
Flow based approach for Dynamic Temporal Causal models with non-Gaussian or Heteroscedastic Noises

Abdellah Rahmani
LTS4, EPFL
Lausanne, Switzerland
abdellah.rahmani@epfl.ch

Pascal Frossard
LTS4, EPFL
Lausanne, Switzerland
pascal.frossard@epfl.ch

Abstract

Understanding causal relationships in multivariate time series is crucial in many scenarios, such as those dealing with financial or neurological data. Many such time series exhibit multiple regimes, i.e., consecutive temporal segments with a priori unknown boundaries, with each regime having its own causal structure. Inferring causal dependencies and regime shifts is critical for analyzing the underlying processes. However, causal structure learning in this setting is challenging due to (1) non-stationarity, i.e., each regime can have its own causal graph and mixing function, and (2) complex noise distributions, which may be non-Gaussian or heteroscedastic. Existing causal discovery approaches cannot address these challenges, since generally assume stationarity or Gaussian noise with constant variance. Hence, we introduce FANTOM, a unified framework for causal discovery that handles non-stationary processes along with non-Gaussian and heteroscedastic noises. FANTOM simultaneously infers the number of regimes and their corresponding indices and learns each regime's Directed Acyclic Graph. It uses a Bayesian Expectation Maximization algorithm that maximizes the evidence lower bound of the data log-likelihood. On the theoretical side, we prove, under mild assumptions, that temporal heteroscedastic causal models, introduced in FANTOM's formulation, are identifiable in both stationary and non-stationary settings. In addition, extensive experiments on synthetic and real data show that FANTOM outperforms existing methods.

1 Introduction

Causal structure learning from multivariate time series (MTS) is a fundamental problem with diverse applications in traffic modeling [9], biology [44], climate science [43], or healthcare [47]. However, identifying causal relationships in MTS poses several challenges. First, real-world time series are often non-stationary, exhibiting multiple unknown regimes, each potentially governed by different causal relationships. Examples include changing dependencies across climate conditions [27], financial markets [21], and epileptic seizure stages [56]. Second, many MTS display complex noises, e.g., non-Gaussian or even heteroscedastic noise, whose variance depends on both instantaneous and lagged causes. This occurs in fMRI data [46], EEG measurements [22], or financial data [18].

Recent causal discovery methods capture linear and nonlinear interactions with instantaneous and lagged effects [41, 35, 51]. More recently, Gong et al. [15] and Wang et al. [55] explored structural equation models for a single stationary regime governed by a one causal graph with historically dependent noise, where noise variance depends solely on time-lagged variables, neglecting heteroscedasticity. Existing multi-regime methods include RPCMCI [45], which identifies only linear, time-lagged interactions and requires prior knowledge of regime numbers and transitions; and CD-NOD [21], which handles causal discovery from non-stationary MTS, but is limited to homoscedastic noise, cannot infer individual causal graph for each regime, and is incapable of

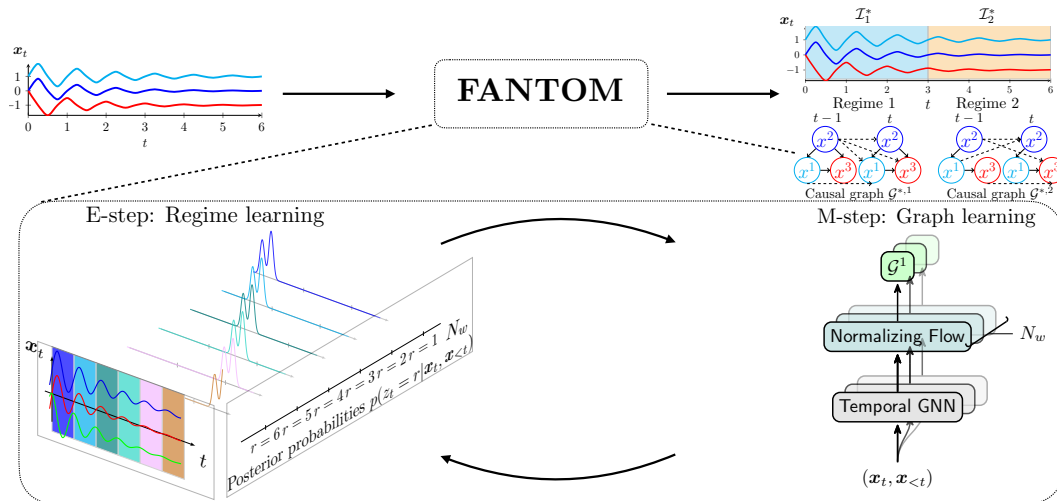


Figure 1: Illustration of FANTOM processing a MTS with two ground truth regimes ($K = 2$). The algorithm recovers the regime indices $\mathcal{T}_1^*$ and $\mathcal{T}_2^*$ and learns a temporal causal graph for each regime (dashed edges represent time-lagged links; solid arrows indicate instantaneous links). In the E-step, posterior probabilities $p(z_t = r \mid \mathbf{x}_t, \mathbf{x}_{<t})$ are estimated, where $z_t = r$ means $\mathbf{x}_t$ belongs to regime r . The M-step then infers causal graphs within each regime. Here, N_w is the number of regimes that converges to $K = 2$.

identifying recurring regimes. CASTOR [39], infers both regime indices and separate causal graphs per regime, accommodating instantaneous and lagged causal relationships, yet it is still restricted to normal noise. Consequently, the previous methods cannot jointly infer the number of regimes, their indices, their causal graphs, nor effectively manage either non-Gaussian or heteroscedastic noises.

To address these limitations, we propose FANTOM, a new framework for Structural Equation Models (SEMs) in multi-regime MTS under either non-Gaussian or heteroscedastic noises. FANTOM is, to the best of our knowledge, the first method to handle heteroscedasticity in both stationary and non-stationary MTS, as well as non-Gaussianity in non-stationary settings. Given a MTS with multiple regimes, FANTOM simultaneously learns each regime's causal graph and mixing function, determines the number of regimes, and infers their indices (Figure 1). It uses a Bayesian Expectation Maximization (BEM) [12] procedure to optimize the evidence lower bound (ELBO), alternatively assigning regime indices (Expectation step) and inferring causal relationships in each regime (Maximization step). Unlike Gaussian-based approaches, FANTOM employs conditional normalizing flows [12] to handle complex distributions and compute regime membership probabilities. It uses Bayesian structure learning that averages across all plausible graphs and naturally filters out spurious edges. Under mild assumptions, we prove that temporal heteroscedastic causal models are identifiable for both stationary and multi-regime MTS. Across extensive comparisons with existing multi-regime causal discovery methods, we show that FANTOM consistently achieves superior performance in structure learning and regime detection. Moreover, it outperforms stationary models, even when they are provided with ground-truth regime partitions, on synthetic and two real-world datasets. The main contributions of this paper can be summarized as follows:

- We introduce FANTOM, a unified framework for causal discovery in multi-regime MTS that handles both homoscedastic non-Gaussian and heteroscedastic noises while simultaneously discovering the number of regimes, their indices, and their corresponding causal graphs.
- Under mild assumptions in causal discovery, we prove identifiability of the temporal heteroscedastic causal models in the stationary case and show that the number of regimes, their indices, and their graphs are identifiable (up to permutation) in the non-stationary setting.
- We demonstrate, via extensive experiments, that FANTOM outperforms state-of-the-art methods on both synthetic and real-world datasets.

Related work. Many works tackle *causal structure learning from stationary MTS*, Granger causality is the primary approach used for this purpose [32, 8]. However, it is unable to accommodate

instantaneous effects. DYNOTEARS [35], learns instantaneous and time lagged structures and leverages the acyclicity constraint, introduced by Zheng et al. in [60], to turn the DAG learning problem to a purely continuous optimization problem. However, DYNOTEARS is limited to linear SEMs. Runge et al. [42] proposed a two-stage algorithm PCMCi+ that can scale to large time series, PCMCi+ is able also to handle non linear relationships. Nevertheless, DYNOTEARS and PCMCi+ are restricted to homoscedastic noises where variance is a constant over time. For this reason, Rhino [15] and SCOTCH [55] introduced models that tackle stationary MTS with historical noise, where the noise variance is a function of solely time lagged parents. However Rhino, DYNOTEARS and PCMCi+, cannot handle heteroscedastic noise and are limited to stationary MTS.

Several studies have sought to tackle the challenge of *causal discovery in non-stationary MTS* [21, 17, 45, 33]. Remarkably, Huang et al. [21] address the setting of time series composed of different regimes by modulating causal relationships through a regime index. CD-NOD detects change points and outputs a single summary causal graph, but it overlooks recurring regimes and provides neither regime-specific graphs nor their count. RPCMCI [45] provides a graph for each regime, yet it assumes prior knowledge of the number of regimes, restricts edges to time-lagged links, and offers no identifiability guarantees. Balsells-Rodas et al. [5] establish identifiability for first-order, regime-dependent causal discovery in multi-regime MTS with Gaussian noise and offer a practical algorithm, but their framework allows only a single time-lagged edge and excludes instantaneous links. Finally, CASTOR [39] jointly infers regime labels and their causal graphs, capturing both instantaneous and lagged links, under an equivariant Gaussian noise assumption. However, none of these models can simultaneously learn the number of regimes, their indices, and their structures under non-Gaussian noise, nor do they offer identifiability guarantees. FANTOM fills this gap and even generalises to richer heteroscedastic noise settings while providing identifiability results. More detailed related work is provided in Appendix A).

2 Problem formulation

In this Section, we introduce our notation, define a temporal causal graph, and then we present a new Structural Equation Models (SEMs) for multi-regime MTS with non-Gaussian/Heteroscedastic noise.

Notation. Matrices, vectors, and scalars are denoted by uppercase bold $\mathbf{G}_\tau$, lowercase bold $\mathbf{x}_t$ and lowercase letters $x_{t-\tau}^i$, respectively. Ground-truth variables are indicated with an asterisk, such as $\mathcal{G}^*$. We assume that all distributions have densities $p(\mathbf{x}_t)$ w.r.t. the Lebesgue measure. The notation $[0 : L]$ represents the set of integers $\{0, \dots, L\}$ and $|\cdot|$ denotes set cardinality. The notation $(\mathbf{x}_t)_{t \in \mathcal{T}} = (x_t^i)_{i \in \mathbf{V}, t \in \mathcal{T}}$ represents a MTS of $|\mathbf{V}| = d$ components and length $|\mathcal{T}|$, $\mathcal{T}$ is the time index set and $\mathbf{x}_{<t}$ refers to $\{\mathbf{x}_{t-L}, \dots, \mathbf{x}_{t-1}\}$.

Definition 2.1 (Temporal Causal Graph [39]). The temporal causal graph, associated with the MTS $(\mathbf{x}_t)_{t \in \mathcal{T}}$, is defined by a Directed Acyclic Graph (DAG) $\mathcal{G} = (\mathbf{V}, E)$, represented by a collection of adjacency matrices $\mathbf{G}_{\tau \in [0:L]} = \{\mathbf{G}_0, \dots, \mathbf{G}_L\}$, and a fixed maximum lag L . Its vertices $\mathbf{V}$ consist of the set of components $x_{t'}^1, \dots, x_{t'}^d$ for each $t' \in [t-L : t]$. The edges E of the graph are defined as follows: $\forall \tau \in [1 : L]$ the variables $x_{t-\tau}^i$ and x_t^j are connected by a lag-specific directed link $x_{t-\tau}^i \rightarrow x_t^j$ 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 $\mathbf{G}_\tau \in \mathcal{M}_d(\mathbb{R})$ will be non-zero and $x^i \in \mathbf{Pa}_{\mathcal{G}}^j(< t)$; the lagged parents of node i in $\mathcal{G}$. For instantaneous links ($\tau = 0$), we cannot have self loops i.e. $i \neq j$. If $\tau = 0$, we have an edge $x_t^i \rightarrow x_t^j$ and $x^i \in \mathbf{Pa}_{\mathcal{G}}^j(t)$; the instantaneous parents of j at the current time t , if and only if x^i at time t causes x^j at time t .

In many real world scenarios, a non-stationary MTS $(\mathbf{x}_t)_{t \in \mathcal{T}}$ may exhibit K distinct, non-overlapping regimes, where non-stationarity is not modeled by continuous changes but rather by sequences of piecewise-constant regimes, as in climate science [27], finance [18], or epileptic recordings [56]. Every regime r is a stationary MTS block, has its own temporal causal graph $\mathcal{G}^r$ (Definition 2.1). At each time $t = 1, 2, \dots, |\mathcal{T}|$ there is a discrete latent state $z_t \in \{1, 2, \dots, K\}$ that models the regime partition, i.e., $z_t = r$ means that $\mathbf{x}_t$ belongs to regime r and we denote $\mathcal{I}_r = \{t | z_t = r, \forall t \in \mathcal{T}\}$ the set of all time indices at which regime r appears. We gather these sets into $\mathcal{I} = (\mathcal{I}_r)_{r \in [1:K]}$, yielding a unique time partition of the MTS $(\mathbf{x}_t)_{t \in \mathcal{T}}$ composed of K regimes. In addition, the observation $\mathbf{x}_t$ follows, a novel and general SEM that takes into account non stationarity, and handles both

non-Gaussian case and heteroscedastic setting, that we introduce as follows, $\forall r \in [1 : K], \forall t \in \mathcal{I}_r$:

$$x_t^i = f^{i,r}(\mathbf{Pa}_{\mathcal{G}^r}^i(< t), \mathbf{Pa}_{\mathcal{G}^r}^i(t)) + g^{i,r}(\mathbf{Pa}_{\mathcal{G}^r}^i(< t), \mathbf{Pa}_{\mathcal{G}^r}^i(t)) \cdot \epsilon_t^{i,r}, \quad (1)$$

where $f^{i,r}$ and $g^{i,r}$ are general differentiable functions, with $g^{i,r}$ strictly positive, and $\epsilon_t^{i,r}$ following an arbitrary probability density. We assume $\mathbb{E}(\epsilon_t^{i,r}) = 0$ and $\mathbb{E}((\epsilon_t^{i,r})^2) = 1$ without loss of generality. We denote the set of these temporal causal graphs as $\mathcal{G} = (\mathcal{G}^r)_{r \in [1:K]}$. In the case of non-stationary MTS, regimes appear sequentially with at least a minimum duration ζ , and a subsequent regime v (where $v = r + 1$) begins only after at least ζ time units from the start of regime r . Additionally, if regime r reoccurs, its duration in the second appearance is also no less than ζ samples. We refer to the phenomenon, where each regime persists for at least ζ consecutive time steps, as the *regime persistent dynamics* assumption and we define it as follows:

Assumption 2.2. We consider a MTS with K multiple regimes $(\mathbf{x}_t)_{t \in \mathcal{T}}$ where the SEM is defined in Eq.(1). Given variable i , we assume that a regime $r \in [1 : K]$ is ζ -persistent if the parents $(\mathbf{Pa}_{\mathcal{G}^r}^i(< t), \mathbf{Pa}_{\mathcal{G}^r}^i(t))$ and functional dependencies $(f^{i,r}, g^{i,r}, \epsilon_t^{i,r})$ are stationary for ζ consecutive time steps t .

The persistence assumption permits to capture different regime dynamics, whether arising from changes in the causal graph across regimes, commonly observed in climate science [27] and epileptic recordings [56], or from shifts in functional dependencies, which correspond to soft interventions in causal discovery. Our newly introduced SEM in Eq.(1) generalizes several existing approaches in three novel aspects: (1) when $K = 1$, if $g^{i,1}(\mathbf{Pa}_{\mathcal{G}^1}^i(< t), \mathbf{Pa}_{\mathcal{G}^1}^i(t)) = 1$ for all $i \in [1 : d]$, we recover the classical additive noise models in causal discovery. Thus, allowing $g^{i,1}$ to be a strictly positive and differentiable function, *not only extends Rhino's SEM [15] but also yields a new, general SEM for stationary multivariate time series with heteroscedastic noise*. (2) When $K > 1$, if $g^{i,1}(\mathbf{Pa}_{\mathcal{G}^r}^i(< t), \mathbf{Pa}_{\mathcal{G}^r}^i(t)) = 1$, and $\epsilon_t^{i,r} \sim \mathcal{N}(0, \sigma_r)$ for all $i \in [1 : d]$, we recover the setting introduced in [39, 4]. Then, allowing $\epsilon_t^{i,r}$ to follow an arbitrary probability density then yields the first *SEM for non-stationary MTS composed of multiple regimes with non-Gaussian noise*. Finally, (3) permitting $g^{i,r}$ to be a strictly positive, differentiable function leads, to the best of our knowledge, to the first *general SEM for non-stationary MTS with heteroscedastic noise*.

3 Flow based approach for Dynamic Temporal Causal models

3.1 ELBO formulation

Many real-world time series e.g., EEG data [22, 40], climate data [27, 16], and financial data [18] are non-stationary and exhibit complex noise distributions. Existing causal discovery methods cannot jointly recover the number of regimes, their indices, and their structures under either non-Gaussian or heteroscedastic noises. With FANTOM, our objective is to close this gap, simultaneously learning the number of regimes K , their indices $\mathcal{I} = (\mathcal{I}_r)_{r \in [1:K]}$, and DAGs $\mathcal{G} = (\mathcal{G}^r)_{r \in [1:K]}$ in both homoscedastic non-Gaussian and general heteroscedastic settings. Because integrating over latent regimes makes the data log-likelihood intractable, we instead maximise its Evidence Lower Bound (ELBO). Proposition 3.1, proved in Appendix G, formalises this ELBO for $N_w > K$ provisional regimes. Section 3.2 outlines the initialization trick that instantiates these N_w regimes, and the E-step that progressively merges them until N_w settles at K .

Proposition 3.1. Let $(\mathbf{x}_t)_{t \in \mathcal{T}}$ be a MTS composed of multiple regimes and following the SEM described in Eq.(1). The data likelihood admits the following evidence lower bound (ELBO):

$$\begin{aligned} \log p_{\Theta}(\mathbf{x}_{t \in \mathcal{T}}) &\geq \sum_{t=1}^{|\mathcal{T}|} \mathbb{E}_{q_{\phi}(\mathcal{G})} [\mathbb{E}_{p(z_t | \mathbf{x}_t, \mathbf{x}_{<t})} [\log p_{\theta^{z_t}}(\mathbf{x}_t | \mathbf{x}_{<t}, \mathcal{G}^{z_t}) + \log p(z_t)]] \\ &\quad + H(p(z_t | \mathbf{x}_t, \mathbf{x}_{<t})) + \sum_{r=1}^{N_w} \mathbb{E}_{q_{\phi^r}(\mathcal{G}^r)} [\log p(\mathcal{G}^r)] + H(q_{\phi^r}(\mathcal{G}^r)) \equiv \text{ELBO}(\Theta), \end{aligned} \quad (2)$$

where $\forall t \in \mathcal{T} : z_t \in [1 : N_w]$ are the discrete latent variables and N_w is the number of regimes.

Here, Θ are all the learnable parameters of our model (detailed in Section 3.2), $\log p_{\theta^{z_t}}(\mathbf{x}_t | \mathbf{x}_{<t}, \mathcal{G}^{z_t})$ represents the observational log-likelihood of $\mathbf{x}_t$ belonging to regime $z_t \in [1, N_w]$, $q_{\phi^r}(\mathcal{G}^r)$ represents the variational distribution that approximates the intractable posterior $p_{\theta^r}(\mathcal{G}^r | \mathbf{x}_{t \in \mathcal{I}_r})$, and $p(z_t | \mathbf{x}_t, \mathbf{x}_{<t})$ represents the posterior distribution of the latent variables z_t . The distribution $p(\mathcal{G}^r)$ is the graph prior and $p(z_t)$ represents our prior belief about the membership of samples to the causal models; typically we model it as a time varying function.

3.2 Model parametrization

In this Section, we will define and motivate FANTOM design choices. FANTOM maximises the ELBO (Eq.(2)) with a Bayesian Expectation Maximization (BEM) scheme. Because BEM normally needs the number of regimes a priori, we first describe an initialisation trick that removes this requirement. Next, we motivate our prior over the latent regime indicator z_t and show how a Temporal Graph Neural Network, combined with CNFs, models the regime-specific likelihood $p_{\theta^{z_t}}(\mathbf{x}_t | \mathbf{x}_{<t}, \mathcal{G}^{z_t})$.

Initialization trick. FANTOM 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). After initialization, FANTOM alternates between two phases. The E-step (subsection 3.4) updates the regime indices $\mathcal{I} = (\mathcal{I}_r)_{r \in [1 : N_w]}$ and removes regimes with too few samples, updating N_w . The M-step (subsection 3.3.1) learns the graphs $(\mathcal{G}_r)_{r \in [1 : N_w]}$ and models heteroscedastic noise with a Bayesian structure-learning method that uses conditional normalizing flows (CNFs). These phases repeat until the algorithm reaches the maximum number of iterations.

BEM choice motivation. We argue that inferring regimes and learning their associated DAGs are interdependent tasks, making the BEM algorithm particularly well-suited for this problem. A two-step alternative, that detects change points with KCP [1], then runs causal discovery on each segment breaks down: First, change point detection methods like KCP [1] fail to detect regime shifts driven by changes in causal mechanisms, because those involve shifts in conditional distributions (See Appendix F.3.3). It also treats recurring regimes as distinct, forcing redundant model fits and raising computation costs. Second, heteroscedastic noise further degrades existing causal methods (Table 2). FANTOM addresses all three issues by coupling CNFs, which capture heteroscedasticity, with Bayesian structure learning that prunes spurious edges.

Time varying weight modeling. We use time-varying weights, initially proposed for financial data modeling [59, 57], as priors for the discrete latent variables z_t . To support smooth regime transitions consistent with our persistence assumption (Assumption 2.2), we adopt a flexible functional form based on the softmax transformation of learnable parameter $\omega^r \in \mathbb{R}^2$ and time index t : $p(z_t = r) = \pi_t(\omega^r) = \frac{\exp(\omega_1^r \cdot t + \omega_0^r)}{\sum_{j=1}^{N_w} \exp(\omega_1^j \cdot t + \omega_0^j)}$. This formulation encourages that, if $\mathbf{x}_t$ belongs to regime r in the current iteration, it is only allowed to remain in the same regime r or smoothly transition to neighboring regimes $(r-1, r+1)$ in the next iteration. See Section 3.4 and Appendix B for details.

Bayesian structure learning. Following [14, 15, 55], FANTOM employs Bayesian structure learning. We approximate the intractable posterior $p_{\theta}(\mathcal{G} | \mathbf{x}_{t \in \mathcal{T}})$ using the variational distribution $q_{\phi}(\mathcal{G}) = \prod_{r=1}^{N_w} q_{\phi^r}(\mathcal{G}^r) \delta(\theta^r)$, where δ denotes the Dirac delta function. Following [14, 15, 55], we model $q_{\phi^r}(\mathcal{G}^r)$ as a product of independent Bernoulli variables and compute its expectation with a single Monte Carlo sample using the Gumbel-Softmax trick [26]. Additional details are in Appendix C.

