Causal Representation Meets Stochastic Modeling under Generic Geometry

Jiaxu Ren*
Courant Instititute of Mathematics
New York University
UCSD

Yixin Wang
Department of Statistics
University of Michigan

Biwei Huang
Halıcıoğlu Data Science Institute
UCSD

Abstract

Learning meaningful causal representations from observations has emerged as a crucial task for facilitating machine learning applications and driving scientific discoveries in fields such as climate science, biology, and physics. This process involves disentangling high-level latent variables and their causal relationships from low-level observations. Previous work in this area that achieves identifiability typically focuses on cases where the observations are either i.i.d. or follow a latent discrete-time process. Nevertheless, many real-world settings require the identification of latent variables that are continuous-time stochastic processes (e.g., a multivariate point process). To this end, we develop identifiable causal representation learning for continuous-time latent stochastic point processes. We study its identifiability by analyzing the geometry of the parameter space. Furthermore, we develop MUTATE, an identifiable variational autoencoder framework with a time-adaptive transition module to infer stochastic dynamics. Across simulated and empirical studies, we find that MUTATE can effectively answer scientific questions, such as the accumulation of mutations in genomics and the mechanisms driving neuron spike triggers in response to time-varying dynamics.

1 Introduction

Inferring causal relationships among variables from observations capitalizes the potential of machine learning to advance scientific discovery, as it reveals underlying mechanisms that are not identifiable from observational distributions alone [1]. However, we often do not have access to the causal variables but only the high-dimensional perceptual data, and causal variables with their structures are unknown and thus need to be learned. Yet, these latent causal variables are often not identifiable [2, 3]. Recently, a growing number of studies on the disentanglement of latent causal representations have developed identifiability guarantees and proposed methods for estimating latent causal variables. Seminal works among them establish identifiability by leveraging sufficient variability in latent

*Work was done when Jay was an intern at Causality Lab, UCSD

Accepted to the NeurIPS 2025 Workshop on CauScien: Uncovering Causality in Science.

distribution arising from multiple-source data [4, 5], auxiliary variable [2, 6, 7, 8], or intervention to a latent causal graph [9, 10, 11, 12, 13].

Most recent work mentioned above aims to recover the latent causal variables that follow a discrete-time process [4, 5] and that are mixed by an invertible function. However, many latent causal variables of interest are continuous-time processes in practice; and the study of latent continuous-time causal variables driven by stochastic processes or systems of stochastic differential equations has received little attention, especially when mixing functions are non-invertible and more generic. For example, in video surveillance systems, cameras are strategically placed to detect and deter crime, safeguard against potential threats to the public, and manage emergency response situations during natural and man-made disasters [14, 15, 16]. In biology, fatal diseases such as cancer are principally caused by multiple cumulative mutations in driver genes as the colonial expansion proceeds. In neuroscience, the latent event dynamics trigger visible biological signals [17, 18]. Finding cancer-associated mutational genes and tracking their behavior through their representation has been given much more paramount importance in recent few decades [19, 20, 21]. Driven by the practical promise across applications, we study *when continuous-time latent stochastic point processes and their causal structure are identifiable*, and develop *algorithms to learn these latent dynamics from high-dimensional data*.

2 Problem Setup: Causal Representation with Stochastic Point Process

2.1 Preliminaries and notations

Let $O_t \in \mathbb{R}^n$ be observable data, and $Z_t \in \mathbb{R}^p$ be a latent causal process with independent noise $\epsilon_t \in \mathbb{R}^p$. O_t is being generated from latent point processes Z_t through an unknown, arbitrary mixing function f. A multi-way array $A^{\otimes d}$ denotes the tensor/Kronecker product. In a time process, $\Phi \in \mathbb{R}^{p\times p}$ denotes the transition operator (e.g., an autoregressive coefficient matrix or continuous kernel matrix) and the symbol $\star$ represents the convolution operator with kernel effects. We assume a probability space $(S, \mathcal{B}(S), \mathbb{P})$, where S is a Polish space (i.e., a complete separable metric space), $\mathcal{B}(S)$ is the Borel σ-algebra, and $\mathbb{P}$ is the probability measure, with μ a generic measure (e.g., for noise or intensity). $\mathcal{F}_t$ is the natural filtration up to the time t of a process. Let K denote an algebraically closed field of characteristic zero. Throughout, and unless specified otherwise, we work over this field K.

2.2 A Generative model for stochastic point processes

Throughout this paper, we consider a branch of non-homogeneous stochastic processes (Hawkes process) with dynamics governed by a conditional intensity defined as follows.

Definition 2.1 (Conditional intensity, informal [22])**.** Suppose a collection of latent processes that evolve stochastically and exhibit self-exciting dynamics over time. Specifically, let $Z_t := N_t$ denote the cumulative count process up to time t. We write $i \leftarrow j$ to indicate that the process j exerts an influence on i. Accordingly, the conditional intensity of process i at time t is given by

$$\lambda_t^i = \mu_i + \sum_j \int_0^t \phi_{i\leftarrow j}(t-s)\, N_s(\Delta)^j,$$

where $\mu_i \in U$ is the baseline rate and $\phi_{i\leftarrow j} \in \Phi$ characterizes the excitation kernel from process j to i. The counting process N_t^i and the conditional intensity λ_t^t satisfy: $N_{t+\Delta}^i - N_t^i = N_t(\Delta)^i$ and $\lambda_t^i = \frac{\mathbb{E}[dN_t^i|\mathcal{F}_t]}{dt}$.

Now, we formally set up the problem of identifying the generative model of stochastic point processes. We consider a collection of unstructured low-level observations $O = (O_t)_{t\leq T}$ generated from the latent process N_t through an arbitrary mixing function f. Compactly, by absorbing the kernel matrix Φ and the counting process N_t into a standard convolution operator, the generative model can be written as

$$O_t = f(N_t(\Delta)), \qquad \lambda_t = U + \Phi_t \star N_t(\Delta). \tag{1}$$

Thus, the central goal is to recover the parameter space $\Theta := (f, N_t, \lambda_t, \Phi, U)$ given samples or the full distribution of observations O_t. Concerning the theoretical soundness, we adopt the setting where the form of the mixing function, the number of latent causal processes, and their causal structure are fully unknown.

3 Recovery from algebraic signature of mixed manifolds

3.1 Maximally identifiable equivalent classes

We begin by introducing the maximal equivalent class that can be identified from discrete-time observations. Suppose we observe a discrete-time observation sequence $O_{t_0}, O_{t_0+\Delta}, O_{t_0+2\Delta}, \ldots, O_{t_0+k\Delta}$ at times $t_0, t_0+\Delta, t_0+2\Delta, \ldots, t_0+k\Delta$. Given a linear Hawkes-type intensity, we are provided a discretized latent process $Z_t^{(\Delta)}$ under the subsequence Δ, with its associated intensity: $\lambda_t^{(\Delta)} = u + \phi(\Delta) \cdot \Delta dN_t^{(\Delta)}$. The discrepancy between realizations arises due to the mismatch between the continuous-time dynamics and its discrete approximation, i.e., $\lambda_t^{(\Delta\to 0)} \neq \lambda_t^{(\Delta)}$, which implies that only latent processes generated under the same discretization scale Δ as the observation resolution can be recovered from $O_t^{(\Delta)}$. Therefore, the identifiability of the underlying latent dynamics is constrained to a discrete-time equivalence class determined by the resolution of observation. To capture the distribution-level changes and dynamics, we argue that recovering the distribution behavior of the latents suffices in most scientific tasks, and it can be used to generate the latents of any other Δ scale. Accordingly, we are able to identify only an equivalence class, as introduced in the subsequent definition.

Definition 3.1 (Weakly-convergent equivalent class)**.** Let (dN_t, λ_t) denote the ground-truth latent point process and its associated continuous-time intensity function. A pair $(Z^{(\Delta)}, \lambda^{(\Delta)})$ is said to belong to the **Weakly-convergent equivalent class** of (dN_t, λ_t) if it satisfies the following weak convergence condition:

$$(Z^{(\Delta)}, \lambda^{(\Delta)}) \xrightarrow[\Delta\to 0]{d} (dN_t, \lambda_t),$$

i.e., the estimated latent process and its discrete-time intensity converge in distribution to the ground-truth continuous-time process as the resolution parameter $\Delta \to 0$.

Thanks to Def. 3.1, it is sufficient to find such a model belonging to the equivalent class and establish its identifiability, as given in the following lemma.

Lemma 1 (Bounding point process in Variational approximation)**.** *Let $N_t \in \mathbb{R}^p$ be a multivariate point process whose conditional intensity function λ_t is governed by a convolution structure described in Eq. (1) and ϵ_t is a mean-zero and mutually independent noise. Then the intensity model admits the following weak convergence:*

$$\lim_{\Delta\downarrow 0} Z_k^{\Delta} := \lambda_k^{\Delta\to 0} + \epsilon_k^{\Delta\to 0} \xrightarrow{w} N_t, \tag{2}$$

where the subscript k denotes an arbitrary subsequence process and λ_k^{Δ} is the corresponding intensity under the same subsequence.

Lem. 1 establishes the weak convergence of continuous stochastic processes under the corresponding weak topology. Roughly, for any compact time interval $[a, b]$, a subsequence process of the original process under such an interval converges to a *continuous-time* causal point process N_t. This convergence ensures that Z^{Δ} effectively represents N_t and maintains all causal structures. Without loss of generality, we can therefore directly work with Z^{Δ} and study its identifiability by analyzing the geometry of the associated parameter space.

