structural equation models. *In Advances in Neural Information Processing Systems*.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053.

Rahmani, A., Venkitaraman, A., and Frossard, P. (2023). A meta-gnn approach to personalized seizure detection and classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Runge, J. (2018). Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7).

Runge, J. (2020). Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1388–1397. PMLR.

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*.

Saggioro, E., de Wiljes, J., Kretschmer, M., and Runge, J. (2020). Reconstructing regime-dependent causal relationships from observational time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(11).

Shen, X., Ma, S., Vemuri, P., and Simon, G. (2020). Challenges and opportunities with causal discovery algorithms: application to alzheimer's pathophysiology. *Scientific reports*.

Song, L., Kolar, M., and Xing, E. (2009). Time-varying dynamic bayesian networks. *In Advances in Neural Information Processing Systems*, 22.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Triantafillou, S. and Tsamardinos, I. (2015). Constraint-based causal discovery from multiple interventions over overlapping variable sets. *The Journal of Machine Learning Research*, 16(1):2147–2205.

Xu, C., Huang, H., and Yoo, S. (2019). Scalable causal graph learning through a deep neural network. In *ACM international conference on information and knowledge management*.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. *In Advances in neural information processing systems*.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. (2020). Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on mathematical software (TOMS)*.

Zhu, S., Ng, I., and Chen, Z. (2019). Causal discovery with reinforcement learning. *Preprint arXiv:1906.04477*.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes, Appendix E.4 and E.11]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes, our model takes few minutes, we encourage the reviewers to run different experiments (to reproduce the results) using the notebooks provided in supplementary material]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes, Appendix F.1 and F.2]
   (c) Clear explanations of any assumptions. [Yes, main text and Appendix F.1 and F.2]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Table of content

A DETAILED RELATED WORK

Causal structure learning has been a hot research topics, Hasan et al. (2023) propose a survey on causal discovery from IID data and time series. For IID data, Some approaches rely on conditional independence to infer causal relationships from observational data. A classic example of such approach is the PC algorithm (Spirtes et al., 2000). In addition to approaches that work with observational data, there are also methods that support interventional data (COmbINE (Triantafillou and Tsamardinos, 2015) and HEJ (Hyttinen et al., 2014)). These methods offer insights into causal relationships based on data acquired through controlled interventions.
A novel line of research introduced by Zheng et al. (2018) has sought to address the combinatorial problem of structure learning by formulating it as a continuous constrained optimization problem. By adopting this approach, they successfully circumvent the need for computationally intensive combinatorial search methods. Similarly, Zhu et al. (2019) leverage the acyclicity constraint in their work but employ reinforcement learning techniques as a search strategy to estimate the DAG. In contrast, Ke et al. (2019) focus on learning a DAG from interventional data through the optimization of an unconstrained objective function. Brouillard et al. (2020) have undertaken a comprehensive investigation into the application of continuous-constrained approaches in the context of interventions, providing a general framework for their utilization. Another notable approach, DiBS by Lorch et al. (2021), aims to infer a full posterior distribution over Bayesian networks given limited available observations. This approach enables the quantification of the uncertainty and the estimation of confidence levels of the structure learning procedure.

The aforementioned state-of-the-art methods have primarily been applied in the context of independent observations over time. Assaad et al. (2022) offer an extensive survey for learning temporal causal relationships. However, when it comes to modeling time-dependent causal relationships, researchers have introduced and utilized Dynamic Bayesian Networks (DBNs). DBNs allow for the modeling of discrete-time temporal dynamics within directed graphical models. In certain approaches, contemporaneous dependencies are disregarded, and the focus is solely on recovering time-lagged relationships. Examples of such approaches include the works of Haufe et al. (2010), Song et al. (2009), and the algorithm tsFCI (Entner and Hoyer, 2010) adapts the Fast Causal Inference (Spirtes et al., 2000) algorithm (developed for the causal analysis of non-temporal variables) to infer causal relationships from time series data. Runge et al. (2019) proposed a two-stage algorithm PCMCI that can scale to large time series. However, these methods primarily emphasize the identification of relationships between variables at different time points, without explicitly considering contemporaneous relationships. Runge (2020) present an extension of PCMCI called PCMCI+ that learns contemporary or instantaneous causal links. Another line of research targets the model with non-Gaussian instantaneous models, Hyvärinen et al. (2010) propose, VARLINGAM, a model that combines the non-Gaussian instantaneous models with autoregressive models and shows that a non-Gaussian model is identifiable without prior knowledge of network structure. Another approach, called Time-series Models with Independent Noise (TiMINo) (Peters et al., 2013) studies a class of restricted structural equation models (SEMs) for time-series data that include nonlinear and instantaneous effects. Recently, a novel study conducted by Pamfil et al. (2020) has emerged, utilizing the algebraic characterization of acyclicity in directed graphs established by Zheng et al. (2018). Their work focuses on the learning of instantaneous and time-lagged graphs within time series data. To achieve this, they have developed a score-based approach for learning DBNs and employed an augmented lagrangien to optimize the resulting program. The resultant method, known as DYNOTEARS, offers the ability to learn causal graph of time dependent variable, without making implicit assumptions about the underlying graph topologies. By leveraging the algebraic characterization of acyclicity, DYNOTEARS enables the estimation of both instantaneous and time-lagged relationships in time series data. Instead of learning a full temporal causal graph, some methods like NBCB (Assaad et al., 2021) or Noise-based/Constraint-based approach learns a summary causal graph from observational time series data without being restricted to the Markov equivalent class even in the case of instantaneous relations.

While DYNOTEARS, Rhino (Gong et al., 2022), PCMCI+, VARLINGAM successfully learn both instantaneous and time-lagged relationships from time series data, it is important to note that the method assumes stationarity and a single regime for the data. However, in numerous real-world scenarios, time series data may exhibit non-stationarity or be composed of multiple regimes, where the causal relationships are different in each regime. This presents a significant challenge for causal discovery.
Some research have aimed to address this challenge by developing methods for causal discovery in heterogeneous

data. An example of such a method is CD-NOD developed by Huang et al. (2020), tackles time series with various regimes. By using the time stamp IDs as a surrogate variable, CD-NOD output one summary causal graph where the parents of each variable are identified as the union of all its parents in graphs from different regimes. Then it detects the change points by using a non stationary driving force that estimates the variability of the conditional distribution $p(x_i|\text{union parents of } x_i)$ over the time index surrogate. While CD-NOD provides a summary graph capturing behavioral changes across regimes, it falls short in inferring individual causal graphs. The overall summary graph does not effectively highlight changes between regimes. Additionally, CD-NOD detects the change points but fails to determine the regime indices, rendering it incapable of inferring the precise number of regimes. In scenarios involving recurring regimes, CD-NOD is unable to detect this crucial information. Another relevant work dealing with MTS composed of multiple regimes is RPCMCI (Saggioro et al., 2020). In this approach, Saggioro et al. (2020) learn a temporal graph for each regime. However, they focus initially on inferring only time-lagged relationships and require prior knowledge of the number of regimes and transitions between them. Balsells-Rodas et al. (2024) addresses first-order regime-dependent causal discovery from MTS with multiple regimes. They proved that first-order Markov switching models with non-linear Gaussian transitions are identifiable up to permutations. Their work offers also a practical algorithms for regime-dependent causal discovery in time series data. However, its primary limitation is the assumption of solely time-lagged relationships, with the theory being restricted to a single time lag.

B CASTOR FRAMEWORK: INTUITION

CASTOR represents a causal discovery framework tailored for Multivariate Time Series (MTS), composed of different regimes. Each regimes can be treated as an independent MTS. Additionally, it is crucial to note that the number of lags L always remains below the minimum length of the regimes ζ.

Figure (5) illustrates a MTS on its left side comprising three variables and two unknown distinct regimes. Each regime possesses its temporal DAG, with one lag attributed to each in this demonstrative scenario.

Upon receiving the MTS as input, CASTOR engages in the process of discerning the number of regimes, determining the indices associated with each regime (indicating their commencement and conclusion), and inferring the temporal DAGs. The resultant DAGs facilitate the straightforward reconstruction of summary graphs encapsulating the entire MTS (CD-NOD output).

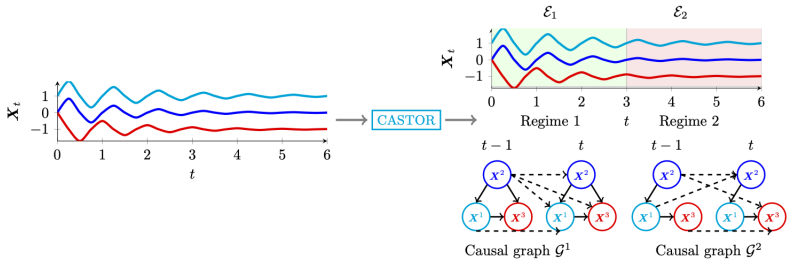


Figure 5: Overview of CASTOR: This illustration demonstrates that CASTOR relies on the MTS to infer the number of regimes (equal to 2 in this figure), the regime partition ($\mathcal{E}_1$ for the first regime and $\mathcal{E}_2$ for the second) and learn the temporal causal graphs ($\mathcal{G}^1$ for the first regime and $\mathcal{G}^2$ for the second). Dashed edges symbolize time-lagged links, while normal arrows represent instantaneous links.

To elucidate the regime learning process, Figure (6) delineates the step-by-step procedure followed by CASTOR in determining the number of regimes and their corresponding indices. The process commences with CASTOR partitioning the MTS into equal windows. In the initial iteration, the length of each regime equals the window size, a user-specified hyperparameter.

Subsequently, CASTOR learns a temporal DAG for each regime. This involves solving an optimization problem, as outlined in Eq (10) for the linear case and Eq (11) for the non-linear scenario. Following graph acquisition, CASTOR updates the regime indices utilizing Eq (7). Notably, CASTOR employs a filtering mechanism to eliminate regimes characterized by an insufficient number of samples. In practical terms, any regime with fewer samples than a defined hyperparameter, denoted as ζ (representing the minimum regime duration), is discarded.

In the event of regime elimination, samples from the discarded regimes are reallocated to the nearest regime in terms of probability. Specifically, if the discarded regime is denoted as u, the sample $\boldsymbol{x}_t$ will be assigned to regime v in the subsequent iteration, where v is the regime with the highest $\gamma_{t,v}$.

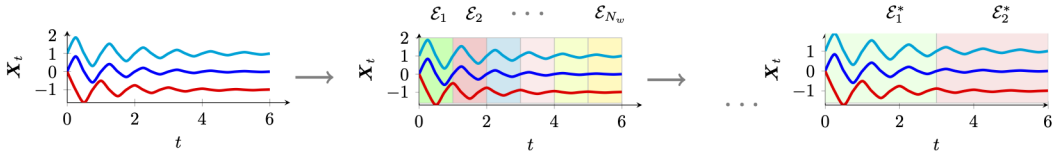


Figure 6: CASTOR initially starts with N_w equal windows, employing the M-step to learn the graph for the initial regimes. The method alternates between updating regime indices in the E-step and inferring temporal causal graphs in the M-step until maximum iterations. Over iterations, some initial regimes disappear, resulting in N_w gradually approaching K (two in this figure).

C EXPECTATION-MAXIMIZATION DERIVATION

In this section, we shall elucidate the computational details surrounding the resolution of our optimization problem. Specifically, we will provide clarity on the various equations introduced in Section 3, namely, Eq (9, 10, 7, 11).

E-step. We model regime participation through a binary latent variable $z_t \in \mathbf{R}^{N_w}$; $\boldsymbol{x}_t$ belongs to regime $u \Rightarrow z_{t,u} = 1$.

$$\begin{aligned}\gamma_{t,u} &= p\left(z_{t,u} = 1 \mid \boldsymbol{x}_t, \boldsymbol{x}_{<t}, \boldsymbol{G}^u_{\{0:L\}}\right) \\ &= \frac{p\left(z_{t,u}=1\right) p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{<t}, z_{t,u}=1, \boldsymbol{G}^u_{\{0:L\}}\right)}{\sum_{j=1}^{N_w} p\left(z_{t,j}=1\right) p\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{<t}, z_{t,j}=1, \boldsymbol{G}^j_{\{0:L\}}\right)} \\ &= \frac{\pi_{t,u}(\alpha) f^u\left(\boldsymbol{x}_t\right)}{\sum_{j=1}^{N_w} \pi_{t,j}(\alpha) f^j\left(\boldsymbol{x}_t\right)}\end{aligned} \tag{12}$$

M-step. Having estimated probabilities $\gamma_{t,u}$ in the E-step, we can now maximise the expected posterior distribution given the MTS $(\boldsymbol{x}_t)_{t\in\mathcal{T}}$ and we have:

$$\begin{aligned}&\sup_{\theta,\alpha} \frac{1}{|\mathcal{T}|} \sum_{u=1}^{N_w} \sum_{t=0}^{|\mathcal{T}|} \gamma_{t,u} \log \pi_{t,u}(\alpha) f^u\left(\boldsymbol{x}_t\right) - \lambda|\mathcal{G}^u|, \text{ s.t } \boldsymbol{G}^u_0 \text{ is a DAG}, \\ \iff &\begin{cases} \max_\alpha \frac{1}{|\mathcal{T}|} \sum_{u=1}^{N_w} \sum_{t=1}^{|\mathcal{T}|} \gamma_{t,u} \ln\left(\pi_{t,u}(\alpha)\right) \\ \sup_\theta \frac{1}{|\mathcal{T}|} \sum_{u=1}^{N_w} \sum_{t\in\mathcal{E}_u} \log f^u\left(\boldsymbol{x}_t\right) - \lambda|\mathcal{G}^u|, \text{ s.t } \boldsymbol{G}^u_0 \text{ is a DAG}, \end{cases}\end{aligned} \tag{13}$$