Likelihood of SEM. Using the functional form Eq.(1), we have

$$\begin{aligned} y_t^{i,r} &= g^{i,r}(\mathbf{Pa}_{\mathcal{G}^r}^i(<t), \mathbf{Pa}_{\mathcal{G}^r}^i(t)) \cdot \epsilon_t^{i,r} \\ &= x_t^i - f^{i,r}(\mathbf{Pa}_{\mathcal{G}^r}^i(<t), \mathbf{Pa}_{\mathcal{G}^r}^i(t)), \end{aligned}$$

then we can write the observational likelihood:

$$p_{\theta^r}(\mathbf{x}_t^i | \mathbf{Pa}_{\mathcal{G}^r}^i(<t), \mathbf{Pa}_{\mathcal{G}^r}^i(t)) = p_{\text{hetero}}(y_t^{i,r} | \mathbf{Pa}_{\mathcal{G}^r}^i(<t), \mathbf{Pa}_{\mathcal{G}^r}^i(t)) \quad (3)$$

where p_{hetero} refers to the density function of the heteroscedastic conditions. To estimate $f^{i,r}$, we build upon the model formulation of [14], which uses neural networks to describe the functional

relationship $f_{\theta^r}^{i,r} : \mathbb{R}^d \rightarrow \mathbb{R}$. Specifically, we propose flexible functional designs for $f_{\theta^r}^{i,r}$, which must respect the relations encapsulated in $\mathcal{G}^r$. Namely, if $x_{t-\tau}^j \notin \mathbf{Pa}_{\mathcal{G}^r}^i(< t) \cup \mathbf{Pa}_{\mathcal{G}^r}^i(t)$, then $\partial f_{\theta^r}^{i,r} / \partial x_{t-\tau}^j = 0$. We design

$$f_{\theta^r}^{i,r}(\mathbf{Pa}_{\mathcal{G}^r}^i(< t), \mathbf{Pa}_{\mathcal{G}^r}^i(t)) = \psi^r \left(\sum_{\tau=0}^L \sum_{j=1}^d G_{\tau,j}^r \vartheta^r(x_{t-\tau}^j, e_{\tau,j}^r), e_{0,i}^r \right), \quad (4)$$

where ψ^r and ϑ^r are neural network blocks illustrated in Figure 2 with all the other colored blocks. Instead of using a neural network block per node, we adopt a weight-sharing mechanism by using a trainable embeddings $e_{\tau,i}$ for $\tau \in [0 : L]$ and $i \in \{1, \dots, d\}$. For the heteroscedastic term, we introduce an invertible mapping $\ell_{\theta^r}^{i,r} : \mathbb{R} \rightarrow \mathbb{R}$ such that:

$$\ell_{\theta^r}^{i,r} \left(g^{i,r}(\mathbf{Pa}_{\mathcal{G}^r}^i(< t), \mathbf{Pa}_{\mathcal{G}^r}^i(t)) \cdot \epsilon_t^{i,r} \right) = n_t^{i,r},$$

where $n_t^{i,r} \sim \mathcal{N}(0, 1)$. The design of $\ell_{\theta^r}^{i,r}$ needs to properly balance the flexibility and tractability of the transformed noise density. We choose a conditional normalizing flows, for heteroscedastic noise, called conditional spline flow [10], that transforms our heteroscedastic noise distribution to a fixed normal noise $n_t^{i,r}$ for regime r . The spline parameters are predicted using a hyper-network with a similar form as Eq.(4) to incorporate heteroscedasticity. Due to the invertibility of $\ell_{\theta^r}^{i,r}$, the noise likelihood conditioned on all parents is:

$$p_{\text{hetero}} \left(g^{i,r}(\mathbf{Pa}_{\mathcal{G}^r}^i(< t), \mathbf{Pa}_{\mathcal{G}^r}^i(t)) \cdot \epsilon_t^{i,r} \mid \mathbf{Pa}_{\mathcal{G}^r}^i(< t), \mathbf{Pa}_{\mathcal{G}^r}^i(t) \right) = p_{n^{i,r}} \left(n_t^{i,r} \mid \frac{\partial \ell_{\theta^r}^{i,r}}{\partial \epsilon_t^{i,r}} \right), \quad (5)$$

where $p_{n^{i,r}}(\cdot)$ is the standard normal density. In the non-Gaussian, *non-heteroscedastic* case, FANTOM learns a base distribution based on a composite affine-spline transformation of a standard Gaussian. Finally, the system parameters comprise the learnable parameters of the time varying weights ω^r , of the variational inference ϕ^r and of the neural networks θ^r . We use Θ , introduced in Eq.(2), to note all the learnable parameters of FANTOM, and we have the set of parameters is: $\Theta = \{(\omega^r, \phi^r, \theta^r)\}_{r=1}^{N_w}$.

3.3 BEM algorithm

3.3.1 M step: graph learning

FANTOM applies BEM to maximize the ELBO described in Proposition 2. It begins by initializing the regime partitions $\beta_t^r = p(z_t = r \mid \mathbf{x}_t, \mathbf{x}_{<t})$ using equally sized windows, selected via a hyperparameter. Note that these binary regime indices β_t^r are updated during the E-step. Then, in the M-step, FANTOM incorporates β_t^r in the ELBO Eq.(2) to estimate the DAGs for each regime and learn the parameters ω^r that align $\pi_t(\omega^r)$ with β_t^r . We have:

$$\arg \max_{\Theta} \left\{ \mathbb{E}_{q_{\phi}(\mathcal{G})} \left[\sum_{t=1}^{|\mathcal{T}|} \sum_{r=1}^{N_w} \beta_t^r \log p_{\theta^r}(\mathbf{x}_t \mid \mathbf{x}_{<t}, \mathcal{G}^r) \right] + \sum_{r=1}^{N_w} \mathbb{E}_{q_{\phi^r}(\mathcal{G}^r)} [\log p(\mathcal{G}^r)] + H(q_{\phi^r}(\mathcal{G}^r)) \right. \\ \left. + \sum_{t=1}^{|\mathcal{T}|} \sum_{r=1}^{N_w} \beta_t^r \log \pi_t(\omega^r) \right\}, \quad (6)$$

The maximization of the above equation can be decomposed into two distinct and separate maximization problems. The first problem, *regime alignment*, focuses on aligning $\pi_t(\omega^r)$ with β_t^r . While the

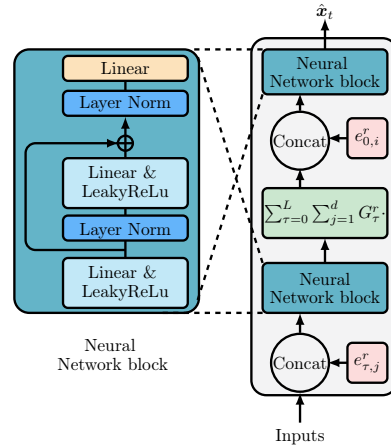


Figure 2: Temporal graph neural network (TGNN) used by FANTOM.

second one, **graph learning**, involves estimating DAGs for every regime.

For the graph prior $p(\mathcal{G}^r)$ for all $r \in [1 : N_w]$ have to combine two components: DAG constraint and graph sparseness prior. Inspired by [60, 15, 14], we propose the following unnormalised prior $p(\mathcal{G}^r) \propto \exp \left(-\lambda_s \|\mathbf{G}_{0:K}^r\|_F^2 - \rho h^2(\mathbf{G}_0^r) - \alpha h(\mathbf{G}_0^r) \right)$. Using Eq.(3), we have $\log p_{\theta^r}(\mathbf{x}_t | \mathbf{x}_{<t}, \mathcal{G}^r) = \sum_{i=1}^d \log p_{\text{hetero}} \left(y_t^{i,r} | \mathbf{Pa}_{\mathcal{G}^r}^i(<t), \mathbf{Pa}_{\mathcal{G}^r}^i(t) \right)$, where p_{hetero} and $y_t^{i,r}$ are defined in Eq.(5). The parameters θ^r, ϕ^r are learned by maximizing the **graph learning** maximization problem Eq.(6), where the Gumbel-softmax gradient estimator is used [26]. We also leverage augmented Lagrangian training similar to [35, 39, 14], to anneal α, ρ .

3.4 E step: Regime learning

In the E-step, FANTOM updates the posterior probability $\beta_t^r = p(z_t = r | \mathbf{x}_t, \mathbf{x}_{<t})$ (see Eq.(7), with derivation provided in Appendix B):

$$\begin{aligned} \beta_t^r &= \frac{p(z_t = r) p(\mathbf{x}_t | \mathbf{x}_{<t}, z_t = r, \mathcal{G}^r)}{\sum_{j=1}^{N_w} p(z_t = j) p(\mathbf{x}_t | \mathbf{x}_{<t}, z_t = j, \mathcal{G}^j)} \\ &\propto \pi_t(\omega^r) p(\mathbf{x}_t | \mathbf{x}_{<t}, z_t = r, \mathcal{G}^r), \end{aligned} \quad (7)$$

where $p(\mathbf{x}_t | \mathbf{x}_{<t}, z_t = r, \mathcal{G}^r)$ denotes the observational likelihood of $\mathbf{x}_t$ being generated by the SEM from Eq (1) for regime r . This probability is computed using the CNFs trained during the M-step, following the same reasoning as in Eq.(3).

The probability of $\mathbf{x}_t$ belonging to regime r is influenced by two main factors: the observation's position within its current regime and whether that regime is designated as **pure** or **impure**. For example, if $\mathbf{x}_t$ is in a **pure** regime r but is near the boundary in the current iteration, $\pi_t(\omega^r)$ and $\pi_t(\omega^{r+1})$ are nearly equal (e.g., $\pi_{t \in [1100, 1500]}(\omega^1)$ vs. $\pi_{t \in [1100, 1500]}(\omega^2)$ in Figure 3). Nonetheless, since regime r was learned from pure data, $p(\mathbf{x}_t | \mathbf{x}_{<t}, z_t = r, \mathcal{G}^r)$ stays high, keeping β_t^r at its maximum value and maintaining $\mathbf{x}_t$ in regime r for the next iteration. In the other hand, if $\mathbf{x}_t$ is in an **impure** regime $r+1$ near the boundary during the current iteration, $\pi_t(\omega^r)$ and $\pi_t(\omega^{r+1})$ are also close in value (e.g., $\pi_{t \in [1501, 1800]}(\omega^1)$ vs. $\pi_{t \in [1501, 1800]}(\omega^2)$ in Figure 3). However, because the causal graph for regime r is more reliable (having been derived from pure data), $p(\mathbf{x}_t | \mathbf{x}_{<t}, z_t = r, \mathcal{G}^r) > p(\mathbf{x}_t | \mathbf{x}_{<t}, z_t = r+1, \mathcal{G}^{r+1})$. As a result, $\mathbf{x}_t$ moves from regime $r+1$ to r in the next iteration.

For simplicity reasons, we explicit these cases from one border but the same thing happens in the other border which accelerates convergence. More details about other cases and Figures that illustrate the idea could be found in Appendix B.

After updating β_t^r , for each sample $\mathbf{x}_t$, FANTOM assigns a value of 1 to the most probable regime r (with the highest β_t^r), and 0 to others. Additionally, FANTOM 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 β_t^r in the subsequent iteration which is in general a neighboring regime ensured by the way we set up the probability $\beta_t^r \propto \pi_t(\omega^r) p(\mathbf{x}_t | \mathbf{x}_{<t}, z_t = r, \mathcal{G}^r)$.

4 Identifiability results

Identifiability is an important statistical property to ensure that the causal discovery problem is meaningful. In causal analysis, the whole point is to find out which variable causes others; if the model is not identifiable, the analysis is not possible at all. This section proves that causal discovery from multi-regime MTS is identifiable in the FANTOM framework namely, when (i) the noise is

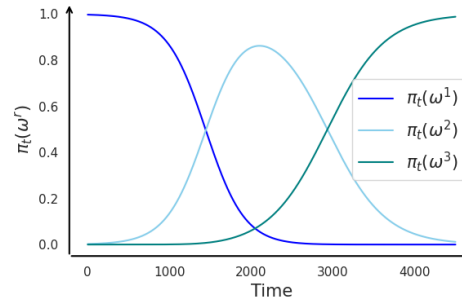


Figure 3: Illustration of $\pi_t(\omega^r)$ after Fantom's first iteration with equal windows of 1500 samples for an MTS of 4500 samples with two ground-truth regimes: $\mathcal{I}_1^* = [0 : 1999]$ and $\mathcal{I}_2^* = [2000 : 4500]$.

non-Gaussian or heteroscedastic and (ii) the latent variable z_t has a time-varying-parent prior. We first formalize identifiability for this setting, then state three theorems covering both stationary and multi-regime MTS under the two noise assumptions.

Definition 4.1. The conditional distribution of multi-regime MTS with a time varying prior is: $p(\mathbf{x}_t|\mathbf{x}_{<t}) = \sum_{r=1}^K \pi_t(\boldsymbol{\omega}^r) p_{\theta^r}(\mathbf{x}_t|\mathbf{x}_{<t}, \mathcal{G}^r)$. We say this model is identifiable up to permutation and translation, if:

- $\forall r \in [1 : K]$ the causal model $(\theta^r, \mathcal{G}^r)$ is identifiable.
- For any two models with parameters $(\boldsymbol{\omega}^r, \theta^r, \mathcal{G}^r)_{r=1}^K$ and $(\tilde{\boldsymbol{\omega}}^r, \tilde{\theta}^r, \tilde{\mathcal{G}}^r)_{r=1}^{\tilde{K}}$, such that for any $t \in \mathcal{T}$: $p(\mathbf{x}_t|\mathbf{x}_{<t}) = \tilde{p}(\mathbf{x}_t|\mathbf{x}_{<t})$, we have $K = \tilde{K}$ and it exists a permutation σ and translation function $\varrho: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ such that $\theta^r = \tilde{\theta}^{\sigma(r)}$ and $\boldsymbol{\omega}^r = \varrho(\tilde{\boldsymbol{\omega}}^{\sigma(r)})$

Following [15, 39, 7], we use the common assumptions of causal discovery settings (Causal Markov property H.2, stationarity H.2, minimality H.2, sufficiency H.2), see Appendix H.1 for precise statements. We present our first theoretical results, Theorem 4.2, states that for any stationary MTS, composed of $K = 1$ regime and following Eq.(1) with $\epsilon_t^i \sim \mathcal{N}(0, 1)$ the ground truth solution $\mathcal{G}^*$ is uniquely identifiable, the detailed proof can be found in Appendix H.

Theorem 4.2 (Identifiability of Temporal Heteroscedastic Gaussian noise model (THGNM)). Assume Causal Markov property, stationarity, minimality, sufficiency and let $(\mathbf{x}_t)_{t \in \mathcal{T}}$ be a MTS following a THGNM, $\forall t \in \mathcal{T}$:

$$\mathbf{x}_t^i = f^i(\mathbf{Pa}_{\mathcal{G}}^i(<t), \mathbf{Pa}_{\mathcal{G}}^i(t)) + g^i(\mathbf{Pa}_{\mathcal{G}}^i(<t), \mathbf{Pa}_{\mathcal{G}}^i(t)) \cdot \epsilon_t^i, \quad (8)$$

where f^i and g^i are differentiable functions, with g_i strictly positive and $\epsilon_t^i \sim \mathcal{N}(0, 1)$ are mutually independent normal noises. The THGNM is identifiable if $\frac{1}{g^i}$ is not a polynomial of degree two.

Identifiability of Temporal Restricted Heteroscedastic noise model (TRHNM). We present and prove in Appendix H.3 our second identifiability results of a TRHNM (Theorem H.10), where ϵ_t^i can follow any arbitrary density distribution. We states the results for bivariate time series, in which we show that, if a backward model exists, a differential equation will always hold. Then inspired from Peters et al. [38], Immer et al. [25], and Strobl et al. [50], we define TRHNM and show its identifiability.

Our last theoretical result states that the mixture of identifiable temporal causal models with either non-Gaussian or heteroscedastic noises is also identifiable as defined in definition 4.1.

Theorem 4.3 (Identifiability of the mixture of identifiable temporal causal models). Let $\mathcal{F}$ be a family of K identifiable temporal causal models, $\mathcal{F} = (p_{\theta^r}(\cdot|\cdot, \mathcal{G}^r))_{r \in \mathbb{N}^*}$ that are linearly independent and let $\mathcal{M}_K$ be the family of all K -finite mixtures of elements from $\mathcal{F}$, i.e.,

$$\left\{ p(\mathbf{x}_t|\mathbf{x}_{<t}) = \sum_{r=1}^K \pi_t(\boldsymbol{\omega}^r) p_{\theta^r}(\mathbf{x}_t|\mathbf{x}_{<t}, \mathcal{G}^r), p_{\theta^r}(\cdot|\cdot, \mathcal{G}^r) \in \mathcal{F}, \forall t \in \mathcal{T} : \pi_t(\boldsymbol{\omega}^r) > 0 \text{ and } \sum_{r=1}^K \pi_t(\boldsymbol{\omega}^r) = 1 \right\}.$$

Then the family $\mathcal{M}_K$ is identifiable as defined in definition 4.1.

Identifiability is important in causal discovery as shown in several papers [7, 15, 35, 5]. We extend these guarantees to FANTOM's settings by proving identifiability for stationary temporal models with heteroscedastic noise, showing that weaker assumptions suffice when the noise ϵ_t^i is simplified, recovering identifiability under explicit restrictions for arbitrary noise, and covering multi-regime MTS scenarios in both cases. Although convergence rates or finite-sample bounds remain elusive because the BEM objective is non-convex, our experiments demonstrate that FANTOM still converges in non-Gaussian and heteroscedastic settings. Further convergence rates or finite-data bounds are however extremely challenging due to the non-convexity of the acyclicity constraint in a BEM procedure. Yet, we empirically demonstrate, in the experiments section, that FANTOM converges in both non-Gaussian and heteroscedastic cases.

Table 1: Average SHD, F1 scores, NHD and Ratio for different models with $d = 10$ nodes and $K = 3$ regimes. *Split* denotes whether regime separation is automatic ($\checkmark$) or manual ($\times$), and *Type* classifies the graph as window (W) or summary (S). *Inst.* refers to instantaneous links, and *Lag* to time-lagged edges.

Model	Split	Type	Homoscedastic non-Gaussian noise												Heteroscedastic noise											
			Inst.				Lag				Regime Acc.	Inst.				Lag				Regime Acc.						
			SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓		SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓							
PCMCi+	×	W	17.5	74.9	0.02	0.24	14.5	88.3	0.01	0.11	×	46.1	11.1	0.05	0.88	46.0	19.0	0.05	0.80	×						
Rhino	×	W	2.50	96.8	0.002	0.03	6.00	95.2	0.006	0.04	×	44.5	5.11	0.06	0.94	53.5	64.7	0.07	0.35	×						
DYNOTEARS	×	W	42.0	54.4	0.06	0.45	21.5	82.1	0.02	0.17	×	89.5	31.5	0.14	0.68	118.0	37.5	0.17	0.61	×						
CASTOR	×	W	22.0	66.2	0.03	0.33	17.0	84.4	0.01	0.15	×	104.0	23.4	0.19	0.76	133.5	34.8	0.24	0.64	×						
RPCMCI	✓	W	-	-	-	-	-	-	-	-	51.6	-	-	-	-	-	-	-	-	-						
CASTOR	✓	W	47.0	34.8	0.05	0.65	59.5	39.4	0.07	0.60	99.4	5.67	93.3	0.006	0.06	12.3	90.9	0.012	0.08	97.1						
FANTOM	✓	W	0.33	99.5	0.00	0.00	12.5	89.1	0.01	0.10	99.4	5.67	93.3	0.006	0.06	12.3	90.9	0.012	0.08	97.1						
			SHD↓				F1↑				Acc.	SHD↓				F1↑				Acc.	SHD↓				Acc.	
			SHD↓				F1↑					SHD↓				F1↑					SHD↓					
CD-NOD	×	S	42.5	31.8	0.42	0.67	4.5	95.6	0.04	0.04	×	42	6.15	0.61	0.93	4.5	96.5	0.04	0.03	×						
FANTOM	×	S	4.5	95.6	0.04	0.04	4.5	95.6	0.04	0.04	96.6	4.5	96.5	0.04	0.03	4.5	96.5	0.04	0.03	96.8						

5 Experiments

5.1 Synthetic data

Data generation. We conduct extensive experiments to evaluate FANTOM’s performance on synthetic datasets. For ground truth graph generation, we use the Barabási-Albert model (degree 4) for instantaneous links and the Erdos–Rényi model (degree 1–2) for time-lagged relationships. For data generation process, f_i^r, g_i^r are chosen to be randomly initialized MLPs with one hidden layer and activation functions randomly chosen from $\{\text{Tanh}, \text{Exp}\}$. We evaluate the different models on multiple complex noise distributions; non-stationary MTS with either (1) heteroscedastic noise or (2) non-Gaussian homoscedastic noise, details in Appendix E.1. We consider $L = 1$, while additional experiments with multiple lags are provided in the Appendix F.1. Regime durations are randomly selected from $\{1000, 1500, 2000, 2500\}$. We test different numbers of nodes $\{5, 10, 20, 40\}$ and varying regime counts ($K \in \{2, 3\}$). Each combination of K and d nodes is repeated three times, resulting in over 24 distinct datasets¹.

Benchmarks. We benchmark our model against several baselines, including causal discovery methods for MTS with multiple regimes, such as CASTOR [39], CD-NOD [21] and RPCMCI [45]. Since CD-NOD returns a summary graph (see Appendix F.1), we compute a comparable summary graph from FANTOM’s output for fair evaluation. FANTOM is also compared with models for single-regime MTS, including Rhino [15], PCMCi+ [42], DYNOTEARS [35]. *Given that these models cannot deal with multiple regimes, we put them in a more favorable position than ours 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 causal model).*

Evaluation Metrics. We assess the performance of our proposed method for learning the DAGs using four key metrics: 1) F1 score, representing the harmonic mean of precision and recall; 2) Structural Hamming Distance (SHD), which counts discrepancies (e.g., reversed, missing, or redundant edges) between two DAGs; 3) Normalized Hamming Distance (NHD) measures how many edges differ normalized by the total number of possible edges; 4) Ratio NHD computes the ratio between the NHD and the baseline NHD of an output with the same number of edges but with all of them incorrect. For the regime detection task, we use Accuracy (Reg Acc) metric.

Results and discussion. Table 1 shows results on MTS with multiple regimes under heteroscedastic noise (right part of the table) and homoscedastic non-Gaussian noise (left part of the table). **In the homoscedastic, non-Gaussian scenario**, baselines generally perform better in the graph learning task, yet FANTOM still surpasses them on regime detection, with 99.4% accuracy, and instantaneous links, 99.5% F1. For the regime detection task, CASTOR succeeds to detect the regimes but with low accuracy (51.6%) compared to FANTOM (99.4%) and this due to the fact that CASTOR assumes equivariance normal noise with, however RPCMCI does not converge in this case too. For time-lagged connections, Rhino (95.2% F1), that has access to the ground-truth regime labels by training it on pure regime separately, slightly outperforms FANTOM (89.1% F1) that learns simultaneously the number of regimes, their indices and structures. **In the heteroscedastic setting**, FANTOM consistently outperforms both multi-regime baselines (CASTOR, RPCMCI, CD-NOD) and stationary approaches. It achieves the top scores on all metrics: for instantaneous links, an F1 of 93.3%, a 60%

¹<https://github.com/arahmani1/fantom.git>

improvement over the second-best, and a ratio of 0.06, 0.64 lower than the next-best DYNOTEARS. For time-lagged links, an F1 of 90.9%, 25% higher than Rhino, and a ratio of 0.08. FANTOM also detects the correct number of regimes and their indices with over 96% accuracy. By contrast, RPCMCI struggles to converge due to its homoscedastic assumption and time-lag-only dependencies, CASTOR relies on Gaussian noise and cannot detect regime labels in the absence of ground truth, DYNOTEARS, PCMCI+, CD-NOD, and Rhino likewise fail in this heteroscedastic scenario, even when regime labels are given a priori. Although Rhino models history-dependent noise, it does not handle general heteroscedasticity. Overall, the table shows that FANTOM is the only method that remains robust across both noise scenarios, matching or surpassing specialised baselines while simultaneously discovering regimes and graphs.

Appendices F.2.1 and F.2.2 report additional results with standard deviations, confirming that FANTOM sustains its performance when scaled to graphs of 20 and 40 nodes. Ablation study in the Appendix F.1 further shows FANTOM's robustness towards the choice of initialized window and ζ .

5.2 Real world data

Wind Tunnel. We use the wind tunnel datasets from Gamella et al. [13], featuring two controllable fans pushing air through a chamber, barometers measuring air pressure at various locations, and a hatch controlling an external opening. The dataset comprises 16 variables across two regimes of 10,000 samples each: the first is observational, while the second involves soft interventions on five variables (see Appendix E.4.1). We compare FANTOM to the aforementioned baselines, with results in Table 2. FANTOM is the only model that detects the regime with 99.9% accuracy and outperforms all baselines on the causal graph learning task, achieving 38.5% on F1 score. Notably, FANTOM surpasses all baselines even when they are given the ground-truth regime partitions.

Table 2: Performance on Wind Tunnel data evaluated on summary causal graph.

	Split	SHD↓	F1↑	Ratio↓	Reg Acc.
PCMCI+	×	37	22.9	0.77	×
DYNOTEARS	×	34	0	1	×
CASTOR	×	104	17.2	0.82	×
CD-NOD	×	40	20.0	0.80	×
Rhino	×	47	32.0	0.68	×
CASTOR	✓	120	19.5	0.80	49.9
FANTOM	✓	29	38.5	0.61	99.9

Epilepsy detection. Huizenga et al. [22] show that scalp potential fields are contaminated by heteroscedastic noise in EEG measurements. We evaluate FANTOM's performance in detecting epileptic regimes using EEG signals from 10 different patients in the Temple University Hospital EEG Seizure Corpus (TUSZ) dataset [40, 52]. We treat this as an unsupervised regime detection problem, analyzing roughly 100 seconds of recordings at a 250 Hz sampling rate for each patient, capturing both normal and seizure states (see Appendix E.4.2). The recordings consist of 19 electrodes, each considered a causal variable. FANTOM detects the correct regime partitions with an average 82.7% accuracy across all patients. The seizure regime's learned graph is denser and more connected than that of the normal state, which aligns with the generalized seizures affecting multiple brain regions. Full details and illustrations are provided in Appendix E.4.2.