3.2 Algebraic structure for stochastic causal representation

Cumulant is an important algebraic signature of its geometric property, as a full order cumulant precisely encodes the entire distribution, including the component-wise and time-wise dependency among variables. This enables the fine-grained mathematical nature of intervention effects beyond traditional mean and variance shifts. Under generic (non-Gaussian) conditions, the d-th order cumulant of a random variable $X = As$ admits closed-form expressions as $\kappa_d(X) = \sum_{j=1}^{p} \kappa_d(s)(A_j)^{\otimes d}$, which is a secant variety $\sigma_k(X)$ that views each tensor factor $(A_j)^{\otimes d}$ as indeterminate. If the matrix A is generic in an open set, the Kruskal rank condition is satisfied and thus $\sum_{j=1}^{p} \kappa_d(s)(A_j)^{\otimes d}$ has a unique decomposition. This means the idea $\mathcal{I} : \langle \kappa_d(X) - \sum_{j=1}^{p} \kappa_d(s)(A_j)^{\otimes d} \rangle$ has dimension zero.

We connect our reasoning to this and illustrate how high-order cumulants can capture sufficient statistical variability in the system, even under temporally independent Gaussian noise. In particular,

we adapt the setting in [23], modifying it to allow additive noise that is independent over time. To this end, we establish a result for full recovery of INAR processes in the asymptotic regime $p \to \infty$, by formulating identifiability conditions through algebraic-geometric constraints imposed on the latent space.

3.3 Generic identification from algebraic structure

Let J_f be the Jacobian matrix of f and $K = J_f(\mathbb{I}_p - \Phi)^{-1}$. We present the following assumption.

Assumption 1.

1. *f is a generic C^d map with a full rank J_f almost surely.*
2. *There exist at least p nonzero tensors $\bar{\kappa}_d(\Delta O_t)$ for $d \in D := \{d \mid \bar{\kappa}_{d+1}(\Delta O_t) = 0\}$, where $\bar{\kappa}_d(\Delta O_t)$ is the difference cumulant truncated at the first-order Taylor expansion of $O_t := f(Z_t)$.*
3. *The ideal $\mathcal{I} := \langle J_f - K^{(k)}(\mathbb{I}_p - \Phi^{(k)}),\ k = 1, 2, \ldots, p \rangle$ is such that the space Θ is zero-dimensional.*

Theorem 1. *Under Assum. 1, the latent sources with their causal structure are identifiable up to permutation and component-wise scaling.*

The C^d assumption is strictly weaker than requiring f to be a diffeomorphism, since even in the presence of directional collapse within the latent space, the cumulants may still faithfully transmit the non-redundant dependency structure to the observed domain. The core idea of our proof leverages the propagation behavior of algebraic structure under nonlinear transformations, allowing us to identify latent structure via the observed partial geometric-algebraic information on the mixed manifold O_t. We distinguish two cases: access to the full observational distribution $P(O_t)$, or access only to realizations drawn from a restricted distribution $P^R(O_t)$. In either case, the geometry of the cumulant can be checked: allowing tolerance of loss in distribution information up to a certain order, the cumulant admits a unique linear degeneration. This degeneration canonically determines a projective embedding of the Veronese variety, identical up to a component-wise scaling and permutation. Additionally, given multiple environments (interventions or variability) in condition (2), it follows that the full generative model cannot be contained in any hypersurface of positive dimension. That is, the cumulant hierarchy of each observed component has finite depth d, and the non-vanishing cumulants up to this order are sufficiently rich to ensure identifiability via tensor decomposition. The proposed rank condition (1) is classic and results in a generically unique decomposition of d order tensor $\kappa_d(O_t)$, which uniquely recovers the component $v_i := J_f(\mathbb{I}_p - \Phi)^{-1}$ up to permutation and rescaling. Condition (2) ensures we can find such p different generic points so that J_f and $(\mathbb{I}_p - \Phi)^{-1}$ can be further disentangled up to the same indeterminacy.

The proof strategy converts the identifiability problem into the precise geometry of latent manifolds associated with the time-dependent process. See the proof in B.1.

4 Recovery from MUTATE

4.1 Architecture of MUTATE

Building upon our identifiability theory, we formally introduce MUTATE (**MU**lti -**T**ime **A**daptive **T**ransition **E**ncoder), a novel Variational Auto-encoding framework for estimation of latent multivariate stochastic point processes. Importantly, the framework is modular and can be readily adapted to other types of stochastic processes with suitable modifications. We highlight two core features of MUTATE, each addressing a key challenge related to identifiability. First, the central objective of our method is to recover the latent realization sequence $\{Z^{\Delta}_{t_0}, Z^{\Delta}_{t_1}, \ldots, Z^{\Delta}_{T}\}$ from multiple unstructured observational sources $\{O_t\}_{t=0}^{T}$. Unlike prior frameworks that rely primarily on time-stamp conditional independence to enforce latent structure, our approach accounts for the nature of progressively adaptive stochastic processes. In such systems, the filtration $\mathcal{F}_T$, which captures the intrinsic history of the process, is defined as $\sigma\left(\bigcup_{0<t<T} \sigma(Z^{\Delta}_t)\right)$ and grows strictly over time.

As shown in Figure 1, this dynamically expanding information structure poses unique challenges for both identifiability and representation learning, which MUTATE is explicitly designed to address. In addition, to leverage mutually independent noise, we employ a power spectral density (PSD) decomposition module in the joint optimization of parameters, which automatically enforces global whiteness of the noise.

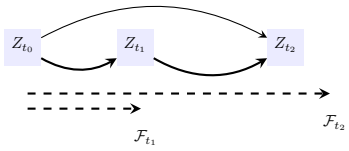

Figure 1: visualization of information loss in increasing filtration

Time adaptive transition module. We first employ an encoder $q_\phi(Z_t^\Delta|O_t^\Delta)$ to learn the estimated latents $Z_t^{(\Delta)} \sim q_\phi(Z_t^\Delta \mid O_t^\Delta)$ as commonly applied in representation learning frameworks. Recall that the latent process is modeled as $Z_t = \Phi \star Z_t + R_t$, where Φ denotes a global convolution kernel and $R_t = U + \epsilon_t$ is a residual process. Under our weak convergence condition, Z_t receives a well-structured representation $Z_t = (\mathbb{I} - \Phi)^{-1} \star (U + \epsilon_t)$, which is also a $(U + \epsilon_t)$-measurable process with tractable Power Spectrum Density (PSD): $S_{Z_t^{(\Delta)}}(w) = (I - \Phi)^{-1}\Sigma_{R_t}(I - \Phi)^{-H}$, where the baseline U is treated as a learnable parameter in the model and A^H is the Hamilton conjugate transpose of A. w is a continuous frequency variable. The inverse mapping f_Z^{-1} is an encoder with training parameters of neural networks. Importantly, the learned functional f maps the observation Z_t to a space of independent varying noise through the designated PSD decomposition that enforces Σ to be diagonal and recursively infers $H^\dagger = (\mathbb{I} - \Phi)^{-1}$. Then, the evaluated prior from the PSD module is sent to calculate the KL divergence. Decomposing $S(w)$, each of the transitions satisfies $\log p(Z_t^{(\Delta)}|\mathcal{F}_{t-}) = \log p[(I - \Phi) \star R_t^\Delta]$, which is the main part of the latent prior estimation. Our model is trained based on the Variational Auto-encoding framework. Therefore, we aim to maximize the log likelihood of observation $\log p_{data}(O)$ through the rule of the ELBO lower bound: $ELBO = -\mathcal{L}_{recon} - \alpha\mathcal{L}_{KL}$.

4.2 Simulation Study

To validate our identifiability results, we evaluate against several representative baselines, including TDRL [4], BetaVAE [24], SlowVAE [25], and PCL [7]. Among them, PCL and TDRL incorporate temporal dependencies by leveraging historical information and explicitly enforcing conditional independence among latent variables to recover underlying dynamics. In contrast, BetaVAE and SlowVAE assume independent latent components and disregard any time-delayed mechanisms. A detailed simulation procedure is included in D.1.

Performance of all baselines and our model is shown in Table 1 and extended results are reported in Table A2. During training, both BetaVAE and SlowVAE tend to converge prematurely, typically reaching a local optimum within the first epoch and triggering early stopping. This behavior highlights their limitations in modeling temporal structures essential for identifying latent event-driven processes. TDRL performs reasonably when the lag module is set to a longer one (we use $L = 9$ in experiments) since it can harness shorter temporary contextual information. It is noticed that our identifiability can be readily applied to the prior framework by either adding the domain index in synthetic datasets or modulating the distribution shifts that change pairs of edges in the latent space. However, we also realize that the fully non-parametric setting is hard to interpret since our identifiability avoids such a case.

Table 1: MCC Scores with standard deviations for five kernels

Method	Ave.(↑ better)	Exponential	Powerlaw	Rectangular	nonlinear	nonparametric
TDRL	0.599	0.593±0.028	0.609±0.043	0.618±0.056	0.556±0.016	**0.616**±0.043
BetaVAE	0.141	0.153±0.863	0.128± 0.077	0.128±0.078	0.146±0.108	0.149±0.096
SlowVAE	0.115	0.108±0.075	0.104±0.073	0.104±0.073	0.126±0.074	0.131±0.076
PCL	0.375	0.395±0.034	0.330±0.029	0.330±0.029	0.414±0.028	0.404±0.028
MUTATE(ours)	**0.837**	**0.853**±0.218	**0.938**±0.036	**0.879**±0.102	**0.921**±0.029	0.598±0.013

References

[1] Judea Pearl. *Causality*. Cambridge university press, 2009.

[2] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, April 1999.

[3] Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by Nonlinear ICA with General Incompressible-flow Networks (GIN), January 2020.

[4] Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally Disentangled Representation Learning. *Advances in Neural Information Processing Systems*, 35:26492–26503, December 2022.

[5] Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric Xing, and Kun Zhang. Temporally Disentangled Representation Learning under Unknown Nonstationarity. *Advances in Neural Information Processing Systems*, 36:8092–8113, December 2023.

[6] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. *Advances in Neural Information Processing Systems*, 29, 2016.

[7] Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ICA of Temporally Dependent Stationary Sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, April 2017. ISSN: 2640-3498.

[8] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, April 2019. ISSN: 2640-3498.