We know $f^u\left(\boldsymbol{x}_t\right) = \mathcal{N}\left(\boldsymbol{x}_t \boldsymbol{G}^u_0 + \sum_{\tau=1}^{L} \boldsymbol{x}_{t-\tau} \boldsymbol{G}^u_\tau, I\right)$, hence:

$$\begin{aligned}&\iff \sup_\theta \frac{1}{|\mathcal{T}|} \sum_{u=1}^{N_w} \sum_{t\in\mathcal{E}_u} \log f^u\left(\boldsymbol{x}_t\right) - \lambda|\mathcal{G}^u|, \text{ s.t } \boldsymbol{G}^u_0 \text{ is a DAG}, \\ &\iff \min_\theta \frac{1}{|\mathcal{T}|} \sum_{u=1}^{N_w} \sum_{t\in\mathcal{E}_u} \left\| \boldsymbol{x}_t - \left(\boldsymbol{x}_t \boldsymbol{G}^u_0 + \sum_{\tau=1}^{L} \boldsymbol{x}_{t-\tau} \boldsymbol{G}^u_\tau\right)\right\|^2_F + \lambda|\mathcal{G}^u|, \text{ s.t } \boldsymbol{G}^u_0 \text{ is a DAG} \\ &\iff \min_\theta \frac{1}{|\mathcal{T}|} \sum_{u=1}^{N_w} \sum_{t=1}^{|\mathcal{T}|} \gamma_{t,u} \left\| \boldsymbol{x}_t - \left(\boldsymbol{x}_t \boldsymbol{G}^u_0 + \sum_{\tau=1}^{L} \boldsymbol{x}_{t-\tau} \boldsymbol{G}^u_\tau\right)\right\|^2_F + \lambda|\mathcal{G}^u|, \text{ s.t } \boldsymbol{G}^u_0 \text{ is a DAG} \\ &\iff \min_\theta \frac{1}{|\mathcal{T}|} \sum_{u=1}^{N_w} \sum_{t=1}^{|\mathcal{T}|} \gamma_{t,u} \left\| \boldsymbol{x}_t - \left(\boldsymbol{x}_t \boldsymbol{G}^u_0 + \sum_{\tau=1}^{L} \boldsymbol{x}_{t-\tau} \boldsymbol{G}^u_\tau\right)\right\|^2_F + \lambda|\mathcal{G}^u| + \frac{\rho}{2} h\left(\boldsymbol{G}^u_0\right)^2 + \alpha h\left(\boldsymbol{G}^u_0\right),\end{aligned} \tag{14}$$

The only difference between the linear and the non linear cases is how we estimate the mean of the normal distribution f^u for every regime u. As we mentioned in section 3.2, we estimate these means using NNs and we have $f^u_i\left(\boldsymbol{x}_t\right) = \mathcal{N}\left(\psi^u_i\left(\phi^u_i(\boldsymbol{x}_t), \phi^{u,\text{lag}}_i(\boldsymbol{x}^{\text{lag}}_t)\right), 1\right)$, Hence, our M-step for non-linear CASTOR:

$$\min_{\theta,\mathcal{G}} \frac{1}{|\mathcal{T}|} \sum_{u=1}^{N_w} \sum_{t=1}^{|\mathcal{T}|} \sum_{i=1}^{d} \gamma_{t,u} \mathcal{L}(x^i_t, \psi^u_i\left(\phi^u_i(\boldsymbol{x}_t), \phi^{u,\text{lag}}_i(\boldsymbol{x}^{\text{lag}}_t)\right)) + \lambda|\mathcal{G}^u| + \frac{\rho}{2} h\left(\boldsymbol{G}^u_0\right)^2 + \alpha h\left(\boldsymbol{G}^u_0\right) \tag{15}$$

D COMPLEXITY, CONVERGENCE DISCUSSION AND LIMITATIONS

Convergence discussion. We provided theoretical results on identifiability in Section 4, which is a key statistical property to ensure that the causal discovery problem is well-defined, as is standard in most causal discovery papers (Brouillard et al., 2020; Gong et al., 2022; Pamfil et al., 2020; Balsells-Rodas et al., 2024). Although obtaining convergence rates or finite-data bounds is undoubtedly interesting, it is extremely challenging due to the non-convexity of the acyclicity constraint in an EM procedure. In fact, we are not aware of any results applicable to the case. However, we empirically demonstrate that CASTOR converges in both linear

and non-linear cases, correctly identifying the number of regimes and their indices, while also learning the corresponding DAGs.

Complexity. We know that the time complexity of one regime is $\mathcal{O}\left(d^3\right)$, where d is the number of nodes, because of the computation of the acyclicity constraints. The complexity of CASTOR per iteration is $\mathcal{O}\left(|\mathcal{T}|N_w d^3\right)$ where $|\mathcal{T}|$ the number of samples and N_w is the number of regimes at each iteration.

Limitations. A fundamental limitation of this work is the assumption that Gaussian additive noise with equal variance. Many real world scenarios have either a non Gaussian noise and different variance for different MTS components. Future research could address this by extending the analysis to non-Gaussian noise models. Another valuable direction would be to conduct a statistical analysis to understand how the number of samples affects the convergence of CASTOR to the ground truth regimes. Additionally, extending the framework to account for the presence of confounders represents an important avenue for future work.

E FURTHER EXPERIMENTAL RESULTS

E.1 Synthetic data

We employ the Erdos–Rényi (ER) (Newman, 2018) model with mean degrees of 1 or 2 to generate lagged graphs, and the Barabasi–Albert (BA) (Barabási and Albert, 1999) model with mean degrees 4 for instantaneous graphs. The maximum number of lags, L, is set at 1 or 2. We experiment with varying numbers of nodes $\{5, 10, 20, 40\}$ and different numbers of regimes $\{2, 3, 4, 5\}$, each representing diverse causal graphs. The length of each regime is randomly sampled from the set $\{300, 400, 500, 600\}$.

- **Linear case.** Data is generated as follows:

$$\forall u \in \{1, ..., K\}, \forall t \in \mathcal{E}_u : \boldsymbol{x}_t = \boldsymbol{x}_t \boldsymbol{G}_0^u + \sum_{\tau=1}^{L} \boldsymbol{x}_{t-\tau} \boldsymbol{G}_\tau^u + \boldsymbol{\epsilon}_t,$$

  with $\boldsymbol{G}_0^u$ is adjacency matrix of the generated graph by BA model, $\forall \tau \in \{1, .., L\} : \boldsymbol{G}_\tau^u$ are the adjacency of the time lagged graphs generated by ER and $\epsilon_t \sim \mathcal{N}(0, I)$, follows to a normal distribution.

- **Non-linear case.** The formulation used to generated the data is:

$$\forall u \in \{1, ..., K\}, \forall t \in \mathcal{E}_u : x_t^i = g_i^u \left(\mathbf{Pa}^i_{\mathcal{G}^u}(< t), \mathbf{Pa}^i_{\mathcal{G}^u}(t)\right) + \epsilon_t^i,$$

  where g_i^u is a general differentiable linear/non-linear function and $\epsilon_t^i \sim \mathcal{N}(0, 1)$, follows a normal distribution. The function g_i^u is a random combination between a linear transformation and a randomly chosen function from the set: {`Tanh, LeakyReLU, ReLU`}.

E.2 Web activity data

The web activity dataset https://github.com/ckassaad/causal_discovery_for_time_series has the following variables: NetIn that represents the data received by the network interface card in Kbytes/second; NPH represents the number of HTTP processes; NPP represents the number of PHP processes; NCM represents the number of open MySql connections which are started by PHP processes, CpuH represents the percentage of CPU used by all HTTP processes; CpuP represents the percentage of CPU used by all PHP processes; CpuG represents the percentage of global CPU usage.

This dataset is challenging because it is collected from multiple sources resulting in a misaligned time series. Also the presence of partially sleeping time series (NetIn and NPP) due to inactivity of certain servers. The low sampling rate and also the presence of missing values (CpuG) can complicate the inference of causal relationships. Also, the experts in the field of IT systems are uncertain whether IT data, such as web activity data, satisfy the causal sufficiency assumption.

(Aït-Bachir et al., 2023) presented a study about the challenges presented by IT monitoring time series, their results shows that the causal discovery method suffers in these scenarios due to aforementioned challenges.

E.3 Baselines

All used benchmarks for the synthetic experiments are run by using publicly available libraries: VARLINGAM (Hyvärinen et al., 2010) is implemenented in the `lingam`[1] python package. PCMCI+ (Runge, 2020) and RPCMCI (Saggioro et al., 2020) are implemented in `Tigramite`[2] and DYNOTEARS (Pamfil et al., 2020) on `causalnex`[3] package. For Rhino we use the publicly available GitHub shared by the authors[4]. We fine tuned the parameters to achieve the optimal graph for each model.
For CASTOR, an edge threshold of 0.4 is selected. In the linear scenario, we establish $\zeta = 100$ as the minimum regime duration, while in the non-linear context, ζ is set at 200. To demonstrate the model's robustness to the choice of the window size, we train CASTOR using diverse window sizes, specifically $w = 200$ or $w = 300$. For the sparsity coefficient, we use $\lambda = 0.05$. In order to optimise our M-step, we use L-BFGS-B algorithm (Zhu et al., 1997).

E.4 Ablation studies

Table 3: CASTOR ablation study (Linear case) with fixed window size $w = 300$, fixed number of nodes $d = 10$ and regime durations randomly sampled from {400, 500, 600}. We report running time and iterations per regime and per minimum regime duration ζ. CASTOR consistently achieves a 100% F1 score in graph learning and 100% regime accuracy.

	$K = 2$		$K = 3$		$K = 4$	
ζ	Running time (R.T)	Iter.	R.T	Iter.	R.T	Iter.
100	0 min 42s	3	1 min 55s	4	2 min 32s	3
200	0 min 38s	3	1 min 50s	4	2 min 22s	3
300	0 min 37s	3	2 min 10s	4	2 min 27s	3

Table 4: CASTOR ablation study (Linear case) with fixed minimum regime duration $\zeta = 90$, fixed number of nodes $d = 10$ and regime durations randomly sampled from {400, 500, 600}. We report running time and iterations per regime and per window size w. CASTOR consistently achieves a 100% F1 score in graph learning and 100% regime accuracy.

	$K = 2$		$K = 3$		$K = 4$	
w	R.T	Iter.	Running time	Iter.	R.T	Iter.
100	1 min 31s	8	6 min 18s	8	7 min 57s	8
150	0 min 43s	5	6 min 04s	10	5 min 40s	5
200	0 min 45s	5	2 min 17s	5	5 min 35s	5
250	0 min 40s	5	1 min 59s	4	2 min 4s	4

CASTOR demonstrates robustness in handling both linear, Table 3, and nonlinear causal relationships, Ta ble6, regardless of the choice of minimum regime duration or window size. The primary impact of these parameters is on the number of iterations and the overall running time.

[1] https://lingam.readthedocs.io/en/latest/
[2] https://jakobrunge.github.io/tigramite/
[3] https://causalnex.readthedocs.io/en/latest/
[4] https://github.com/microsoft/causica/tree/v0.0.0

Table 5: CASTOR ablation study (Non-linear case) with fixed window size $w = 300$, fixed number of nodes $d = 10$ and regime durations randomly sampled from {400, 500, 600}. We report running time and iterations per regime and per minimum regime duration ζ.

	$K = 2$					$K = 3$				
ζ	R.T	Iter.	F1 Inst.	F1 Lag	Reg Acc.	R.T	Iter.	F1 Inst.	F1 Lag	Reg Acc.
100	2'45s	5	92.5	73.5	95.1	4'27s	6	93.9	87.1	92.0
200	1'50s	4	92.5	73.5	96.2	2'44s	5	95.2	84.2	92.3
300	1'31s	4	92.5	73.5	96.2	2'48s	5	93.9	87.1	91.7

Table 6: CASTOR ablation study (Non-linear case) with fixed minimum regime duration $\zeta = 90$, fixed number of nodes $d = 10$ and regime durations randomly sampled from {400, 500, 600}. We report running time, iterations, F1 score for instantaneous and time lagged links and regime accuracy per regime and per minimum regime duration ζ.

	$K = 2$					$K = 3$				
w	R.T	Iter.	F1 Inst.	F1 Lag	Reg Acc.	R.T	Iter.	F1 Inst.	F1 Lag	Reg Acc.
100	11'30s	10	89.7	100.	97.4	11'50s	10	94.1	81.1	93.1
200	8'50s	10	85.1	90.1	96.7	5'58s	6	92.6	79.1	93.2
250	4'46s	7	89.7	100.	97.2	3'24s	4	92.1	78.0	92.1

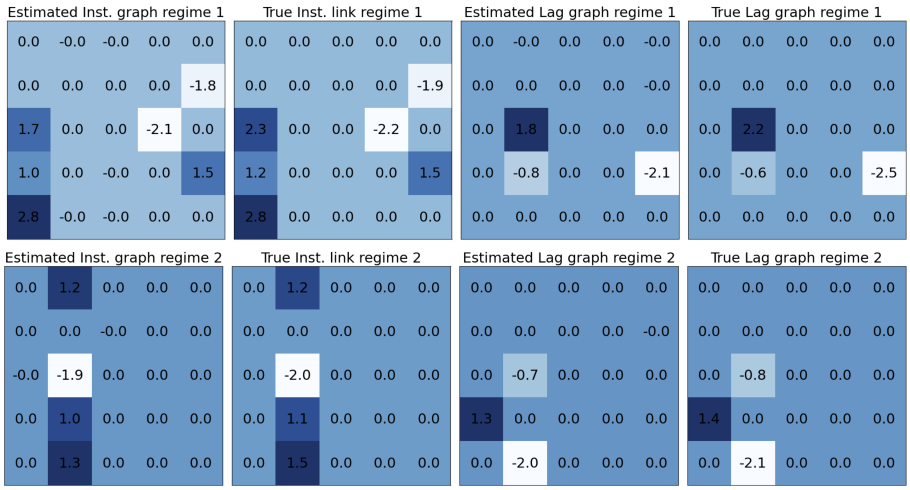


Figure 7: Illustration of CASTOR's estimated graphs compared to ground truth ones, for data generated with weighted DAGs.

Figure 3 illustrates that CASTOR effectively handles weighted adjacency matrices, accurately estimating both the edge weights and the presence of edges.

E.5 Table 1 with standard deviation and $K = 2$

Table 7: Average F1 Scores on regimes for different models and settings for linear causal relationships. Here, d denotes the number of nodes, K the number of regimes, "Split" specifies whether the regime split is automatic (A) or manual (M), and "Type" categorizes the graph as either a window (W) or summary (S). "Inst." refers to instantaneous links and "Lag" to time-lagged edges.