6 Conclusion

We introduced FANTOM, a unified framework for multi-regime MTS that jointly infers (i) the number of regimes, (ii) their boundaries, and (iii) their corresponding causal DAG, under either non-Gaussian or heteroscedastic noises. Under mild assumptions in causal discovery, we prove identifiability of the temporal heteroscedastic causal models in the stationary case and show that the number of regimes, their indices, and their graphs are identifiable (up to permutation) in the non-stationary setting. Extensive experiments on synthetic and real-world data show consistent gains over strong baselines. FANTOM offers a principled means to uncover regime-specific causal dynamics, enhancing regime detection, and causal discovery with potential applications in various domains such as finance, climate science, and neuroscience.

Acknowledgment

We thank Alessandro Favero, Nikolaos Dimitriadis, and Ortal Senouf 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

- [1] Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of machine learning research*, 2019.
- [2] Charles K Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 2022.
- [3] Karim Assaad, Emilie Devijver, Eric Gaussier, and Ali Ait-Bachir. A mixed noise and constraint-based approach to causal inference in time series. In *ECML PKDD*, 2021.
- [4] Carles Balsells-Rodas, Yixin Wang, and Yingzhen Li. On the identifiability of switching dynamical systems. *arXiv preprint arXiv:2305.15925*, 2023.
- [5] Carles Balsells-Rodas, Yixin Wang, and Yingzhen Li. On the identifiability of switching dynamical systems. In *Forty-first International Conference on Machine Learning*, 2024.
- [6] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 1999.
- [7] Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems*, 2020.
- [8] Bart Bussmann, Jannes Nys, and Steven Latré. Neural additive vector autoregression models for causal discovery in time series. In *Discovery Science*, 2021.
- [9] Yuxiao Cheng, Ziqian Wang, Tingxiong Xiao, Qin Zhong, Jinli Suo, and Kunlun He. Causalttime: Realistically generated time-series for benchmarking of causal discovery. *arXiv preprint arXiv:2310.01753*, 2023.
- [10] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 2019.
- [11] Doris Entner and Patrik O Hoyer. On causal discovery from time series data using fci. *Probabilistic graphical models*, 2010.
- [12] Nir Friedman. The bayesian structural em algorithm. *arXiv preprint arXiv:1301.7373*, 2013.
- [13] Juan L Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for ai methodology. *Nature Machine Intelligence*, 7(1):107–118, 2025.
- [14] Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, et al. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*, 2022.
- [15] Wenbo Gong, Joel Jennings, Cheng Zhang, and Nick Pawlowski. Rhino: Deep causal temporal relationship learning with history-dependent noise. *Preprint arXiv:2210.14706*, 2022.
- [16] Wiebke Günther, Urmi Ninad, Jonas Wahl, and Jakob Runge. Conditional independence testing with heteroskedastic data and applications to causal discovery. *Advances in Neural Information Processing Systems*, 2022.
- [17] Wiebke Günther, Oana-Iuliana Popescu, Martin Rabel, Urmi Ninad, Andreas Gerhardus, and Jakob Runge. Causal discovery with endogenous context variables. *Advances in Neural Information Processing Systems*, pages 36243–36284, 2024.

- [18] James D Hamilton and Raul Susmel. Autoregressive conditional heteroskedasticity and changes in regime. *Journal of econometrics*, 1994.
- [19] Uzma Hasan, Emam Hossain, and Md Osman Gani. A survey on causal discovery methods for iid and time series data. *arXiv preprint arXiv:2303.15027*, 2023.
- [20] Stefan Haufe, Klaus-Robert Müller, Guido Nolte, and Nicole Krämer. Sparse causal discovery in multivariate time series. In *causality: objectives and assessment*. PMLR, 2010.
- [21] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. In *Journal of Machine Learning Research*, 2020.
- [22] Hilde M Huizenga and Peter CM Molenaar. Equivalent source estimation of scalp potential fields contaminated by heteroscedastic and correlated noise. *Brain topography*, 1995.
- [23] Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, 2014.
- [24] Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 2010.
- [25] Alexander Immer, Christoph Schultheiss, Julia E Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models. In *International Conference on Machine Learning*. PMLR, 2023.
- [26] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.
- [27] Soufiane Karmouche, Evgenia Galytska, Jakob Runge, Gerald A Meehl, Adam S Phillips, Katja Weigel, and Veronika Eyring. Regime-oriented causal model evaluation of atlantic–pacific teleconnections in cmip6. *Earth System Dynamics*, 2023.
- [28] Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *Preprint arXiv:1910.01075*, 2019.
- [29] Ilyes Khemakhem, Ricardo Monti, Robert Leech, and Aapo Hyvarinen. Causal autoregressive flows. In *International conference on artificial intelligence and statistics*. PMLR, 2021.
- [30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [31] Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. Dibs: Differentiable bayesian structure learning. In *Advances in Neural Information Processing Systems*, 2021.
- [32] Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, 2022.
- [33] Sarah Mameche, Lénaïg Cornanguer, Urmi Ninad, and Jilles Vreeken. Spacetime: Causal discovery from non-stationary time series. *arXiv preprint arXiv:2501.10235*, 2025.
- [34] Mark Newman. *Networks*. Oxford university press, 2018.
- [35] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, 2020.
- [36] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. *Advances in neural information processing systems*, 2013.
- [37] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

- [38] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 2014.
- [39] Abdellah Rahmani and Pascal Frossard. Causal temporal regime structure learning. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.
- [40] Abdellah Rahmani, Arun Venkitaraman, and Pascal Frossard. A meta-gnn approach to personalized seizure detection and classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
- [41] Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 2018.
- [42] Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020.
- [43] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 2019.
- [44] Karen Sachs, Omar Perez, Dana Pe’er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 2005.
- [45] Elena Saggioro, Jana de Wiljes, Marlene Kretschmer, and Jakob Runge. Reconstructing regime-dependent causal relationships from observational time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 2020.
- [46] Christof Seiler and Susan Holmes. Multivariate heteroscedasticity models for functional brain connectivity. *Frontiers in neuroscience*, 2017.
- [47] Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, and Gyorgy Simon. Challenges and opportunities with causal discovery algorithms: application to alzheimer’s pathophysiology. *Scientific reports*, 2020.
- [48] Le Song, Mladen Kolar, and Eric Xing. Time-varying dynamic bayesian networks. In *Advances in Neural Information Processing Systems*, 2009.
- [49] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.
- [50] Eric V Strobl and Thomas A Lasko. Identifying patient-specific root causes with the heteroscedastic noise model. *Journal of Computational Science*, 2023.
- [51] Xiangyu Sun, Oliver Schulte, Guiliang Liu, and Pascal Poupart. Nts-notears: Learning nonparametric dbns with prior knowledge. *arXiv preprint arXiv:2109.04286*, 2021.
- [52] Siyi Tang, Jared A Dunnmon, Khaled Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel L Rubin, and Christopher Lee-Messer. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. *arXiv preprint arXiv:2104.08336*, 2021.
- [53] Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *The Journal of Machine Learning Research*, 2015.
- [54] Sumanth Varambally, Yi-An Ma, and Rose Yu. Discovering mixtures of structural causal models from time series data. *arXiv preprint arXiv:2310.06312*, 2023.
- [55] Benjie Wang, Joel Jennings, and Wenbo Gong. Neural structure learning with stochastic differential equations. *arXiv preprint arXiv:2311.03309*, 2023.
- [56] Xiaojia Wang, Yanchao Liu, and Chunfeng Yang. Ictal-onset localization through effective connectivity analysis based on rnn-gc with intracranial eeg signals in patients with epilepsy. *Brain Informatics*, 2024.
- [57] Chun S Wong and Wai K Li. On a logistic mixture autoregressive model. *Biometrika*, 2001.

- [58] Sidney J Yakowitz and John D Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 1968.
- [59] Shuguang Zhang, Minjing Tao, Xu-Feng Niu, and Fred Huffer. Time-varying gaussian-cauchy mixture models for financial risk management. *arXiv preprint arXiv:2002.06102*, 2020.
- [60] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.
- [61] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. *Preprint arXiv:1906.04477*, 2019.

Table of content

A Detailed related works	16
B Expectation step: Derivation, intuition and illustration	17
C Maximization step:	18
C.1 Mathematical Derivation: Equation (2) to (6)	18
C.2 Variational Inference Details	19
D Limitations & Risk of spurious causality	19
E Data generation and Baselines	20
E.1 Synthetic data	20
E.2 Baselines	20
E.3 Optimization parameters	21
E.4 Real world data	21
E.4.1 Causal Chambers data	21
E.4.2 Epilepsy data	22
F Additional Experiments	24
F.1 Ablation studies	24
F.1.1 Impact of Heteroscedastic Modeling and Flow Architecture	24
F.1.2 Robustness to the pruning step, minimum regime duration, and regime initialization	25
F.1.3 Robustness to data standardization	26
F.2 Additional results on synthetic data	27
F.2.1 Heteroscedastic noise with different number of nodes and regimes	27
F.2.2 Non-Gaussian noise with different number of nodes and regimes	29
F.3 Additional experiments for $L=2$	31
F.3.1 Illustrations of learned graphs	32
F.3.2 Time complexity analysis	35
F.3.3 Regime detection experiments	36
G Proof of proposition	37
H Proofs of our theoretical contributions	38
H.1 Assumptions	38
H.2 Proof of theorem 4.2	39
H.3 Identifiability results in the case of Temporal General Heteroscedastic Noise Models	40
H.4 Proof of theorem 4.3	42

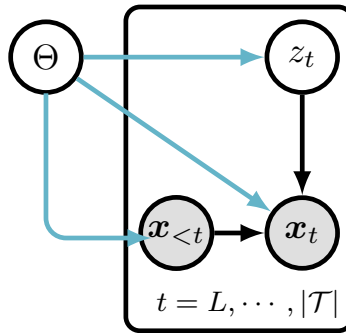


Figure 4: Graphical model of FANTOM. Observed variables (x_t) are in gray, while latent variables (z_t) and parameters (Θ) are in white. Blue edges represent parameter-variable interactions.

A Detailed related works

Causal structure learning from IID data. Causal structure learning has become an active area of research. hasan et al. [19] recently presented a comprehensive review of causal discovery methods for IID data and time series. For IID data, several approaches rely on conditional independence to infer causal relationships from observational data, such as the classical PC algorithm [49]. Additionally, some methods extend beyond observational data, incorporating interventional data to enhance causal inference, including COMBINE [53] and HEJ [23]. These approaches utilize data collected from controlled interventions to uncover causal relationships. A novel research direction introduced in [60] addresses the combinatorial challenges of structure learning by formulating it as a continuous constrained optimization problem, thus avoiding computationally costly combinatorial searches. Similarly, Zhu et al. [61] utilize the acyclicity constraint but employ reinforcement learning techniques for estimating directed acyclic graphs (DAGs). In contrast, Ke et al. [28] propose an approach that learns DAGs from interventional data by optimizing an unconstrained objective function. [7] provide a comprehensive analysis of continuous-constrained methods, offering a generalized framework applicable to interventional data scenarios. Another significant method, DiBS [31], estimates the full posterior distribution over Bayesian networks from limited observations, enabling quantification of uncertainty and assessment of confidence in causal discovery.

Causal structure learning from stationary MTS. The previously mentioned state-of-the-art methods primarily target independent observations rather than temporal dependencies. Assaad et al. [2] provide a comprehensive review of approaches specifically designed for causal discovery from MTS. To model causal relationships involving time dependencies, researchers frequently employ Dynamic Bayesian Networks (DBNs), which effectively capture discrete-time temporal dynamics within directed graphical frameworks. Some methods neglect contemporaneous (instantaneous) dependencies and focus exclusively on recovering time-lagged causal links [20, 48], and tsFCI [11], the latter adapting the Fast Causal Inference algorithm [49] for time series data. Runge et al. [43] introduced PCMCI, a scalable two-stage algorithm for time series, initially focusing only on time-lagged relationships. They subsequently extended it to PCMCI+ [42], enabling the identification of contemporaneous causal connections. Additionally, models addressing non-Gaussian instantaneous effects have been developed, such as VARLINGAM [24], which integrates non-Gaussian instantaneous models with autoregressive components. Another significant method is Time-series Models with Independent Noise (TiMINo) [36], which studies nonlinear and instantaneous effects using constrained SEMs. Pamfil et al. [35] recently proposed DYNOTEARS, leveraging an algebraic characterization of graph acyclicity from [60] to estimate both instantaneous and time-lagged relationships from time series data. DYNOTEARS utilizes a score-based DBN learning approach optimized via an augmented Lagrangian framework, enabling causal graph inference without assumptions on the underlying topology. In contrast, methods like NBCB [3], a noise-based/constraint-based approach, aim to learn a summary causal graph directly from observational time series data, going beyond Markov equivalence constraints even in the presence of instantaneous relationships. Rhino [15] introduced the first model that tackle stationary MTS with historical noise, where they assume that the noise variance changes over time as a function of solely time lagged

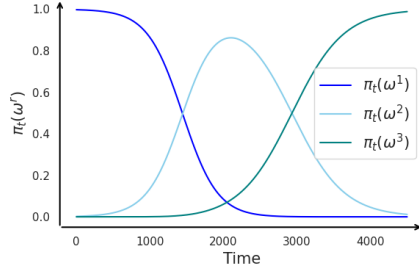


Figure 5: Illustration of $\pi_t(\omega^r)$ after Fantom's first iteration with equal windows of 1500 samples for an MTS of 4500 samples with two ground-truth regimes: $\mathcal{I}_1^* = [0 : 1999]$ and $\mathcal{I}_2^* = [2000 : 4500]$.

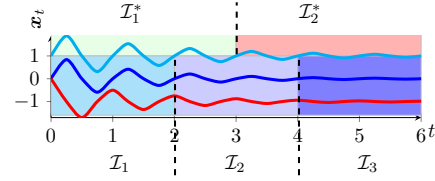


Figure 6: Example of initialization with $N_w = 3$ windows, $\mathcal{I}_1$ and $\mathcal{I}_3$ are pure regimes while $\mathcal{I}_2$ is impure (composed of samples from two ground-truth regimes $\mathcal{I}_1^*$ and $\mathcal{I}_2^*$, $K = 2$).

parents. All the aforementioned methods does not handle heteroscedastic setting and also fail short in the case of MTS with multiple regimes.

Causal structure learning from MTS with multiple regimes. 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 [21], 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, also CD-NOD cannot handle either non-Gaussian or heteroscedastic noise. 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 [45]. In this paper, the model [45] learns 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. [5] 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. CASTOR [39] learns regime indices and their corresponding causal graphs, including instantaneous and time-lagged relationships, under the assumption of normally distributed noise with equivariance. However, like other causal discovery methods for non-stationary MTS, it does not address non-Gaussian or heteroscedastic noise. [54] tackle a different setting in which they aim to discover a mixtures of Structural Causal Models from a datasets of MTS. They assume that they have different stationary MTS in the same dataset, one regime per MTS, and each one could be explained by one causal model in the mixture. In their case, they assume that every MTS in the dataset is stationary but the whole dataset is a mixture. In our case, we assume that we have only one non-stationary MTS and it is composed of different regimes where we do not know when the regime starts and ends, and our goal is to identify the regimes and the corresponding causal graphs.

B Expectation step: Derivation, intuition and illustration

$$\begin{aligned} \beta_t^r &= p(z_r = 1 \mid \mathbf{x}_t, \mathbf{x}_{<t}, \mathbf{G}_{\{0:L\}}^r) \\ &= \frac{p(z_t = r) p(\mathbf{x}_t \mid \mathbf{x}_{<t}, z_t = r, \mathcal{G}^r)}{\sum_{j=1}^{N_w} p(z_t = j) p(\mathbf{x}_t \mid \mathbf{x}_{<t}, z_t = j, \mathcal{G}^j)} \\ &\propto \pi_t(\omega^r) p(\mathbf{x}_t \mid \mathbf{x}_{<t}, z_t = r, \mathcal{G}^r), \end{aligned}$$

where $p(\mathbf{x}_t \mid \mathbf{x}_{<t}, z_t = r, \mathcal{G}^r)$ denotes the likelihood of $\mathbf{x}_t$ being generated by the SEM from Equation (1) for regime r . This probability is computed using the normalizing flows trained during the M-step, following the same reasoning as in Eq(3).

The probability of $\mathbf{x}_t$ belonging to regime r is influenced by two main factors: the observation's position within its current regime and whether that regime is designated as **pure** or **impure**. In order to clarify the intuition behind **pure** and **impure** regimes, Figure 6 shows an example of such case. It presents an initialization of three equal windows while the MTS is composed of two ground truth regimes presented by green color $\mathcal{I}_1^*$ and red one $\mathcal{I}_2^*$. In such case, the regimes $\mathcal{I}_1$ and $\mathcal{I}_3$ are pure, because they are composed of samples coming from the same ground truth regime ($\mathcal{I}_1^*$ for the regime $\mathcal{I}_1$ and $\mathcal{I}_2^*$ for the regime $\mathcal{I}_3$), while $\mathcal{I}_2$ is an impure regime and has samples from the two ground truth regimes.

We highlight all the different cases for a sample $\mathbf{x}_t$ either near to the border or not of regime r and also either the causal graph learned for r is on **pure** or **impure** data:

Case 1: If $\mathbf{x}_t$ is in a **pure** regime r and is far from the boundary in the current iteration, $\pi_t(\omega^r)$ takes a high value (for example, $\pi_{t \in [0, 1000]}(\omega^1)$ in Figure 3). Because regime r was trained on pure data, its causal graph is more accurate, leading to a high likelihood $p(\mathbf{x}_t \mid \mathbf{x}_{<t}, z_t = r, \mathcal{G}^r)$. Consequently, $\beta_t^r \propto \pi_t(\omega^r) p(\mathbf{x}_t \mid \mathbf{x}_{<t}, z_t = r, \mathcal{G}^r)$ remains dominant, causing $\mathbf{x}_t$ to stay in regime r at the next iteration.

Case 2: If $\mathbf{x}_t$ is in a **pure** regime r but is near the boundary in the current iteration, $\pi_t(\omega^r)$ and $\pi_t(\omega^{r+1})$ are nearly equal (e.g., $\pi_{t \in [1100, 1500]}(\omega^1)$ vs. $\pi_{t \in [1100, 1500]}(\omega^2)$ in Figure 3). Nonetheless, since regime r was learned from pure data, $p(\mathbf{x}_t \mid \mathbf{x}_{<t}, z_t = r, \mathcal{G}^r)$ stays high, keeping β_t^r at its maximum value and maintaining $\mathbf{x}_t$ in regime r for the next iteration.

Case 3: If $\mathbf{x}_t$ is in an **impure** regime $r + 1$ near the boundary during the current iteration, $\pi_t(\omega^r)$ and $\pi_t(\omega^{r+1})$ are also close in value (e.g., $\pi_{t \in [1501, 1800]}(\omega^1)$ vs. $\pi_{t \in [1501, 1800]}(\omega^2)$ in Figure 3). However, because the causal graph for regime r is more reliable (having been derived from pure data), $p(\mathbf{x}_t \mid \mathbf{x}_{<t}, z_t = r, \mathcal{G}^r) > p(\mathbf{x}_t \mid \mathbf{x}_{<t}, z_t = r + 1, \mathcal{G}^{r+1})$. As a result, $\mathbf{x}_t$ moves from regime $r + 1$ to r in the next iteration.

Case 4: $\mathbf{x}_t$ belongs to impure regime $r + 1$ and is far from the border (e.g., $t \in [1801, 2500]$ in Figure 3). In this case, it's uncertain whether $\mathbf{x}_t$ will switch regimes in the next iteration. However, as the pure regime r expands with each iteration, $\mathbf{x}_t$ will eventually be near the border of regime $r + 1$, bringing us back to Case 3.

C Maximization step:

C.1 Mathematical Derivation: Equation (2) to (6)

We perform the maximization of ELBO presented in proposition 3.1, using a BEM procedure, where we alternate between E-step (updating posterior probabilities while fixing all the parameters $\Theta = \{\theta^r, \phi^r, \omega^r\}_{r=1}^{N_w}$) and M step (Updating the parameters while using the posteriors learned in the E-step), we can summarize the process as follows:

- In the Estep: we learn $\beta_t^r = p(z_t \mid \mathbf{x}_t, \mathbf{x}_{<t}, \Theta_{\text{old}})$, where Θ_{old} is Θ of the previous iteration.
- In the M-step: we fix the learned posterior probabilities to the values β_t^r and we update Θ .
- By fixing the posterior probabilities in the M-step, the entropy of these probabilities $H(p(z_t \mid \mathbf{x}_t, \mathbf{x}_{<t}, \Theta_{\text{old}}))$ is a constant, which allow us to discard it.

The detail derivation from Eq.(2) to Eq.(6) is the following:

$$\begin{aligned}
\text{ELBO}(\Theta) &= \sum_{t=1}^{|T|} \mathbb{E}_{q_{\phi}(\mathcal{G})} \left[\mathbb{E}_{p(z_t | x_t, x_{<t})} \left(\log p_{\theta_{z_t}}(x_t | x_{<t}, \mathcal{G}^{z_t}) + \log p(z_t) \right) + H(p(z_t | x_t, x_{<t})) \right] \\
&\quad + \sum_{r=1}^{N_w} \mathbb{E}_{q_{\phi_r}(\mathcal{G}^r)} [\log p(\mathcal{G}^r)] + H(q_{\phi_r}(\mathcal{G}^r)) \\
&= \sum_{t=1}^{|T|} \mathbb{E}_{q_{\phi}(\mathcal{G})} \left[\sum_{z_t} p(z_t | x_t, x_{<t}, \Theta_{\text{old}}) \left(\log p_{\theta_{z_t}}(x_t | x_{<t}, \mathcal{G}^{z_t}) + \log p(z_t) \right) \right] \\
&\quad + \underbrace{H(p(z_t | x_t, x_{<t}, \Theta_{\text{old}}))}_{\text{Cte}} + \sum_{r=1}^{N_w} \mathbb{E}_{q_{\phi_r}(\mathcal{G}^r)} [\log p(\mathcal{G}^r)] + H(q_{\phi_r}(\mathcal{G}^r))
\end{aligned} \tag{9}$$

We replace $p(z_t | x_t, x_{<t}, \Theta_{\text{old}})$ by β_t^r , $p(z_t)$ by our prior choice $\pi_t(\omega^r)$, and we discard the entropy of the posterior because it is a constant in these steps. Hence, we got the Eq.(6):

$$\begin{aligned}
\text{ELBO}(\Theta) &= \mathbb{E}_{q_{\phi}(\mathcal{G})} \left[\sum_{t=1}^{|T|} \sum_{r=1}^{N_w} \beta_t^r \left(\log p_{\theta_r}(x_t | x_{<t}, \mathcal{G}^r) + \log \pi_t(\omega^r) \right) \right] \\
&\quad + \sum_{r=1}^{N_w} \mathbb{E}_{q_{\phi_r}(\mathcal{G}^r)} [\log p(\mathcal{G}^r)] + H(q_{\phi_r}(\mathcal{G}^r)).
\end{aligned} \tag{10}$$

After the maximization of Eq.(6), FANTOM update the posteriors $p(z_t | x_t, x_{<t}, \theta_{\text{old}})$ following Eq.(7).

C.2 Variational Inference Details

We provide the detailed formulations for $q_{\phi^r}(\mathcal{G}^r)$. In order to model the temporal adjacency matrices $\mathbf{G}_{\tau}^r$ where $\tau \in [1 : K]$, we use two learnable matrices $\mathbf{U}_{\tau}, \mathbf{Q}_{\tau} \in \mathbf{R}^{d \times d}$ such that:

$$p_{k,ij} = \frac{\exp(u_{k,ij})}{\exp(u_{k,ij}) + \exp(q_{k,ij})} \tag{11}$$

For instantaneous graphs $\mathbf{G}_0^r$, we used the same trick as in [15, 54], in which we employ three lower triangular learnable matrices $\mathbf{U}_0, \mathbf{Q}_0, \mathbf{E}_0 \in \mathbf{R}^{d \times d}$ to characterise three scenarios: (1) $i \rightarrow j$; (2) $j \rightarrow i$; (3) no edge between them. For node $i > j$:

$$\begin{aligned}
p(i \rightarrow j) &= \frac{\exp(u_{ij})}{\exp(u_{ij}) + \exp(q_{ij}) + \exp(e_{ij})} \\
p(j \rightarrow i) &= \frac{\exp(q_{ij})}{\exp(u_{ij}) + \exp(q_{ij}) + \exp(e_{ij})} \\
p(\text{no edge}) &= \frac{\exp(e_{ij})}{\exp(u_{ij}) + \exp(q_{ij}) + \exp(e_{ij})}.
\end{aligned} \tag{12}$$

With this formulation, the instantaneous adjacency matrix is free of self-loops, eliminating any length-1 cycles.