[9] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional Causal Representation Learning. In *International Conference on Machine Learning*, pages 372–407. PMLR, July 2023. ISSN: 2640-3498.

[10] Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32540–32560. PMLR, 23–29 Jul 2023.

[11] Yibo Jiang and Bryon Aragam. Learning Nonparametric Latent Causal Graphs with Unknown Interventions. *Advances in Neural Information Processing Systems*, 36:60468–60513, December 2023.

[12] Simon Bing, Urmi Ninad, Jonas Wahl, and Jakob Runge. Identifying Linearly-Mixed Causal Representations from Multi-Node Interventions. In *Causal Learning and Reasoning*, pages 843–867. PMLR, March 2024. ISSN: 2640-3498.

[13] Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning Linear Causal Representations from Interventions under General Nonlinear Mixing. *Advances in Neural Information Processing Systems*, 36:45419–45462, December 2023.

[14] Vinicius Lima, Fernanda Jaiara Dellajustina, Renan O. Shimoura, Mauricio Girardi-Schappo, Nilton L. Kamiji, Rodrigo F. O. Pena, and Antonio C. Roque. Granger causality in the frequency domain: derivation and applications. *Revista Brasileira de Ensino de Física*, 42:e20200007, 2020. arXiv:2106.03990 [physics, q-bio, stat].

[15] Emmanuel Bacry and Jean-François Muzy. Hawkes model for price and trades high-frequency dynamics. *Quantitative Finance*, 14(7):1147–1166, July 2014.

[16] E. Bacry, K. Dayri, and J. F. Muzy. Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. *The European Physical Journal B*, 85(5):157, May 2012. arXiv:1112.1838 [physics, q-fin].

[17] Patricia Reynaud-Bouret and Sophie Schbath. Adaptive estimation for Hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5), October 2010.

[18] Lars Lorch, Andreas Krause, and Bernhard Schölkopf. Causal modeling with stationary diffusions. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1927–1935. PMLR, 02–04 May 2024.

[19] Ali Torkamani and Nicholas J. Schork. Identification of rare cancer driver mutations by network reconstruction. *Genome Research*, 19(9):1570–1578, September 2009.

[20] Matthew H. Bailey, Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, and *et al.* Bertrand. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, 173(2):371–385.e18, April 2018.

[21] Mona Nourbakhsh, Kristine Degn, Astrid Saksager, Matteo Tiberti, and Elena Papaleo. Prediction of cancer driver genes and mutations: the potential of integrative computational frameworks. *Briefings in Bioinformatics*, 25(2):bbad519, January 2024.

[22] Emmanuel Bacry and Jean-Francois Muzy. Second order statistics characterization of Hawkes processes and non-parametric estimation, January 2014.

[23] Paula Leyes Carreno, Chiara Meroni, and Anna Seigal. Linear causal disentanglement via higher-order cumulants. *arXiv preprint arXiv:2407.04605*, 2024.

[24] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

[25] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding, March 2021. arXiv:2007.10930 [stat].

[26] Patrick Billingsley. *Convergence of probability measures*. Wiley series in probability and statistics Probability and statistics section. Wiley, New York Weinheim, 2. ed edition, 1999.

[27] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes Processes in Finance. *Market Microstructure and Liquidity*, 01(01):1550005, June 2015.

[28] Matthias Kirchner. Hawkes and INAR(∞) processes. *Stochastic Processes and their Applications*, 126(8):2494–2525, August 2016. arXiv:1509.02007 [math].

[29] Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. 1971.

[30] Daryl J. Daley and David Vere-Jones. *An introduction to the theory of point processes. 1: Elementary theory and methods*. Springer, New York NY, 2. ed., 2. corr. print edition, 2005.

[31] Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning Temporally Causal Latent Processes from General Temporal Data, February 2022. arXiv:2110.05428 [cs, stat].

[32] Pierre Brémaud and Laurent Massoulié. Stability of nonlinear Hawkes processes. *The Annals of Probability*, 24(3), July 1996.

[33] Erhan Cinlar and RA Agnew. On the superposition of point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(3):576–581, 1968.

[34] Susan L Albin. On poisson approximations for superposition arrival processes in queues. *Management Science*, 28(2):126–137, 1982.

[35] Payam Dibaeinia and Saurabh Sinha. Sergio: a single-cell expression simulator guided by gene regulatory networks. *Cell systems*, 11(3):252–271, 2020.

[36] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. *Advances in Neural Information Processing Systems*, 34:22419–22430, December 2021.

[37] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting, March 2024. arXiv:2310.06625 [cs].

[38] Ruichu Cai, Zhifang Jiang, Zijian Li, Weilin Chen, Xuexin Chen, Zhifeng Hao, Yifan Shen, Guangyi Chen, and Kun Zhang. From Orthogonality to Dependency: Learning Disentangled Representation for Multi-Modal Time-Series Sensing Signals, May 2024.

[39] Xiangchen Song, Zijian Li, Guangyi Chen, Yujia Zheng, Yewen Fan, Xinshuai Dong, and Kun Zhang. Causal temporal representation learning with nonstationary sparse transition. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 77098–77131. Curran Associates, Inc., 2024.

[40] Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal Representation Learning from Multiple Distributions: A General Setting, February 2024.

[41] Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering Causality from Multivariate Hawkes Integrated Cumulants. 2018.

[42] Vanessa Didelez. Graphical Models for Marked Point Processes Based on Local Independence. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(1):245–264, February 2008.

[43] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. Causal Representation Learning for Instantaneous and Temporal Effects in Interactive Systems. 2023.

[44] Dingling Yao, Caroline Muller, and Francesco Locatello. Marrying Causal Representation Learning with Dynamical Systems for Science. 2024.

[45] Amir Mohammad Karimi Mamaghan, Andrea Dittadi, Stefan Bauer, Karl Henrik Johansson, and Francesco Quinzan. Diffusion-Based Causal Representation Learning. 26(7):556, 2024.

[46] Alexander Sokol. Intervention in Ornstein-Uhlenbeck SDEs, August 2013.

[47] Stephan Bongers, Tineke Blom, and Joris M. Mooij. Causal Modeling of Dynamical Systems, March 2018.

[48] Stephan Bongers, Tineke Blom, and Joris M. Mooij. Causal Modeling of Dynamical Systems, March 2022. arXiv:1803.08784 [cs].

[49] Philip Boeken and Joris M. Mooij. Dynamic Structural Causal Models, June 2024.

[50] Alexander Sokol and Niels Richard Hansen. Causal interpretation of stochastic differential equations. *Electronic Journal of Probability*, 19(none), January 2014. arXiv:1304.0217 [math].

[51] Ivana Bozic, Tibor Antal, Hisashi Ohtsuki, Hannah Carter, Dewey Kim, Sining Chen, Rachel Karchin, Kenneth W. Kinzler, Bert Vogelstein, and Martin A. Nowak. Accumulation of driver and passenger mutations during tumor progression. *PNAS*, oct 2010.

[52] Johannes G. Reiter, Alvin P. Makohon-Moore, Jeffrey M. Gerold, Alexander Heyde, Marc A. Attiyeh, Zachary A. Kohutek, Collin J. Tokheim, Alexia Brown, Rayne M. DeBlasio, Juliana Niyazov, Amanda Zucker, Rachel Karchin, Kenneth W. Kinzler, Christine A. Iacobuzio-Donahue, Bert Vogelstein, and Martin A. Nowak. Minimal functional driver gene heterogeneity among untreated metastases. *Science*, 361(6406):1033–1037, September 2018.

[53] Ali Torkamani and Nicholas J. Schork. Identification of rare cancer driver mutations by network reconstruction. *Genome Research*, 19(9):1570–1578, September 2009.

[54] Benjamin J. Raphael, Jason R. Dobson, Layla Oesper, and Fabio Vandin. Identifying driver mutations in sequenced cancer genomes: Computational approaches to enable precision medicine. *Genome Medicine*, 6(1):1–17, December 2014.

[55] Ruth Nussinov, Chung-Jung Tsai, and Hyunbum Jang. A New View of Activating Mutations in Cancer. *Cancer Research*, 82(22):4114–4123, November 2022.

Supplement to

"Causal Representation Meets Stochastic Modeling"

Appendix organization:

A Useful Lemmata

A.1 Preliminary lemmas

Lemma A.1 (Weak Convergence [26]). *Let $(S, \mathcal{S})$ be a Polish space equipped with its Borel σ-algebra, and let $\{Z_n\}_{n\in\mathbb{N}}$ and Z be S-valued random elements defined on a common probability space. Then the sequence $\{Z_n\}$ converges in distribution (i.e., weakly) to Z, denoted $Z_n \Rightarrow Z$, if and only if*

$$\lim_{n\to\infty} \mathbb{E}[f(Z_n)] = \mathbb{E}[f(Z)]$$

for all bounded continuous functions $f : S \to \mathbb{R}$.

The proof and demonstration of this lemma is classic in basic probability that we omit here. The weak convergence, in most cases, corresponds to the convergence of finite dimension distribution of a process or a variable.

Lemma A.2 (Tightness of the Measure $\mathbb{P}_{Z^{(\Delta)}}$). *Let $\{Z_n^{(\Delta)}\}_{n\in\mathbb{N}}$ be a sequence of S-valued random elements (e.g., stochastic processes or path evaluations) indexed by Δ and defined on a Polish space S with Borel σ-algebra. Then the sequence of corresponding probability measures $\{\mathbb{P}_{Z_n^{(\Delta)}}\}$ is tight. In particular, any subsequence admits a further weakly convergent subsequence.*

Tightness of a sequence of probability measures ensures the existence of well-behaved subsequences: every subsequence admits a further weakly convergent subsequence. This property is particularly useful in Polish spaces, where tightness is equivalent to relative compactness (precompactness) under the weak topology. However, it is important to note that precompactness does not imply full compactness; in general, a tight sequence need not converge without an additional uniqueness or limit identification argument. Thus, tightness provides necessary control over subsequential behavior, but does not guarantee full convergence of the entire sequence.