			$d = 10$						$d = 40$					
			$K = 2$		$K = 3$		$K = 4$		$K = 2$		$K = 3$		$K = 4$	
Model	Split	Type	Inst.	Lag	Inst.	Lag	Inst.	Lag	Inst.	Lag	Inst.	Lag	Inst.	Lag
VARLINGAM	M	W	$18.2_{\pm 3.8}$	$10.4_{\pm 7.8}$	$11.0_{\pm 6.4}$	$5.01_{\pm 1.4}$	$9.70_{\pm 1.8}$	$5.10_{\pm 3.1}$	$8.40_{\pm 1.2}$	$1.2_{\pm 0.1}$	$9.83_{\pm 0.9}$	$1.13_{\pm 0.8}$	$10.9_{\pm 3.1}$	$1.43_{\pm 0.6}$
Rhino	M	W	$62.4_{\pm 10.}$	$40.7_{\pm 4.4}$	$54.8_{\pm 3.9}$	$44.4_{\pm 5.0}$	$50.7_{\pm 2.5}$	$43.0_{\pm 6.6}$	$0.00_{\pm 0.0}$	$21.1_{\pm 2.1}$	$0.00_{\pm 0.0}$	$20.8_{\pm 1.3}$	$0.00_{\pm 0.0}$	
Rhino w/o hist	M	W	$90.2_{\pm 3.4}$	$58.5_{\pm 2.3}$	$87.5_{\pm 3.6}$	$53.5_{\pm 2.2}$	$88.2_{\pm 1.7}$	$61.8_{\pm 2.9}$	$0.00_{\pm 0.0}$	$44.1_{\pm 1.2}$	$0.00_{\pm 0.0}$	$38.4_{\pm 3.6}$	$0.00_{\pm 0.0}$	$39.1_{\pm 1.3}$
PCMCI+	M	W	$81.1_{\pm 7.4}$	$89.5_{\pm 3.8}$	$80.4_{\pm 4.5}$	$83.6_{\pm 3.5}$	$77.7_{\pm 1.8}$	$77.6_{\pm 3.2}$	$55.2_{\pm 3.6}$	$85.4_{\pm 4.1}$	$54.1_{\pm 5.4}$	$84.6_{\pm 2.6}$	$53.7_{\pm 3.2}$	$86.1_{\pm 1.8}$
DYNOTEARS	M	W	$\underline{100._{\pm 0.0}}$	$\underline{100._{\pm 0.0}}$	$\underline{96.9_{\pm 2.5}}$	$\underline{100._{\pm 0.0}}$	$\mathbf{99.6}_{\pm 0.9}$	$\mathbf{99.8}_{\pm 1.4}$	$\mathbf{100.}_{\pm 0.0}$	$\underline{100._{\pm 0.0}}$	$\underline{97.4_{\pm 1.7}}$	$\underline{98.8_{\pm 0.2}}$	$\underline{97.3_{\pm 0.8}}$	$\underline{97.9_{\pm 0.6}}$
RPCMCI	A	W	-	$42.3_{\pm 11.}$	-	$18.8_{\pm 2.5}$	-	-	-	$42.1_{\pm 3.5}$	-	$18.4_{\pm 14.}$	-	-
CASTOR	A	W	$\mathbf{100.}_{\pm 0.0}$	$\mathbf{100.}_{\pm 0.0}$	$\mathbf{97.3}_{\pm 2.5}$	$\mathbf{100.}_{\pm 0.0}$	$\mathbf{99.3}_{\pm 0.9}$	$\underline{98.0_{\pm 1.4}}$	$\underline{98.3_{\pm 1.7}}$	$\mathbf{100.}_{\pm 0.0}$	$\mathbf{98.2}_{\pm 1.2}$	$\mathbf{99.8}_{\pm 0.2}$	$\mathbf{98.3}_{\pm 0.4}$	$\mathbf{98.9}_{\pm 0.9}$
CD-NOD	A	S	20.2		11.4		38.8		0		11.3		5.57	
CASTOR	A	S	**100**		**100**		**97.9**		**100**		**99.8**		**99.2**	

Table 7 presents the complete results for linear case. CASTOR and DYNOTEARS outperform all the baselines, either those designed for MTS with multiple regimes, such as RPCMCI and CD-NOD or the other methods that assumes stationarity (Table 1). We emphasize that CASTOR performs similarly to DYNOTEARS, even though the latter benefits from prior access to ground truth regime partitions by being trained separately on each pure regime.

E.6 Violations

In the main text, we introduced our method for multivariate time series that exhibit multiple regimes, each characterized by a distinct causal graph. To challenge this assumption, we created two scenarios:

- **Scenario 1,** *Causal Graph fixed, only the nonlinear mixing function changes*: CASTOR showed the same causal graph learning performance and also robustness to such scenarios. However, when the number of regime increases, CASTOR regime accuracy drops in this situation. But we want to clarify that this is a violation of the setting that CASTOR is designed to tackle.

- **Scenario 2,** *Causal graph fixed , changes in nonlinear function and the noise distribution (case of soft intervention)*: CASTOR showed the same causal graph learning performance and also robustness to such scenarios. In this case, when the number of regimes increases CASTOR succeeded to identify the different regimes, we want to highlight that in this setting we violate our assumption.

Table 8: Comparison of performance metrics for different regime configurations.

	K = 2			**K = 3**		
	F1 Inst	**F1 lag**	**Reg Acc**	**F1 Inst**	**F1 lag**	**Reg Acc**
Scenario 1	80.0	81.1	97.1	73.4	84.6	62.0
Scenario 2	81.8	88.8	98.1	86.5	74.9	89.1

E.7 Further experiments and evaluation using SHD: Linear case

Table 9: We report the average F1 Score on regimes for different models and settings for linear causal relationships. Here, d denotes the number of nodes, K indicates the number of regimes, "Split" specifies whether the algorithm automatically splits the regimes ("A") or if the split was done manually beforehand ("M"), and "Type" categorizes the type of returned graph as either window graph ("W") or summary graph ("S"). Inst. refers to instantaneous links and Lag to time-lagged edges.

			$d = 10$						$d = 40$					
Model	Split	Type	$K = 2$		$K = 3$		$K = 4$		$K = 2$		$K = 3$		$K = 4$	
			Inst.	Lag	Inst.	Lag	Inst.	Lag	Inst.	Lag	Inst.	Lag	Inst.	Lag
Rhino	M	W	22	25	35	43	53	60	136	372	199	539	265	693
Rhino w/o hist	M	W	5	14	9	28	11	32	136	189	208	340	273	420
PCMCI+	M	W	16	9	18	11	24	17	29	8	53	8	72	12
DYNOTEARS	M	W	0	0	3	0	**1**	**1**	**0**	0	11	6	9	4
RPCMCI	A	W	-	38	-	48	-	-	-	155	-	451	-	-
CASTOR	A	W	0	0	**1**	**0**	**1**	2	2	**0**	**9**	**1**	**7**	**2**

The evaluation using the SHD metric demonstrates consistent findings: CASTOR outperforms the baselines (Rhino, PCMCI+) even when we place these baselines in more favorable scenarios by providing them with the ground truth regime partition.
RPCMCI encounters challenges in our settings due to its assumption of only inferring time-lagged relations. This assumption makes our scenario more complex for this algorithm, impacting both regime learning and graph inference tasks.
Although our settings are identifiable, PCMCI+ infers a Markov equivalent class for the instantaneous links, which explains its performance deterioration, particularly in instantaneous relations.
Rhino, which is a state-of-the-art causal discovery method for nonlinear relationships, faces challenges in the absence of historical dependent noises (as confirmed by Figure 4 on page 24 of the Rhino paper). Moreover, Rhino utilizes ConvertibleGNN with Normalizing flows to learn the causal graphs. To train this model, a minimum of 50 time series of length 200 (10000 samples), all sharing the same causal graph. In contrast, our dataset consists of regimes that do not exceed 600 samples. This difference in dataset characteristics poses challenges for Rhino's performance in our scenario.

Table 10: We report the average F1 Score on regimes for different Models and Settings for linear causal relationships: d denotes the number of nodes, K indicates the number of regimes, "Split" specifies whether the algorithm automatically splits the regimes ("A") or if the split was done manually beforehand ("M") for the models. And "Type" categorizes the type of returned graph as either window graph ("W") or summary graph ("S"). Inst. refers to instantaneous links and Lag to time-lagged edges.

			$d = 10$						$d = 40$					
Model	Split	Type	$K = 2$		$K = 3$		$K = 4$		$K = 2$		$K = 3$		$K = 4$	
			Inst.	Lag	Inst.	Lag	Inst.	Lag	Inst.	Lag	Inst.	Lag	Inst.	Lag
VARLINGAM	M	W	$22.9_{\pm 5.6}$	$8.01_{\pm 12}$	$18.4_{\pm 3.6}$	$15.9_{\pm 4.2}$	$24.1_{\pm 6.5}$	$8.20_{\pm 6.5}$	$8.83_{\pm 1.7}$	$2.96_{\pm 0.2}$	$10.3_{\pm 2.4}$	$1.66_{\pm 1.2}$	$14.0_{\pm 2.4}$	$2.30_{\pm 0.7}$
PCMCI+	M	W	$98.5_{\pm 2.1}$	$86.1_{\pm 11.1}$	$99.0_{\pm 1.4}$	$88.6_{\pm 9.4}$	$96.2_{\pm 1.8}$	$85.9_{\pm 6.9}$	$59.2_{\pm 2.9}$	$79.8_{\pm 5.1}$	$61.5_{\pm 1.4}$	$81.9_{\pm 5.1}$	$60.2_{\pm 4.0}$	$81.4_{\pm 3.6}$
RPCMCI	A	W	-	-	-	-	-	-	-	$46.6_{\pm 9.1}$	-	$14.1_{\pm 2.3}$	-	-
CASTOR	A	W	$\mathbf{100}_{\pm 0.0}$	$\mathbf{100}_{\pm 0.0}$	$\mathbf{100}_{\pm 0.0}$	$\mathbf{100}_{\pm 0.0}$	$\mathbf{97.3}_{\pm 1.9}$	$\mathbf{97.2}_{\pm 2.0}$	$\mathbf{97.0}_{\pm 2.7}$	$\mathbf{100}_{\pm 0.0}$	$\mathbf{88.1}_{\pm 6.2}$	$\mathbf{89.9}_{\pm 5.1}$	$\mathbf{98.3}_{\pm 1.4}$	$\mathbf{99.7}_{\pm 0.4}$

We examine varying numbers of nodes, specifically $\{5, 20\}$, and generated time series with different regime counts $\{2, 3, 4\}$. Our model's performance is benchmarked against multiple baselines, namely Rhino (Gong et al., 2022), PCMCI+ (Runge, 2020), RPCMCI (Saggioro et al., 2020) and VARLINGAM (Hyvärinen et al., 2010) and the results are presented in Table 10.

RPCMCI represents the sole baseline tailored to address a similar setting. RPCMCI necessitates prior knowledge of the number of regimes and the maximum number of transitions, and with this input, it only infers time-lagged relations. Even with this detailed information, RPCMCI struggles to achieve convergence, particularly in settings with more than 3 different regimes. The absence of the inference of instantaneous relationships in RPCMCI poses

a challenge for learning regime indices within our setting, as we assume the presence of instantaneous relationships. In contrast, CASTOR does not only surpass RPCMCI in performance but also converges consistently, correctly identifying the number of regimes and recovering both the regime indices and the underlying causal graphs of each regime. We can notice that CASTOR successively infers the regime indices and learns as well the instantaneous links as well as time lagged relations.

When we compare CASTOR and PCMCI+ with VARLINGAM and Rhino. CASTOR and PCMCI+ also demonstrate markedly superior performance (Specially for $d = 5$ for PCMCI+, the model struggles when the number of nodes become greater). To provide context, we manually partition our generated data into K regimes to facilitate the evaluation of VARLINGAM, PCMCI+. We then execute these aforementioned models on each segmented regime, infer the graphs, and compare these composite structures against their true counterparts. Even when executing VARLINGAM, PCMCI+ separately on each regime, CASTOR still outperforms these models without access to any prior information, such as the number of regimes or the indices of the regimes. PCMCI+, while excelling in capturing time-lagged relations, faces challenges with instantaneous links, particularly when dealing with a higher number of nodes. It can only identify the graph up to MECs without explicit functional relations.

E.8 Further experiments and evaluation using SHD: nonlinear case

Table 11: We report the SHD on regimes for different Models and Settings for linear causal relationships: d denotes the number of nodes, K indicates the number of regimes, “Split" specifies whether the algorithm automatically splits the regimes (“A") or if the split was done manually beforehand (“M") for the models. And “Type" categorizes the type of returned graph as either window graph (“W") or summary graph (“S"). Inst. refers to instantaneous links and Lag to time-lagged edges.

			$d = 10$				$d = 40$			
Model	Split	Type	$K = 2$		$K = 4$		$K = 2$		$K = 4$	
			Inst.	Lag	Inst.	Lag	Inst.	Lag	Inst.	Lag
Rhino	M	W	21	29	43	53	71	106	128	207
Rhino w/o hist	M	W	<u>15</u>	10	<u>32</u>	**7**	57	10	123	<u>30</u>
PCMCI+	M	W	19	**6**	36	<u>17</u>	<u>49</u>	**6**	<u>91</u>	**13**
DYNOTEARS	M	W	16	**6**	42	16	57	31	99	47
CASTOR	A	W	**3**	**6**	**27**	22	**18**	<u>7</u>	**58**	33

In the SHD evaluation for non-linear settings, it's evident that CASTOR outperforms all the baselines in predicting instantaneous links. However, when it comes to time-lagged relationships, PCMCI+ excels and outperforms CASTOR. It's important to highlight that PCMCI+ operates in a more favorable scenario, as it is provided with ground truth regime partitions. Nevertheless, CASTOR stands out for its ability to learn the graphs, determine the number of regimes, and identify their respective indices.

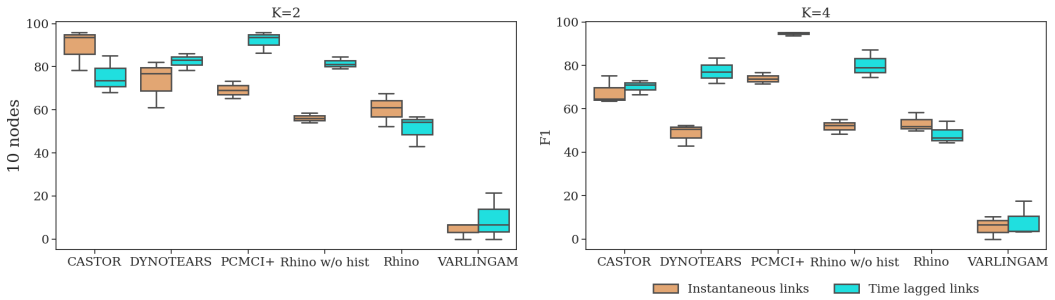


Figure 8: F1 scores by Models and Settings: Orange indicates performance on instantaneous links, and sky-blue signifies performance on time-lagged relationships. Notably, **CASTOR is the only model capable of learning the number of regimes and regime indices; for other baselines, a manual split was performed beforehand.** Number of nodes is 10 nodes

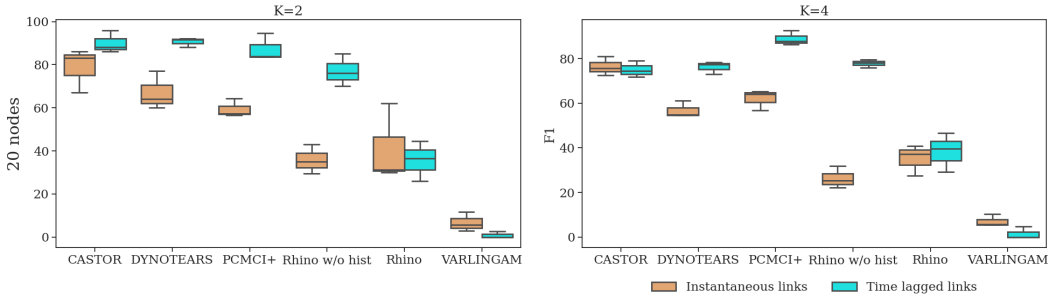


Figure 9: F1 scores by Models and Settings: Orange indicates performance on instantaneous links, and sky-blue signifies performance on time-lagged relationships. Notably, **CASTOR is the only model capable of learning the number of regimes and regime indices; for other baselines, a manual split was performed beforehand.** Number of nodes is 20 nodes

In this section, we present also some additional results using non-linear synthetic data with varying numbers of nodes, specifically $\{10, 20\}$, and diverse numbers of regimes $\{2, 3, 4\}$. In this case, we compare our model against various baseline models, namely DYNOTEARS (Pamfil et al., 2020), Rhino (Gong et al., 2022), PCMCI+ (Runge, 2020) and VARLINGAM (Hyvärinen et al., 2010). For setting with 10 nodes, we can see from Figure (9) that CASTOR, Rhino, PCMCI+ and DYNOTEARS exhibit superior performance to VARLINGAM. It is important to outline that Rhino, PCMCI+, DYNOTEARS and VARLINGAM are each applied to individual regimes separately; neither is designed to learn or infer the number or indices of regimes. PCMCI+ outperforms CASTOR, Rhino and DYNOTEARS in modeling time-lagged relations for non-linear scenarios. PCMCI+ demonstrates robustness in capturing non-linear relationships. Benefiting from a manual split performed beforehand, PCMCI+ exhibits comparable performance (for $K = 2$) or outperforms (for $K = 4$) CASTOR in capturing time-lagged relations. These results are understandable due to the fact that CASTOR has to learn more aspects, including the number of regimes and their indices. For instantaneous links, CASTOR consistently outperforms all models. In the case of $K = 3$ regimes, CASTOR outperforms all models in instantaneous links and achieves comparable results in time-lagged relations as DYNOTEARS and PCMCI+. Notably, Rhino, a state-of-the-art causal discovery model, exhibits lower performance than DYNOTEARS and PCMCI+, **consistent with the findings reported by Rhino authors in their appendix when testing Rhino in settings with non historical dependent noise**. Moreover, Rhino requires a substantial amount of data for training (50 MTS with 200 time steps), while our setting only includes regimes with 600 time steps.