D Limitations & Risk of spurious causality

FANTOM's performance deteriorates when a regime contains only a handful of samples or is recorded at an extremely low sampling rate. This shortcoming is not surprising, estimating a separate causal graph for each regime is intrinsically difficult in the presence of multiple regimes and heteroscedastic noise. Yet this ability to pinpoint which edges vanish or emerge from one regime to the next is what makes FANTOM valuable in domains such as healthcare and climate science, where regime shifts

carry substantive meaning. Importantly, in realistic settings where each regime offers sufficient data, for example, epileptic seizures that last several minutes at 250 Hz, FANTOM delivers strong results and provides insights unattainable with existing methods.

FANTOM requires a suitable initial segmentation to effectively learn the regime indices. Our ablation studies demonstrate that selecting a reasonable initial window size establishes a basis for accurate regime detection, which is crucial for achieving high performance. However, we highlight a critical limitation regarding the subsequent pruning step: if the initial window is too high (over-pruning), it can drastically reduce the number of initial regimes below the ground truth. This initialization leads to a loss of essential information and a significant deterioration of FANTOM's final performance.

Spurious causality is a practical risk, so we advocate using the learned graphs as decision support rather than as ground truth. Before acting on them, especially in medical or financial settings, practitioners should adopt safeguards: (i) expert vetting of proposed edges and mechanisms; (ii) robustness checks (e.g., stability under resampling or perturbations, alternative specifications); and (iii) interventional validation when feasible.

E Data generation and Baselines

E.1 Synthetic data

We employ the Erdos–Rényi (ER) [34] model with mean degrees of 1 or 2 to generate lagged graphs, and the Barabasi–Albert (BA) [6] model with mean degrees 4 for instantaneous graphs. The maximum number of lags, L , is set at 1. We experiment with varying numbers of nodes $\{10, 20, 40\}$ and different numbers of regimes $\{2, 3\}$, each representing diverse causal graphs or mixing functions. The length of each regime is randomly sampled from the set $\{1000, 1500, 2000, 2500, 3000\}$.

- **Heteroscedastic case.** In heteroscedastic settings, noise variance shifts across both variables and observations, making the underlying DAG much harder to recover from data. Given a random set of directed acyclic graphs $\mathcal{G} = (\mathcal{G}^r)_{r \in [1:K]}$, we generate observations from the SEMs in Eq 1 as follows:

$$\forall r \in \{1, \dots, K\}, \forall t \in \mathcal{I}_r : x_t^i = f^{i,r}(\text{Pa}_{\mathcal{G}^r}^i(<t), \text{Pa}_{\mathcal{G}^r}^i(t)) + \exp(g^{i,r}(\text{Pa}_{\mathcal{G}^r}^i(<t), \text{Pa}_{\mathcal{G}^r}^i(t))) \cdot \epsilon_t^{i,r},$$

where $f^{i,r}, g^{i,r}$ are chosen to be randomly initialized MLPs with one hidden layer of size number of nodes and $\tanh$ activation functions. $\epsilon_t^{i,r}$ follows either a normal distribution $\mathcal{N}(0, 1)$ or a more complex one obtained by transforming samples from a standard Gaussian with an MLP with random weights and $\sin$ activation function.

- **Homoscedastic non-Gaussian case.** The formulation used to generate the data is:

$$\forall r \in \{1, \dots, K\}, \forall t \in \mathcal{I}_r : x_t^i = f^{i,r}(\text{Pa}_{\mathcal{G}^r}^i(<t), \text{Pa}_{\mathcal{G}^r}^i(t)) + \epsilon_t^{i,r},$$

where $f^{i,r}$ is a general differentiable non-linear function. The function $f^{i,r}$ is a random combination between a linear transformation and a randomly chosen function from the set: $\{\tanh, \exp\}$. $\epsilon_t^{i,r}$ follows either a Triangular distribution or a more complex one obtained by transforming samples from a standard Gaussian with an MLP with random weights and $\sin$ activation function.

E.2 Baselines

DYNOTEARS [35]. DYNOTEARS formulates causal discovery for multivariate time series through a linear vector autoregressive (VAR) model that simultaneously captures lagged and instantaneous causal effects. Its key innovation is the *DAGness* penalty a smooth, continuously differentiable relaxation of the acyclicity constraint optimized via an augmented Lagrangian scheme alongside a mean squared error loss. DYNOTEARS emerges as the special case of FANTOM obtained by setting $K = 1$, using linear component functions $f^{i,1}$, fixing the noise scaling to $g^{i,1} = 1$ and $\epsilon_t^{i,1} \sim \mathcal{N}(0, 1)$ in Eq(1). For comparing with this model, we use publicly available package `causalnex`².

PCMCI+ [42], a scalable two-stage algorithm for time series, enabling the identification of contemporaneous causal connection. As DYNOTEARS, PCMCI+ is a special case of FANTOM, obtained

²<https://causalnex.readthedocs.io/en/latest/>

by setting $K = 1$, using linear or non linear component functions $f^{i,1}$, fixing the noise scaling to $g^{i,1} = 1$ and allowing $\epsilon_t^{i,1}$ to follow any distribution. For the comparison, we use publicly available package Tigramite³.

Rhino [15]. Gong et al. propose the first structural equation models with historically dependent noise, where noise variance depends solely on time-lagged variables, the Rhino’s SEM is as follow:

$$x_t^i = f^i(\mathbf{Pa}_G^i(< t), \mathbf{Pa}_G^i(t)) + g^i(\mathbf{Pa}_G^i(< t), \epsilon_t^i), \quad (13)$$

where $\epsilon_t^i \sim \mathcal{N}(0, 1)$. Rhino neglects heteroscedasticity, and assumes a single stationary regime governed by a one causal graph. By our SEM proposed in Eq(1), we can recover the Rhino’s SEM by setting $K = 1$ and making $g^{i,1}$ a function of only time lagged parents. In Rhino, they took a non linear transformation of normal noise which is equivalent in our case to allowing $\epsilon_t^{i,1}$ to follow any distribution. To compare with Rhino, we used the open package *causica*⁴.

RPCMCI [45]. RPCMCI learns regime indices and time lagged causal relationships from multi-regime MTS. FANTOM’s SEM in Eq(1) generalize it. We can recover RPCMCI settings by making $f^{i,r}$ depends only on time lagged relations and $g^{i,r} = 1$. For the comparison, we use publicly available package Tigramite⁵.

CASTOR [39]. CASTOR learns number of regimes their indices and also their corresponding DAGs including instantaneous and time lagged causal relationships from multi-regime MTS. But they assume that they only have gaussian noise with equivariance. FANTOM’s SEM in Eq(1) generalize it. We can recover CASTOR settings by making $g^{i,r} = 1$ and $\epsilon_t^{i,r} \sim \mathcal{N}(0, 1)$. For the comparison, we use publicly available code CASTOR⁶.

E.3 Optimization parameters

Heteroscedastic settings. Unless noted (i.e., in the synthetic-data study), we set the model lag to the true value of 1 and allow FANTOM to capture instantaneous effects. The variational posterior $q_{\phi^r}(\mathcal{G}^r)$ is initialized to prefer sparse graphs (edge probability < 0.5). Heteroscedastic noise is modeled with conditional normalizing flows (CNFs). Every neural block is a two-layer MLP with 32 hidden units, residual connections, and layer normalization.

Gradients for discrete edges are estimated via the Gumbel–Softmax trick, using a hard forward pass and a soft backward pass with temperature 0.25. All spline flows employ 128 bins, and each transformation uses an embedding dimension equal to the number of nodes.

The sparsity penalty is fixed at $\lambda_s = 50$. For graphs with 10 or 20 nodes we use $\rho = 1$ and $\alpha = 0$, whereas for 40 nodes we set $\rho = 0.001$. Models are optimized with Adam [30] at a learning rate of 0.005. We establish $\zeta = 900$ as the minimum regime duration, and we use 1000 as initial window size.

Homoscedastic non-Gaussian settings. Unless noted (i.e., in the synthetic-data study), we set the model lag to the true value of 1 and allow FANTOM to capture instantaneous effects. We use the same parameters as the heteroscedastic settings. The main difference is the use of a composite of affine-spline transformation. All spline flows employ 16 bins, and each transformation uses an embedding dimension equal to the number of nodes.

The sparsity penalty is fixed at $\lambda_s = 5$. For graphs with 10 or 20 nodes we use $\rho = 1$ and $\alpha = 0$, whereas for 40 nodes we set $\rho = 0.001$. Models are optimized with Adam [30] at a learning rate of 0.005.

E.4 Real world data

E.4.1 Causal Chambers data

We use the wind tunnel datasets from Gamella et al. [13], featuring two controllable fans pushing air through a chamber, barometers measuring air pressure at various locations, and a hatch controlling an

³<https://jakobrunge.github.io/tigramite/>

⁴<https://github.com/microsoft/causica/blob/main/README.md>

⁵<https://jakobrunge.github.io/tigramite/>

⁶<https://github.com/arahmani1/CASTOR>

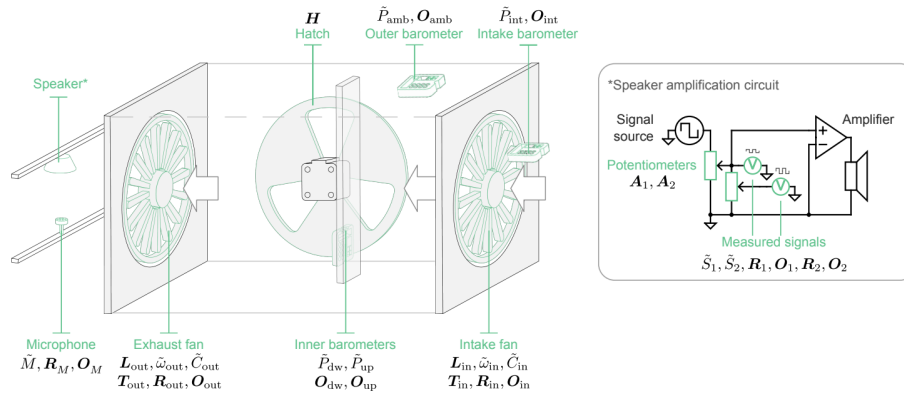


Figure 7: Figure taken from Gamella et al. [13]. Diagrams of the wind tunnel causal chamber and its main components, including the amplification circuit that drives the speaker of the wind tunnel. The variables measured by the chamber are displayed in black math print. Sensor measurements are denoted by a tilde. Manipulable variables, that is, actuators and sensor parameters, are shown in bold symbols.

external opening. The tunnel is a chamber with two controllable fans that push air through it and barometers that measure air pressure at different locations. A hatch precisely controls the area of an additional opening to the outside (see Figure 7). The dataset comprises 16 variables: controllable load of the two fans L_{in}, L_{out} , their measurable speed $(\tilde{\omega}_{in}, \tilde{\omega}_{out})$, the current draw by the fans $(\tilde{C}_{in}, \tilde{C}_{out})$ ($\tilde{C}_{in}, \tilde{C}_{out}$), the resulting air pressure inside the chamber ($\tilde{P}_{dw}, \tilde{P}_{up}$) or at its intake ($\tilde{P}_{int}$), and the hatch H . In the circuit that drives the speaker, we can manipulate the potentiometers (A_1, A_2) that control the amplification, monitoring the resulting signal at different points of the circuit ($\tilde{S}_1, \tilde{S}_2$) and through the microphone output ($\tilde{M}$). $\tilde{P}_{amb}$ is the ambient pressure measure by the outer barometer.

We evaluate all the models on two regimes of 10,000 samples each: the first is observational, while the second involves soft interventions on five variables:

- T_{out} , the resolution of the tachometer timer that measures the elapsed time between successive revolutions of the fan. Choosing microseconds yields a higher resolution in the fan-speed measurement. Hence intervention on T_{out} yields to a change on $\tilde{\omega}_{out}$.
- O_{up} the oversampling rates when taking measurements of the current ($\tilde{C}_{in}, \tilde{C}_{out}$), amplifier ($\tilde{S}_1, \tilde{S}_2$) and microphone signals ($\tilde{M}$), and of air pressure at the different barometers ($\tilde{P}_{up}, \tilde{P}_{dw}, \tilde{P}_{amb}, \tilde{P}_{int}$).
- R_{in}, R_{out}, R_2 the reference voltages, in volts, of the sensors used to measure the current ($\tilde{C}_{in}, \tilde{C}_{out}$), and amplifier ($\tilde{S}_2$), respectively.

For the training procedure, we start by a window of 6000 samples, which gives us three different initial regimes then FANTOM converges smoothly to the exact number of regimes $K = 2$. We set our lag to 8 time lagged and we allow the presence of instantaneous parents. Regarding the parameters, we use a sparsity coefficient equal to 50, spline of 8 bins and MLPs of size 32.

We compare FANTOM to the baselines mentioned in the main text, with results in Table 2. FANTOM is the only model that detects the regime with 99.9% accuracy and outperforms all baselines on the graph learning task, achieving 38.5% on F1 score. Notably, FANTOM surpasses all the models tailored for stationary MTS, even when they are given the ground-truth regime partitions.

E.4.2 Epilepsy data

Huizenga et al. [22] show that scalp potential fields are contaminated by heteroscedastic noise in EEG measurements. We evaluate FANTOM's performance in detecting epileptic regimes using EEG

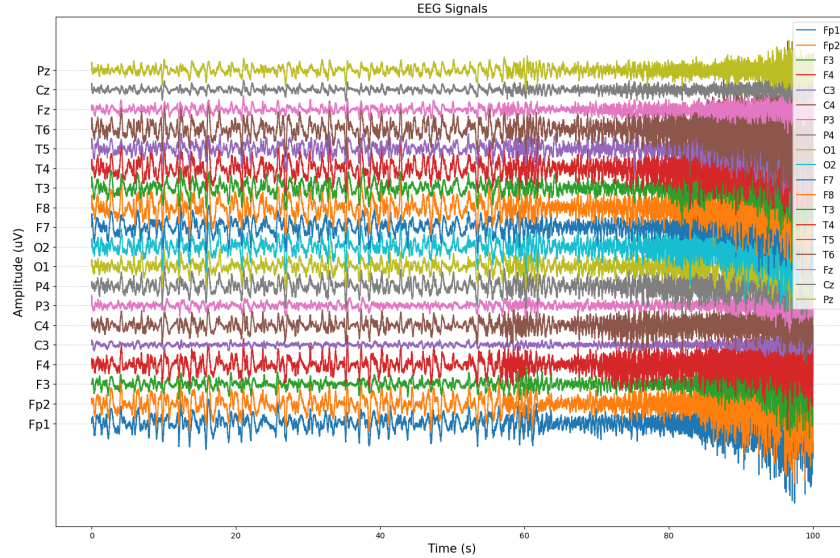


Figure 8: EEG signals for patient of id 7170, session 1 in the TUSZ data.

signals from 10 different patients in the Temple University Hospital EEG Seizure Corpus (TUSZ) dataset [52]. The dataset encompasses multiple seizure types; in this study, we focus on generalized seizures, which engage the entire brain. Each patient's record contains scalp EEG signals from 19 channels (Figure 8), each considered a causal variable.

State-of-the-art methods [52] use graph neural networks (GNNs) for seizure detection, typically building a fixed distance graph and feeding Fast Fourier Transform (FFT) coefficients from 10–12 seconds EEG windows as node features. These approaches (i) train a single model for all patients, offering no personalization; (ii) cannot operate in zero-shot or unsupervised settings; and (iii) reuse an identical graph for both seizure and normal periods. FANTOM addresses these limitations: it detects seizures in a personalized manner without any training data and learns distinct temporal causal graphs for normal and seizure states.

Preprocessing. We apply FANTOM to 10 different patients from TUSZ dataset, we treat this as an unsupervised regime detection problem, analyzing roughly 100 seconds of recordings at a 250 Hz sampling rate for each patient, capturing both normal and seizure states. Before running FANTOM, we filter out the EEG signals by a band pass filter of order 6, the lower frequency is 0.5Hz while the highest frequency is 50Hz.

Table 3: Seizure detection accuracy using FANTOM for 10 different patients

Patient id	Regime Acc
0002	84.8
0021	80.8
0302	76.6
0492	86.6
6440	86.1
6520	80.2
7128	94.1
7170	81.1
7936	82.0
8303	75.1
Avg	82.7 ± 5.2

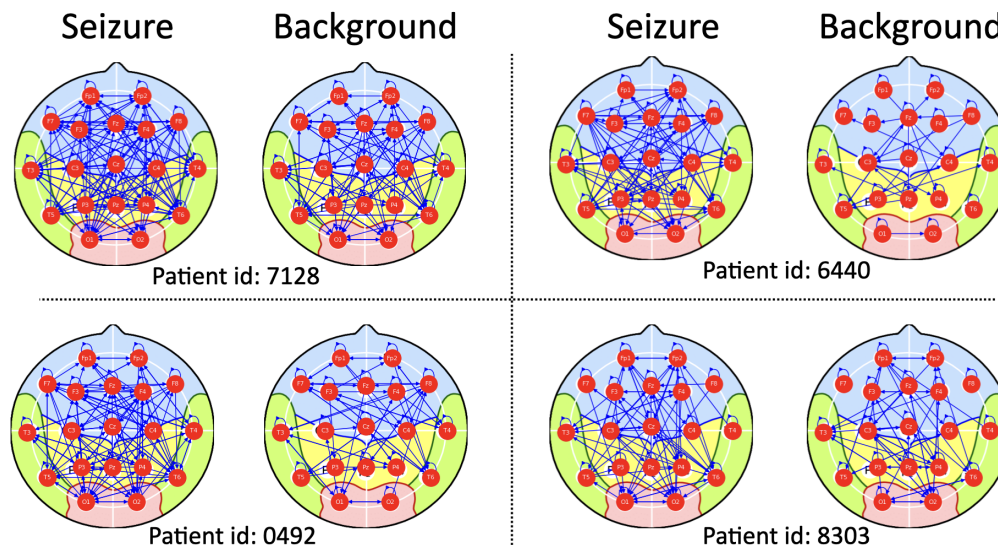


Figure 9: Figure illustrates the summary causal graph per regime learned by FANTOM for different patients

We segment each EEG recording into fixed 12 s windows (3000 samples). FANTOM is initialized with eight initial regimes and reliably converges to the two ground-truth states; quantitative results appear in Table 3. We employ a temporal lag of eight samples and allow instantaneous parental links. Averaged over all patients, FANTOM achieves 82.7% regime-assignment accuracy. The graph learned for the seizure state is substantially denser and more interconnected than that of the normal state (Figure 9), consistent with the widespread neural involvement of generalized seizures.

F Additional Experiments

F.1 Ablation studies

F.1.1 Impact of Heteroscedastic Modeling and Flow Architecture

To assess the contribution of heteroscedastic modelling and the flow architecture, we ran an ablation on MTS comparing three FANTOM variants: (i) Gaussian output (no flow, FANTOM-Gauss), (ii) a simple coupling-spline flow (FANTOM-Spline), and (iii) the full model with a conditional normalizing flow (CNF, FANTOM). In the single-regime setting, removing or simplifying the CNF consistently degrades performance. In the multi-regime setting, only the CNF variant remains stable and converges to the two ground truth regimes; the alternatives collapse to one regime. Results are reported in the table below:

Table 4: Importance of heteroscedastic modelling and flow architecture for 10 nodes. - means the method collapses to 1 regime.

Heteroscedastic noise						
K	model	Inst		Lag		Regime
		SHD	F1	SHD	F1	Accuracy
1	FANTOM-Gauss	39	0.0	44	42.5	-
	FANTOM-Spline	17	0.0	18	0	-
	FANTOM	0	100	0	100	-
2	FANTOM-Gauss	-	-	-	-	-
	FANTOM-Spline	-	-	-	-	-
	FANTOM	1	97.8	1	98.9	98.6

F.1.2 Robustness to the pruning step, minimum regime duration, and regime initialization

We evaluated robustness to the pruning step, minimum regime duration (ζ), and regime initialization. Across these variations, FANTOM’s predictive performance remains stable. We find that changing (window size, ζ) primarily affects runtime, especially at long horizons, without materially impacting graph and regime learning performances. To show efficiency under long-horizon conditions, we used regimes with lengths [2000, 3000, 2000, 2500].

Table 5: Performance of FANTOM with varying hyperparameters (window size and ζ) on 10 node graphs with 2, 3, and 4 regimes.

Heteroscedastic noise								
(window, ζ)	K	Numb	Running	Inst.		Lag		Regime
				SHD	F1	SHD	F1	Acc.
(1000,900)	2	4	17’8s	5	91.2	13	85.7	96.5
(1500,1200)	2	4	10’12s	1	97.9	9	89.8	95.1
(1000,900)	3	4	30’33s	5	94.2	8	94.4	96.4
(1500,1200)	3	4	16’2s	3	96.7	10	91.9	91.8
(1000,900)	4	4	40’2s	5	95.8	12	92.7	92.9
(1500,1200)	4	4	25’38s	3	97.5	16	90.5	91.5

Table 6: Performance of FANTOM with varying hyperparameters (window size and ζ) on 40 node graphs with 2, 3, and 4 regimes.

Heteroscedastic noise								
(window, ζ)	K	Numb	Running	Inst.		Lag		Regime
				SHD	F1	SHD	F1	Acc.
(1000,900)	2	4	22’15s	33	81.9	27	91.9	100
(1500,1200)	2	4	10’20s	35	83.8	25	91.7	100
(1000,900)	3	4	34’25s	49	85.6	58	88.6	99.8
(1500,1200)	3	4	15’52s	43	87.3	72	87.3	99.9
(1000,900)	4	5	66’30s	34	86.5	53	91.9	99.7
(1500,1200)	4	4	21’38s	32	87.1	54	91.8	99.6

From the table, FANTOM scales to 40 nodes and converges to four regimes under different initializations. With a window initialization of 1500 and $\zeta = 1200$, it starts with six initial regimes and converges smoothly to four with 99.6% accuracy in 21min38s. With a window initialization of

1000 and $\zeta = 900$, it starts with nine initial regimes and converges to four with 99.7% accuracy in 66min30s. All rebuttal experiments were run on a Tesla V100-SXM2 (26GB). These results indicate that FANTOM scales without requiring large GPU memory.

To test scalability, we ran FANTOM on a 40 node dataset with four regimes and long horizons (lengths 2000, 3000, 2000, 2500). As expected, additional regimes and longer horizons increase runtime. With 40 nodes and two regimes, training finishes in 10 min 20 s, whereas the four regime setting requires 21 min 38 s.

We showed in the table above (Table 5 and 6) that FANTOM is robust to the initialization parameter (window size that impacts directly the initial regime count and the pruning threshold), e.g. in the case of 40 nodes with 4 regimes, with a window initialization of 1500 and $\zeta = 1200$, it starts with six initial regimes and converges smoothly to four in 21min38s. With a window initialization of 1000 and $\zeta = 900$, it starts with nine initial regimes and converges to four with 99.6% accuracy in 66 min 30s.

F.1.3 Robustness to data standardization

We investigated FANTOM’s robustness to data standardization, a common challenge for most optimization-based causal discovery models. The results in Table 7 clearly demonstrate FANTOM’s resilience. This robustness stems from its use of Bayesian structure learning to estimate a distribution over plausible graphs and conditional normalizing flows to effectively model complex data distributions. In contrast, models such as CASTOR are sensitive to standardization and even fail when provided with ground-truth regime labels.

Table 7: Robustness of FANTOM to data standardization (10 nodes).