Lemma A.3 (Higher-Order Moment Bound Implies Lower-Order Bounds). *Let $\{Z_n\}_{n\in\mathbb{N}}$ be a sequence of real-valued random variables defined on a common probability space. Fix an integer $d > 0$. Suppose there exists a constant $C > 0$ such that*

$$\sup_{n\in\mathbb{N}} \mathbb{E}[|Z_n|^d] \le C.$$

Then for any $0 < p < d$, there exists a constant $C_p > 0$ such that

$$\sup_{n\in\mathbb{N}} \mathbb{E}[|Z_n|^p] \le C_p.$$

A.2 Background of Hawkes Process

For a point process to be well-defined, some non-trivial constraints are required, one of which is the stationary condition, an assumption widely adopted in most stochastic process literature to ensure the uniqueness of the process.

Assumption 2.

1. *(Stationary increments) The process $N_t^{(\Delta)}$ is wide-sense stationary, i.e., its first and second moments exist and are time-invariant. In particular, the intensity process $\mathbb{E}[\Lambda_t]$ is uniformly bounded and $dN_t^{(\Delta)}$ has stationary increments.*
2. *(Kernel Integrability) The convolutional causal kernel $\Phi_t \in \mathbb{R}^{p\times p}$ is square-integrable, i.e.,*

$$\int_0^\infty \|\Phi_t\|_F^2 \, dt < \infty,$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Assumption A.4 (Stability and stationary Increment, Proposition 1 in [27]). *The process N_t has asymptotically stationary increments, and intensity λ_t is asymptotically stationary if the kernel satisfies the assumption:*

$$\rho_{\Phi(t)} = \|\Phi(t)\| = \int_0^t |\Phi(t)| \, dt \text{ has spectral radium smaller than 1} \tag{A1}$$

Asm. A.4 gives a necessary condition so that the point process has stable, stationary increments in its intensity. In particular, it means the entire process tends to be stable with an unknown but fixed expectation of the conditional intensity $\mathbb{E}[\lambda_t^i] = \Lambda^i$. Restricted by the stationary increment assumption, the existence of the corresponding process is ensured by Lem. 2. To illustrate those conditions, we show a simpler version kernel in Example 1.

Lemma 2 (Proposition 6 in [28]). *If all conditions and results in Asm. A.4 hold almost everywhere, there exists only one determined process whose dynamics match observations with regard to* Λ^i.

Example 1. Consider a point process whose kernel functions relay causal influence with an exponential decay to other processes. The generating process thus be accordingly

$$\lambda_t^i = u^i + \sum_{j=1}^{p} \int_0^t \alpha^{ij} e^{-\beta(t-t')} \, dN_{t'}^j$$

shows the exponential kernel triggers influences that are sustaining but decaying as time proceeds. Technically, the induced causal influences, although decaying from inside the system dynamics, will not disappear unless the causal strength $\alpha = 0$ for all j.

A.2.1 Remarks on the filtration

In probability theory, the filtration $\mathcal{F}_t$ is defined as the smallest σ-algebra that renders the intensity process λ_t to be $\mathcal{F}_t$-adapted and measurable. This filtration is constructed by the minimal closure under set operations (e.g., union, intersection) over past events, ensuring that λ_t evolves consistently with the observable history [29, 30]. Therefore, for any filtration as its internal history, we have $\mathcal{F}_s \subseteq \mathcal{F}_t$, for $s \leq t$. Note that the filtration $\mathcal{F}_t$ may theoretically differ from the intrinsic history $\mathcal{H}_t$, which introduces additional challenges in the evaluation and modeling of point processes. For a comprehensive discussion on scenarios where $\mathcal{F}_t$ and $\mathcal{H}_t$ are defined differently, we refer the interested reader to [30]. We occasionally overload the notation dN_t^i, which represents an integral element in stochastic calculus, to distinguish it from its deterministic counterpart. Despite potential similarities in notation, they are fundamentally different: while standard calculus considers infinitesimal increments over fixed mesh widths (e.g., $dg(x)$ as $\Delta t \to 0$), the increment dN_t^i is a random variable governed by the stochastic process. Specifically, its realization at each infinitesimal interval is drawn from a Bernoulli process with intensity λ_t^i, such that $\mathbb{P}(dN_t^i > 0 \mid \mathcal{F}_t) = \lambda_t^i, dt$. In contrast to deterministic differentials, dN_t^i encapsulates the uncertainty of event occurrences within each interval. The kernel matrix Φ_t consists of time-decaying kernel functions that transmit the influence of past events across processes. It captures both time-delayed and causal dependencies, and plays a central role in modeling self-exciting or mutually-exciting dynamics.

A.3 Cumulants and Tensors

Cumulant tensor notation. The d-th order cumulant tensor of a random vector $X \in \mathbb{R}^p$ is denoted $\kappa_d(X) \in \mathbb{R}^{p \times \cdots \times p}$, and is symmetric in all modes. In ICA and CRL settings, cumulants of independent components often admit a CP form:

$$\kappa_d(X) = \sum_{r=1}^{R} \lambda_r \cdot v_r^{\otimes d},$$

where $v_r \in \mathbb{R}^p$ and $\lambda_r \in \mathbb{R}$. This structure enables identifiability of latent sources from cumulant information.

Tensor notation and operations. We denote an order d tensor as $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_d}$. The outer product $u^{(1)} \otimes \cdots \otimes u^{(d)} \in \mathbb{R}^{I_1 \times \cdots \times I_d}$ produces a rank-1 tensor with entries:

$$\mathcal{T}_{i_1,\ldots,i_d} = u_{i_1}^{(1)} \cdots u_{i_d}^{(d)}.$$

Given a tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ and a matrix $U \in \mathbb{R}^{J \times I_n}$, the *mode-n* product $\mathcal{T} \times_n U \in \mathbb{R}^{I_1 \times \cdots \times I_{n-1} \times J \times I_{n+1} \times \cdots \times I_N}$ is defined as:

$$(\mathcal{T} \times_n U)_{i_1,\ldots,i_{n-1},j,i_{n+1},\ldots,i_N} = \sum_{i_n=1}^{I_n} \mathcal{T}_{i_1,\ldots,i_N} \cdot U_{j,i_n}.$$

B Proof of Identifiability Theory

B.1 Proof of Thm. 1

B.1.1 Useful Lemmas

To potentially identify any latent components of dynamics, we must introduce tensor algebra beyond our current setting as we present the following important results.

Corollary B.1 (CP decomposition). *Let $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ be an order-N tensor. We say that $\mathcal{T}$ admits an exact rank-R* Canonical Polyadic (CP) *decomposition if there exist component vectors $a_r^{(n)} \in \mathbb{R}^{I_n}$ for each $r = 1, \ldots, R$, $n = 1, \ldots, N$, such that:*

$$\mathcal{T} = \sum_{r=1}^{R} a_r^{(1)} \otimes a_r^{(2)} \otimes \cdots \otimes a_r^{(N)} = [\![A^{(1)}, A^{(2)}, \ldots, A^{(N)}]\!],$$

Where $A^{(n)} = [a_1^{(n)}\ a_2^{(n)}\ \cdots\ a_R^{(n)}] \in \mathbb{R}^{I_n \times R}$ are the factor matrices.

Corollary B.2. *Let $X^{(1)}, X^{(2)}, \ldots, X^{(n)} \in \mathbb{R}^p$ be independent random vectors with nonzero d-th order cumulants, such that each admits the form*

$$\kappa_d(X^{(i)}) = \lambda_i \cdot v_i^{\otimes d}, \quad \text{for } i = 1, \ldots, n,$$

with $v_i \in \mathbb{R}^p$ and $\lambda_i \in \mathbb{R} \setminus \{0\}$. Let $\mathcal{T} := \kappa_d(X^{(1)} + \cdots + X^{(n)}) \in \mathbb{R}^{p \times \cdots \times p}$ be the d-th order cumulant tensor of their sum.

Assume that the matrix $V = [v_1\ v_2\ \cdots\ v_n] \in \mathbb{R}^{p \times n}$ satisfies

$$\operatorname{krank}(V) \geq \left\lceil \frac{2n + (d-1)}{d} \right\rceil.$$

Then the CP decomposition

$$\mathcal{T} = \sum_{i=1}^{n} \lambda_i \cdot v_i^{\otimes d}$$

is unique up to scaling and permutation.

B.1.2 Proof of Thm. 1

We prove Thm. 1 by showing that the tuple (f, Φ, U) is identifiable up to a component-wise transformation and permutation. Our proof is based on the dimension of the associated variety defining special hypersurfaces in a polynomial ring K^n. We study the nonlinear propagation of the cumulant structure to find the identifiability conditions for INAR processes. Given a generic nonlinear f, its exact cumulant $\kappa_d(O_t)$ follows an order d expansion of with Bell polynomial coefficients. For a smooth map $f : Z_t \to f(Z_t)$, we can construct $O_: = f(Z_{t+\Delta t})$ using Taylor expansion:

$$f(Z_{t+\Delta t}) = f(Z_t) + \frac{\partial}{\partial Z_t} f(Z_t) \Delta Z_t + \frac{1}{2} \frac{\partial^2}{\partial Z_t^2} f(Z_t)(\Delta Z_t)^2 + o(\Delta Z_t^3)$$

At this time, the expansion has rather abnormal behavior, as the order can be extremely large. However, the truncated expansion at order 1 has intriguing theoretical attractions. Let higher-order components be $\mathcal{R}(1)$, and recall $Z_t = H \star \epsilon_t$, we obtain the truncated differential process $\Delta \tilde{f}(Z_t)$, denoted as:

$$\Delta f(Z_t) - \mathcal{R}(1) = J_f \sum_{k=1}^{t} H_{t-s} \epsilon_s \tag{A1}$$