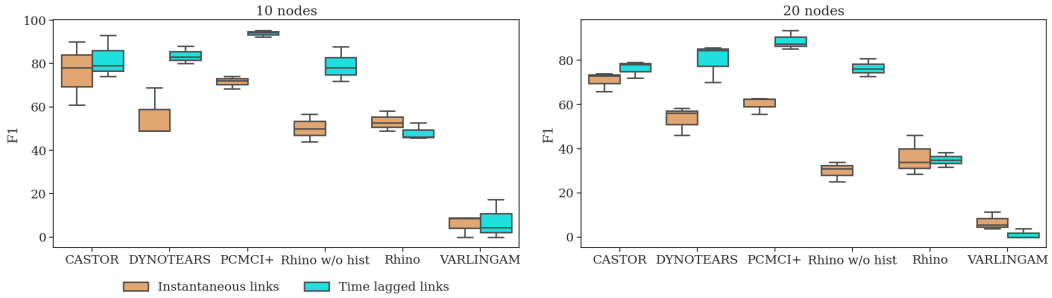


Figure 10: F1 by Models for K=3 Setting: Gray indicates performance on instantaneous links, and pink signifies performance on time-lagged relationships. Notably, **CASTOR is the only model capable of learning the number of regimes and regime indices; for other baselines, a manual split was performed beforehand.**

E.9 Further experiments: Comparison with CD-NOD

We conducted a comparative analysis with CD-NOD, a causal discovery model specifically designed for heterogeneous data and non-stationary time series (Huang et al., 2020). In the context of MTS with multiple regimes, CD-NOD learns a summary causal graph encapsulating the entire MTS.

It's worth noting that in the work by Huang et al. (2020), CD-NOD demonstrates the capability to learn both the change points and the summary causal graph. We use the open-source code of CD-NOD implemented in Matlab by the authors[5].

Our experimental setup involves linear causal relations and diverse configurations, including 2, 3, and 4 regimes, each with varying numbers of nodes (10, 20, 40). For independence test of CD-NOD, we chose Fisher's Z conditional independence test for linear causal relationships and KCI independent test for non linear relations. We systematically compared the performance of CD-NOD against CASTOR, with the evaluation centered on the summary causal graphs as the basis for comparison.

Definition 3. *(Summary causal graph, (Assaad et al., 2022)) Let* $(\boldsymbol{x_t})_{t\in\mathcal{T}}$ *be a MTS and* $\mathcal{G} = (V, E)$ *the associated summary causal graph. The set of vertices in that graph consists of the set of components* $x^1, \ldots, x^d$ *at each time* $t \in \mathbb{N}$*. The edges* E *of the graph are defined as follows: variables* x^p *and* x^q *are connected if and only if there exists some time* t *and some time lag* τ *such that* $x^p_{t-\tau}$ *causes* x^q_t *at time* t *with a time lag of* $0 \leq \tau$ *for* $p \neq q$ *and with a time lag of* $0 < \tau$ *for* $p = q$*.*

d	Method	$K=2$	$K=3$	$K=4$
10	CD-NOD	20.2	11.4	38.8
	CASTOR	**100**	**100**	**97.9**
20	CD-NOD	25.2	23.7	12.7
	CASTOR	**100**	**97.2**	**93.4**
40	CD-NOD	0	11.3	5.57
	CASTOR	**100**	**99.8**	**99.2**

Table 12: F1 Scores by Models and Settings: d indicates number of nodes and K refers to the number of regimes. The comparison is made for linear relations. **CASTOR detects the regimes automatically**.

From Table 12, we can notice that CD-NOD does not manage to outperform CASTOR in various settings (with an F1 score that does not exceed 26%). Additionally, a clear trend emerges where CD-NOD's performance declines when the number of nodes is 40. On the contrary, CASTOR exhibits consistent performance across different settings, achieving a F1 score of over 93% in all scenarios.

[5] https://github.com/Biwei-Huang/Causal-Discovery-from-Nonstationary-Heterogeneous-Data

	$d = 10$	
	$K = 2$	$K = 3$
CD-NOD	28.5	25.3
CASTOR	**86.1**	**85.2**

Table 13: F1 Scores by Models and Settings: d indicates number of nodes and K refers to the number of regimes. The comparison is made for non linear relations. **CASTOR detects the regimes automatically**.

From Table 13, CD-NOD with KCI independent does not manage to outperform CASTOR in non linear causal relationships settings (with an F1 score that does not exceed 28%). This is understandable, because firstly, CD-NOD learns one summary graph. The model assumes smooth changes in graphs between regimes i.e. it expects only a few variables of the graph to be affected by the regime switch (Assumption may not hold true in scenarios such as epileptic seizures or climate science). However, CASTOR did not have this assumption and also learn one causal graph per regime.

E.10 Further experiments: Regime detection experiment

We compare CASTOR to CD-NOD (Huang et al., 2020) and KCP (Arlot et al., 2019) in the task of regime detection. KCP is a multiple change-point detection method designed to handle univariate, multivariate, or complex data. Being non-parametric, KCP does not necessitate knowing the true number of change points in advance. It detects abrupt changes in the complete distribution of the data by employing a characteristic kernel.

As we described in the previous section, CD-NOD is a causal discovery model specifically designed for heterogeneous data and non-stationary time series. In phase III of CD-NOD, a method is proposed to learn change points (that occur when causal relationships changes) for MTS with multiple regimes. CD-NOD estimates a non-stationary driving force for each component (node, considering its parents), where this driving force is a function of the time index. If the driving force changes with the time index, it indicates a change in the regime for that component; otherwise, it signifies that the component is within the same regime.

As previously explained, CD-NOD learns a driving force for components that change behavior. To detect the regimes, one approach is to learn the change points for all components exhibiting changing behavior and then form the union of these change points, yielding the regime partition.

In contrast to CD-NOD, CASTOR directly learns the regime indices. Consequently, for every sample, CASTOR assigns it to a specific regime, directly yielding the regime partition. Additionally, CASTOR conducts regime detection at the graph level rather than the node level which makes it faster than CD-NOD in the task of regime detection.

We opted to perform regime detection in two settings: one with 10 nodes and four different regimes, and another with 20 nodes and four distinct regimes, the causal relationships are non linear in this scenario. For a fair comparison, we chose four regimes without re-occurrence, as KCP and CD-NOD only detect change points and cannot identify the re-occurrence of a specific regime. CASTOR excels in this regard since it detects regime indices. For instance, if regime u occurs from $t = 1$ to $t = 400$ and then reoccurs from $t = 1000$ to $t = 1300$, CASTOR encompasses all these indices in the previously defined regime partition $\mathcal{E}_u$. It utilizes all these indices collectively to learn a more accurate graph.

Regarding the models employed, we use the open-source code of CD-NOD implemented in Matlab by the authors[6]. For KCP, we employ the `Rupture` package[7].

[6] https://github.com/Biwei-Huang/Causal-Discovery-from-Nonstationary-Heterogeneous-Data
[7] https://centre-borelli.github.io/ruptures-docs/

	Regime accuracy $K = 4$ and $d = 10$	Regime accuracy $K = 4$ and $d = 20$
KCP	33.4	66.8
CD-NOD	70.1	88.2
CASTOR	**95.9**	**98.7**

Table 14: Regime detection accuracy by models: CASTOR and CD-NOD outperform the state-of-the-art change-point detection method KCP. CASTOR achieves maximum accuracy, surpassing 95% for both settings.

From the table, it is evident that causal models (CASTOR and CD-NOD) outperform the change-point detection method KCP. This outcome can be attributed to the limitation of KCP in detecting changing points within causal mechanisms that are represented by conditional distributions. CASTOR outperforms CD-NOD in detecting regime indices. This result can be explained by the fact that CASTOR learns regime indices based on graph-level change points, while CD-NOD detects change points at the node level. The node-level approach in CD-NOD may not effectively detect simultaneous changes in behavior for components that actually change behavior simultaneously.
From this analysis, we can conclude that in scenarios involving MTS with multiple regimes and unknown regime indices, CASTOR offers a robust solution. Additionally, employing other methods to split the regimes and learn the causal graph through traditional causal discovery methods may not be an optimal solution:

- We demonstrate that regime indices are not well recoverable by CD-NOD and other state-of-the-art change point detection method KCP. Therefore, employing CD-NOD or KCP to learn the regimes and subsequently using methods like DYNOTEARS, PCMCI, or Rhino to learn the graph may not constitute an optimal solution.
- In cases of regime recurrence, the aforementioned methods are unable to accurately detect the exact number of regimes. Therefore, if a user employs CD-NOD and subsequently uses the regime partitions revealed by CD-NOD as an input to a causal discovery method (such as PCMCI+, DYNOTEARS, Rhino, etc.), the running time will be significantly high.
  Example: To elaborate further, let's consider the epilepsy setting. Imagine we have a recording from an epileptic patient where the sequence involves a non-seizure phase, followed by a seizure phase, and then a reappearance of the non-seizure phase. Employing CD-NOD, particularly with a KCI independence test, to detect regimes in such a scenario can be computationally expensive (For the table above scenario with 20 nodes CD-NOD takes more than 24h compared to CASTOR that learns the regimes and the graph in less than 1h). Subsequently, applying algorithms like Rhino, DYNOTEARS, or Lingam to learn temporal causal graphs would also be resource-intensive. This is primarily because the user would need to run the chosen algorithm at least three times (twice for non-seizure and once for seizure) since CD-NOD does not clarify that a the non-seizure regime is reappearing.

E.11 Models running time

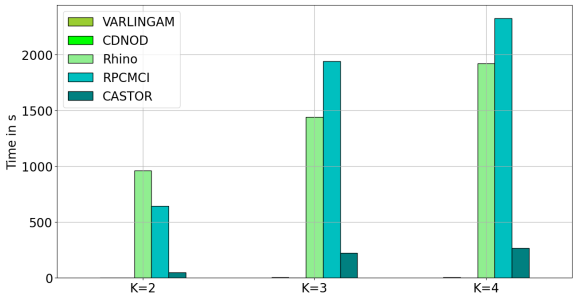


Figure 11: Running time per model, the y axis represents the running time in s and the x axis the number of regime

We compute the running time of every model in different settings, that includes 10 nodes and 2,3 or 4 different regimes, Figure 11 summarizes the results.

VARLINGAM and CD-NOD (employing Fisher's Z conditional independence test for faster runs; note that KCI CD-NOD takes over 2000 seconds for 2 regimes, with F1 scores in a similar range) exhibit remarkable speed compared to other methods. However, in terms of scores, both models encounter challenges in effectively learning causal graphs. Notably, CASTOR runs faster than Rhino, even though CASTOR learns both temporal causal graphs and regime indices.

A fair comparison arises when comparing RPCMCI and CASTOR, as both models learn regime indices and temporal causal graphs. It's essential to mention that RPCMCI necessitates specifying the number of regimes and the maximum number of transitions, producing only time-lagged relations. From Figure 11, it is apparent that CASTOR converges more rapidly than RPCMCI. This difference in convergence time becomes more pronounced as the settings become more complex.

E.12 Biosphere–Atmosphere data

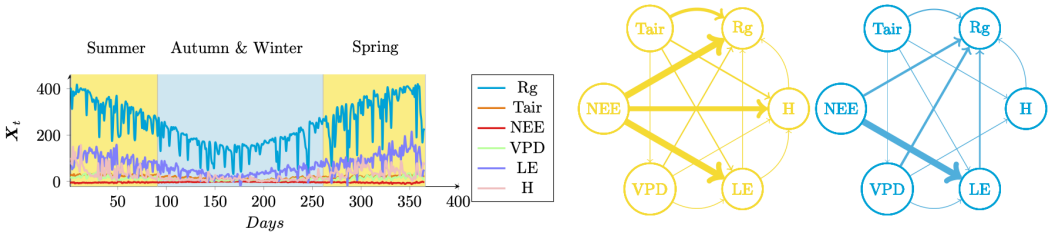


Figure 12: Applying CASTOR to Biosphere atmosphere data, which lacks a ground-truth regime partition or causal graph, aims to emphasize the practical utility of CASTOR. CASTOR identifies two regimes with distinct graphs. Here, only instantaneous links are shown, with the blue graph corresponding to the blue regime and the yellow one to the yellow regime.

We apply CASTOR to Biosphere atmosphere data[8]. The objective is to demonstrate the practical utility of CASTOR in a real-world scenario.

Krich et al. (2021) employ this type of data to learn causal relations between six variables (global radiation (R_g), air temperature (T_{air}), net ecosystem exchange (NEE), vapor pressure deficit (VPD), sensible heat (H), latent heat flux (LE)) under climate change conditions. They utilize the causal discovery model PCMCI+ for windows of three months (with one overlapping month) to learn causal relations. The objective of our experiments is to automate the learning of causal graphs and regime partitions. In other words, instead of assuming that the causal graph changes every three months and applying a causal discovery method at that interval, we provide MTS of one year's length to CASTOR and we let the model (CASTOR) decide when the causal graph changes and which months are more similar. Although we initially set a window length of three months, CASTOR stabilizes after a few iterations in a state with two regimes.
It is essential to note that assuming causal sufficiency in this setting is a difficult assumption, as the biosphere-atmosphere relationship is considerably more complex in reality. Consequently, validating the causal sufficiency assumption is challenging, which explains the appearance of some suspicious relationships. Nevertheless, we will interpret some of the relations and explain why certain causal relations appear and why others are inverted.