Heteroscedastic noise							
model	K	data	Inst.		Lag		Regime
			SHD	F1	SHD	F1	Acc.
CASTOR	2	Raw	90	26.9	28	29.7	x
		standardized	28	0.0	33	0.0	x
FANTOM	2	Raw	4	93.3	7	90.9	97.1
		standardized	5	91.5	7	90.9	91.1
CASTOR	3	Raw	135	24.1	135	29.6	x
		standardized	37	0.0	48	0.0	x
FANTOM	3	Raw	3	96.1	10	91.5	92.2
		standardized	2	97.3	9	92.3	92.2

From the Table 7, FANTOM achieves the same results on raw and standardized data while CASTOR fails in both cases even when we give it access to the ground truth regime labels. In the case of standardized data, CASTOR predicts adjacency matrices full of zeros due to its incapability of handling such scenarios.

F.2 Additional results on synthetic data

F.2.1 Heteroscedastic noise with different number of nodes and regimes

Table 8: Average SHD, F1 scores, NHD and Ratio for different models with $d = 10$ nodes and $K = 2$ regimes. *Split* denotes whether regime separation is automatic ($\checkmark$) or manual ($\times$). *Inst.* refers to instantaneous links, and *Lag* to time-lagged edges.

Heteroscedastic noise, $K = 2$ and $d = 10$										
Model	Split	Inst.				Lag				Regime
		SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓	Acc.
PCMCI+	$\times$	34.3 $\pm$ 3.7	14.2 $\pm$ 2.5	0.09 $\pm$ 0.0	0.85 $\pm$ 0.1	26.0 $\pm$ 4.3	25.6 $\pm$ 4.8	0.07 $\pm$ 0.0	0.74 $\pm$ 0.0	$\times$
Rhino	$\times$	28.0 $\pm$ 5.6	8.56 $\pm$ 4.9	0.09 $\pm$ 0.0	0.92 $\pm$ 0.0	38.5 $\pm$ 6.3	58.4 $\pm$ 11.7	0.13 $\pm$ 0.0	0.43 $\pm$ 0.0	$\times$
DYNOTEARS	$\times$	58.5 $\pm$ 4.9	31.2 $\pm$ 2.8	0.22 $\pm$ 0.0	0.68 $\pm$ 0.0	79.5 $\pm$ 3.5	33.9 $\pm$ 4.4	0.27 $\pm$ 0.0	0.66 $\pm$ 0.0	$\times$
CASTOR	$\times$	90.0 $\pm$ 0.0	28.1 $\pm$ 2.9	0.38 $\pm$ 0.0	0.74 $\pm$ 0.0	90.5 $\pm$ 3.2	35.6 $\pm$ 3.4	0.42 $\pm$ 0.0	0.64 $\pm$ 0.0	$\times$
RPCMCI	$\checkmark$	-	-	-	-	-	-	-	-	-
CASTOR	$\checkmark$	-	-	-	-	-	-	-	-	-
FANTOM	$\checkmark$	4.67 $\pm$ 1.5	91.7 $\pm$ 3.2	0.005 $\pm$ 0.0	0.04 $\pm$ 0.0	11.6 $\pm$ 1.1	88.2 $\pm$ 2.1	0.02 $\pm$ 0.0	0.08 $\pm$ 0.0	96.6 $\pm$ 1.1

Table 9: Average SHD, F1 scores, NHD and Ratio for different models with $d = 10$ nodes and $K = 3$ regimes. *Split* denotes whether regime separation is automatic ($\checkmark$) or manual ($\times$). *Inst.* refers to instantaneous links, and *Lag* to time-lagged edges.

Heteroscedastic noise, $K = 3$ and $d = 10$										
Model	Split	Inst.				Lag				Regime
		SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓	Acc.
PCMCI+	$\times$	46.1 $\pm$ 4.7	11.1 $\pm$ 2.1	0.05 $\pm$ 0.0	0.88 $\pm$ 0.0	46.0 $\pm$ 0.8	19.0 $\pm$ 3.3	0.05 $\pm$ 0.0	0.80 $\pm$ 0.0	$\times$
Rhino	$\times$	44.5 $\pm$ 6.5	5.11 $\pm$ 1.9	0.06 $\pm$ 0.0	0.94 $\pm$ 0.0	53.5 $\pm$ 1.5	64.7 $\pm$ 4.6	0.07 $\pm$ 0.0	0.35 $\pm$ 0.0	$\times$
DYNOTEARS	$\times$	89.5 $\pm$ 3.5	31.5 $\pm$ 0.4	0.14 $\pm$ 0.0	0.68 $\pm$ 0.0	118.0 $\pm$ 4.0	37.5 $\pm$ 1.2	0.17 $\pm$ 0.0	0.61 $\pm$ 0.0	$\times$
CASTOR	$\times$	104 $\pm$ 3.7	23.4 $\pm$ 0.9	0.19 $\pm$ 0.0	0.76 $\pm$ 0.0	133.5 $\pm$ 2.8	34.8 $\pm$ 1.2	0.24 $\pm$ 0.0	0.64 $\pm$ 0.0	$\times$
RPCMCI	$\checkmark$	-	-	-	-	-	-	-	-	-
CASTOR	$\checkmark$	-	-	-	-	-	-	-	-	-
FANTOM	$\checkmark$	5.67 $\pm$ 3.3	93.3 $\pm$ 4.2	0.006 $\pm$ 0.0	0.06 $\pm$ 0.0	12.3 $\pm$ 6.3	90.9 $\pm$ 1.0	0.012 $\pm$ 0.0	0.08 $\pm$ 0.0	97.1 $\pm$ 0.0

Under heteroscedastic conditions, $d = 10$ node graphs with either two or three regimes, FANTOM outperforms every baseline. Among methods explicitly designed for multi-regime MTS, it is the only one that converges to the true regime partitions, attaining 97.1% (Table 9) in regime detection accuracy. CASTOR and RPCMCI break down in heteroscedastic settings.

To give stationary MTS baselines the best possible chance, we supply them with the ground truth regime partitions. This is done by training the aforementioned models on each pure regime separately (regime governed by the same causal model). Yet, FANTOM still dominates the structure learning task, while learning the number of regime and their indices as well, achieving an F1 of 93.3 % and an NHD of 0.006 (Table 9). Because NHD penalizes every missing, extra, or mis-oriented edge, a value of 0.006 implies that FANTOM not only recovers the graph skeleton but orients edges with high precision.

Table 10: Average SHD, F1 scores, NHD and Ratio for different models with $d = 20$ nodes and $K = 2$ regimes. *Split* denotes whether regime separation is automatic ($\checkmark$) or manual ($\times$). *Inst.* refers to instantaneous links, and *Lag* to time-lagged edges.

Heteroscedastic noise, $K = 2$ and $d = 20$										
Model	Split	Inst.				Lag				Regime
		SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓	Acc.
PCMCI+	$\times$	84.0 $\pm$ 17.	39.7 $\pm$ 2.6	0.03 $\pm$ 0.0	0.6 $\pm$ 0.0	42.5 $\pm$ 12.	76.9 $\pm$ 1.0	0.14 $\pm$ 0.0	0.22 $\pm$ 0.0	$\times$
Rhino	$\times$	70.0 $\pm$ 6.0	1.26 $\pm$ 1.2	0.04 $\pm$ 0.0	0.98 $\pm$ 0.0	188.0 $\pm$ 30.	26.3 $\pm$ 9.8	0.14 $\pm$ 0.0	0.73 $\pm$ 0.0	$\times$
DYNOTEARS	$\times$	221.5 $\pm$ 9.5	26.9 $\pm$ 1.2	0.22 $\pm$ 0.0	0.72 $\pm$ 0.0	45.0 $\pm$ 7.0	61.5 $\pm$ 3.9	0.02 $\pm$ 0.0	0.38 $\pm$ 0.0	$\times$
CASTOR	$\times$	377.5 $\pm$ 1.5	14.9 $\pm$ 0.1	0.41 $\pm$ 0.0	0.84 $\pm$ 0.0	379.5 $\pm$ 1.5	19.1 $\pm$ 1.0	0.41 $\pm$ 0.0	0.80 $\pm$ 0.0	$\times$
RPCMCI	$\checkmark$	-	-	-	-	-	-	-	-	-
CASTOR	$\checkmark$	-	-	-	-	-	-	-	-	-
FANTOM	$\checkmark$	9.00 $\pm$ 3.0	89.1 $\pm$ 5.7	0.006 $\pm$ 0.0	0.10 $\pm$ 0.0	26.0 $\pm$ 5.0	85.4 $\pm$ 1.3	0.01 $\pm$ 0.0	0.14 $\pm$ 0.0	97.8 $\pm$ 0.2

Table 11: Average SHD, F1 scores, NHD and Ratio for different models with $d = 20$ nodes and $K = 3$ regimes. *Split* denotes whether regime separation is automatic ($\checkmark$) or manual ($\times$). *Inst.* refers to instantaneous links, and *Lag* to time-lagged edges.

Heteroscedastic noise, $K = 3$ and $d = 20$										
Model	Split	Inst.				Lag				Regime
		SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓	
PCMCi+	$\times$	83.5 $\pm$ 15.	39.9 $\pm$ 1.6	0.03 $\pm$ 0.0	0.59 $\pm$ 0.0	38.0 $\pm$ 2.0	77.8 $\pm$ 0.9	0.01 $\pm$ 0.0	0.22 $\pm$ 0.0	$\times$
Rhino	$\times$	108.5 $\pm$ 11.	1.6 $\pm$ 1.2	0.02 $\pm$ 0.0	0.98 $\pm$ 0.0	292.0 $\pm$ 62.	25.8 $\pm$ 8.9	0.09 $\pm$ 0.0	0.74 $\pm$ 0.0	$\times$
DYNOTEARS	$\times$	325. $\pm$ 14.	27.6 $\pm$ 1.0	0.14 $\pm$ 0.0	0.72 $\pm$ 0.0	55.5 $\pm$ 0.5	69.8 $\pm$ 7.6	0.01 $\pm$ 0.0	0.29 $\pm$ 0.0	$\times$
CASTOR	$\times$	412.3 $\pm$ 2.5	13.8 $\pm$ 0.0	0.20 $\pm$ 0.0	0.86 $\pm$ 0.0	570. $\pm$ 2.0	17.5 $\pm$ 0.4	0.28 $\pm$ 0.0	0.81 $\pm$ 0.0	$\times$
RPCMCI	$\checkmark$	-	-	-	-	-	-	-	-	-
CASTOR	$\checkmark$	-	-	-	-	-	-	-	-	-
FANTOM	$\checkmark$	13.5$\pm$6.5	88.2$\pm$6.2	0.05$\pm$0.0	0.11$\pm$0.0	46.5$\pm$9.5	84.6$\pm$2.5	0.01$\pm$0.0	0.14$\pm$0.0	97.8$\pm$1.4

Under heteroscedastic conditions, $d = 20$ node graphs with either two or three regimes, FANTOM outperforms all the baselines. Among methods explicitly designed for multi-regime MTS, it is the only one that converges to the true regime partitions, attaining 97.8% in regime detection accuracy (Table 10). CASTOR and RPCMCI break down in heteroscedastic settings, also in the case of 20 nodes.

To give stationary MTS baselines the best possible chance, we supply them with the ground truth regime partitions. This is done by training the aforementioned models on each pure regime separately (regime governed by the same causal model). Yet, FANTOM still dominates the structure learning task, while learning the number of regime and their indices as well, achieving an F1 of 89.1 % and an NHD of 0.006 for instantaneous links and an F1 of 85.4% and an NHD 0.01% for time lagged (Table 10).

Table 12: Average SHD, F1 scores, NHD and Ratio for different models with $d = 40$ nodes and $K = 2$ regimes. *Split* denotes whether regime separation is automatic ($\checkmark$) or manual ($\times$). *Inst.* refers to instantaneous links, and *Lag* to time-lagged edges.

Heteroscedastic noise, $K = 2$ and $d = 40$										
Model	Split	Inst.				Lag				Regime
		SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓	
PCMCi+	$\times$	146.5 $\pm$ 5.0	25.8 $\pm$ 0.1	0.02 $\pm$ 0.0	0.74 $\pm$ 0.0	109.0 $\pm$ 15.	58.0 $\pm$ 4.5	0.01 $\pm$ 0.0	0.42 $\pm$ 0.0	$\times$
Rhino	$\times$	137.0 $\pm$ 7.1	0.0 $\pm$ 0.0	0.02 $\pm$ 0.0	1.00 $\pm$ 0.0	700.0 $\pm$ 87.	28.2 $\pm$ 5.7	0.12 $\pm$ 0.0	0.71 $\pm$ 0.0	$\times$
DYNOTEARS	$\times$	137.5 $\pm$ 9.2	17.1 $\pm$ 1.4	0.020 $\pm$ 0.0	0.82 $\pm$ 0.0	129.0 $\pm$ 15.	36.6 $\pm$ 4.9	0.01 $\pm$ 0.0	0.63 $\pm$ 0.0	$\times$
CASTOR	$\times$	151.0 $\pm$ 11.	0.0 $\pm$ 0.0	0.03 $\pm$ 0.0	1.00 $\pm$ 0.0	333.0 $\pm$ 9.3	19.6 $\pm$ 5.6	0.050 $\pm$ 0.0	0.80 $\pm$ 0.0	$\times$
RPCMCI	$\checkmark$	-	-	-	-	-	-	-	-	-
CASTOR	$\checkmark$	-	-	-	-	-	-	-	-	-
FANTOM	$\checkmark$	27.00$\pm$6.0	85.6$\pm$3.7	0.005$\pm$0.0	0.14$\pm$0.0	26.5$\pm$0.5	91.9$\pm$1.0	0.004$\pm$0.0	0.08$\pm$0.0	99.9$\pm$0.1

Table 13: Average SHD, F1 scores, NHD and Ratio for different models with $d = 40$ nodes and $K = 3$ regimes. *Split* denotes whether regime separation is automatic ($\checkmark$) or manual ($\times$). *Inst.* refers to instantaneous links, and *Lag* to time-lagged edges.

Heteroscedastic noise, $K = 3$ and $d = 40$										
Model	Split	Inst.				Lag				Regime
		SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓	
PCMCi+	$\times$	222.0 $\pm$ 1.4	25.2 $\pm$ 0.4	0.01 $\pm$ 0.0	0.75 $\pm$ 0.0	142.0 $\pm$ 17.	62.4 $\pm$ 1.0	0.01 $\pm$ 0.0	0.37 $\pm$ 0.0	$\times$
Rhino	$\times$	210.5 $\pm$ 9.2	0.0 $\pm$ 0.0	0.01 $\pm$ 0.0	1.00 $\pm$ 0.0	1005. $\pm$ 146.	29.3 $\pm$ 5.4	0.08 $\pm$ 0.0	0.7 $\pm$ 0.0	$\times$
DYNOTEARS	$\times$	203.0 $\pm$ 14.	15.2 $\pm$ 2.2	0.01 $\pm$ 0.0	0.85 $\pm$ 0.0	161.0 $\pm$ 1.4	51.3 $\pm$ 15.	0.01 $\pm$ 0.0	0.49 $\pm$ 0.1	$\times$
CASTOR	$\times$	224.0 $\pm$ 10.2	0.00 $\pm$ 0.0	0.01 $\pm$ 0.0	1.00 $\pm$ 0.0	501.0 $\pm$ 21.	23.1 $\pm$ 1.2	0.03 $\pm$ 0.0	0.76 $\pm$ 0.0	$\times$
RPCMCI	$\checkmark$	-	-	-	-	-	-	-	-	-
CASTOR	$\checkmark$	-	-	-	-	-	-	-	-	-
FANTOM	$\checkmark$	52.0$\pm$9.0	82.2$\pm$5.1	0.004$\pm$0.0	0.19$\pm$0.0	65.0$\pm$7.0	87.9$\pm$0.6	0.004$\pm$0.0	0.11$\pm$0.0	99.8$\pm$0.0

Under heteroscedastic conditions, $d = 40$ node graphs with either two or three regimes, FANTOM outperforms every baseline and shows that it can scale to large graphs even in this complex setting. Among methods explicitly designed for multi-regime MTS, it is the only one that converges to the true regime partitions, attaining 99.9% in regime detection accuracy. CASTOR and RPCMCI break

down in heteroscedastic settings (Table 12). Large scale graphs helps FANTOM to differentiate between the different regimes and increases the regime detection accuracy by 2% compared to 20 node graphs settings.

To give stationary MTS baselines the best possible chance, we supply them with the ground truth regime partitions. This is done by training the aforementioned models on each pure regime separately (regime governed by the same causal model). Yet, FANTOM still dominates the structure learning task, while learning the number of regime and their indices as well, achieving an F1 of 85.6 % and an NHD of 0.14 (Table 12).

F2.2 Non-Gaussian noise with different number of nodes and regimes

Table 14: Average SHD, F1 scores, NHD and Ratio for different models with $d = 10$ nodes and $K = 2$ regimes. *Split* denotes whether regime separation is automatic ($\checkmark$) or manual ($\times$). *Inst.* refers to instantaneous links, and *Lag* to time-lagged edges.

Homoscedastic non-Gaussian noise, $K = 2$ and $d = 10$										
Model	Split	Type	Inst.				Lag			
			SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓
PCMCi+	$\times$	W	6.00 $\pm$ 0.0	88.1 $\pm$ 0.5	0.01 $\pm$ 0.0	0.12 $\pm$ 0.0	8.00 $\pm$ 5.0	90.7 $\pm$ 6.1	0.01 $\pm$ 0.0	0.09 $\pm$ 0.0
Rhino	$\times$	W	2.50 $\pm$ 0.5	96.0 $\pm$ 0.1	0.004 $\pm$ 0.0	0.04 $\pm$ 0.0	4.00 $\pm$ 1.0	96.0 $\pm$ 0.4	0.006 $\pm$ 0.0	0.04 $\pm$ 0.0
DYNOTEARS	$\times$	W	42.0 $\pm$ 11.	51.8 $\pm$ 7.9	0.12 $\pm$ 0.0	0.48 $\pm$ 0.0	8.00 $\pm$ 1.0	86.8 $\pm$ 0.5	0.02 $\pm$ 0.0	0.12 $\pm$ 0.0
CASTOR	$\times$	W	17.0 $\pm$ 3.0	62.6 $\pm$ 8.5	0.055 $\pm$ 0.0	0.37 $\pm$ 0.0	11.0 $\pm$ 1.0	85.2 $\pm$ 0.7	0.02 $\pm$ 0.0	0.15 $\pm$ 0.0
RPCMCI	$\checkmark$	W	-	-	-	-	-	-	-	-
CASTOR	$\checkmark$	W	34.0 $\pm$ 20.	42.5 $\pm$ 27.	0.09 $\pm$ 0.0	0.57 $\pm$ 0.2	45.0 $\pm$ 31.	47.0 $\pm$ 25.	0.13 $\pm$ 0.1	0.52 $\pm$ 0.2
FANTOM	$\checkmark$	W	1.00 $\pm$ 0.0	98.2 $\pm$ 0.1	0.002 $\pm$ 0.0	0.01 $\pm$ 0.0	8.00 $\pm$ 4.0	91.0 $\pm$ 4.9	0.02 $\pm$ 0.0	0.08 $\pm$ 0.05
			SHD↓		F1↑		NHD↓		Ratio↓	
									Acc.	
CD-NOD	$\times$	S	33.0 $\pm$ 3.0		47.0 $\pm$ 5.9		0.41 $\pm$ 0.0		0.52 $\pm$ 0.0	$\times$
FANTOM	$\checkmark$	S	7.5 $\pm$ 2.5		93.1 $\pm$ 2.2		0.07 $\pm$ 0.0		0.06 $\pm$ 0.0	99.6 $\pm$ 0.0

Table 15: Average SHD, F1 scores, NHD and Ratio for different models with $d = 10$ nodes and $K = 3$ regimes. *Split* denotes whether regime separation is automatic ($\checkmark$) or manual ($\times$). *Inst.* refers to instantaneous links, and *Lag* to time-lagged edges.

Homoscedastic non-Gaussian noise, $K = 3$ and $d = 10$										
Model	Split	Type	Inst.				Lag			
			SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓
PCMCi+	$\times$	W	17.5 $\pm$ 2.1	74.9 $\pm$ 5.0	0.02 $\pm$ 0.0	0.24 $\pm$ 0.0	14.5 $\pm$ 2.1	88.3 $\pm$ 1.6	0.01 $\pm$ 0.0	0.11 $\pm$ 0.0
Rhino	$\times$	W	2.50 $\pm$ 0.7	96.8 $\pm$ 1.3	0.002 $\pm$ 0.0	0.03 $\pm$ 0.0	6.00 $\pm$ 1.4	95.2 $\pm$ 1.6	0.006 $\pm$ 0.0	0.04 $\pm$ 0.0
DYNOTEARS	$\times$	W	42.0 $\pm$ 33.	54.4 $\pm$ 11.2	0.06 $\pm$ 0.0	0.45 $\pm$ 0.1	21.0 $\pm$ 1.4	82.1 $\pm$ 1.6	0.02 $\pm$ 0.0	0.17 $\pm$ 0.0
CASTOR	$\times$	W	22.0 $\pm$ 2.8	66.2 $\pm$ 2.5	0.030 $\pm$ 0.0	0.33 $\pm$ 0.0	17.0 $\pm$ 4.2	84.4 $\pm$ 1.8	0.01 $\pm$ 0.0	0.15 $\pm$ 0.0
RPCMCI	$\checkmark$	W	-	-	-	-	-	-	-	-
CASTOR	$\checkmark$	W	47.0 $\pm$ 17.	34.8 $\pm$ 30.	0.05 $\pm$ 0.0	0.65 $\pm$ 0.2	59.5 $\pm$ 31.	39.4 $\pm$ 19.	0.07 $\pm$ 0.0	0.60 $\pm$ 0.2
FANTOM	$\checkmark$	W	0.33 $\pm$ 0.4	99.5 $\pm$ 0.6	0.00 $\pm$ 0.0	0.00 $\pm$ 0.0	12.5 $\pm$ 6.8	89.1 $\pm$ 3.7	0.01 $\pm$ 0.0	0.10 $\pm$ 0.0
			SHD↓		F1↑		NHD↓		Ratio↓	
									Acc.	
CD-NOD	$\times$	S	42.5 $\pm$ 2.5		31.8 $\pm$ 1.1		0.42 $\pm$ 0.0		0.67 $\pm$ 0.0	$\times$
FANTOM	$\checkmark$	S	4.5 $\pm$ 2.0		95.6 $\pm$ 1.5		0.04 $\pm$ 0.0		0.04 $\pm$ 0.0	96.6 $\pm$ 0.2

Under homoscedastic, non-Gaussian noise with $d = 10$ node graphs and either two or three regimes, FANTOM outperforms every baseline on instantaneous-link inference, reaching an F1 of 98.2 % and an NHD of 0.002. Among methods explicitly designed for multi-regime MTS, both FANTOM and CASTOR recover the exact number of regimes, whereas RPCMCI fails to converge to the true partitions. FANTOM further surpasses CASTOR in regime detection (98.6 % vs. 77.0 %) and in DAG learning (F1 = 98.2 % vs. 42.5 %).

To give the stationary-MTS baselines their best chance, we supply them with the ground-truth regime labels and train each model on the corresponding pure regime. Even in this favorable setting, only Rhino exceeds FANTOM on time-lagged links, achieving an F1 of 96.0 % compared with FANTOM's 91.0 %.

Table 16: Average SHD, F1 scores, NHD and Ratio for different models with $d = 20$ nodes and $K = 2$ regimes. *Split* denotes whether regime separation is automatic ($\checkmark$) or manual ($\times$). *Inst.* refers to instantaneous links, and *Lag* to time-lagged edges.