We treat all quantities appearing in Eq. (A1) as indeterminates in a polynomial ring

$$R = k\big[\{\Delta f(Z_t)\}_t,\ \mathcal{R}(1),\ J_f,\ \{H_{t-s}\}_s,\ \{\epsilon_s\}_s\big],$$

where k is a base field such as $\mathbb{R}$ or $\mathbb{C}$. For each time index t, the defining polynomial is $g_t := \Delta f(Z_t) - \mathcal{R}(1) - J_f \sum_{s=1}^{t} H_{t-s} \epsilon_s \in R$. This polynomial generates the principal ideal $\mathcal{I}_t =$

$\langle g_t \rangle \subset R$, and considering all time indices $t = 1, 2, \ldots, T$, we obtain the global ideal $\mathcal{I} = \langle g_1, g_2, \ldots, g_T \rangle \subset R$. The corresponding affine variety is then

$$V(\mathcal{I}) = \Big\{ (Z_t, \Delta f(Z_t), J_f, H_{t-s}, \epsilon_s, \mathcal{R}(1)) \in k^N \;\Big|\; g_t = 0 \text{ for all } t \Big\}.$$

It is evident that $V(\mathcal{I})$ is positive-dimensional, since the defining relations do not specify finitely many points. To obtain more structure, we consider higher-order statistics. In particular, the d-th order cumulant tensor of the transformed increments takes the form

$$\kappa_d\Big(\Delta \tilde{f}(Z_t)\Big) = \sum_{s=1}^{t} \sum_{j=1}^{p} \kappa_d^{(j)}(\epsilon) \cdot \Big(J_f H_{t-s}^{(:,j)}\Big)^{\otimes d}.$$

This expression shows that the cumulant naturally defines a point in the projective tensor space

$$\mathbb{P}(V^n \otimes V^n \otimes \cdots \otimes V^n),$$

where the number of tensor factors equals d. Hence, while the affine variety $V(\mathcal{I})$ is too large to give identifiability, the cumulant tensors lift the problem into a projective geometric setting, where connections to secant varieties of the Veronese embedding provide a natural framework for studying uniqueness and decomposition. So far, the generic mixing f is preserved by its Jacobian matrix J_f; hence, identifying J_f is equivalent to the recovery of f up to a constant. Now, we are ready to prove our main theorem. Without loss of generality, we write J_f as F since they behave the same way in an algebraically closed field.

Step 1: Uniqueness of mixing kernel $F(\mathbb{I}_p - \Phi)^{-1}$ We prove this supporting result via an extension of Proposition 3.1 in [23]. In the classical linear source decomposition (LSD) setting, the d-th order cumulant of X admits the following tensor decomposition: $\kappa_d(X) = \sum_{i=1}^{q} \kappa_d(\epsilon_i) \cdot (B_i)^{\otimes d}$ under the assumption that the components of ϵ are non-Gaussian with non-vanishing d-order cumulants, and that multiple interventions are available. The sufficient order d cumulant of each Z_t for a fixed $t = t_i$ is

$$\kappa_d(Z_t) = \kappa_d[(I - \Phi)^{-1} \star \epsilon_t] = \kappa_d[\sum_{k=1}^{t} H_{t-s}\epsilon_s]$$

for each s, the linear transformation H_{t-s} results in a multi-linear transformation of their cumulants

$$\kappa_d(H_{t-s}\epsilon_k) = (H_{t-s})^{\otimes d} \mathcal{C}^d_{\epsilon_s} = \sum_{i=1}^{p} \kappa_d(\epsilon_s^i)(H_{t-s})_j^{\otimes d}$$

The full order d cumulant of the mixed manifold O_t is

$$\begin{aligned}
\kappa_d(O_t) &= \kappa_d(\underbrace{FZ_t, FZ_t, \ldots, FZ_t}_{d \text{ times}}) \\
&= F^{\otimes d} \cdot \kappa_d(\underbrace{Z_t, Z_t, \ldots, Z_t}_{d \text{ times}}) \qquad \text{(A2)} \\
&= \underbrace{F \otimes F \otimes \cdots \otimes F}_{d \text{ times}} \cdot \kappa_d \left(\underbrace{\sum_{s_1=1}^{t} H_{t-s_1}\epsilon_{s_1}, \sum_{s_2=1}^{t} H_{t-s_2}\epsilon_{s_2}, \ldots, \sum_{s_d=1}^{t} H_{t-s_d}\epsilon_{s_d}}_{d \text{ times}} \right) \\
&= \underbrace{F \otimes F \otimes \cdots \otimes F}_{d \text{ times}} \cdot \sum_{s_1=1}^{t} \cdots \sum_{s_d=1}^{t} \kappa_d \left(H_{t-s_1}\epsilon_{s_1}, H_{t-s_2}\epsilon_{s_2}, \ldots, H_{t-s_d}\epsilon_{s_d} \right) \\
&= \underbrace{F \otimes F \otimes \cdots \otimes F}_{d \text{ times}} \cdot \sum_{s_1=1}^{t} \cdots \sum_{s=1}^{t} \sum_{j=1}^{p} \kappa_d \left(H_{t-s}^{(:,j)}\epsilon_s^{(j)}, H_{t-s}^{(:,j)}\epsilon_s^{(j)}, \ldots, H_{t-s}^{(:,j)}\epsilon_s^{(j)} \right) \\
&= \underbrace{F \otimes F \otimes \cdots \otimes F}_{d \text{ times}} \cdot \left(\sum_{s=1}^{t} \sum_{j=1}^{p} \kappa_d^{(j)}(\epsilon) \cdot \underbrace{H_{t-s}^{(:,j)} \otimes H_{t-s}^{(:,j)} \otimes \cdots \otimes H_{t-s}^{(:,j)}}_{d \text{ times}} \right)
\end{aligned}$$

$$= \left(\sum_{s=1}^{t} \sum_{j=1}^{p} \kappa_d^{(j)}(\epsilon) \cdot \underbrace{F \otimes F \otimes \cdots \otimes F}_{d \text{ times}} \cdot \left(H_{t-s}^{(:,j)} \right)^{\otimes d} \right)$$

$$= \sum_{s=1}^{t} \sum_{j=1}^{p} \kappa_d^{(j)}(\epsilon) \cdot \left(F H_{t-s}^{(:,j)} \right)^{\otimes d} \tag{A3}$$

Unlike in a time-free process, the joint cumulant of a time process is of order d that is coupled with the number of time lags:

$$\kappa_d(O_{t_1}, \ldots, O_{t_d}) = \sum_{s=1}^{\min(t_1,\ldots,t_d)-1} \sum_{j=1}^{p} \kappa_d^{(j)}(\epsilon) \cdot \bigotimes_{\ell=1}^{d} \left(F H_{t_\ell - s}^{(:,j)} \right)$$

$$= \sum_{j=1}^{p} \sum_{s=1}^{\min(t_1,\ldots,t_d)-1} \kappa_d^{(j)}(\epsilon) \cdot \bigotimes_{\ell=1}^{d} \left(F H_{t_\ell - s}^{(:,j)} \right) \tag{A4}$$

We denote the Fourier transform of $x(t)$ with respect to time t as $\mathcal{F}[x](\omega)$. Using the convolution theorem and linearity of the Fourier transform, we have:

$$\mathcal{F}[\kappa_d(O_t)](\omega) = \mathcal{F}\left[\sum_{s \geq 0}^{t} \sum_{j=1}^{p} \kappa_d^{(j)}(\epsilon) \cdot \left(F H_{t-s}^{(:,j)} \right)^{\otimes d} \right]$$

$$= \mathcal{F}\left[\int_0^t \sum_{j=1}^{p} \kappa_d^{(j)}(\epsilon) \cdot \left(F H(t-s)^{(:,j)} \right)^{\otimes d} ds \right]$$

$$= \left(\sum_{j=1}^{p} \kappa_d^{(j)}(\epsilon) \cdot \mathcal{F}\left[\left(F H^{(:,j)} \right)^{\otimes d} \star \mathcal{U}(t-s) \right] \right)$$

$$= \left(\sum_{j=1}^{p} \kappa_d^{(j)}(\epsilon) \cdot \mathcal{F}\left[\left(F H^{(:,j)} \right)^{\otimes d} \right](\omega) \cdot \left(\pi\delta(\omega) + \frac{1}{i\omega} \right) \right) \tag{A5}$$

Since $\delta(\omega)$ vanishes everywhere except at $\omega = 0$, multiplying it by ω gives zero, Eq. (A5) yields

$$i\omega \mathcal{F}[\kappa_d(O_t)](\omega) = \left(i\omega \sum_{j=1}^{p} \kappa_d^{(j)}(\epsilon) \cdot \mathcal{F}\left[\left(F H^{(:,j)} \right)^{\otimes d} \right](\omega) \cdot \left(\pi\delta(\omega) + \frac{1}{i\omega} \right) \right)$$

$$= \left(\sum_{j=1}^{p} \kappa_d^{(j)}(\epsilon) \cdot \mathcal{F}\left[\left(F H^{(:,j)} \right)^{\otimes d} \right](\omega) \cdot i\omega \cdot \left(\pi\delta(\omega) + \frac{1}{i\omega} \right) \right).$$

which reads

$$\left(\sum_{j=1}^{p} \kappa_d^{(j)}(\epsilon) \cdot \mathcal{F}\left[\left(F H^{(:,j)} \right)^{\otimes d} \right](\omega) \cdot i\omega \cdot \left(\pi\delta(\omega) + \frac{1}{i\omega} \right) \right) = \left(\sum_{j=1}^{p} \kappa_d^{(j)}(\epsilon) \cdot \mathcal{F}\left[\left(F H^{(:,j)} \right)^{\otimes d} \right](\omega) \cdot 1 \right) \tag{A6}$$

By assuming non-Gaussianity in ϵ_t, for each j, Eq. (A6), hence $i\omega\mathcal{F}[\kappa_d(O_t)](\omega)$ has a unique decomposition of the summation of rank-1 tensor. Therefore, each column of the sub-linear mixing transferring matrix $\mathcal{F}\left[\left(F H^{(:,j)}\right)\right]$ is theoretically recovered up to a scaling and permutation π if all assumptions made are satisfied in Φ. This indicates that even if one needs to calculate the tensor decomposition unnecessarily, such uniqueness guarantees the possibility of further disentanglement. In the sequel, $\mathcal{F}\left[\left(F H^{(:,j)}\right)\right] DP$ is available; we thus obtain the unique indeterminacy as an immediate result of Lemma. B.3.