In our application, we use CASTOR on this MTS data of six variables, aiming to automatically learn the regime partition and the causal graphs. Initiating with non-overlapping windows of three months (4 initialed different regimes) and a minimum regime duration of two months, CASTOR splits the data into two regimes: the first corresponds to the cold regime grouping Autumn and Winter (From April to September), while the second encompasses Summer and Spring. The accuracy of this partition is 85.4%, with some days in the second regime

[8]FLUXNET Dataset San Luis site, Argentina published by (Garcia et al., 2015)

occasionally misclassified as the first one.

Regarding the causal links, an initial observation is that the strength (represented by thickness) of certain causal relations is greater in the hot regime (yellow) compared to the cold regime (blue). Furthermore, all variables in both regimes are identified as parents of global radiation, which is intuitively incorrect. However, this observation can be explained by our assumption of causal sufficiency, where we assume observation of all causal variables, which is, in reality, not accurate. To explain this result, our model interprets global radiation as the net radiation. Mathematically, we have:

$$R_n = R_g - SW_\uparrow + LW_\downarrow - LW_\uparrow,$$

where R_n is the net radiation, R_g is the global radiation which is also global shortwave radiation, part of the the incoming radiation is reflected at the surface (shortwave upward radiation $SW_\uparrow = \alpha R_g$, $LW_\uparrow$ is the long-wave upward radiation which is the amount of long-wave radiation emitted at the surface and $LW_\uparrow$ is the long-wave downward radiation. Also we have that the net radiation R_n:

$$R_n = H + LE + G,$$

where H is the sensible heat, LE is the latent heat flux and G is the ground heat flux.
In our setting, where we assume causal sufficiency (which is not accurate), CASTOR learns incorrect link directions. Every variable appears to act as a parent of R_g (global radiation), a behavior explained by the two equations above, where CASTOR interprets R_g as the net radiation R_n. Additionally, we observe that temperature influences vapor pressure deficit (VPD), which is understandable. For instance, in hot weather, evaporation can be sustained for a longer period, leading to higher VPD (VPD is the difference between the maximum moisture that could be hold in the air could hold and the actual moisture in the air). Furthermore, we can notice that the temperature T_{air} is a common cause of sensible heat H and latent heat flux LE which is understandable as we know that these two variable can mathematically be described as function temperature. The relationship between NEE and LE is complex and influenced by various factors including plant physiology, climate, soil moisture, and atmospheric conditions. There's no direct relation linking NEE and LE, both variables are influenced by the conductance of stomata which regulates both CO_2 uptake for photosynthesis (affecting NEE) and water vapor release (affecting LE). Thus, having this common confounder variable causes the appearance of suspicious link ($NEE \to LE$). As we mentioned, LE is affected by water vapor release, which mathematically could be written as follows: $LE \propto g_s \cdot VPD$ where (VDP) Vapor pressure deficit and g_s is the conductance of stomata. Hence, this equation explains the link $VPD \to LE$.

The key difference between the two regimes lies in the strength of the causal relationships. This is because the variables take on different values in each regime. This finding demonstrates CASTOR's ability to distinguish between graphs with identical structures but varying edge weights.

F REGIME AND CAUSAL GRAPHS IDENTIFIABILITY

In this section, we concentrate on establishing the identifiability of regimes and causal graphs within the CASTOR framework. Before diving into the details, let us set and clarify the required assumptions.

Definition 4. *(Causal Stationarity Runge (2018)). The time series (that has one regime) process $(\boldsymbol{x}_t)_{t\in\mathcal{T}}$ with a graph $\mathcal{G}$ is called causally stationary over a time index set $\mathcal{T}$ if and only if for all links $x^i_{t-\tau} \to x^j_t$ in the graph*

$$x^i_{t-\tau} \not\perp\!\!\!\perp x^j_t \mid \boldsymbol{x}_{<t} \setminus \left\{x^i_{t-\tau}\right\} \text{ holds for all } t \in \mathcal{T}$$

This elucidates the inherent characteristics of the time-series data generation mechanism, thereby validating the choice of the auto-regressive model. In our setting, we generalize Causal Stationarity as follows:

Assumption 5. *(Causal Stationarity for time series with multiple regimes). The time series process $(\boldsymbol{x}_t)_{t\in\mathcal{T}}$ comprise multiple regimes K, where K is the number of regime, we note $\mathcal{E}_u = \{t|\gamma_{t,u} = 1\}$ the set of time indices where the regime u is active, and $\mathcal{T} = \cup_u \mathcal{E}_u$. $(\boldsymbol{x}_t)_{t\in\mathcal{T}}$ with a graph $\{\mathcal{G}^u\}_{u\in\{1,\dots,K\}}$ is called causally stationary over a time index set $\mathcal{T}$ if and only if for all $u \in \{1, ..., K\}$, $(\boldsymbol{x}_t)_{t\in\mathcal{E}_u}$ is causal stationary with graph $\mathcal{G}^u$ for time index set $\mathcal{E}_u$.*

Definition 5. *(Causal Markov Property, Peters et al. (2017)). Given a DAG $\mathcal{G}$ and a joint distribution p, this distribution is said to satisfy causal Markov property w.r.t. the DAG $\mathcal{G}$ if each variable is independent of its non-descendants given its parents.*

This is a common assumptions for the distribution induced by an SEM. With this assumption, one can deduce conditional independence between variables from the graph.

Assumption 6. *(Causal Markov Property for multiple regimes). Given a set of DAGs $(\mathcal{G}^u)_{u\in\{1,\dots,K\}}$ and a set of joint distribution $(p(\cdot|\mathcal{G}^u))_{u\in\{1,\dots,K\}}$, we say that this set of distributions satisfies causal Markov property w.r.t. the set of DAGs $(\mathcal{G}^u)_{u\in\{1,\dots,K\}}$ if for every u: $p(\cdot|\mathcal{G}^u)$ satisfy causal Markov property w.r.t the DAG $\mathcal{G}^u$.*

Definition 6. *(Causal Minimality, Gong et al. (2022)). Consider a distribution p and a DAG $\mathcal{G}$, we say this distribution satisfies causal minimality w.r.t. $\mathcal{G}$ if it is Markovian w.r.t. $\mathcal{G}$ but not to any proper subgraph of $\mathcal{G}$.*

Assumption 7. *(Causal Minimality for multiple regimes). Given a set of DAGs $(\mathcal{G}^u)_{u\in\{1,\dots,K\}}$ and a set of joint distribution $(p(\cdot|\mathcal{G}^u))_{u\in\{1,\dots,K\}}$, we say that this set of distributions satisfies causal minimality w.r.t. the set of DAGs $(\mathcal{G}^u)_{u\in\{1,\dots,K\}}$ if for every u: $p(\cdot|\mathcal{G}^u)$ satisfy causal minimality w.r.t the DAG $\mathcal{G}^u$.*

Assumption 8. *(Causal Sufficiency). A set of observed variables V is causally sufficient for a process $\boldsymbol{x}_t$ if and only if in the process every common cause of any two or more variables in $\boldsymbol{V}$ is in $\boldsymbol{V}$ or has the same value for all units in the population.*

This assumption implies there are no latent confounders present in the time-series data. The table 15 illustrates that most assumptions (causal sufficiency, causal Markov, faithfulness/minimality) are commonly shared among various state-of-the-art models in causal discovery.

However, CASTOR, RPCMCI, and CD-NOD relax the assumption of stationarity and instead assume that the MTS (Multivariate Time Series) are composed of different regimes. While CD-NOD predicts only a summary causal graph, CASTOR and RPCMCI predict a window causal graph, which can subsequently be used to reconstruct a summary graph.

One notable difference in assumptions between CASTOR and RPCMCI is that RPCMCI assumes only time-lagged relations, whereas CASTOR incorporates the presence of instantaneous links.

F.1 Proof of theorem 1

Assuming the aforementioned assumptions we want to prove the theorem 1.
We consider $\mathcal{G} = (\mathcal{G}^u)_{u\in\{1,..,N_w\}}$, $\mathcal{E} = \cup_{u=1}^{N_w}\mathcal{E}_u$ where N_w is the number of window, $\mathcal{G}^* = (\mathcal{G}^{*,u})_{u\in\{1,..,K\}}$, K is

	Causal graph	Causal Markov	Causal sufficiency	Faithfulness / Minimality	Linear model	Stationarity per regime
DYNOTEARS	W	✓	✓		✓	×
PCMCI+	W	✓	✓	F	×	×
RPCMCI	W	✓	✓	F	×	✓
Rhino	W	✓	✓	M	×	×
VARLINGAM	W	✓	✓		✓	×
CD-NOD	S	✓	✓	F	×	✓
CASTOR	W	✓	✓	M	✓	✓

Table 15: Summary of the main assumptions of algorithms considered in the paper. For causal graphs, S means that the algorithm provides a summary causal graph and W means that the algorithm provides a window causal graph; F corresponds to faithfulness and M to minimality. An empty cell mean that the information given in the corresponding column was not discussed by the authors of the corresponding algorithm.

the exact number of true regimes and $\mathcal{E}^* = \cup_{u=1}^{K} \mathcal{E}_u^*$. We denote $\mathcal{E}_c \mathcal{E}_\ell^*$ the set of time indices that is shared between regime c of our model estimation and the true regime ℓ and $q_\ell^* := \frac{|\mathcal{E}_\ell^*|}{T}, q_c := \frac{|\mathcal{E}_c|}{T}$, $q_{c\ell} := \frac{|\mathcal{E}_c \mathcal{E}_\ell^*|}{T}$. Our objective is to prove that for any estimation $(\hat{\mathcal{G}}, \hat{\mathcal{E}})$: if $\exists u \in \{1, ..., K\}$ s.t. $\hat{\mathcal{G}}^u$ disagree with $\mathcal{G}^{*,u}$ on instantaneous or/and time lagged link, or any regime $\hat{\mathcal{E}}_u \in \hat{\mathcal{E}}$ is close to none of the true regimes in the sense of Kullback–Leibler divergence: $\mathcal{S}(\mathcal{G}^*, \mathcal{E}^*) > \mathcal{S}(\hat{\mathcal{G}}, \hat{\mathcal{E}})$. We have by Eq (9):

$$\mathcal{S}(\mathcal{G}, \mathcal{E}) := \sup_{\theta, \mathcal{G}} \frac{1}{T} \sum_{u=1}^{N_u} \sum_{t \in \mathcal{E}_u} \log f^u(\boldsymbol{x}_t) - \lambda |\mathcal{G}^u|,$$

where λ is the sparsity penalty coefficient and $f^u(\boldsymbol{x}_t) := \prod_{j=1}^{d} f_i^u \left(\mathbf{Pa}^i_{\mathcal{G}^u}(< t), \mathbf{Pa}^i_{\mathcal{G}^u}(t) \right)$ with $f_i^u \left(\mathbf{Pa}^i_{\mathcal{G}^u}(< t), \mathbf{Pa}^i_{\mathcal{G}^u}(t) \right)$ the function used to describe the distribution family in Eq (5).
We will structure the proof as follows:

- Prove that if the score is optimized, then all the estimated regimes will be pure (have only elements of the same true regime).

- Prove that, when the regimes are pure and $N_w = K$, we have $\mathcal{S}(\mathcal{G}^*, \mathcal{E}^*) > \mathcal{S}(\hat{\mathcal{G}}, \hat{\mathcal{E}})$ for any estimation $\hat{\mathcal{G}}$ where $\exists u \in \{1, ..., K\}$ s.t $\hat{\mathcal{G}}^u$ disagrees with $\mathcal{G}^{*,u}$ on instantaneous or/and time lagged link.

F.1.1 Optimizing the score will lead to pure regimes

We denote by $p^{(u)}$ the distribution $p(.|\mathcal{G}^u)$, ignoring penalty terms, we have:

$$
\begin{aligned}
-\mathcal{S}(\mathcal{G},\mathcal{E}) &= -\sup_{\theta}\sum_{c=1}^{N_w}\sum_{\ell=1}^{K} q_{c\ell}\frac{1}{|\mathcal{E}_e\mathcal{E}^*_\ell|}\sum_{t\in\mathcal{E}_e\mathcal{E}^*_\ell}\left[\log f^c(\boldsymbol{x}_t)\right] \\
&\to -\sup_{\phi}\sum_{c=1}^{N_w}\sum_{\ell=1}^{K} q_{c\ell}\mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}}\left[\log f^c\right] \\
&= -\sup_{\theta}\sum_{c=1}^{N_w}\sum_{\ell=1}^{K} q_{c\ell}\mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}}\left[\sum_{i=1}^{d}\log f_i^c\left(\mathbf{Pa}^i_{\mathcal{G}^c}(<t),\mathbf{Pa}^i_{\mathcal{G}^c}(t)\right)\right] \\
&= -\sup_{\theta}\sum_{c=1}^{N_w}\sum_{\ell=1}^{K}\sum_{i=1}^{d} q_{c\ell} \\
&\mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}}\left[-\log\frac{p\left(x_t^i\mid\left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t),\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)}{f_i^c\left(\mathbf{Pa}^i_{\mathcal{G}^c}(<t),\mathbf{Pa}^i_{\mathcal{G}^c}(t)\right)}+\log p\left(x_t^i\mid\left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t),\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\right] \\
&= -\sup_{\theta}\sum_{c=1}^{N_w}\sum_{\ell=1}^{K}\sum_{i=1}^{d} q_{c\ell} \\
&\mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}}\Big[-\mathrm{D_{KL}}\left(p\left(x_t^i\mid\left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t),\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\|f_i^c\left(\mathbf{Pa}^i_{\mathcal{G}^c}(<t),\mathbf{Pa}^i_{\mathcal{G}^c}(t)\right)\right) \\
&-\mathrm{H}\left(p\left(x_t^i\mid\left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t),\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\right)\Big] \\
&= -\sup_{\theta}\sum_{c=1}^{N_w}\sum_{\ell=1}^{K}\sum_{i=1}^{d} q_{c\ell}\mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}}\left[-\mathrm{D_{KL}}\left(p\left(x_t^i\mid\left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t),\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\|f_i^c\left(\mathbf{Pa}^i_{\mathcal{G}^c}(<t),\mathbf{Pa}^i_{\mathcal{G}^c}(t)\right)\right)\right] \\
&-\sup_{\theta}\sum_{c=1}^{N_c}\sum_{\ell=1}^{K}\sum_{i=1}^{d} q_{c\ell}\mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}}\left[-\mathrm{H}\left(p\left(x_t^i\mid\left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t),\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\right)\right] \\
&= \inf_{\theta}\sum_{c=1}^{N_w}\sum_{\ell=1}^{K}\sum_{j=1}^{d} q_{c\ell}\mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}}\left[\mathrm{D_{KL}}\left(p\left(x_t^i\mid\left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t),\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\|f_i^c\left(\mathbf{Pa}^i_{\mathcal{G}^c}(<t),\mathbf{Pa}^i_{\mathcal{G}^c}(t)\right)\right)\right] \\
&+\sum_{c=1}^{N_w}\sum_{\ell=1}^{K}\sum_{j=1}^{d} q_{c\ell}\mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}}\left[\mathrm{H}\left(p\left(x_t^i\mid\left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t),\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\right)\right] \\
&= \inf_{\theta}\sum_{c=1}^{N_w}\sum_{\ell=1}^{K}\sum_{i=1}^{d} q_{c\ell}\mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}}\left[\mathrm{D_{KL}}\left(p\left(x_t^i\mid\left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t),\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\|f_i^c\left(\mathbf{Pa}^i_{\mathcal{G}^c}(<t),\mathbf{Pa}^i_{\mathcal{G}^c}(t)\right)\right)\right] \\
&+\sum_{\ell=1}^{K}\sum_{i=1}^{d}\left(\sum_{c=1}^{N_c} q_{c\ell}\right)\mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}}\left[\mathrm{H}\left(p\left(x_t^i\mid\left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t),\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\right)\right] \\
&= \inf_{\theta}\sum_{c=1}^{N_w}\sum_{\ell=1}^{K}\sum_{i=1}^{d} q_{c\ell}\mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}}\left[\mathrm{D_{KL}}\left(p\left(x_t^i\mid\left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t),\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\|f_i^c\left(\mathbf{Pa}^i_{\mathcal{G}^c}(<t),\mathbf{Pa}^i_{\mathcal{G}^c}(t)\right)\right)\right] \\
&+\sum_{\ell=1}^{K}\sum_{i=1}^{d} q^*_{\ell}\mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}}\left[\mathrm{H}\left(p\left(x_t^i\mid\left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t),\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\right)\right]
\end{aligned}
\tag{16}
$$

Note that θ could be the parameters of the neural networks used in Eq (11) for non linear causal relationship or $\theta = (\mathcal{G}^u)_{u\in\{1,..,N_w\}}$ for linear case Eq 10.