Homoscedastic non-Gaussian noise, $K = 2$ and $d = 20$											
Model	Split	Type	Inst.				Lag				Regime
			SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓	Acc.
PCMCI+ Rhino DYNOTEARS CASTOR	×	W	46.0 $\pm$ 2.8	54.9 $\pm$ 0.6	0.03 $\pm$ 0.0	0.45 $\pm$ 0.0	17.0 $\pm$ 4.2	88.7 $\pm$ 0.5	0.009 $\pm$ 0.0	0.11 $\pm$ 0.0	×
	×	W	17.5 $\pm$ 14.	82.5 $\pm$ 16.	0.007 $\pm$ 0.0	0.17 $\pm$ 0.1	29.5 $\pm$ 3.5	83.5 $\pm$ 2.1	0.01 $\pm$ 0.0	0.16 $\pm$ 0.0	×
	×	W	44.5 $\pm$ 3.5	44.9 $\pm$ 3.1	0.03 $\pm$ 0.0	0.55 $\pm$ 0.0	46.5 $\pm$ 12.	55.8 $\pm$ 5.2	0.025 $\pm$ 0.0	0.44 $\pm$ 0.0	×
	×	W	139.5 $\pm$ 13.	41.8 $\pm$ 5.3	0.10 $\pm$ 0.0	0.58 $\pm$ 0.0	186.0 $\pm$ 28.	40.1 $\pm$ 2.3	0.13 $\pm$ 0.0	0.60 $\pm$ 0.0	×
RPCMCI	✓	W	-	-	-	-	-	-	-	-	-
CASTOR	✓	W	68.0 $\pm$ 8.5	37.2 $\pm$ 12.	0.04 $\pm$ 0.0	0.62 $\pm$ 0.1	86.0 $\pm$ 9.9	37.4 $\pm$ 7.8	0.06 $\pm$ 0.0	0.64 $\pm$ 0.0	46.6 $\pm$ 19.
FANTOM	✓	W	3.50 $\pm$ 1.5	97.2 $\pm$ 1.2	0.002 $\pm$ 0.0	0.02 $\pm$ 0.0	11.5 $\pm$ 2.5	93.1 $\pm$ 1.6	0.006 $\pm$ 0.0	0.06 $\pm$ 0.0	100. $\pm$ 0.0
			SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓	Acc.
CD-NOD	×	S	106. $\pm$ 4.0		26.0 $\pm$ 4.1		0.31 $\pm$ 0.0		0.73 $\pm$ 0.0		×
FANTOM	✓	S	4.0 $\pm$ 0.0		98.3 $\pm$ 0.0		0.01 $\pm$ 0.0		0.01 $\pm$ 0.0		100. $\pm$ 0.0

Under homoscedastic, non-Gaussian noise with $d = 20$ node graphs and two regimes, FANTOM outperforms every baseline on instantaneous-link and time lagged link inference, reaching an F1 of 97.2 % and an NHD of 0.002 for instantaneous links and an F1 of 93.1% on time lagged relationships. Among methods explicitly designed for multi-regime MTS, both FANTOM and CASTOR recover the exact number of regimes, whereas RPCMCI fails to converge to the true partitions. FANTOM further surpasses CASTOR in regime detection (100. % vs. 46.6 %) and in DAG learning (F1 = 97.2 % vs. 37.2 %).

To give the stationary-MTS baselines their best chance, we supply them with the ground-truth regime labels and train each model on the corresponding pure regime. Even in this favorable setting, FANTOM outperforms all the baselines, achieving an F1 of 97.2 %.

Table 17: Average SHD, F1 scores, NHD and Ratio for different models with $d = 20$ nodes and $K = 3$ regimes. *Split* denotes whether regime separation is automatic ($\checkmark$) or manual ($\times$). *Inst.* refers to instantaneous links, and *Lag* to time-lagged edges.

Homoscedastic non-Gaussian noise, $K = 3$ and $d = 20$											
Model	Split	Type	Inst.				Lag				Regime
			SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓	Acc.
PCMCI+ Rhino DYNOTEARS CASTOR	×	W	66.5 $\pm$ 2.1	55.1 $\pm$ 2.5	0.02 $\pm$ 0.0	0.45 $\pm$ 0.0	24.5 $\pm$ 12.	88.9 $\pm$ 3.0	0.007 $\pm$ 0.0	0.11 $\pm$ 0.0	×
	×	W	27.5 $\pm$ 14.	82.1 $\pm$ 11.	0.007 $\pm$ 0.0	0.17 $\pm$ 0.1	50.5 $\pm$ 3.5	81.8 $\pm$ 1.4	0.01 $\pm$ 0.0	0.18 $\pm$ 0.0	×
	×	W	69.0 $\pm$ 4.2	43.6 $\pm$ 3.7	0.02 $\pm$ 0.0	0.56 $\pm$ 0.0	54.5 $\pm$ 10.6	66.3 $\pm$ 2.4	0.01 $\pm$ 0.0	0.34 $\pm$ 0.0	×
	×	W	167.0 $\pm$ 9.9	38.3 $\pm$ 4.6	0.05 $\pm$ 0.0	0.61 $\pm$ 0.0	264.5 $\pm$ 34.	41.2 $\pm$ 1.5	0.08 $\pm$ 0.0	0.58 $\pm$ 0.0	×
RPCMCI	✓	W	-	-	-	-	-	-	-	-	-
CASTOR	✓	W	121.5 $\pm$ 27.	34.3 $\pm$ 9.6	0.03 $\pm$ 0.0	0.65 $\pm$ 0.1	181.5 $\pm$ 17.	33.7 $\pm$ 7.8	0.05 $\pm$ 0.0	0.66 $\pm$ 0.0	79.8 $\pm$ 4.3
FANTOM	✓	W	7.00 $\pm$ 0.0	96.1 $\pm$ 0.0	0.001 $\pm$ 0.0	0.03 $\pm$ 0.0	45.5 $\pm$ 23.0	85.2 $\pm$ 6.1	0.01 $\pm$ 0.0	0.14 $\pm$ 0.0	100 $\pm$ 0.0
			SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓	Acc.
CD-NOD	×	S	133 $\pm$ 1.5		31.5 $\pm$ 1.8		0.43 $\pm$ 0.0		0.61 $\pm$ 0.0		×
FANTOM	✓	S	6.0 $\pm$ 1.0		98.2 $\pm$ 0.2		0.01 $\pm$ 0.0		0.01 $\pm$ 0.0		100 $\pm$ 0.0

Under homoscedastic, non-Gaussian noise with $d = 20$ node graphs and three regimes, FANTOM outperforms every baseline on instantaneous-link inference, reaching an F1 of 96.1 % and an NHD of 0.001. Among methods explicitly designed for multi-regime MTS, both FANTOM and CASTOR recover the exact number of regimes, whereas RPCMCI fails to converge to the true partitions. FANTOM further surpasses CASTOR in regime detection (100. % vs. 79.8 %) and in DAG learning (F1 = 96.1 % vs. 34.3 %).

To give the stationary-MTS baselines their best chance, we supply them with the ground-truth regime labels and train each model on the corresponding pure regime. For this scenario of 20 nodes and 3 regimes and benefiting from regime labels, PCMCI+ exceeds FANTOM and Rhino on time-lagged links, achieving an F1 of 88.9 % compared with FANTOM's 85.2%.

Table 18: Average SHD, F1 scores, NHD and Ratio for different models with $d = 40$ nodes and $K = 2$ regimes. *Split* denotes whether regime separation is automatic ($\checkmark$) or manual ($\times$). *Inst.* refers to instantaneous links, and *Lag* to time-lagged edges.

Homoscedastic non-Gaussian noise, $K = 2$ and $d = 40$											
Model	Split	Type	Inst.				Lag				Regime
			SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓	Acc.
PCMCI+ Rhino DYNOTEARS CASTOR	×	W	91.0 $\pm$ 2.8	56.7 $\pm$ 0.1	0.01 $\pm$ 0.0	0.43 $\pm$ 0.0	25.0 $\pm$ 1.4	91.9 $\pm$ 0.8	0.003 $\pm$ 0.0	0.08 $\pm$ 0.0	×
	×	W	24.5 $\pm$ 2.1	90.7 $\pm$ 0.9	0.004 $\pm$ 0.0	0.09 $\pm$ 0.0	34.5 $\pm$ 5.0	89.8 $\pm$ 0.9	0.005 $\pm$ 0.0	0.10 $\pm$ 0.0	×
	×	W	100.0 $\pm$ 5.7	40.8 $\pm$ 3.7	0.015 $\pm$ 0.0	0.59 $\pm$ 0.0	58.5 $\pm$ 17.	76.4 $\pm$ 5.4	0.009 $\pm$ 0.0	0.21 $\pm$ 0.0	×
	×	W	94.0 $\pm$ 46.	67.3 $\pm$ 5.0	0.010 $\pm$ 0.0	0.33 $\pm$ 0.0	100.0 $\pm$ 49.	78.6 $\pm$ 3.4	0.01 $\pm$ 0.0	0.21 $\pm$ 0.0	×
RPCMCI	✓	W	-	-	-	-	-	-	-	-	-
CASTOR	✓	W	-	-	-	-	-	-	-	-	-
FANTOM	✓	W	29.5 $\pm$ 1.5	88.1 $\pm$ 0.2	0.004 $\pm$ 0.0	0.11 $\pm$ 0.0	32.3 $\pm$ 0.5	90.8 $\pm$ 0.2	0.005 $\pm$ 0.0	0.08 $\pm$ 0.0	100. $\pm$ 0.0
			SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓	Acc.
CD-NOD	×	S	260 $\pm$ 5.0	17.8 $\pm$ 3.6	0.18 $\pm$ 0.0	0.82 $\pm$ 0.0					×
FANTOM	✓	S	39.5 $\pm$ 4.5	92.4 $\pm$ 0.6	0.02 $\pm$ 0.0	0.07 $\pm$ 0.0					100. $\pm$ 0.0

Under homoscedastic, non-Gaussian noise with $d = 40$ -node graphs and two regimes, FANTOM is the only method that recovers the correct number of regimes, whereas CASTOR and RPCMCI fail to converge. FANTOM achieves 100% regime-detection accuracy and an 88.1% F1 score in DAG recovery.

For a fair comparison with stationary-MTS baselines, we provide these models with the ground-truth regime labels and train them separately on each pure regime. In this advantaged setting, Rhino surpasses FANTOM on instantaneous links (F1 = 90.7% vs. 88.1%), while PCMCI+ leads on time-lagged links (F1 = 91.9% vs. 90.8%).

Table 19: Average SHD, F1 scores, NHD and Ratio for different models with $d = 40$ nodes and $K = 3$ regimes. *Split* denotes whether regime separation is automatic ($\checkmark$) or manual ($\times$). *Inst.* refers to instantaneous links, and *Lag* to time-lagged edges.

Homoscedastic non-Gaussian noise, $K = 3$ and $d = 40$											
Model	Split	Type	Inst.				Lag				Regime
			SHD↓	F1↑	NHD↓	Ratio↓	SHD↓	F1↑	NHD↓	Ratio↓	Acc.
PCMCI+ Rhino DYNOTEARS CASTOR	×	W	141.5 $\pm$ 12.	56.5 $\pm$ 4.5	0.008 $\pm$ 0.0	0.43 $\pm$ 0.0	37.5 $\pm$ 9.2	91.6 $\pm$ 2.8	0.003 $\pm$ 0.0	0.08 $\pm$ 0.0	×
	×	W	53.5 $\pm$ 26.	85.5 $\pm$ 7.4	0.004 $\pm$ 0.0	0.14 $\pm$ 0.0	52.0 $\pm$ 5.7	89.2 $\pm$ 2.2	0.004 $\pm$ 0.0	0.11 $\pm$ 0.0	×
	×	W	152.0 $\pm$ 4.2	40.8 $\pm$ 3.7	0.01 $\pm$ 0.0	0.59 $\pm$ 0.0	91.0 $\pm$ 17.	75.8 $\pm$ 3.4	0.006 $\pm$ 0.0	0.24 $\pm$ 0.0	×
	×	W	93.5 $\pm$ 2.1	65.9 $\pm$ 7.6	0.009 $\pm$ 0.0	0.34 $\pm$ 0.0	94.0 $\pm$ 8.5	78.6 $\pm$ 4.5	0.009 $\pm$ 0.0	0.21 $\pm$ 0.0	×
RPCMCI	✓	W	-	-	-	-	-	-	-	-	-
CASTOR	✓	W	-	-	-	-	-	-	-	-	-
FANTOM	✓	W	35.3 $\pm$ 0.7	90.9 $\pm$ 0.4	0.002 $\pm$ 0.0	0.09 $\pm$ 0.0	46.3 $\pm$ 0.3	90.7 $\pm$ 1.3	0.003 $\pm$ 0.0	0.09 $\pm$ 0.0	100. $\pm$ 0.0
			SHD↓	F1↑	NHD↓	Ratio↓					Acc.
CD-NOD	×	S	342 $\pm$ 4.3	17.1 $\pm$ 3.2	0.24 $\pm$ 0.0	0.82 $\pm$ 0.0					×
FANTOM	✓	S	57.0 $\pm$ 2.5	92.8 $\pm$ 0.7	0.03 $\pm$ 0.0	0.07 $\pm$ 0.0					100. $\pm$ 0.0

F.3 Additional experiments for L=2

Table 20: Performance of FANTOM on $L = 2$ time series under varying node counts $d \in \{10, 20, 40\}$ and regime settings $K = 2$ and 3.

Heteroscedastic noise						
k	d	Inst.		Lag		Regime
		SHD	F1	SHD	F1	
2	10	1	97.7	0	100.0	94.3
	20	14	87.0	4	98.7	99.0
	40	15	94.5	3	99.5	98.9
3	10	1	98.5	1	99.5	91.5
	20	5	97.5	11	97.4	97.9
	40	43	88.6	3	99.0	99.0

FANTOM maintains similar performance with a larger lag ($L = 2$). It achieves an F1 score above 87% on instantaneous links for 10, 20, and 40 nodes in MTS with 2 or 3 regimes. For time-lagged links, we evaluate using the summary graph of lagged relations; the table shows that FANTOM detects them effectively.

F.3.1 Illustrations of learned graphs

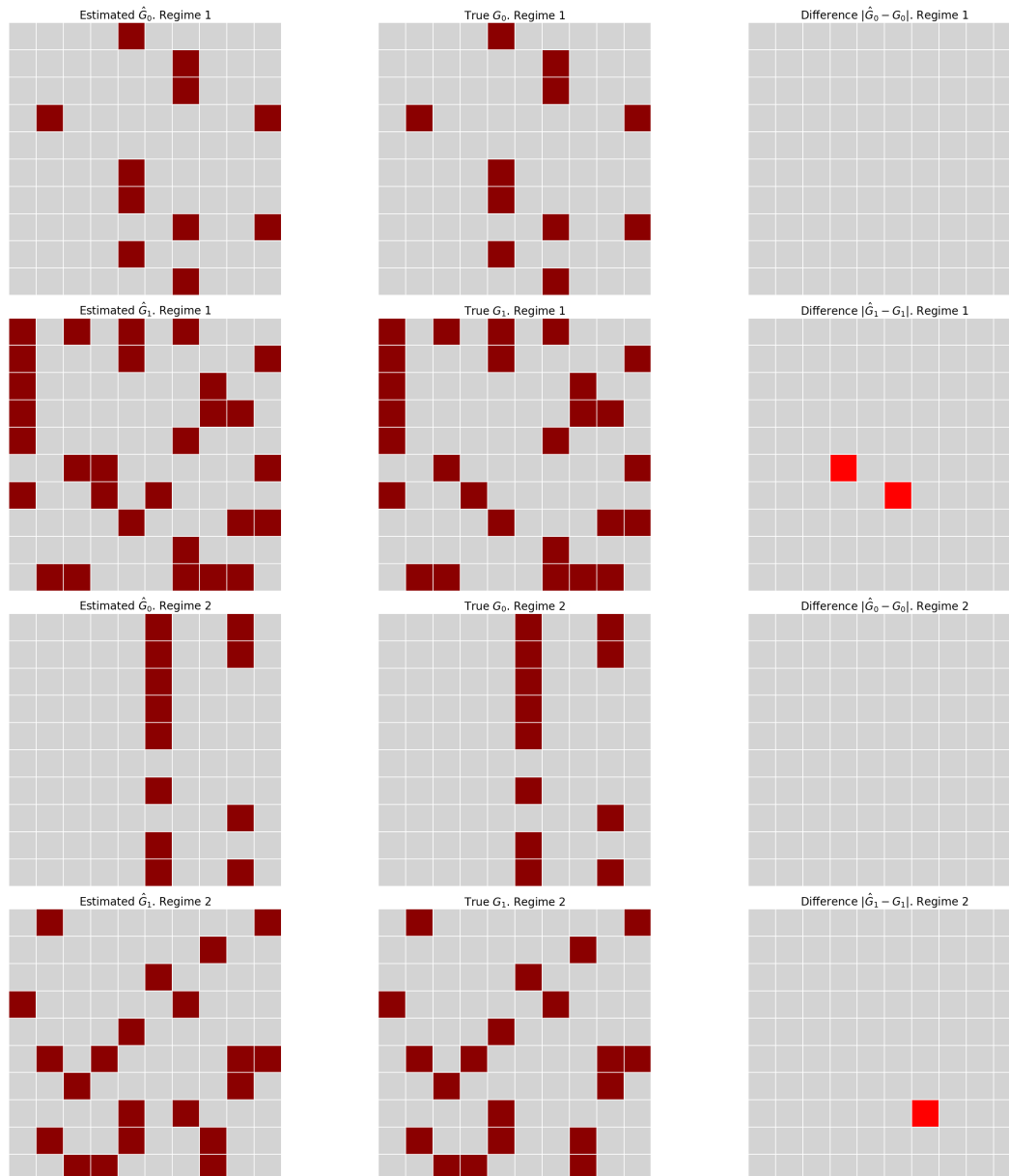


Figure 10: The estimated temporal causal graphs for two regimes, in **Heteroscedastic case**, consist of one matrix of 10 rows and 10 columns representing instantaneous links and another of 10 rows and 10 columns delineating time-lagged relations (with a maximum lag $L = 1$ in this case). Dark red indicates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

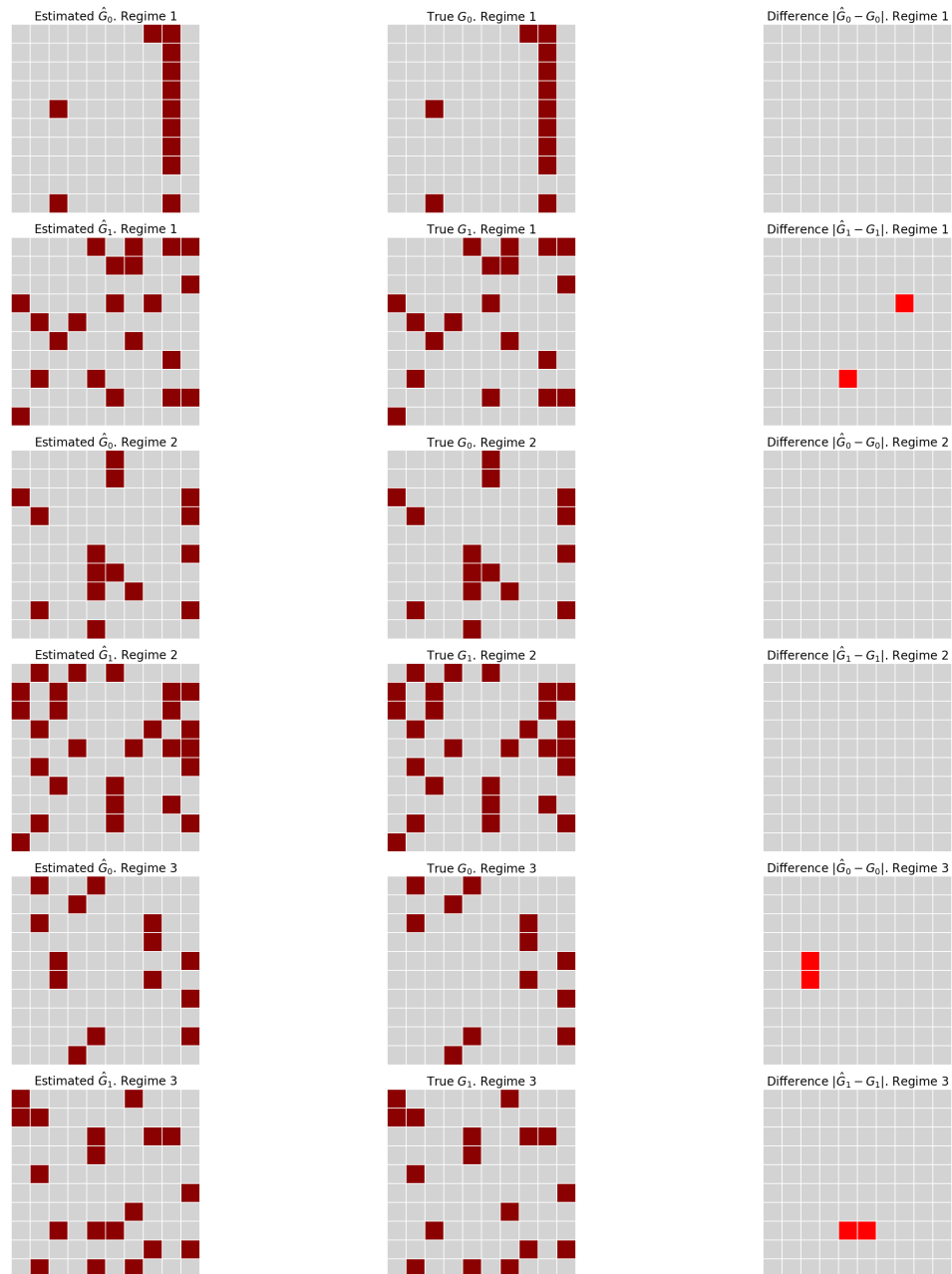


Figure 11: The estimated temporal causal graphs for three regimes, in **Heteroscedastic case**, consist of one matrix of 10 rows and 10 columns representing instantaneous links and another of 10 rows and 10 columns delineating time-lagged relations (with a maximum lag $L = 1$ in this case). Dark red indicates a value of one (presence of an edge), while gray symbolizes a value of 0 (absence of an edge). The second column displays the ground-truth causal graphs, and the final column highlights the difference between the estimated and true graphs.

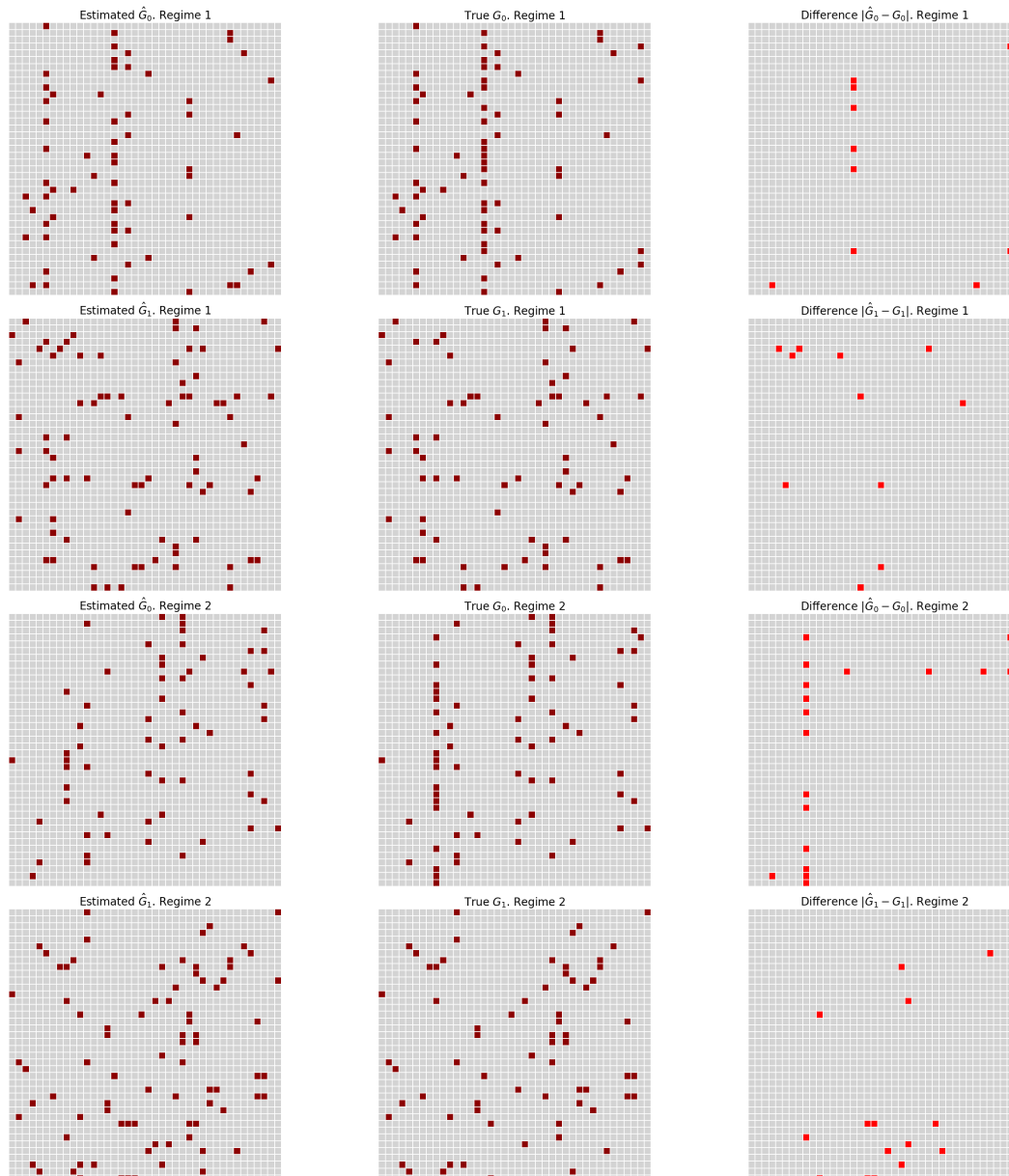


Figure 12: The estimated temporal causal graphs for three regimes, in **Heteroscedastic case**, consist of one matrix of 40 rows and 40 columns representing instantaneous links and another of 40 rows and 40 columns delineating time-lagged relations (with a maximum lag $L = 1$ in this case). Dark red indicates a value of one (presence of an edge), while gray symbolizes a value of 0 (absence of an edge). The second column displays the ground-truth causal graphs, and the final column highlights the difference between the estimated and true graphs.