Lemma B.3. *Consider $\mathbb{F}$ is an algebraically closed field, the unknown indeterminacy DP is preserved in $\mathbb{F}$, that is the following relation*

$$F H_\tau^{(:,j)} = \hat{F} \hat{H}_\tau^{(:,j)} D_{j,\tau} P_{j,\tau},$$

$$\text{with } (P_{j,\tau}, D_{j,\tau}) \in \{(P_j, D_j) \mid \mathcal{F}[F H^{(:,j)}]\, D_j P_j = \mathcal{F}[\hat{F} \hat{H}^{(:,j)}]\}.$$

Proof. Let $\mathbb{F}$ be a field and $n \geq 1$ an integer. The *general linear group* of degree n over $\mathbb{F}$ is

$$\mathrm{GL}_n(\mathbb{F}) \;:=\; \{\, A \in M_n(\mathbb{F}) \;\mid\; \det(A) \neq 0 \,\}.$$

Equivalently, if V is a n-dimensional vector space over $\mathbb{F}$,

$$\mathrm{GL}(V) \;:=\; \mathrm{Aut}_{\mathbb{F}}(V) \;=\; \{\, T : V \to V \text{ linear isomorphisms} \,\},$$

and any choice of basis identifies $\mathrm{GL}(V)$ with $\mathrm{GL}_n(\mathbb{F})$. It is obvious that $\mathcal{F}[\mathbb{F}]$ is exactly a subgroup of $\mathrm{GL}(\mathbb{F})$ as the group $\mathrm{GL}_n(\mathbb{F})$ satisfies: i) It is precisely the set of all invertible linear transformations (invertible matrices). ii) If $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, then $\mathrm{GL}_n(\mathbb{F})$ is an open subset of $M_n(\mathbb{F})$ since $\mathrm{GL}_n(\mathbb{F}) = \det^{-1}(\mathbb{F} \setminus \{0\})$, and it is a Lie group. By the definition of kernel matrix, one notes i) trivially holds due to the maximal spectrum being less than 1. For ii), $\det^{-1}$ denotes the preimage of the open set $\mathbb{F} \setminus \{0\}$ under $\det$. Cutting the one-dimensional line at 0 produces two open intervals (for $\mathbb{F} = \mathbb{R}$) or a punctured plane (for $\mathbb{F} = \mathbb{C}$), hence the preimage is open in $M_n(\mathbb{F})$. Therefore, the permutation and scaling must be preserved in $M \in \mathbb{R}^{p \times p}$. □

So far, the original kernel mixing matrix FH_τ is recovered up to the same permutation and scaling for any τ. In the sequel, what needs to be proved is the recovery of the causal structure as well as its full parameter space.

Step 2: Uniqueness of causal kernel Φ Once FH is recovered as unique generic points, the full identifiability is obtained by bounding the dimension of the associated variety $V : \langle F - K^{(k)}(\mathbb{I} - \Phi) = 0, k = 1, 2, \cdots, p \rangle$ to be zero. Geometrically, $\dim(V)$ coincides the fundamental identifiability of the entire parameter space unless other constraints are imposed. Applying Theorem 1.5 of [23] to both upper and lower triangular parts of Φ, the identifiability is achieved by the minimum P generic points with $P = p$, the number of processes. This follows the idea $F - K^{(k)}(\mathbb{I} - \Phi)$ leads to a simper ideal lying in a lower-dimension ambient space such as $\langle F - K^{(0)}(\mathbb{I} - \Phi) - (F - K^{(k)}(\mathbb{I} - \Phi)) \rangle$ degenerates as a linear system.

Step 3: Independent conditions for each Φ under the generic F When Φ_{t-s} is fully recovered, the full matrix F can be obtained as

$$F = K(\mathbb{I}_p - \Phi).$$

If F is injective, identification of F is equivalent to identifying each Φ individually, due to the direct multiplication of F^{-1} (or the pseudo-inverse $F^\dagger$) with K. However, identification of the full generating model is hindered by the genericity of F. Even if K is unique up to the usual indeterminacies, recovering other kernel matrices $\Phi_{t-s'}$ requires analogous identifiability conditions for each individual kernel matrix.

Under $k \in 1, 2, ..., p$ distinct contexts—each introducing sufficient variability in the distribution, or ensuring that each lag k receives at least one intervention that shifts the downstream mechanism—the full latent structure is identifiable up to the same indeterminacy.

Recovery of baseline U Once the full causal structure is recovered up to a scaling and permutation matrix, lower level moments of $\mathbb{E}[F(\mathbb{I}_p - \Phi)^{-1} \star (U + \epsilon_t)]$ are can be directly computed to find U up to the same indeterminacy.

B.1.3 Identifiability under Gaussian Noise

Now we focus on the case where non-Gaussianity does not hold for the entire time process. When the noise is Gaussian, $\kappa_d(O_t) = 0$ for all $d \geq 3$, which leads to a $\mathbf{0} \in \mathbb{R}^{p^{\times d}}$ (the d-th order zero tensor). The solution of decomposition is infinite, thus FH cannot be recovered up to a column scaling and permutation, nor can the latent transition graph $\mathcal{G}$. We argue that preserving only d- order cumulant of order $d \leq 2$ is a minimal building block for identification. $\kappa_2(O_t)$ is an order 2 variance-covariance matrix. Let $M_1, M_2, \ldots, M_T \in \mathbb{R}^{p \times p}$ be a collection of order-2 tensors. We define the order-3 tensor $\mathcal{X} \in \mathbb{R}^{T \times p \times p}$ via concatenation along the first mode (tensor slices):

$$\mathcal{K}_c = \mathrm{con}\,(M_1, M_2, \ldots, M_T)\,, \quad \text{where } \mathcal{X}_{t,:,:} = M_t. \tag{A7}$$

By standard tensor algebra, an order-3 tensor $\mathcal{X} \in \mathbb{R}^{T\times p\times p}$ can be reshaped or flattened into a higher-order tensor, under a specific indexing scheme. More generally, given a desired tensor order d, and assuming $T = p^{d-2}$, we define a transformation:

$$\mathcal{T} : \mathbb{R}^{T\times p\times p} \to \mathbb{R}^{p^d},\ \kappa_d(\varepsilon_t) \in \mathbb{R}^{\underbrace{p\times\cdots\times p}_{d \text{ times}}}$$

$$\kappa_2(\varepsilon_t) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \in \mathbb{R}^{p\times p}, \kappa_3(X_t) = \begin{bmatrix} \mathbf{0}_{p\times p} \\ \mathbf{0}_{p\times p} \\ \vdots \\ \mathbf{0}_{p\times p} \end{bmatrix} \in \mathbb{R}^{p\times p\times p}$$

that re-indexes the tensor slices M_t to fill the missing indices of an order d cumulant tensor. The replacing and re-indexing rule is illustrated in an order 3 tensor as a real plane, assuming $d-1$ is the maximal order such that $\kappa_d = 0$.

Under this transformation, each slice M_t is interpreted as contributing to a specific mode configuration of the higher-order tensor. That is, the tensor $\mathcal{X}$ is "lifted" into a d-way tensor by embedding each $p \times p$ matrix slice as filling in the cumulant entries with fixed positions in the first $d-2$ indices corresponding to $t \in \{1, \ldots, T\}$, and varying the remaining two indices over $p \times p$. This leads to the same form as Eq. (A4) where all $\kappa_d(\epsilon_t) \neq 0$. Under Corollary B.1 and B.2, the new tensor has a unique decomposition of rank-1 tensor summation. To be specific, we assume v_i has no pair of columns to be collinear. This ensures the identification of $F(I-\Phi)^{-1}$ and restricts F to be injective to only $\operatorname{span}(H_j)$.

The sequential steps are the same for non-Gaussian noise since the construction of the ideal $\mathcal{I}^\star$ associated with its variety is not influenced by ϵ once FH is fixed up to a permutation and scaling.

Discussion of noise. Consequently, when Gaussianity is assumed, the distribution of O_t is fully characterized by the first two cumulants. This property implies that the entire cumulant expansion, and hence any higher-order dependency, collapses at second order. In this sense, the Gaussian distribution is the unique fixed point of the cumulant hierarchy at order two. In temporal parametric transition [31], a widely known condition to ensure the component-wise identifiability of the latent process Z_t is to require that the driving noise ϵ is not a fully isotropic Gaussian. That is, the Gaussian noise distribution must shift under either intervention [13] or exhibit heterogeneity in its variance. This is because all cumulants of order $p > 2$, which encode the exact causal dependencies, vanish for Gaussian noise. As a result, sufficient variability can only arise from changes in the second-order cumulant. Our results reflect that non-isotropic Gaussian noise is a *necessary* but not a *sufficient* condition for full identifiability of the time-delayed generative model and parameters.

B.1.4 Identifying causal structure

Our proof is constructive: with the minimal hierarchy satisfied by any type of variability (i.e., soft intervention) such that no fixed value of entry $(j \to i)$ in $\Phi(w)$ induces a dependence removal, the transitive closure $\bar{\mathcal{G}}$ of the ground truth process with its causal structure can be recovered up to the trivial transformation aforementioned. If a time process has no instantaneous influence it must have $TC(\mathcal{G}) = \mathcal{G}$, we can recover the original process and its causal structure up to a scaling and permutation π.

C Proof of Supportive Results

C.1 Proof of Lem. 1

This lemma significantly constitutes the reasoning chains that lead to our identifiability results. We restate the original statement to cover more details and background in point process and theory of weak topology and convergence.