For the score of ground truth (ignoring penalty terms):

$$
-\mathcal{S}\left(\mathcal{G}^*,\mathcal{E}^*\right)\to 0+\sum_{\ell=1}^{K}\sum_{i=1}^{d} q^*_{\ell}\mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}}\left[\mathrm{H}\left(p\left(x_t^i\mid\left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t),\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\right)\right] \tag{17}
$$

Combining Equation (16) and Equation (17) , we have (considering penalty terms):

$$\begin{aligned}\mathcal{S}\left(\mathcal{G}^*, \mathcal{E}^*\right) - \mathcal{S}(\mathcal{G}, \mathcal{E}) = \inf_{\theta} & \sum_{c=1}^{N_w} \sum_{\ell=1}^{K} \sum_{i=1}^{d} q_{c\ell} \\ & \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(\ell)}} \left[\mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^{*,\ell}}(<t), \mathbf{Pa}^i_{\mathcal{G}^{*,\ell}}(t)\right)\right) \| f_i^c\left(\mathbf{Pa}^i_{\mathcal{G}^c}(<t), \mathbf{Pa}^i_{\mathcal{G}^c}(t)\right)\right)\right] \\ & + \lambda\left(\sum_{c=1}^{N_w} |\mathcal{G}^c| - \sum_{\ell=1}^{K} |\mathcal{G}^{*,\ell}|\right)\end{aligned} \tag{18}$$

The first term in Equation (18) is the score term, others are penalty term.

In the following lines, our goal is to demonstrate that optimizing the score term ensures that all identified regimes will accurately match the real regimes. In other words, each estimated regime will be a true representation of an actual one. Additionally, by shifting samples from less significant regimes (regimes with few samples) to the most similar significant regimes, our variable N_w will eventually stabilize at the value of K. To do this, we will proceed by contradiction:

Suppose the score term in Eq (18) is optimized and there exists a regime e that is **not pure**, i.e., there exist $a, b \in [K]$ with $a \neq b$ but $q_{ea} > 0$ and $q_{eb} > 0$. Since they are different distributions for two different regimes with two different causal graphs, there exists $i \in \{1, ..., d\}$ such that $p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^{*,a}}(<t), \mathbf{Pa}^i_{\mathcal{G}^{*,a}}(t)\right)\right) \neq p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^{*,b}}(<t), \mathbf{Pa}^i_{\mathcal{G}^{*,b}}(t)\right)\right)$. Then the score term in Equation (18) has the following lower bound:

$$\begin{aligned}& \inf_{\theta} \sum_{\ell=1}^{K} \sum_{i=1}^{d} q_{e\ell} \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(\ell)}} \left[\mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^{*,\ell}}(<t), \mathbf{Pa}^i_{\mathcal{G}^{*,\ell}}(t)\right)\right) \| f_i^e\left(\mathbf{Pa}^i_{\mathcal{G}^e}(<t), \mathbf{Pa}^i_{\mathcal{G}^e}(t)\right)\right)\right] \\ \geq & \inf_{\theta} \Big\{ q_{ea} \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(a)}} \left[\mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^{*,a}}(<t), \mathbf{Pa}^i_{\mathcal{G}^{*,a}}(t)\right)\right) \| f_i^e\left(\mathbf{Pa}^i_{\mathcal{G}^e}(<t), \mathbf{Pa}^i_{\mathcal{G}^e}(t)\right)\right)\right] \\ & + q_{eb} \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(b)}} \left[\mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^{*,b}}(<t), \mathbf{Pa}^i_{\mathcal{G}^{*,b}}(t)\right)\right) \| f_i^e\left(\mathbf{Pa}^i_{\mathcal{G}^e}(<t), \mathbf{Pa}^i_{\mathcal{G}^e}(t)\right)\right)\right] \Big\}\end{aligned} \tag{19}$$

As we assumed that the score term in Eq (18) is optimized, it means that:

$$\begin{aligned}0 = & \inf_{\theta} \sum_{c=1}^{N_w} \sum_{\ell=1}^{K} \sum_{i=1}^{d} q_{c\ell} \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(\ell)}} \left[\mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^{*,\ell}}(<t), \mathbf{Pa}^i_{\mathcal{G}^{*,\ell}}(t)\right)\right) \| f_i^c\left(\mathbf{Pa}^i_{\mathcal{G}^c}(<t), \mathbf{Pa}^i_{\mathcal{G}^c}(t)\right)\right)\right] \\ \Rightarrow 0 = & \inf_{\theta} \sum_{\ell=1}^{K} \sum_{i=1}^{d} q_{e\ell} \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(\ell)}} \left[\mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^{*,\ell}}(<t), \mathbf{Pa}^i_{\mathcal{G}^{*,\ell}}(t)\right)\right) \| f_i^e\left(\mathbf{Pa}^i_{\mathcal{G}^e}(<t), \mathbf{Pa}^i_{\mathcal{G}^e}(t)\right)\right)\right] \\ \Rightarrow & \begin{cases} \mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^{*,a}}(<t), \mathbf{Pa}^i_{\mathcal{G}^{*,a}}(t)\right)\right) \| f_i^e\left(\mathbf{Pa}^i_{\mathcal{G}^e}(<t), \mathbf{Pa}^i_{\mathcal{G}^e}(t)\right)\right) = 0 \\ \mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^{*,b}}(<t), \mathbf{Pa}^i_{\mathcal{G}^{*,b}}(t)\right)\right) \| f_i^e\left(\mathbf{Pa}^i_{\mathcal{G}^e}(<t), \mathbf{Pa}^i_{\mathcal{G}^e}(t)\right)\right) = 0 \end{cases} \\ \Rightarrow & \forall i \in \{1, ..., d\} : p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^{*,a}}(<t), \mathbf{Pa}^i_{\mathcal{G}^{*,a}}(t)\right)\right) = p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^{*,b}}(<t), \mathbf{Pa}^i_{\mathcal{G}^{*,b}}(t)\right)\right)\end{aligned} \tag{20}$$

and the last line, Eq (20), is a contradiction because the two distributions represent two different regimes with two different graphs. Hence, if the score term of Eq (18) is optimized all the estimated regimes will be pure.

First case: If we matched the samples of less significant regimes to the wrong regimes, the regime is not pure and then the score term is not optimized (contradiction).

Second case: If we eliminate a lot of regimes such that $N_w \leq K - 1$, at least one of our estimated regimes will not be pure and this contradicts the assumption of optimized score term (same reasoning).
Based on this reasoning, optimizing the score term of Equation (18) will ensure convergence to the true number of regimes and also every regime will be pure.

F.1.2 In case of edge disagreement $\mathcal{S}(\mathcal{G}^*, \mathcal{E}^*) > \mathcal{S}(\hat{\mathcal{G}}, \hat{\mathcal{E}})$

Now we will show that Eq (18)is positive, if $\exists u \in \{1, ..., K\}$ s.t $\hat{\mathcal{G}}^u$ disagrees with $\mathcal{G}^{*,u}$ on instantaneous or/and time lagged link.
To simplify the notation, we denote by $p^{(u)}$ the distribution $p(.|\mathcal{G}^u)$ the optimal distribution that describes the CGM of regime u. We assume that each estimated regime $\hat{\mathcal{E}}_c\,(c \in \{1, \ldots, N_w\}, N_w \geq K)$ contains samples from same true regime. Then Equation (18) has lower bound:

$$\begin{aligned} &\inf_\theta \textstyle\sum_{\ell=1}^K q_\ell^* \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(\ell)}} \mathrm{D}_{\mathrm{KL}}\left(p^{(\ell)} \| f^\ell\right) \\ \geq &\left(\min_\ell q_\ell^*\right) \inf_\theta \textstyle\sum_{\ell=1}^K \mathrm{D}_{\mathrm{KL}}\left(p^{(\ell)} \| f^\ell\right) \end{aligned} \tag{21}$$

Equation (21) is positive if and only if $\eta(\mathcal{G})$ is positive.

$$\eta(\mathcal{G}) := \inf_\theta \sum_{\ell=1}^K \mathrm{D}_{\mathrm{KL}}\left(p^{(\ell)} \| f^\ell\right) \tag{22}$$

Let assume that $\exists r \in\in \{1, ..., K\}$ s.t $\hat{\mathcal{G}}^r$ disagrees with $\mathcal{G}^{*,r}$ on instantaneous or/and time lagged link. We follow the same intuition as Gong et al. (2022); Peters et al. (2013, 2017), we will show that $\mathrm{D}_{\mathrm{KL}}\left(p^{(r)} \| f^r\right)$ is positive in two cases:

- **Disagreement on lagged parents only.** This means that for all $t \in [\mathcal{S}+1, T]$, the instantaneous connections at t for $\hat{\mathcal{G}}^r$ and $\mathcal{G}^{*,r}$ are the same, and $\exists t \in [\mathcal{S}+1, T]$ and $i \in \{1, ..., d\}$ such that $\mathbf{Pa}^{x_t^i}_{\mathcal{G}^{*,r}}(< t) \neq \overline{\mathbf{Pa}}^{x_t^i}_{\hat{\mathcal{G}}^r}(< t)$. We can use a similar argument as the theorem 1 in Peters et al. (2013). Without loss of generality, we assume under $\hat{\mathcal{G}}^r$, we have $x_{t-\tau}^j \to x_t^i$ and there is no connections between them under $\mathcal{G}^{*,r}$. Thus, from Markov conditions, we have

  $$x_t^i \perp\!\!\!\perp x_{t-\tau}^j \mid \mathbf{Pa}^{x_t^i}_{\mathcal{G}^{*,r}}(< t) \cup \mathrm{ND}_t^{x^i} \setminus \left\{x_t^i, x_{t-\tau}^j\right\}$$

  under $\mathcal{G}^{*,r}$, where $\mathrm{ND}_t^{x^i}$ are the non-descendants of node x_t^i at some time t. However, from the causal minimality and Proposition 6.16 in Peters et al. (2017), we have

  $$x_t^i \not\perp\!\!\!\perp x_{t-\tau}^j \mid \overline{\mathbf{Pa}}^{x_t^i}_{\hat{\mathcal{G}}^r}(< t) \cup \mathrm{ND}_t^{x^i} \setminus \left\{x_t^i, x_{t-\tau}^j\right\}$$

  under $\hat{\mathcal{G}}^r$, and we have $\mathrm{D}_{\mathrm{KL}}\left(p^{(r)} \| f^r\right) \neq 0$

- **Disagreement on instantaneous parents.** In this Section we will use two different results one for the linear and the other one for the non linear case.
  - *Linear case.* For this case, we will use Theorem 1 in Peters and Bühlmann (2014). In this theorem, the author confirms that the graph is identifiable for linear models with Gaussian additive noise, if for each $j \in \{1, \ldots, d\}$, the weights of the causal relations $\beta_{jk} \neq 0$ for all $k \in \mathbf{PA}_j^{\mathcal{G}_0}$. For our instantaneous links, we have all the weights of the parents are non null. Hence, the instantaneous links are identifiable. Otherwise if $\mathrm{D}_{\mathrm{KL}}\left(p^{(r)} \| f^r\right) \neq 0$
  - *Non linear case.* Using Corollary 30 from Peters et al. (2014), in which they state that in the case of Gaussian independent noise and non linear mixing functions, the graphs are identifiable. Hence, our instantaneous links are identifiable, otherwise, $\mathrm{D}_{\mathrm{KL}}\left(p^{(r)} \| f^r\right) \neq 0$.

Based on the above reasoning, we can show that if $\exists r \in\in \{1, ..., K\}$ s.t., $\hat{\mathcal{G}}^r$ disagree with $\mathcal{G}^{*,r}$ on instantaneous or/and time lagged links, $\mathrm{D}_{\mathrm{KL}}\left(p^{(r)} \| f^r\right) \neq 0$.
Thus, $\eta(\mathcal{G}) > 0$. Then as we assume in Theorem 1 that λ is sufficient small would implies Equation (20) is positive.
If $|\hat{\mathcal{G}}^r| \geq |\mathcal{G}^{*,r}|$ then clearly Eq 20 is positive. Let $\mathbb{G}^+ := \left\{\hat{\mathcal{G}}^r \in \mathbb{G} || \hat{\mathcal{G}}^r| < |\mathcal{G}^{*,r}|\right\}$. To make sure that we have $\mathcal{S}(\mathcal{G}^*, \mathcal{E}^*) - \mathcal{S}(\mathcal{G}, \mathcal{E}) > 0$ for all $\mathcal{G} \in \mathbb{G}^+$, we need to pick λ sufficiently small. Choosing $0 < \lambda < \min_{\mathcal{G} \in \mathbb{G}^+} \frac{\eta(\mathcal{G})}{\left(\sum_{c=1}^{N_w} |\mathcal{G}^c| - \sum_{\ell=1}^K |\mathcal{G}^{*,\ell}|\right)}$ is sufficient.