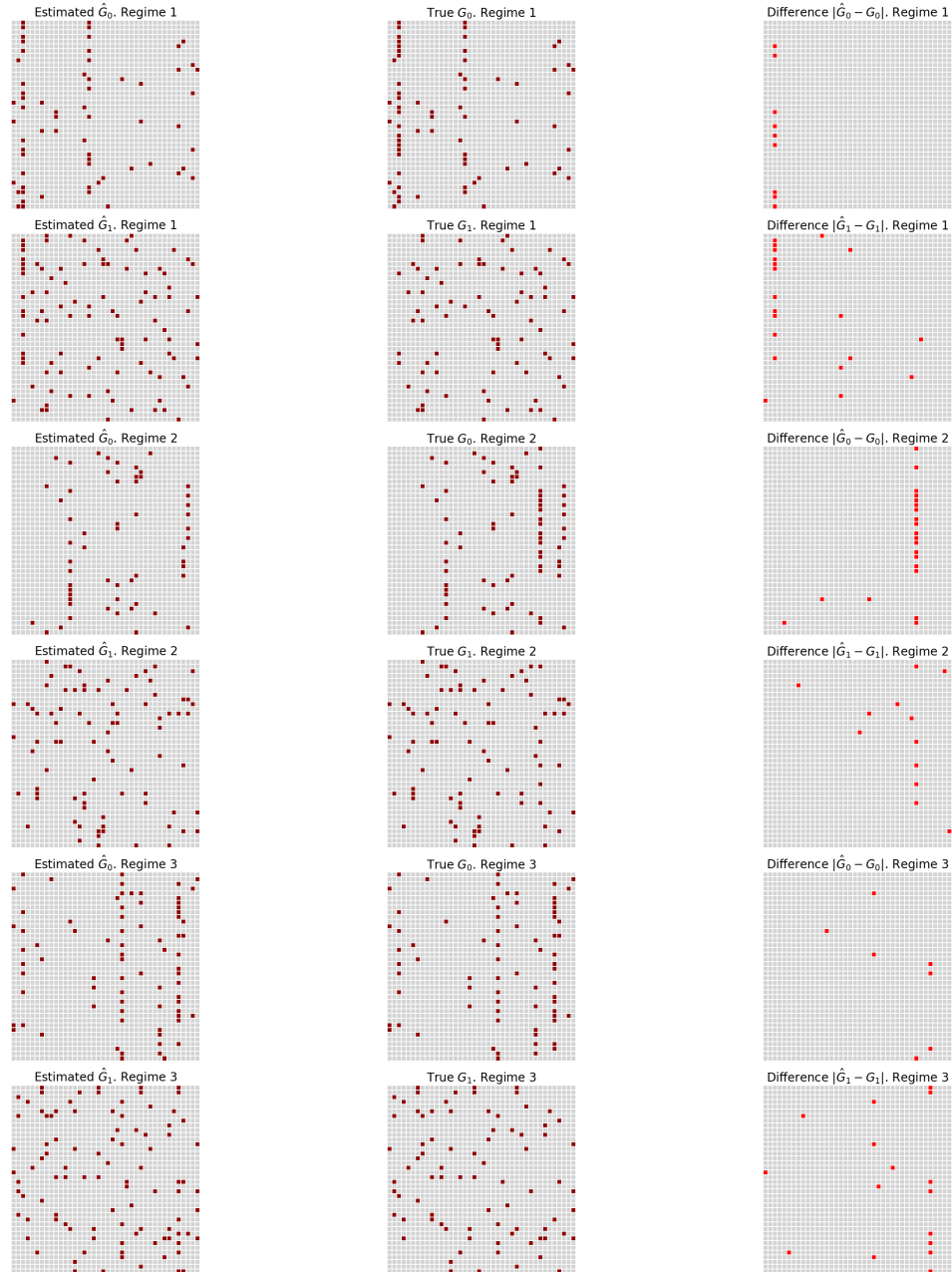


Figure 13: The estimated temporal causal graphs for three regimes, in **Non-Gaussian case**, consist of one matrix of 40 rows and 40 columns representing instantaneous links and another of 40 rows and 40 columns delineating time-lagged relations (with a maximum lag $L = 1$ in this case). Dark red indicates a value of one (presence of an edge), while gray symbolizes a value of 0 (absence of an edge). The second column displays the ground-truth causal graphs, and the final column highlights the difference between the estimated and true graphs.

F.3.2 Time complexity analysis

We start first by computing the time complexity of our Temporal Graph Neural network illustrated in Figure 2. We note d the input size, e the embedding size used for e_{ij} and h hidden layer size. After the first NN block,

$$T_{\text{NN}-1} = \mathcal{O}((d + e)h + hd) + \mathcal{O}(d) = \mathcal{O}(h(d + e))$$

then after the matrix multiplication block and the second NN block, we have :

$$T_{\text{forward}} = \mathcal{O}(Ld^2 + 2h(d + e)),$$

where L is the maximum lag.

Using the same architecture for a Conditional normalizing flow has a time complexity of :

$$T_{\text{CNF}} = \mathcal{O}(Ld^2 + h(e + dK)).$$

The complexity of FANTOM per iteration is $\mathcal{O}(N_w|\mathcal{T}|(2Ld^2 + h(e + dK)))$, where K is the number of bins, N_w is the number of regimes, and $|\mathcal{T}|$ is the number of samples.

F.3.3 Regime detection experiments

We compare FANTOM to CASTOR [39] and KCP [1] 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.

CASTOR is a causal discovery model specifically designed for multi-regime MTS. CASTOR is learns number of regime and their indices and their corresponding causal graphs without any prior knowledge. But it is limited to normal noise with equivariance. FANTOM learns the regime indices, while handling heteroscedastic noises.

We opted to perform regime detection with 10 nodes and three different regimes. For a fair comparison, we chose three regimes without re-occurrence, as KCP only detect change points and cannot identify the re-occurrence of a specific regime.

Regarding the models employed, we use the open-source code of CASTOR implemented in Python by the authors⁷. For KCP, we employ the Rupture package⁸.

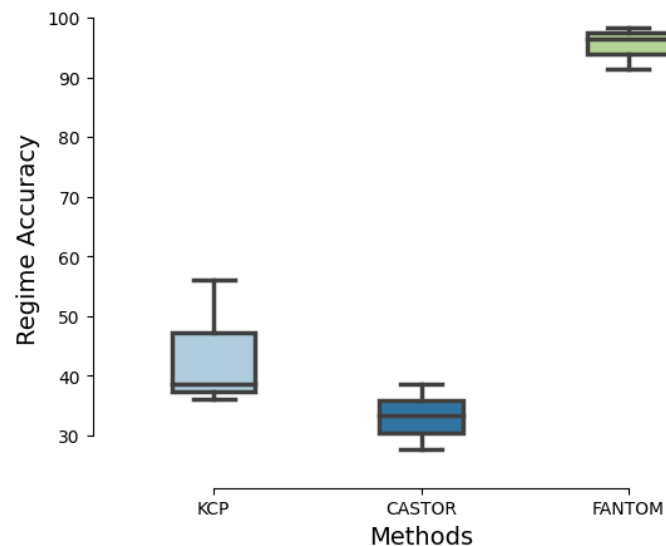


Figure 14: Comparison between FANTOM, CASTOR and KCP on regime detection for a MTS with heteroscedastic noise and composed of 3 regimes using accuracy metric. The number of nodes is $d = 10$.

From Figure 14, it is evident that FANTOM outperform CASTOR and 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. FANTOM outperforms

⁷<https://github.com/arahmani1/CASTOR>

⁸<https://centre-borelli.github.io/ruptures-docs/>

CASTOR in detecting regime indices. This result can be explained by the fact that CASTOR fails in handling heteroscedastic noises and fails to learn meaningful graphs which also lead to poor performance in regime detection.

From this analysis and the other experiments shown in the different tables, we can conclude that in scenarios involving MTS with multiple regimes with non-Gaussian or Heteroscedastic noises, FANTOM 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 other state-of-the-art change point detection method KCP. Therefore, employing 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 KCP and subsequently uses the regime partitions revealed by KCP as an input to a causal discovery method (such as PCMCi+, DYNOTEARS, Rhino, etc.), the running time will be significantly high.
- We show throughout all our experiments, that FANTOM outperforms all causal discovery method in DAG learning task, even when these models are in more favorable scenarios, by having access to the regime labels beforehand.

G Proof of proposition

In this section, we are going to provide the entire proof of our proposition 3.1. As we state before, we note $\mathcal{G} = (\mathcal{G}^r)_{r \in [1:N_w]}$.

Proof. we have:

$$\begin{aligned} \log p_{\Theta}(\mathbf{x}_{t \in \mathcal{T}}) &= \sum_{t=1}^{|\mathcal{T}|} \log p_{\Theta}(\mathbf{x}_t | \mathbf{x}_{<t}) \\ &= \sum_{t=1}^{|\mathcal{T}|} \log \sum_{\mathcal{G}} p_{\Theta}(\mathbf{x}_t | \mathbf{x}_{<t}, \mathcal{G}) p(\mathcal{G}) \frac{q_{\phi}(\mathcal{G})}{q_{\phi}(\mathcal{G})} \\ &\geq \sum_{t=1}^{|\mathcal{T}|} \mathbb{E}_{q_{\phi}(\mathcal{G})} [\log p_{\Theta}(\mathbf{x}_t | \mathbf{x}_{<t}, \mathcal{G}) + \log p(\mathcal{G}) - \log q_{\phi}(\mathcal{G})] \end{aligned}$$

Let's focus on the first term of our last inequality $\log p_{\Theta}(\mathbf{x}_t | \mathbf{x}_{<t}, \mathcal{G})$, we have:

$$\begin{aligned} \log p_{\Theta}(\mathbf{x}_t | \mathbf{x}_{<t}, \mathcal{G}) &= \log \sum_{z_t} p_{\theta^{z_t}}(\mathbf{x}_t | \mathbf{x}_{<t}, \mathcal{G}, z_t) p(z_t) \frac{p(z_t | \mathbf{x}_t, \mathbf{x}_{<t})}{p(z_t | \mathbf{x}_t, \mathbf{x}_{<t})} \\ &\geq \mathbb{E}_{p(z_t | \mathbf{x}_t, \mathbf{x}_{<t})} [\log p_{\theta^{z_t}}(\mathbf{x}_t | \mathbf{x}_{<t}, \mathcal{G}^{z_t}) + \log p(z_t) - \log p(z_t | \mathbf{x}_t, \mathbf{x}_{<t})] \\ &\geq \mathbb{E}_{p(z_t | \mathbf{x}_t, \mathbf{x}_{<t})} [\log p_{\theta^{z_t}}(\mathbf{x}_t | \mathbf{x}_{<t}, \mathcal{G}^{z_t}) + \log p(z_t)] + H(p(z_t | \mathbf{x}_t, \mathbf{x}_{<t})) \end{aligned}$$

Including this result in the previous equation gives us the following:

$$\begin{aligned} \log p_{\Theta}(\mathbf{x}_{t \in \mathcal{T}}) &\geq \sum_{t=1}^{|\mathcal{T}|} \mathbb{E}_{q_{\phi}(\mathcal{G})} [\mathbb{E}_{p(z_t | \mathbf{x}_t, \mathbf{x}_{<t})} [\log p_{\theta^{z_t}}(\mathbf{x}_t | \mathbf{x}_{<t}, \mathcal{G}^{z_t}) + \log p(z_t)] + H(p(z_t | \mathbf{x}_t, \mathbf{x}_{<t}))] \\ &\quad + \sum_{r=1}^{N_w} \mathbb{E}_{q_{\phi^r}(\mathcal{G}^r)} [\log p(\mathcal{G}^r)] + H(q_{\phi^r}(\mathcal{G}^r)) \\ &\equiv \text{ELBO}(\Theta), \end{aligned}$$

we note that the priors $p(\mathcal{G}^r)$ and the variational estimations $q_{\phi^r}(\mathcal{G}^r)$ are independents.

H Proofs of our theoretical contributions

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

H.1 Assumptions

Definition H.1 (Causal Stationarity, [41]). A stationary time series process $(\mathbf{x}_t)_{t \in \mathcal{T}}$ with graph $\mathcal{G}$ is called causally stationary over a time index set $\mathcal{T}$ if and only if for all links $\mathbf{x}_{t-\tau}^i \rightarrow \mathbf{x}_t^j$ in the graph

$$\mathbf{x}_{t-\tau}^i \not\perp\!\!\!\perp \mathbf{x}_t^j \mid \mathbf{x}_{<t} \setminus \{\mathbf{x}_{t-\tau}^i\}.$$

This elucidates the inherent characteristics of the time-series data generation mechanism, thereby validating the choice of the auto-regressive model.

Assumption H.2 (Causal Stationarity for MTS with multiple regime). A MTS $(\mathbf{x}_t)_{t \in \mathcal{T}}$ 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 $(\mathbf{x}_t)_{t \in \mathcal{E}_u}$ is causally stationary with graph $\mathcal{G}^u$ as defined in definition H.1.

Definition H.3. (Causal Markov Property, [37]). 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 H.4 (Causal Markov Property (CMP)). A set of joint distributions $(p(\cdot|\mathcal{G}^r))_{r \in [1:K]}$ satisfies the **CMP** with respect to the DAGs $(\mathcal{G}^r)_{r \in [1:K]}$ if, for each $r \in [1:K]$, the distribution $p(\cdot|\mathcal{G}^r)$ satisfies the CMP relative to the DAG $\mathcal{G}^r$. Specifically, in every regime r , each variable is independent of its non-descendants given its parents.

Assumption H.5 (Causal Minimality). Given a set of DAGs $(\mathcal{G}^r)_{r \in [1:K]}$ and a set of joint distribution $(p(\cdot|\mathcal{G}^r))_{r \in [1:K]}$, we say that this set of distributions satisfies causal minimality w.r.t. the set of DAGs $(\mathcal{G}^r)_{r \in [1:K]}$ if for every r : $p(\cdot|\mathcal{G}^r)$ is Markovian w.r.t the DAG $\mathcal{G}^r$ but not to any proper subgraph of $\mathcal{G}^r$.

Assumption H.6 (Causal Sufficiency). A set of observed variables $\mathbf{V}$ is causally sufficient for a process $\mathbf{x}_t$ if and only if in the process every common cause of any two or more variables in $\mathbf{V}$ is in $\mathbf{V}$ 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.

	Causal graph	Causal Markov	Causal sufficiency	Faithfulness / Minimality	Heteroscedastic noise	Stationarity per regime
DYNOTEARS	W	✓	✓		×	×
PCMCi+	W	✓	✓	F	×	×
RPCMCI	W	✓	✓	F	×	✓
Rhino	W	✓	✓	M	×	×
CD-NOD	S	✓	✓	F	×	✓
CASTOR	W	✓	✓	M	×	✓
FANTOM	W	✓	✓	M	✓	✓

Table 21: 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 table 21 illustrates that most assumptions (causal sufficiency, causal Markov, faithfulness/minimality) are commonly shared among various state-of-the-art models in causal discovery.

However, FANTOM, 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, FANTOM, CASTOR and RPCMCI predict a window causal graph, which can subsequently be used to reconstruct a summary graph. FANTOM is the only model that can handle heteroscedastic noise.

H.2 Proof of theorem 4.2

We start first by proving theorem 4.2. To do so, we will prove identifiability in the case of bivariate time series, Lemma H.7. Then we will prove identifiability in the case of MTS.

Lemma H.7. Assume Causal Markov property, minimality, stationarity, sufficiency and $(x_t^1, x_t^2)_{t \in \mathcal{T}}$ be a bivariate time series such that $(x_t^1, x_t^2)_{t \in \mathcal{T}}$ following Eq(1) where $K = 1$ and $\epsilon_t^i \sim \mathcal{N}(0, 1)$. We have x_t^1, x_t^2 follow Gaussian distribution, if f^1, f^2 are non linear and $\frac{1}{g^1}, \frac{1}{g^2}$ are not a polynomial of degree two then the bivariate Temporal heteroscedastic Gaussian noise (THGNM) model is identifiable.

Proof of the Lemma. Let's assume we have two temporal causal graph $\mathcal{G}$ and $\mathcal{G}'$ for the bivariate THGNM.

Disagreement in time lagged relationships. Assume that $\mathcal{G}$ and $\mathcal{G}'$ do not differ in the instantaneous effects. $\forall i \in \{1, 2\}$: $\text{Pa}_{\mathcal{G}}^i(t) = \text{Pa}_{\mathcal{G}'}^i(t)$. Hence and Wlog, there is some $k > 0$ and an edge $x_{t-k}^1 \rightarrow x_t^2$ in $\mathcal{G}$ but not in $\mathcal{G}'$. From $\mathcal{G}'$ and the Causal Markov property, we have that $x_{t-k}^1 \perp\!\!\!\perp x_t^2 \mid \mathcal{S}$, where $\mathcal{S} = (\{x_{t-l}^i, 1 \leq l \leq L, i \in \{1, 2\}\} \cup \text{ND}_t) \setminus \{x_{t-k}^1, x_t^2\}$, and ND_t are all X_t^i that are non-descendants (wrt instantaneous effects) of x_t^2 . Applied to $\mathcal{G}$, causal minimality leads to a contradiction because $x_{t-k}^1 \not\perp\!\!\!\perp x_t^2 \mid \mathcal{S}$ in $\mathcal{G}$, and the above reasoning shows that it exists a subgraph $\mathcal{G}'$ of $\mathcal{G}$ that is Markovian to the joint distribution of the data.

Disagreement on instantaneous parents. Now, let's assume we have a forward model, $\forall t \in \mathcal{T}$:

$$x_t^1 = f^1(\text{Pa}_{\mathcal{G}}^1(< t), x_t^2) + g^1(\text{Pa}_{\mathcal{G}}^1(< t), x_t^2) \cdot \epsilon_t^1,$$

We will prove by contradiction that a backward model

$$x_t^2 = f^2(\text{Pa}_{\mathcal{G}}^2(< t), x_t^1) + g^2(\text{Pa}_{\mathcal{G}}^2(< t), x_t^1) \cdot \epsilon_t^2,$$

can not exists.

We note $\mathbf{h}_t = \text{Pa}_{\mathcal{G}}^1(< t) \cup \text{Pa}_{\mathcal{G}}^2(< t)$. We know that $\epsilon_t^2 \perp\!\!\!\perp (x_t^1, \text{Pa}_{\mathcal{G}}^1(< t), \text{Pa}_{\mathcal{G}}^2(< t))$ and $\epsilon_t^1 \perp\!\!\!\perp (x_t^2, \text{Pa}_{\mathcal{G}}^1(< t), \text{Pa}_{\mathcal{G}}^2(< t))$, using Lemma 36 in Peters et al. [38], we have:

$$x_t^1 | \mathbf{h}_t = f^1(pa_{< t}^1, x_t^2 | \mathbf{h}_t) + g^1(pa_{< t}^1, x_t^2 | \mathbf{h}_t) \cdot \epsilon_t^1,$$

$$x_t^2 | \mathbf{h}_t = f^2(pa_{< t}^2, x_t^1 | \mathbf{h}_t) + g^2(pa_{< t}^2, x_t^1 | \mathbf{h}_t) \cdot \epsilon_t^2,$$

where $x_t^i | \mathbf{h}_t$ is x_t^i conditioned on $\mathbf{h}_t$. This last result contradicts the theorem states by Khemakhem et al. [29]. Hence, our bivariate THGNM is identifiable.

Let's prove this results in the case of MTS (Theorem 4.2). In the case of Disagreement in time lagged relationships, we can use the same proof for the bivariate case.

Disagreement on instantaneous parents. Let's assume we have two temporal causal graph $\mathcal{G}$ and $\mathcal{G}'$ such that $\mathcal{G} \neq \mathcal{G}'$. According to the Proposition 28 in Peters et al [38], for $\mathcal{G}$ and $\mathcal{G}'$ be two different DAGs over a set of variables $\mathbf{V}$, such that $\mathbf{x}_t$ is generated by our HNM and satisfies the Markov condition and causal minimality with respect to $\mathcal{G}$ and $\mathcal{G}'$. Then there are variables $x_t^1, x_t^2 \in \mathbf{V}$ such that for the set $\mathbf{Q} := \text{Pa}_{\mathcal{G}}^1(t) \setminus \{x_t^2\}$, $\mathbf{Y} := \text{Pa}_{\mathcal{G}'}^2(t) \setminus \{x_t^1\}$ and $\mathbf{S} := \mathbf{Q} \cup \mathbf{Y}$, we have: 1) $x_t^2 \rightarrow x_t^1$ in $\mathcal{G}$ and $x_t^1 \rightarrow x_t^2$ in $\mathcal{G}'$. 2) $\mathbf{S} \subseteq \text{ND}_{x_t^1}^{\mathcal{G}} \setminus \{x_t^2\}$ and $\mathbf{S} \subseteq \text{ND}_{x_t^2}^{\mathcal{G}'} \setminus \{x_t^1\}$. $\text{Pa}_{\mathcal{G}}^1(t)$ is the set of parent variables of x_t^1 in graph $\mathcal{G}$. $\text{ND}_{x_t^1}^{\mathcal{G}}$ is the set of non-descendant (wrt instantaneous effects) of x_t^1 in graph $\mathcal{G}$.

We consider $\mathbf{S} = \mathbf{s}$ with $p(\mathbf{s}) > 0$. Denote $x_t^{1,*} := x_t^1 \mid \mathbf{S} = \mathbf{s}$ and $x_t^{2,*} := x_t^2 \mid \mathbf{S} = \mathbf{s}$. Lemma 37 in Peters et al. [38] states that if $p(\mathbf{x}_t)$ is generated according to the SEM models as follows:

$$x_t^i = f_i(\mathbf{Pa}_G^i(< t), \mathbf{Pa}_G^i(t), \epsilon_t^i), i \in \{1, 2, \dots, d\}, x_t^i \in V$$

with corresponding DAG $\mathcal{G}$, then for a variable $x_t^i \in V$, if $\mathbf{S} \subseteq \mathbf{ND}_{x_t^i}^{\mathcal{G}}$ then $\epsilon_t^i \perp\!\!\!\perp \mathbf{S}$. Our TGHNM can be viewed one specific class of the SEM in the aforementioned equation. Hence, Lemma 37 holds under our TGHNM and renders $\epsilon_t^1 \perp\!\!\!\perp (x_t^2, \mathbf{S})$ and $\epsilon_t^2 \perp\!\!\!\perp (x_t^1, \mathbf{S})$, using Lemma 36 in Peters et al. [38], we have:

$$\begin{aligned} x_t^{1,*}|_{\mathbf{h}_t} &= f^1(\mathbf{q}, pa_{<t}^1, x_t^{2,*}|_{\mathbf{h}_t}) + g^1(\mathbf{q}, pa_{<t}^1, x_t^{2,*}|_{\mathbf{h}_t}) \cdot \epsilon_t^1, \\ x_t^{2,*}|_{\mathbf{h}_t} &= f^2(\mathbf{y}, pa_{<t}^2, x_t^{1,*}|_{\mathbf{h}_t}) + g^2(\mathbf{y}, pa_{<t}^2, x_t^{1,*}|_{\mathbf{h}_t}) \cdot \epsilon_t^2, \end{aligned}$$

where $x_t^i|_{\mathbf{h}_t}$ is x_t^i conditioned on $\mathbf{h}_t$. This results contradict our previous proved Lemma, then THGNM is identifiable model under the conditions stated in the theorem.