Lemma C.1 (Bounding Point Process in intensity, *constructive*)**.** *For a measurable mapping* $N^\Delta : (\Omega, \mathcal{F}) \to (M_p, \mathcal{M})$ *such that* $\omega \mapsto N(\omega)$ *is a point process at scale* Δ*. Let* Δ *be the control operator*

for any subsequence of its point process. Consider $A \in \mathcal{B}$ generated by the topology $\mathcal{M}_p := \mathcal{B}(M_p)$. λ and ϵ is defined on this metric space. If λ satisfies the stationary increment condition, then we can establish the weak convergence of the constructed equivalent class:

$$\sum_{k:k\Delta\in A} \lambda_k^{(\Delta)} + \epsilon_k^{(\Delta)} \overset{w}{\Rightarrow} N(A) \quad \textit{for } \lambda_k^{\Delta} = \lim_{\Delta\to 0}\lim_{\delta\to 0} \frac{\mathbb{E}[dN^{\Delta}|\mathcal{F}]}{\delta}$$

Proof. We start the proof with a trivial case. If $\delta = \Delta_1 = 1$, the condition always trivially holds. In this case, we only need to show $N_t^{(\Delta_1)} = \lambda_t^{(\Delta_1)} + R_t$ by simply using the tower rule. Therefore, our proof gives more attention to the non-trivial case for $\delta \neq 1$.

Case 2: $\delta \in (0, \Delta_1)$

The reasoning of this case becomes more complicated if the time step operator used for generating subsequences proportionally shrinks to a sufficiently small unit in $(0, \Delta_1)$. We rewrite the approximating sequence N to leverage the metricizability of the space. Since we work in a Polish space, the Borel δ-algebra is countably generated and the space is separable and metrizable. Given a measurable set $A \in \mathcal{B}$, and a metric ρ, define the open δ-neighborhood as:

$$\mathcal{A} = A^{\delta} := \{x \in \mathbb{R}^d : \rho(x, A) < \delta\}$$

By outer regularity of Borel probability measures on Polish spaces, for every $\epsilon > 0$, there exists a countable collection of open sets $\{A_i\}_{i\in\mathbb{N}}$ such that $\bigcup_i A_i \supset \mathcal{A}$ and $\sum_i \mu(A_i \setminus A) < \epsilon$. This allows us to approximate any compact subset from outside using open sets with arbitrarily small excess mass and ensures the approximating sequence is defined on a non-decreasing base. We paraphrase the convergence as

$$\sum_{k:k\Delta\in A} \lim_{i\to\infty} \frac{\mathbb{E}[N^{\Delta}(A_i)|\mathcal{F}]}{|A_i|} + \epsilon_k^{(\Delta)} \overset{w}{\Rightarrow} N(A) \quad \text{for } \Delta \to 0 \tag{A1}$$

The equation above is adapted from the continuous-time intensity for point processes. However, it requires us to work with two limit conditions for A_i with the $1/k$ closed ball shrinking to zero measure and for the subsequence operator Δ approaching to 0. A common method is to ensure dominated and uniform convergence of the limit. To harness information regarding the intensity in our convergence to a more generalized process, we work with only Δ to induce the same time scale of intensity function. Therefore, we have the equivalent condition

$$\sum_{k:k\Delta\in A} \frac{\mathbb{E}[\sum_{k=1} Z_k^{\Delta} - \sum_{k=1} Z_{k-1}^{\Delta}|\mathcal{F}]}{|A_i|} + \epsilon_k^{(\Delta)} = \sum_{k:k\Delta\in A} \frac{\mathbb{E}[Z^{\Delta}(\Delta)|\mathcal{F}]}{\Delta} + \epsilon_k^{(\Delta)} \overset{w}{\Rightarrow} N(A) \quad \text{for } \Delta \to 0 \tag{A2}$$

We remove the limit condition as it is clear that $|A_i|$ is of measure zero when $\Delta = 0$, which ensures the alignment between our topological property and plausibility to analyze only subsequences in the sequel. According to Lemma.2 by [28], for any compact interval $[a, b]^{(\delta)}$ with the number of bins $[b - a]/\delta$, $\mathbb{E}[N^{(\delta)}([a, b])] < (b - a + 2)(I - G^{(\delta)}(a, b))^{-1}\Lambda$ where $G(a, b) = \int_a^b \Phi(s)ds$ is a solution of the stochastic differential equation systems

$$\mathbb{E}[\lambda([a, b])] = \mathbb{E}[u + G(a, b)\Lambda], \text{ for } \mathbb{E}[\lambda(a, b)] = \Lambda$$

Note that, by reapplying tower rule, Eq. (A1) implies:

$$\lim_{\delta\to 0} \mathbb{E}[N_A^{(\delta\in(0,1))}] \to \lim_{\delta\to 0} \mathbb{E}[\frac{\mathbb{E}[N_A^{(\delta)}|\mathcal{F}_t]}{\delta}]$$

Next, we show the necessity of tightness of the corresponding probability measure $\mathbb{P}^{\Delta}$ for the left-hand of Eq. (A2) to achieve the desired convergence. Without loss of generality, we consider a nonparametric intensity function $\lambda_t = \psi(u + \int \phi(t - s)Z^{\Delta}(s)\, ds)$. Consequently, $\mathbb{E}[\lambda_t] = \Lambda$ and $\mathbb{E}[\lambda_t] = \mathbb{E}[\psi(u + \int \phi(t - s)Z^{\Delta}(s)\, ds)]$. Assume that ψ is α-Lipschitz and $\alpha\|\phi\|_1 < 1$ [32], so the mapping $F(\Lambda) = \psi(u + \|\phi\|_1\Lambda)$ is a contraction on $\mathbb{R}_+$. By Banach's fixed-point theorem, there exists a unique solution $\Lambda^{\star}$ to the equation:

$$\Lambda^{\star} = \psi(u + \|\phi\|_1\Lambda^{\star})$$

Formally, this can be rearranged as:

$$\psi^{-1}(\Lambda^\star) - \|\phi\|_1 \Lambda^\star = u \quad \Longrightarrow \quad \Lambda^\star = \left(\mathrm{id} - \|\phi\|_1 \cdot \psi^{-1}\right)^{-1}(-u)$$

provided that $\mathrm{id} - \|\phi\|_1 \cdot \psi^{-1}$ is invertible on the image of ψ.

To control the tail probability, we apply Markov's inequality:

$$\mathbb{P}\left(\sum_{k:k\Delta\in A} \frac{\mathbb{E}[Z^\Delta(\Delta)|\mathcal{F}]}{\Delta} + \epsilon_k^{(\Delta)} > M_\varepsilon\right) \leq \frac{\mathbb{E}[\sum_k \Lambda^\Delta]}{M_\varepsilon} \leq \frac{(b-a+2\delta)\cdot\Lambda^\star}{M_\varepsilon}$$

Here, we define:

$$M_\varepsilon := \frac{(b-a+2\delta)\cdot\Lambda^\star}{\varepsilon} \quad \text{where } \Lambda^\star = \psi(u + \|\phi\|_1\Lambda^\star)$$

This choice ensures the upper bound remains within the prescribed ε-level for all $\Delta \in (0, \Delta_1)$. Since the only thing we need is the precompactness, we will not establish any tighter bound. Tightness of measure indicates we can always find a subsequence $\lambda_{k_n}^\Delta + \epsilon_{k_n}^\Delta$ in $\lambda_k^\Delta + \epsilon_k^\Delta$ converges weakly to a sequence $\lambda^\star + \epsilon^\star$. This weak convergence of subsequences, however, cannot control the limit uniqueness for each sequence. Therefore, we also should further control the limiting behavior of each sequence by uniform convergence of the characteristic functional defined by the approximating process and the target process. □

D Detailed MUTATE Configuration

D.1 Simulation Regime

We simulate multivariate point processes and their converging equivalent class Z_t extensively studied in our identifiability theory. We sample all point processes using the Poisson Superposition method (rejection sampling from the upper bound of conditional intensity [33, 34]) in order to mimic highly dynamic changes in conditional intensity, to capture denser information contained in stochastic processes. Then we create corresponding converging classes as a proof-of-concept validation: A total of 20,000 latent trajectories are sampled for each of the five kernel functions—exponential, power-law, rectangular, simple nonlinear, and flexible mixing—under two noise regimes: heterogeneous noise and Gaussian mixture noise. To illustrate the latent events underlying the unstructured data, we also simulate stochastic dynamics for biological data using SERGIO [35], a GRN-guided gene expression simulator used in Lorch et al.'s [18] causal modeling as well. All observation O_t is obtained from latents Z_t through MLP and LeakyReLU nonlinearity mixing.