F.2 Proof of theorem 2

Our objective is to prove that for any estimations $(\mathcal{G}, \mathcal{E})$ and $(\mathcal{G}', \mathcal{E}')$ such that $(\mathcal{G}, \mathcal{E})$ is closer in terms of Kullback–Leibler to the optimal solution $(\mathcal{G}^*, \mathcal{E}^*)$ than $(\mathcal{G}', \mathcal{E}')$: $\mathcal{S}(\mathcal{G}, \mathcal{E}) > \mathcal{S}(\mathcal{G}', \mathcal{E}')$. Our goal is to demonstrate that, for any estimation $(\mathcal{G}, \mathcal{E})$, one that is closer in terms of Kullback-Leibler (KL) divergence to the optimal solution $(\mathcal{G}^*, \mathcal{E}^*)$ will have a higher score estimated by CASTOR compared to another estimation $(\mathcal{G}', \mathcal{E}')$. To clarify, when we mention "closer to $(\mathcal{G}^*, \mathcal{E}^*)$ in terms of KL," we are referring to the degree of similarity to the optimal solution $(\mathcal{G}^*, \mathcal{E}^*)$. In other terms for every regime ℓ:

$$\begin{aligned} &\mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right) \| f_i^\ell\left(\mathbf{Pa}^i_{\mathcal{G}^\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^\ell}(t)\right)\right) \\ \leq &\mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right) \| f_i^\ell\left(\mathbf{Pa}^i_{\mathcal{G}'^\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}'^\ell}(t)\right)\right) \end{aligned} \tag{23}$$

First case: Let's assume, in this first scenario, that both suboptimal estimations $((\mathcal{G}', \mathcal{E}^*)$ and $(\mathcal{G}, \mathcal{E}^*))$ accurately detect the regimes, with the only distinction lying in the estimation of the graph.

We know that the score function of each estimation after the optimization procedure could be written as the following:

$$\begin{aligned} -\mathcal{S}(\mathcal{G}, \mathcal{E}^*) &= \sum_{\ell=1}^{K}\sum_{i=1}^{d} q_\ell^* \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(\ell)}}\left[\mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right) \| f_i^\ell\left(\mathbf{Pa}^i_{\mathcal{G}^\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^\ell}(t)\right)\right)\right] \\ &+ \sum_{\ell=1}^{K}\sum_{i=1}^{d} q_\ell^* \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(\ell)}}\left[\mathrm{H}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\right)\right] \\ -\mathcal{S}(\mathcal{G}', \mathcal{E}^*) &= \sum_{\ell=1}^{K}\sum_{i=1}^{d} q_\ell^* \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(\ell)}}\left[\mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right) \| f_i^\ell\left(\mathbf{Pa}^i_{\mathcal{G}'^\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}'^\ell}(t)\right)\right)\right] \\ &+ \sum_{\ell=1}^{K}\sum_{i=1}^{d} q_\ell^* \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(\ell)}}\left[\mathrm{H}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\right)\right] \\ \mathcal{S}(\mathcal{G}, \mathcal{E}^*) - \mathcal{S}(\mathcal{G}', \mathcal{E}^*) &= \sum_{\ell=1}^{K}\sum_{i=1}^{d} q_\ell^* \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(\ell)}}[\mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right) \| f_i^\ell\left(\mathbf{Pa}^i_{\mathcal{G}'^\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}'^\ell}(t)\right)\right) \\ &- \mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right) \| f_i^\ell\left(\mathbf{Pa}^i_{\mathcal{G}^\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^\ell}(t)\right)\right)] \\ \mathcal{S}(\mathcal{G}, \mathcal{E}^*) - \mathcal{S}(\mathcal{G}', \mathcal{E}^*) &\geq 0 \end{aligned} \tag{24}$$

In the first case, we demonstrated that if two estimations differ in the graph learning component, the one closer to the optimal solution will have a higher score. The last inequality is correct even when we add the sparsity term because, we can pick λ sufficiently small, to ensure that we have $\mathcal{S}\left(\mathcal{G}, \mathcal{E}^*\right) - \mathcal{S}(\mathcal{G}', \mathcal{E}^*) > 0$.

Second case: Let's assume that both suboptimal estimations $((\mathcal{G}, \mathcal{E}')$ and $(\mathcal{G}, \mathcal{E}))$ learn identical causal graphs for all regimes. However, for the estimation $\mathcal{E}'$, there exists at least one regime a that misclassifies more S samples and incorrectly assigns them to regime b.

We know that the score function of each estimation after the optimization procedure could be written as the

following:

$$
\begin{aligned}
-\mathcal{S}(\mathcal{G},\mathcal{E}) &= \sum_{c=1}^{N_w}\sum_{\ell=1}^{K}\sum_{i=1}^{d} q_{c\ell} \mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}} \left[\mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right) \| f_i^c\left(\mathbf{Pa}^i_{\mathcal{G}^c}(<t), \mathbf{Pa}^i_{\mathcal{G}^c}(t)\right)\right)\right] \\
&\quad + \sum_{\ell=1}^{K}\sum_{i=1}^{d} q^*_{\ell} \mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}} \left[\mathrm{H}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\right)\right] \\
-\mathcal{S}(\mathcal{G},\mathcal{E}') &= \sum_{c=1}^{N_w}\sum_{\ell=1}^{K}\sum_{i=1}^{d} q'_{c\ell} \mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}} \left[\mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right) \| f_i^c\left(\mathbf{Pa}^i_{\mathcal{G}^c}(<t), \mathbf{Pa}^i_{\mathcal{G}^c}(t)\right)\right)\right] \\
&\quad + \sum_{\ell=1}^{K}\sum_{i=1}^{d} q^*_{\ell} \mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}} \left[\mathrm{H}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\right)\right]
\end{aligned}
$$

As previously explained, for the estimation $\mathcal{E}'$, S samples from regime a are misclassified as belonging to regime b. This implies: $q'_{aa} = q_{aa} - \frac{S}{|\mathcal{T}|}$ and $q'_{ba} = q_{ba} + \frac{S}{|\mathcal{T}|}$, and we have the difference between the 2 scores is written as follows:

$$
\begin{aligned}
\mathcal{S}(\mathcal{G},\mathcal{E}) - \mathcal{S}(\mathcal{G},\mathcal{E}') &= \sum_{i=1}^{d} \frac{S}{|\mathcal{T}|} \mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(a)}} [D_{\mathrm{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,a}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,a}(t)\right)\right) \| f_i^b\left(\mathbf{Pa}^i_{\mathcal{G}^b}(<t), \mathbf{Pa}^i_{\mathcal{G}^b}(t)\right)\right) \\
&\quad - D_{\mathrm{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,a}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,a}(t)\right)\right) \| f_i^a\left(\mathbf{Pa}^i_{\mathcal{G}^a}(<t), \mathbf{Pa}^i_{\mathcal{G}^a}(t)\right)\right)] \\
\mathcal{S}(\mathcal{G},\mathcal{E}^*) - \mathcal{S}(\mathcal{G}',\mathcal{E}^*) &\geq 0
\end{aligned}
\tag{25}
$$

The aforementioned equation is positive because, by definition, the distribution estimated by CASTOR for regime a is closer to the true distribution of regime b than the estimation of the distribution for regime b.

Third case: Let's assume that there exists at least one regime a that misclassifies more samples S and incorrectly assigns them to regime b. In these two regimes, the two suboptimal estimations yield different causal graphs. However, as previously described, the regime indices learned by $\mathcal{E}$ are closer to $\mathcal{E}^*$. In other words, $\mathcal{E}'$ misclassifies more samples S from regime a and incorrectly assigns them to regime b. Additionally, the graph inferred by $\mathcal{G}$ is closer to the optimal solution than the graph $\mathcal{G}'$ estimated by the second suboptimal method.

We assume that $q_{ba} \ll q_{aa}$, signifying that the number of well-classified regime samples is greater than the number of misclassified samples. This assumption is reasonable because altering graphs between regimes will cause changes in the mean of our mixtures f^u. Furthermore, given the higher dimension, making slight modifications such as deleting a few edges and adding new ones will result in distinct mixtures. Consequently, the number of samples that could be misclassified will be low.

We know that the score function of each estimation after the optimization procedure could be written as the following:

$$
\begin{aligned}
-\mathcal{S}(\mathcal{G},\mathcal{E}) &= \sum_{c=1}^{N_w}\sum_{\ell=1}^{K}\sum_{i=1}^{d} q_{c\ell} \mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}} \left[\mathrm{D_{KL}}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right) \| f_i^c\left(\mathbf{Pa}^i_{\mathcal{G}^c}(<t), \mathbf{Pa}^i_{\mathcal{G}^c}(t)\right)\right)\right] \\
&\quad + \sum_{\ell=1}^{K}\sum_{i=1}^{d} q^*_{\ell} \mathop{\mathbb{E}}_{\boldsymbol{x}_t\sim p^{(\ell)}} \left[\mathrm{H}\left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\right)\right]
\end{aligned}
\tag{26}
$$

$$
\begin{aligned}
-\mathcal{S}(\mathcal{G}', \mathcal{E}') &= \sum_{c=1}^{N_w} \sum_{\ell=1}^{K} \sum_{i=1}^{d} q'_{c\ell} \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(\ell)}} \left[\mathrm{D_{KL}} \left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right) \| f_i^c \left(\mathbf{Pa}^i_{\mathcal{G}'^c}(<t), \mathbf{Pa}^i_{\mathcal{G}'^c}(t)\right)\right)\right] \\
&+ \sum_{\ell=1}^{K} \sum_{i=1}^{d} q^*_{\ell} \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(\ell)}} \left[\mathrm{H} \left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,\ell}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,\ell}(t)\right)\right)\right)\right] \\
\mathcal{S}(\mathcal{G}, \mathcal{E}) - \mathcal{S}(\mathcal{G}', \mathcal{E}') &= \sum_{i=1}^{d} \frac{S}{|\mathcal{T}|} \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(a)}} [D_{\mathrm{KL}} \left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,a}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,a}(t)\right)\right) \| f_i^b \left(\mathbf{Pa}^i_{\mathcal{G}'^b}(<t), \mathbf{Pa}^i_{\mathcal{G}'^b}(t)\right)\right) \\
&- D_{\mathrm{KL}} \left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,a}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,a}(t)\right)\right) \| f_i^a \left(\mathbf{Pa}^i_{\mathcal{G}'^a}(<t), \mathbf{Pa}^i_{\mathcal{G}'^a}(t)\right)\right)] \\
&+ \sum_{i=1}^{d} q_{aa} \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(a)}} [D_{\mathrm{KL}} \left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,a}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,a}(t)\right)\right) \| f_i^a \left(\mathbf{Pa}^i_{\mathcal{G}'^a}(<t), \mathbf{Pa}^i_{\mathcal{G}'^a}(t)\right)\right) \\
&- D_{\mathrm{KL}} \left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,a}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,a}(t)\right)\right) \| f_i^a \left(\mathbf{Pa}^i_{\mathcal{G}^a}(<t), \mathbf{Pa}^i_{\mathcal{G}^a}(t)\right)\right)] \\
&+ \sum_{i=1}^{d} q_{ba} \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(a)}} [D_{\mathrm{KL}} \left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,a}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,a}(t)\right)\right) \| f_i^b \left(\mathbf{Pa}^i_{\mathcal{G}'^b}(<t), \mathbf{Pa}^i_{\mathcal{G}'^b}(t)\right)\right) \\
&- D_{\mathrm{KL}} \left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,a}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,a}(t)\right)\right) \| f_i^b \left(\mathbf{Pa}^i_{\mathcal{G}^b}(<t), \mathbf{Pa}^i_{\mathcal{G}^b}(t)\right)\right)] \\
\mathcal{S}(\mathcal{G}, \mathcal{E}) - \mathcal{S}(\mathcal{G}', \mathcal{E}') &\approx \sum_{i=1}^{d} \frac{S}{|\mathcal{T}|} \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(a)}} [D_{\mathrm{KL}} \left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,a}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,a}(t)\right)\right) \| f_i^b \left(\mathbf{Pa}^i_{\mathcal{G}'^b}(<t), \mathbf{Pa}^i_{\mathcal{G}'^b}(t)\right)\right) \\
&- D_{\mathrm{KL}} \left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,a}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,a}(t)\right)\right) \| f_i^a \left(\mathbf{Pa}^i_{\mathcal{G}'^a}(<t), \mathbf{Pa}^i_{\mathcal{G}'^a}(t)\right)\right)] \\
&+ \sum_{i=1}^{d} q_{aa} \mathop{\mathbb{E}}_{\boldsymbol{x}_t \sim p^{(a)}} [D_{\mathrm{KL}} \left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,a}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,a}(t)\right)\right) \| f_i^a \left(\mathbf{Pa}^i_{\mathcal{G}'^a}(<t), \mathbf{Pa}^i_{\mathcal{G}'^a}(t)\right)\right) \\
&- D_{\mathrm{KL}} \left(p\left(x_t^i \mid \left(\mathbf{Pa}^i_{\mathcal{G}^*,a}(<t), \mathbf{Pa}^i_{\mathcal{G}^*,a}(t)\right)\right) \| f_i^a \left(\mathbf{Pa}^i_{\mathcal{G}^a}(<t), \mathbf{Pa}^i_{\mathcal{G}^a}(t)\right)\right)] \\
\mathcal{S}(\mathcal{G}, \mathcal{E}^*) - \mathcal{S}(\mathcal{G}', \mathcal{E}^*) &\geq 0
\end{aligned}
\tag{27}
$$

The last inequality is correct for these two last cases, even when we add the sparsity term because, we can pick λ sufficiently small, to ensure that we have $\mathcal{S}\left(\mathcal{G}, \mathcal{E}\right) - \mathcal{S}(\mathcal{G}', \mathcal{E}') > 0$.