Theorem H.8 (Identifiability of Temporal Non Gaussian noise model (TNGNM)). *Assume Causal Markov property, stationarity, minimality, sufficiency and let $(x_t)_{t \in \mathcal{T}}$ be a MTS following a TNGNM, $\forall t \in \mathcal{T}$:*

$$x_t^i = f^i(\mathbf{Pa}_G^i(< t), \mathbf{Pa}_G^i(t)) + \epsilon_t^i, \quad (14)$$

where f^i is a differentiable function, and ϵ_t^i are mutually independent noises and follow a non Gaussian distribution. The TNGNM is identifiable.

Proof. The proof of this theorem could be concluded from theorem 1 in Rhino [15]. Eq(14) is a special case of Rhino SEMs.

H.3 Identifiability results in the case of Temporal General Heteroscedastic Noise Models

In this section, we will present our identifiability results for the case of Temporal General Heteroscedastic Noise, where a MTS has the following SEM $\forall t \in \mathcal{T}$:

$$x_t^i = f^i(\mathbf{Pa}_G^i(< t), \mathbf{Pa}_G^i(t)) + g^i(\mathbf{Pa}_G^i(< t), \mathbf{Pa}_G^i(t)) \cdot \epsilon_t^i, \quad (15)$$

where f^i and g^i are differentiable functions, with g_i strictly positive and ϵ_t^i are mutually independent normal noises and can have any arbitrary density distribution. We assume $\mathbb{E}(\epsilon_t^i) = 0$ and $\mathbb{E}((\epsilon_t^i)^2) = 1$ without loss of generality.

We will start first by showing that if backward model, respects to instantaneous links, exists in the bivariate case then, the data generating mechanism must fulfill the a Partial Differential Equation (PDE). Then, following Peters et al. [38] and Strobl et al. [50] for defining Restricted SEM on iid, we will define a Temporal Restricted Heteroscedastic Noise model and show its identifiability.

Lemma H.9. *Assume Causal Markov property, minimality, stationarity, sufficiency and $(x_t^1, x_t^2)_{t \in \mathcal{T}}$ be a bivariate time series. Then we have time lagged parents are identifiable, and a backward model with respect to instantaneous links can be fit i.e. $\forall t \in \mathcal{T}$:*

$$\begin{cases} \tilde{x}_t^1 = x_t^1|_{\mathbf{h}_t} = f^1(\tilde{x}_t^2) + g^1(\tilde{x}_t^2) \cdot \epsilon_t^1 \\ \tilde{x}_t^2 = x_t^2|_{\mathbf{h}_t} = f^2(\tilde{x}_t^1) + g^2(\tilde{x}_t^1) \cdot \epsilon_t^2. \end{cases} \quad (16)$$

We note $\mathbf{h}_t = \mathbf{Pa}^1(< t) \cup \mathbf{Pa}^2(< t)$, and let $\nu_1(\cdot)$ and $\nu_2(\cdot)$ be the twice differentiable log densities of $\tilde{X}_t^1$ and $\tilde{\epsilon}_t^2$ respectively. For compact notation, define

$$\begin{aligned} \nu_{\tilde{X}_t^2|\tilde{X}_t^1}(\tilde{x}_t^2 | \tilde{x}_t^1) &= \log(p_{\tilde{X}_t^2|\tilde{X}_t^1}(\tilde{x}_t^2 | \tilde{x}_t^1)) \\ &= \log\left(p_{\epsilon_t^2}\left(\frac{\tilde{x}_t^2 - f^2(\tilde{x}_t^1)}{g^2(\tilde{x}_t^1)}\right) / g^2(\tilde{x}_t^1)\right) \\ &= \nu_2\left(\frac{\tilde{x}_t^2 - f^2(\tilde{x}_t^1)}{g^2(\tilde{x}_t^1)}\right) - \log(g^2(\tilde{x}_t^1)) \quad \text{and} \\ G(\tilde{x}_t^2, \tilde{x}_t^1) &= g^1(\tilde{x}_t^2)(f^1)'(\tilde{x}_t^2) + (g^1)'(\tilde{x}_t^2)[\tilde{x}_t^1 - f^1(\tilde{x}_t^2)]. \end{aligned}$$

Assume that f^1, g^1, f^2 , and g^2 are twice differentiable. Then, the data generating mechanism must fulfill the following PDE for all $(\tilde{x}_t^2, \tilde{x}_t^1)$ with $G(\tilde{x}_t^2, \tilde{x}_t^1) \neq 0$.

$$0 = \nu_1''(\tilde{x}_t^1) + \frac{(g^1)'(\tilde{x}_t^2)}{G(\tilde{x}_t^2, \tilde{x}_t^1)} \nu_1'(\tilde{x}_t^1) + \frac{\partial^2}{\partial (\tilde{x}_t^1)^2} \nu_{\tilde{x}_t^2 | \tilde{x}_t^1}(\tilde{x}_t^2 | \tilde{x}_t^1) + \frac{g^1(\tilde{x}_t^2)}{G(\tilde{x}_t^2, \tilde{x}_t^1)} \frac{\partial^2}{\partial \tilde{x}_t^1 \partial \tilde{x}_t^2} \nu_{\tilde{x}_t^2 | \tilde{x}_t^1}(\tilde{x}_t^2 | \tilde{x}_t^1) + \frac{(g^1)'(\tilde{x}_t^2)}{G(\tilde{x}_t^2, \tilde{x}_t^1)} \frac{\partial}{\partial \tilde{x}_t^1} \nu_{\tilde{x}_t^2 | \tilde{x}_t^1}(\tilde{x}_t^2 | \tilde{x}_t^1). \quad (17)$$

We drop the time-lagged parent in Eq (16) to simplify the notation, since conditioning on the history makes it redundant.

Proof of the Lemma. Let's assume we have two temporal causal graph $\mathcal{G}$ and $\mathcal{G}'$ for the bivariate temporal heteroscedastic causal models where the noise distribution could follow any arbitrary distribution.

Disagreement in time lagged relationships. Assume that $\mathcal{G}$ and $\mathcal{G}'$ do not differ in the instantaneous effects. $\forall i \in \{1, 2\}$: $\text{Pa}_{\mathcal{G}}^i(t) = \text{Pa}_{\mathcal{G}'}^i(t)$. Hence and Wlog, there is some $k > 0$ and an edge $x_{t-k}^1 \rightarrow x_t^2$ in $\mathcal{G}$ but not in $\mathcal{G}'$. From $\mathcal{G}'$ and the Causal Markov property, we have that $x_{t-k}^1 \perp\!\!\!\perp x_t^2 \mid \mathcal{S}$, where $\mathcal{S} = (\{x_{t-l}^i, 1 \leq l \leq L, i \in \{1, 2\}\} \cup \text{ND}_t) \setminus \{x_{t-k}^1, x_t^2\}$, and ND_t are all X_t^i that are non-descendants (wrt instantaneous effects) of x_t^2 . Applied to $\mathcal{G}$, causal minimality leads to a contradiction because $x_{t-k}^1 \not\perp\!\!\!\perp x_t^2 \mid \mathcal{S}$ in $\mathcal{G}$, and the above reasoning shows that it exists a subgraph $\mathcal{G}'$ of $\mathcal{G}$ that is Markovian to the joint distribution of the data.

Disagreement in instantaneous parents. Now, let's assume we have a forward model, after conditioning on $\mathbf{h}_t = \text{Pa}_{\mathcal{G}}^1(< t) \cup \text{Pa}_{\mathcal{G}}^2(< t)$, $\forall t \in \mathcal{T}$:

$$\tilde{x}_t^1 = x_t^1 | \mathbf{h}_t = f^1(\tilde{x}_t^2) + g^1(\tilde{x}_t^2) \cdot \epsilon_t^1$$

We want to prove that if a backward model

$$\tilde{x}_t^2 = x_t^2 | \mathbf{h}_t = f^2(\tilde{x}_t^1) + g^2(\tilde{x}_t^1) \cdot \epsilon_t^2$$

exists then the PDE in Eq(17) is fulfilled.

Our conditioning trick on time lagged parents makes the use of Immer et al. [25] theorem 1 feasible in our case. We employ the change of variables from $\{\tilde{x}_t^2, \epsilon_t^1\}$ to $\{\tilde{x}_t^1, \epsilon_t^2\}$ and the proof will be the same as Immer et al.. Hence, we leverage Theorem 1 of Immer et al. and we conclude that if a backward model exists the PDE Eq(17) is verified.

Theorem H.10 (Identifiability of Temporal Restricted Heteroscedastic noise model (TRHNM)). Assume Causal Markov property, minimality, sufficiency and let $(\mathbf{x}_t)_{t \in \mathcal{T}}$ be a MTS following Eq(1) where $K = 1$. The graph $\mathcal{G}$ is uniquely identified if $\forall i \in [1 : d], \forall j : x_t^j \in \text{Pa}_{\mathcal{G}}^i(t)$ and $\mathcal{S}$ such that $(\text{Pa}_{\mathcal{G}}^i(t) \setminus x_t^j) \subseteq \mathcal{S} \subseteq (\text{Nd}(x_t^i) \setminus x_t^j)$, there exists $\mathcal{S} = \mathbf{s}$ where $p(\mathbf{s}) > 0$ $\mathbf{h}_t = \text{Pa}_{\mathcal{G}}^i(< t) \cup \text{Pa}_{\mathcal{G}}^j(< t)$ and $p(x_t^i, x_t^j \mid \mathbf{s}, \mathbf{h}_t)$ do not satisfy PDE of Equation 17, and we call the model that verify this condition, the Temporal Restricted Heteroscedastic noise model.

Proof of the theorem. We will follow the same steps in the proof of theorem 4.2. Let's assume we have two temporal causal graph $\mathcal{G}$ and $\mathcal{G}'$ for the multivariate TRHNM. We assume also that $\forall i \in [1 : d], \forall j : x_t^j \in \text{Pa}_{\mathcal{G}}^i(t)$ and $\mathcal{S}$ such that $(\text{Pa}_{\mathcal{G}}^i(t) \setminus x_t^j) \subseteq \mathcal{S} \subseteq (\text{Nd}(x_t^i) \setminus x_t^j)$, there exists $\mathcal{S} = \mathbf{s}$ where $p(\mathbf{s}) > 0$ $\mathbf{h}_t = \text{Pa}_{\mathcal{G}}^i(< t) \cup \text{Pa}_{\mathcal{G}}^j(< t)$ and $p(x_t^i, x_t^j \mid \mathbf{s}, \mathbf{h}_t)$ do not satisfy PDE of Equation 17.

We will start by showing that time lagged parents are identifiable. Same reasoning in the bivariate case.

Disagreement in time lagged relationships. Assume that $\mathcal{G}$ and $\mathcal{G}'$ do not differ in the instantaneous effects. $\forall i \in \{1, 2\}$: $\text{Pa}_{\mathcal{G}}^i(t) = \text{Pa}_{\mathcal{G}'}^i(t)$. Hence and Wlog, there is some $k > 0$ and an edge $x_{t-k}^1 \rightarrow x_t^2$ in $\mathcal{G}$ but not in $\mathcal{G}'$. From $\mathcal{G}'$ and the Causal Markov property, we have that $x_{t-k}^1 \perp\!\!\!\perp x_t^2 \mid \mathcal{S}$, where $\mathcal{S} = (\{x_{t-l}^i, 1 \leq l \leq L, i \in \{1, 2\}\} \cup \text{ND}_t) \setminus \{x_{t-k}^1, x_t^2\}$, and ND_t are all X_t^i that are non-descendants (wrt instantaneous effects) of x_t^2 . Applied to $\mathcal{G}$, causal minimality leads to a contradiction

because $x_{t-k}^1 \not\perp\!\!\!\perp x_t^2 \mid \mathcal{S}$ in $\mathcal{G}$, and the above reasoning shows that it exists a subgraph $\mathcal{G}'$ of $\mathcal{G}$ that is Markovian to the joint distribution of the data.

Disagreement on instantaneous parents. Let's now assume we have two temporal causal graph $\mathcal{G}$ and $\mathcal{G}'$ such that $\mathcal{G} \neq \mathcal{G}'$. According to the Proposition 29 in Peters et al [38], for $\mathcal{G}$ and $\mathcal{G}'$ be two different DAGs over a set of variables V , such that x_t is generated by our TRHNM and satisfies the Markov condition and causal minimality with respect to $\mathcal{G}$ and $\mathcal{G}'$. Then there are variables $x_t^1, x_t^2 \in V$ such that for the set $Q := \text{Pa}_{\mathcal{G}}^1(t) \setminus \{x_t^2\}$, $Y := \text{Pa}_{\mathcal{G}'}^2(t) \setminus \{x_t^1\}$ and $S := Q \cup Y$, we have: 1) $x_t^2 \rightarrow x_t^1$ in $\mathcal{G}$ and $x_t^1 \rightarrow x_t^2$ in $\mathcal{G}'$. 2) $S \subseteq \text{ND}_{x_t^1}^{\mathcal{G}} \setminus \{x_t^2\}$ and $S \subseteq \text{ND}_{x_t^2}^{\mathcal{G}'} \setminus \{x_t^1\}$. $\text{Pa}_{\mathcal{G}}^1(t)$ is the set of parent variables of x_t^1 in graph $\mathcal{G}$. $\text{ND}_{x_t^1}^{\mathcal{G}}$ is the set of non-descendant (wrt instantaneous effects) of x_t^1 in graph $\mathcal{G}$.

We consider $S = s$ with $p(s) > 0$. Lemma 37 in Peters et al. [38] states that if $p(x_t)$ is generated according to the SEM models as follows:

$$x_t^i = f_i(\text{Pa}_{\mathcal{G}}^i(< t), \text{Pa}_{\mathcal{G}}^i(t), \epsilon_t^i), i \in \{1, 2, \dots, d\}, x_t^i \in V$$

with corresponding DAG $\mathcal{G}$, then for a variable $x_t^i \in V$, if $S \subseteq \text{ND}_{x_t^i}^{\mathcal{G}}$ then $\epsilon_t^i \perp\!\!\!\perp S$. Our TRHNM can be viewed one specific class of the SEM in the aforementioned equation. Hence, Lemma 37 holds under our TRHNM and applying it to x_t^1 renders $\epsilon_t^1 \perp\!\!\!\perp (x_t^2, S)$ and $x_t^2 \epsilon_t^2 \perp\!\!\!\perp (x_t^1, S)$.

Using now Lemma 36 in Peters et al. [38], and we denote $x_t^{1,*} := x_t^1 \mid S = s$ and $x_t^{2,*} := x_t^2 \mid S = s$. We have:

$$\begin{aligned} x_t^{1,*} |_{h_t} &= f^1(x_t^{2,*} |_{h_t}) + g^1(x_t^{2,*} |_{h_t}) \cdot \epsilon_t^1 \\ x_t^{2,*} |_{h_t} &= f^2(x_t^{1,*} |_{h_t}) + g^2(x_t^{1,*} |_{h_t}) \cdot \epsilon_t^2, \end{aligned} \quad (18)$$

where $x_t^{i,*} |_{h_t}$ is x_t^i conditioned on h_t, S and $h_t = \text{Pa}_{\mathcal{G}}^1(< t) \cup \text{Pa}_{\mathcal{G}}^2(< t)$ in this case, which is also equal to $h_t = \text{Pa}_{\mathcal{G}'}^1(< t) \cup \text{Pa}_{\mathcal{G}'}^2(< t)$, because we proved identifiability of time lagged parents. Eq(18) raise a contradiction because, having these forward and backward models imply the verification of PDE 17. But we chose s such that this PDE is not verified hence contradiction. Then, the identifiability of our Temporal Restricted Heteroscedastic noise.

H.4 Proof of theorem 4.3

In this section, we want to prove the identifiability of mixture of Temporal causal models either in the case of Temporal Heteroscedastic Gaussian noise, Temporal Restricted Heteroscedastic noise and Homoscedastic NonGaussian noise.

Proof. Let $\mathcal{F}$ be a family of identifiable temporal causal models either from TGHNM or TNGNM, $\mathcal{F} = (p_{\theta^r}(\cdot | \cdot, \mathcal{G}^r))_{r \in \mathbb{N}^*}$ that are linearly independent and let $\mathcal{M}_K$ be the family of all K -finite mixtures of elements from $\mathcal{F}$, i.e.

$$\mathcal{M}_K = \left\{ p(x_t | x_{<t}) = \sum_{r=1}^K \pi_t(\omega^r) p_{\theta^r}(x_t | x_{<t}, \mathcal{G}^r), p_{\theta^r}(\cdot | \cdot, \mathcal{G}^r) \in \mathcal{F}, \forall t \in \mathcal{T} : \pi_t(\omega^r) > 0 \text{ and } \sum_{r=1}^K \pi_t(\omega^r) = 1 \right\}$$

First, we introduce a result from Yakowitz & Spragins [58] that established a necessary and sufficient condition for the identifiability of finite mixtures of multivariate distributions.

Theorem H.11 (Identifiability of finite mixtures of distributions, Yakowitz & Spragins [58]). *Let $\mathcal{F} = \{F(x; \alpha), \alpha \in \mathbb{R}^m, x \in \mathbb{R}^n\}$ be a finite mixture of distributions. Then $\mathcal{F}$ is identifiable if and only if $\mathcal{F}$ is a linearly independent set over the field of real numbers.*

We will further assume that it exists two distribution such that $\forall (x_t, x_{<t})$ covering the space value of random variables $(X_t, X_{<t})$:

$$\begin{aligned} p(x_t | x_{<t}) &= \sum_{r=1}^K \pi_t(\omega^r) p_{\theta^r}(x_t | x_{<t}, \mathcal{G}^r) \\ p(x_t | x_{<t}) &= \sum_{r=1}^{\tilde{K}} \pi_t(\tilde{\omega}^r) \tilde{p}_{\tilde{\theta}^r}(x_t | x_{<t}, \tilde{\mathcal{G}}^r) \end{aligned} \quad (19)$$

Our objective is to show first that $K = \tilde{K}$, it exists a permutation σ and a translation function $\varrho : \mathbb{R}^2 \rightarrow \mathbb{R}^2$: $(\theta^r, \mathcal{G}^r) = (\tilde{\theta}^{\sigma(r)}, \tilde{\mathcal{G}}^{\sigma(r)})$ and $\omega^r = \varrho(\tilde{\omega}^{\sigma(r)})$.

Using that $\mathcal{F} = (p_{\theta^r}(\cdot|\cdot, \mathcal{G}^r))_{r=1}^K$ are linearly independent and fixing t :

$$\begin{aligned} \sum_{r=1}^K \underbrace{\pi_t(\omega^r)}_{a_r} p_{\theta^r}(\mathbf{x}_t | \mathbf{x}_{<t}, \mathcal{G}^r) &= \sum_{r=1}^{\tilde{K}} \underbrace{\pi_t(\tilde{\omega}^r)}_{b_r} \tilde{p}_{\tilde{\theta}^r}(\mathbf{x}_t | \mathbf{x}_{<t}, \tilde{\mathcal{G}}^r) \\ \sum_{r=1}^K a_r p_{\theta^r}(\mathbf{x}_t | \mathbf{x}_{<t}, \mathcal{G}^r) &= \sum_{r=1}^{\tilde{K}} b_r \tilde{p}_{\tilde{\theta}^r}(\mathbf{x}_t | \mathbf{x}_{<t}, \tilde{\mathcal{G}}^r), \end{aligned}$$

and this true $\forall(\mathbf{x}_t, \mathbf{x}_{<t})$ covering the space value of the random variable $\mathbf{Y} = (\mathbf{X}_t, \mathbf{X}_{<t})$. By using theorem H.11, we can conclude that: $K = \tilde{K}$ and it exists a permutation σ such that: $(\theta^r, \mathcal{G}^r) = (\tilde{\theta}^{\sigma(r)}, \tilde{\mathcal{G}}^{\sigma(r)})$ and $\forall t \in \mathcal{T} : \pi_t(\omega^r) = \pi_t(\tilde{\omega}^{\sigma(r)})$.

To proof our identifiability as defined in definition 4.1, we still need to prove that $\omega^r = \varrho(\tilde{\omega}^{\sigma(r)})$. We have $\forall t \in \mathcal{T} : \pi_t(\omega^r) = \pi_t(\tilde{\omega}^{\sigma(r)})$, we take two indices $r, s \in [1 : K]$:

$$\begin{cases} \pi_t(\omega^r) = \pi_t(\tilde{\omega}^{\sigma(r)}) \\ \pi_t(\omega^s) = \pi_t(\tilde{\omega}^{\sigma(s)}) \end{cases} \quad (20)$$

To handle time varying weights identifiability, we will consider ratios of mixture weights:

$$\left\{ \begin{aligned} \frac{\pi_t(\omega^r)}{\pi_t(\omega^s)} &= \frac{\frac{\exp(\omega_1^r \cdot t + \omega_0^r)}{\sum_{j=1}^K \exp(\omega_1^j \cdot t + \omega_0^j)}}{\frac{\exp(\omega_1^s \cdot t + \omega_0^s)}{\sum_{j=1}^K \exp(\omega_1^j \cdot t + \omega_0^j)}} \\ \frac{\pi_t(\omega^{\sigma(r)})}{\pi_t(\omega^{\sigma(s)})} &= \frac{\frac{\exp(\omega_1^{\sigma(r)} \cdot t + \omega_0^{\sigma(r)})}{\sum_{j=1}^K \exp(\omega_1^{\sigma(j)} \cdot t + \omega_0^{\sigma(j)})}}{\frac{\exp(\omega_1^{\sigma(s)} \cdot t + \omega_0^{\sigma(s)})}{\sum_{j=1}^K \exp(\omega_1^{\sigma(j)} \cdot t + \omega_0^{\sigma(j)})}} \end{aligned} \right.$$

By Equation 20:

$$\begin{aligned} \frac{\pi_t(\omega^r)}{\pi_t(\omega^s)} &= \frac{\pi_t(\omega^{\sigma(r)})}{\pi_t(\omega^{\sigma(s)})} \\ \Leftrightarrow \exp[(\omega_1^r - \omega_1^s)t + (\omega_0^r - \omega_0^s)] &= \exp[(\omega_1^{\sigma(r)} - \omega_1^{\sigma(s)})t + (\omega_0^{\sigma(r)} - \omega_0^{\sigma(s)})] \\ \Leftrightarrow \forall t \in \mathcal{T} : (\omega_1^r - \omega_1^s)t + (\omega_0^r - \omega_0^s) &= (\omega_1^{\sigma(r)} - \omega_1^{\sigma(s)})t + (\omega_0^{\sigma(r)} - \omega_0^{\sigma(s)}) \end{aligned}$$

As a consequence of the last equation, we have for all the indices:

$$\begin{cases} \omega_1^r - \omega_1^s = \omega_1^{\sigma(r)} - \omega_1^{\sigma(s)} \\ \omega_0^r - \omega_0^s = \omega_0^{\sigma(r)} - \omega_0^{\sigma(s)} \end{cases} \quad \begin{cases} \omega_1^r - \omega_1^{\sigma(r)} = \omega_1^s - \omega_1^{\sigma(s)} = \Delta_1 \\ \omega_0^r - \omega_0^{\sigma(r)} = \omega_0^s - \omega_0^{\sigma(s)} = \Delta_0 \end{cases}$$

Hence it exists a translation function $\varrho : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, such that $\forall r \in [1 : K]$:

$$\omega^r = \varrho(\tilde{\omega}^{\sigma(r)}).$$

Hence our mixture of temporal causal models is identifiable as defined in definition 4.1.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Yes the main claims are clear in the abstract, introduction and in the whole paper

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Yes, we have a section in appendix [D](#) in which we talk about the limitation of our work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Our work offers three principal theoretical contributions. All assumptions and complete proofs appear in Appendix H. Due to space constraint, we state two theorems in the main text and the third in the appendix, with each theorem explicitly citing its underlying assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: Yes, we provide in the main text figures that illustrate the exact architecture used for our model. In the appendices, we provide the exact hyper-parameters needed to reproduce our results in all the presented experiments. [Our code is also provided in the supplementary material.](#)

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code, with all the instruction to reproduce our results, is provided in the supplementary materials

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All the detailed needed to run our code and to reproduce our results are presented in Appendix E.3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In the main text, we presented only average scores for the different metric. But, in the appendix F we provide erro bars for our model and all the other baselines.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: We provide time complexity and running time in appendices [F.3.2](#) and [F.1](#)

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: -

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[Yes\]](#)

Justification: In our conclusion, we talk about paper's broader impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: -

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We presented all the details about our model: parameters, architecture, data. Also the detailed proofs are provided

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: -

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: -

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: -

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.