G Illustration of CASTOR's estimated graphs

G.1 Illustration of the estimated graphs by CASTOR: Linear case, 5 regimes with $L = 1$

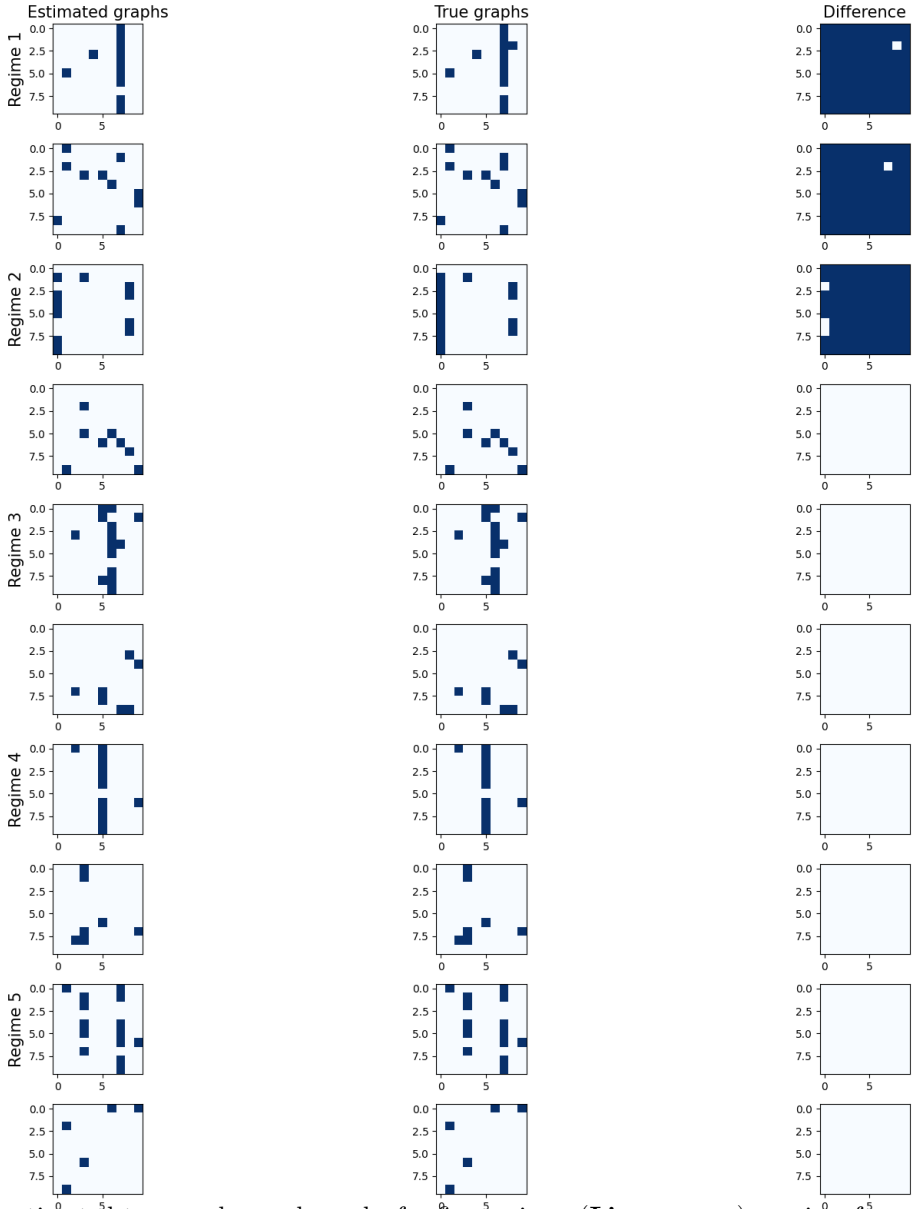


Figure 13: The estimated temporal causal graphs for five regimes (**Linear case**) consist of one matrix of 10 rows and 10 columns representing instantaneous links and another of 10 rows and 10 columns delineating time-lagged relations (with a maximum lag $L = 1$ in this case). Dark blue indicates a value of one (presence of an edge), while sky blue symbolizes a value of 0 (absence of an edge). The second column displays the groundtruth causal graphs, and the final column highlights the difference between the estimated and true graphs.

G.2 Illustration of the estimated graphs by CASTOR: Linear case, 2 regimes with $L = 2$

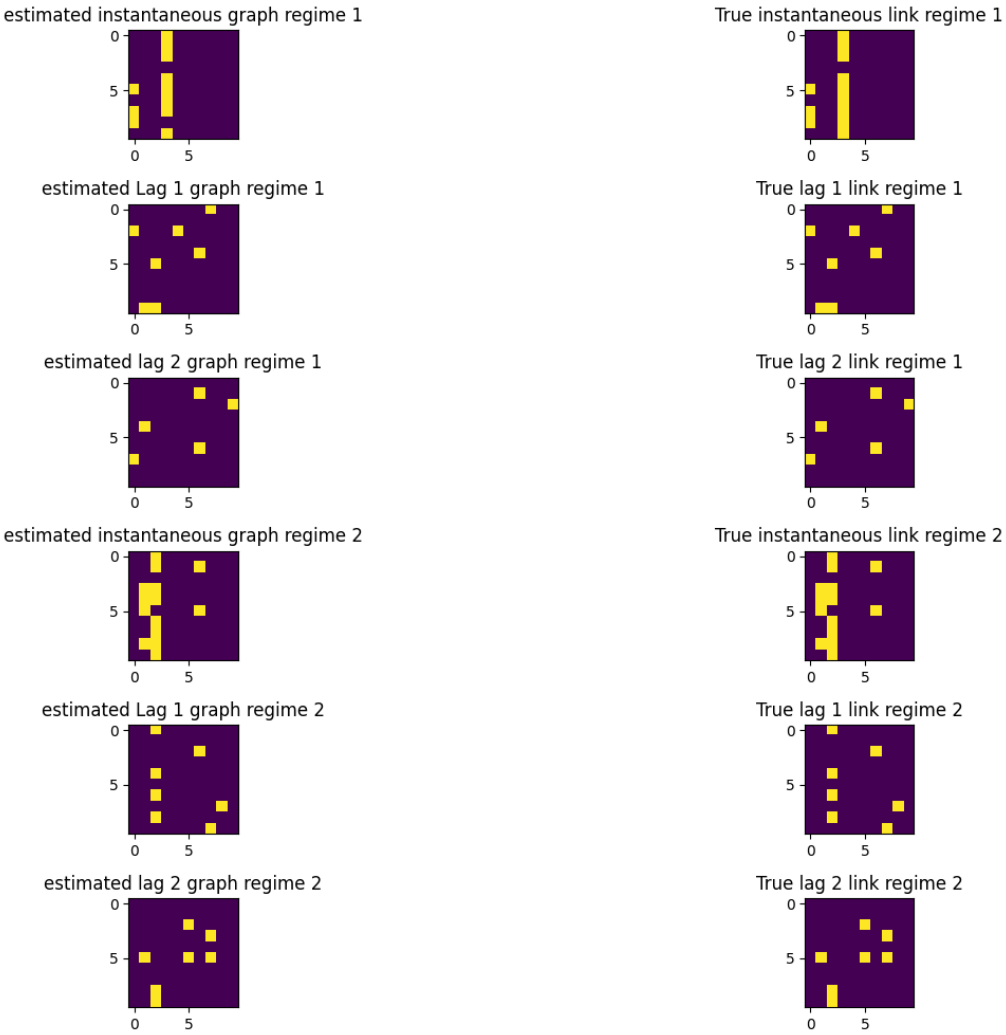


Figure 14: The estimated temporal causal graphs for two regimes **(non Linear case)**, with one matrix representing instantaneous links and another delineating time-lagged relations. The second column showcases the actual causal graphs, while the final column highlights the discrepancies between the estimated and true graphs. Yellow indicates a value of one (presence of an edge), while black symbolizes a value of 0 (absence of an edge).

G.3 Illustration of the estimated graphs by CASTOR: Non-linear case, 3 regimes with $L = 1$

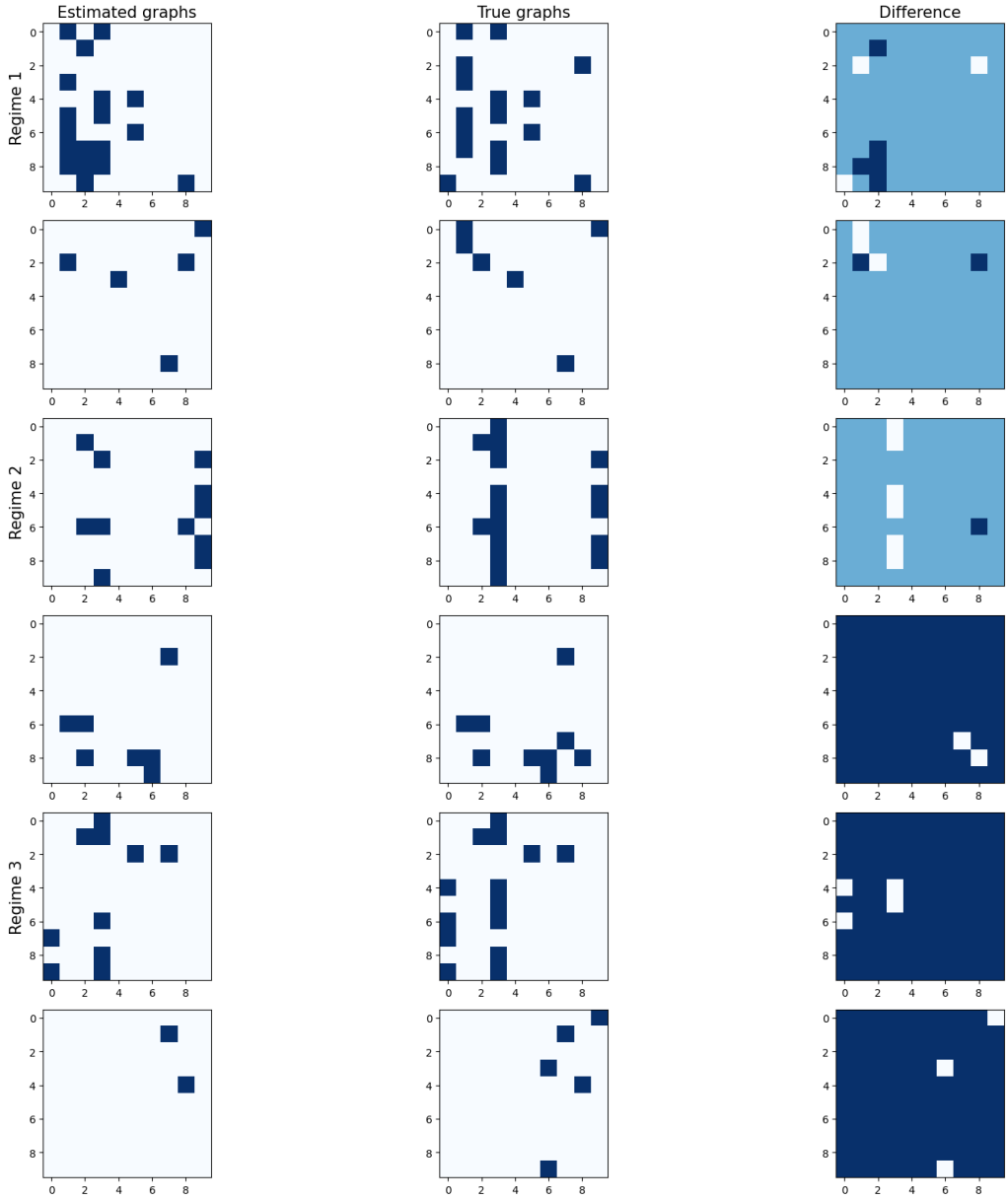


Figure 15: The estimated temporal causal graphs for three regimes (**Non-Linear case**) consist of one matrix of 10 rows and 10 columns representing instantaneous links and another of 10 rows and 10 columns delineating time-lagged relations (with a maximum lag $L = 1$ in this case). Dark blue indicates a value of one (presence of an edge), while sky blue symbolizes a value of 0 (absence of an edge). The second column displays the groundtruth causal graphs, and the final column highlights the difference between the estimated and true graphs.

G.4 Illustration of the estimated graphs by CASTOR: Non-linear case, 5 regimes with $L = 1$

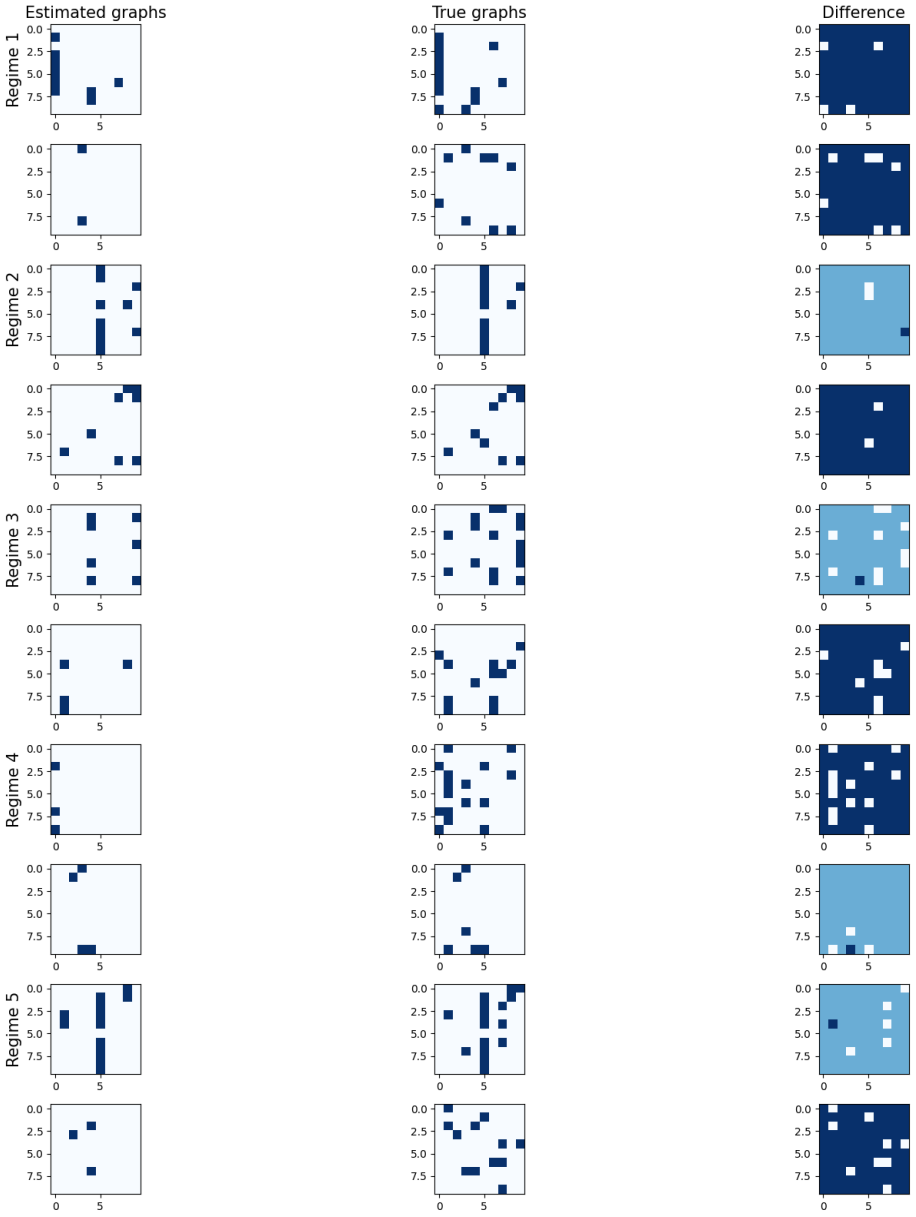


Figure 16: The estimated temporal causal graphs for five regimes (**Non-Linear case**) consist of one matrix of 10 rows and 10 columns representing instantaneous links and another of 10 rows and 10 columns delineating time-lagged relations (with a maximum lag $L = 1$ in this case). Dark blue indicates a value of one (presence of an edge), while sky blue symbolizes a value of 0 (absence of an edge). The second column displays the groundtruth causal graphs, and the final column highlights the difference between the estimated and true graphs.