We demonstrate the generative process for INAR equivalent classes. For a fair comparison to those baselines mainly addressing step-wise conditional independence, we generate for both time-step dynamics and denser dynamics by changing the setup to very short kernel effects with $\tau \in (0.001, 0.01) = t - t'$. We generate stochastic point processes from three basic kernel response functions:

$$\phi_{\text{exponential}}(t) = \alpha e^{-\beta t'}, \alpha \sim \text{uniform}(0.1, 0.5) \text{ and } \beta \sim \text{uniform}[0.5, 2)$$

$$\phi_{\text{powerlaw}}(t) = \frac{\alpha}{(t+c)^\beta} \cdot \mathbf{1}, \alpha \sim \text{uniform}(0.5, 1.2), \beta \sim \text{uniform}[0.1, 0.8) \text{ and } \gamma \in \text{uniform}(1, 3, 1.8)$$

$$\phi_{\text{rectangular}}(t) = \frac{1}{T - T'} \cdot \mathbf{1}_{\{t' \leq T\}}$$

The baseline intensity u_0 is sampled from $\text{uniform}(0, 1, 0.2)$. All parameters of the basic kernel are uniformly sampled by ensuring $\alpha < \beta$ in exponential responses, $\alpha < \gamma$ in power-law response, respectively, to satisfy the stationary increment condition such that $|\phi| < 1$. In the simulation, we also consider two extreme cases for simple nonlinear intensity and nonparametric intensity. We construct the conditional intensity function by mixing latent features through a linear transformation followed by a non-linear activation. Specifically, we first compute a log-linear intensity using the expression

$$\lambda_t = \log(1 + \exp(z_\ell - r_\ell[:, \Delta, :]))$$

that ensures positivity and controls the scale of the output through a smoothed ReLU (i.e., softplus).In an alternative setting (`kernel == "np"`), we learn the intensity function using a small neural network

(MLP): a two-layer perceptron with ReLU activation, ending in a Softplus to maintain positive outputs. This setup enables flexible, data-driven modeling of intensity dynamics beyond purely additive or linear forms. We define the mixing intensity function using a two-layer feedforward neural network with ReLU and Softplus activations. Formally, the architecture is given by:

$$\lambda_t = \sigma_+ \left(W_2 \cdot \mathrm{ReLU}(W_1 \lambda_t(l) + b_1) + b_2 \right), \tag{A1}$$

where

- $\lambda_t(l) \in \mathbb{R}^d$ is the input linear basic intensity at time t,
- $W_1 \in \mathbb{R}^{64 \times d},\ b_1 \in \mathbb{R}^{64}$ are the weights and bias of the first layer,
- $W_2 \in \mathbb{R}^{d \times 64},\ b_2 \in \mathbb{R}^d$ are the weights and bias of the second layer,
- $\sigma_+(x) := \log(1 + e^x)$ denotes the Soft-plus activation.

This design ensures the output λ_t remains strictly positive and can model complex dependencies in the latent dynamics while maintaining numerical stability.

We model the transformation from the latent variable $Z_t \in \mathbb{R}^d$ to the observational space via a multi-layer mixing network. Specifically, for each layer $l = 1, \ldots, L-1$, the transformation is given by $Z_t^{(l)} = \mathbf{A}^{(l)} \cdot \sigma_{\text{leaky}}(Z_t^{(l-1)})$, where $\mathbf{A}^{(l)} \in \mathbb{R}^{d \times d}$ is an orthogonal mixing matrix and σ_{leaky} denotes the leaky ReLU activation with slope $\alpha = 0.2$. The initial input is $Z_t^{(0)} = Z_t$, and the final output $Z_t^{(L-1)}$ represents the observation-space signal.

D.2 Prior decomposition of time-adaptive module

Without loss of generality, we consider non-finite steps for a latent stochastic generative process, as discussed in Lem. 1, where $\Delta t \to 0$. This induces an equivalence that the intrinsic history—the filtration $\mathcal{F}_t := \sigma\left(\bigcup_{0<t<T} \sigma(Z_t^{\Delta})\right)$—ensures that the process $Z_t^{(\Delta)}$ is $\mathcal{F}_t$-adaptive and measurable.

We decompose the ELBO objective as follows:

$$\begin{aligned}
\text{ELBO} &= \log p(O) - D_{\text{KL}}(q_\phi(Z|O) \| p(Z)) \\
&= \mathbb{E}_{z \sim q(Z_t|O_t)}[\log p(O_t|Z_t)] + \mathbb{E}_{z \sim q(Z_t|O_t)} \left[\log \frac{q(Z_t|O_t)}{p(Z_t)} \right] \\
&= \mathbb{E}_{z \sim q(Z_t|O_t)}[\log p(O_t|Z_t)] - \mathbb{E}_{z \sim q(Z_t|O_t)} \left[\log q(Z_t|O_t) - \log p(Z_t) \right] \\
&= \mathbb{E}_{z \sim q(Z_t|O_t)} \left[\log p(O_t|Z_t) - \log q(Z_t|O_t) \right] + \mathbb{E}_{z \sim q(Z_t|O_t)} \left[\log p(Z_t) \right] \\
&= \mathbb{E}_{z \sim q(Z_t|O_t)} \left[\log p(O_t|Z_t) - \log q(Z_t|O_t) \right] + \mathbb{E}_{z \sim q(Z_t|O_t)} \left[\sum_{\mathcal{F}_0^+}^{\mathcal{F}_T} \log p(Z_t^{(\Delta)} | \mathcal{F}_t) \right]
\end{aligned}$$

The reason we can segment the increasing filtration in the last term is due to the nice property of $\mathcal{F}_t$-measurable sequence. We can show that filtration of $Z_t|Z_s, R_t$ and $Z_t|R_s$ is equal because it is well known that any p-order INAR sequence with stationary increments admits a moving average (MA) representation. Further construction of their filtration $\tilde{\mathcal{F}}_t(\text{resp.}R_{s<t})$ and $\mathcal{F}_t(\text{resp.}Z_{s<t}, R_t)$ can show

$$\tilde{\mathcal{F}}_t = \mathcal{F}_t$$

We prove the result in the sequel. For $\tilde{\mathcal{F}}_t$, Z_t is a measurable function for $s < t$. By causality of the convolution kernel $\Psi = (I - \Phi)^{-1}$ satisfying $\Psi_\tau = 0$ for $\tau < 0$, which indicates $Z_t \in \sigma(R_s : s < t)$. Then, we construct another filtration $\tilde{\mathcal{F}}_s : \sigma(R_u : u < s)$. By adaptivity $\tilde{\mathcal{F}}_s : \sigma(R_u : u < s) \subseteq \tilde{\mathcal{F}}_t : \sigma(R_u : u < t)$. Therefore, Z_s is also $\sigma(R_u : u < t)$-measurable. Since the minimal σ-algebra of the original $\mathcal{F}_t$-measurable function must be contained in its σ-algebra, we have $\sigma(Z_s) \subseteq \sigma(R_u : u < t)$ and $\sigma(\bigcup_{s \le t} \sigma(Z_s)) \subseteq \sigma(R_u : u < t)$. For $\mathcal{F}_t$, $R_t = Z_t - \Psi \star Z_t$ so R_t is $\sigma(Z_s : s \le t)$-measurable. Therefore, by a similar construction, it is evident that $\sigma(\bigcup_{s<t} \sigma(R_s)) \subseteq \sigma(Z_s : s \le t)$. Therefore, because $\tilde{\mathcal{F}}_t \subseteq \mathcal{F}_t$ and $\mathcal{F}_t \subseteq \tilde{\mathcal{F}}_t$, there must be $\tilde{\mathcal{F}}_t = \mathcal{F}_t$.

Following this set-up, the prior becomes:

$$Z_t \mid \mathcal{F}_t \sim \mathcal{N}\left(\begin{bmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_p(t) \end{bmatrix} \sum_{t'<t} (I-\Phi)^{-1},\ \sum_{t'<t} (I-\Phi)^{-1}\Sigma_{t'}(I-\Phi)^{-T}\right)$$

The latents are generated by $Z_t = (I-\Phi) \star R_t$, where R_t is modeled as isotropic Gaussian noise with mean U and variance Σ. Note that the variance matrix Σ_{Z_t} is zero for any $t - t' \neq 0$. By the Wiener-Khinchin Theorem, we have the covariance matrix $C_{Z_t}(0) = \frac{1}{N}\sum_{k=0}^{N-1} S_z(w_k)$, we drop the sub-index Z_t whenever no confusion is caused. Now we can derive the decomposition of the convolution prior as

$$\begin{aligned}
&\mathbb{E}_{z\sim q(Z_t|O_t)}\left\{\sum_{\mathcal{F}_0^+, Z_t}^{\mathcal{F}_T} \log p(Z_t^{(\Delta)}|\mathcal{F}_t)\right\} \\
&= \mathbb{E}_{z\sim q(Z_t|O_t)}\left\{\sum_{\mathcal{F}_0^+, Z_t}^{\mathcal{F}_T} \log p\left[(I-\Phi_t)\star \hat{R}_t^{(\Delta)}\right]\right\} = \mathbb{E}_{z\sim q(Z_t|O_t)}\left\{\sum_{\mathcal{F}_0^+, Z_t}^{\mathcal{F}_T} \log p\left[\int_0^t (I-\Phi_{t-t'})\hat{R}_{t'}^{(\Delta)}\, dt\right]\right\} \\
&= \mathbb{E}_{z\sim q(Z_t|O_t)}\left\{\sum_{\mathcal{F}_0^+, Z_t}^{\mathcal{F}_T} \log p\left[\mathcal{N}(\hat{U}_R, \sum H_t^{-1}\Sigma_{\hat{R}_t'}H_t^{-T})\right]\right\} \\
&= \mathbb{E}_{z\sim q(Z_t|O_t)}\left\{\sum_{\mathcal{F}_0^+, Z_t}^{\mathcal{F}_T} \log p\left[\mathcal{N}(\hat{U}_R, \underbrace{\sum H_t^{-1}(PSD_{\hat{Z}_t})\Sigma_{\hat{R}_t'}H_t^{-T}(PSD_{\hat{Z}_t}))}_{C_{p(Z_t)}(0)}\right]\right\} \\
&= \mathbb{E}_{z\sim q(Z_t|O_t)}\left\{\sum_{\mathcal{F}_0^+, Z_t}^{\mathcal{F}_T} \log p\left[\mathcal{N}(\hat{U}\sum_{t'<t} \underbrace{(I-\Phi(t-t'))^{-1}}_{(1-\Phi)(w_k)^{-1}\Sigma(1-\Phi)(w_k)^{-H}=S_Z(w_k)}, \frac{1}{N}\sum_{k=0}^{N-1} S_{Z_t}(w_k))\right]\right\} \qquad \text{(A2)} \\
&= \mathbb{E}_{z\sim q(Z_t|O_t)}\left\{\sum_{\mathcal{F}_0^+, Z_t}^{\mathcal{F}_T} \log p\left[\mathcal{N}(\hat{U}\underbrace{\sum_{\tau>0, w=0} (I-\Phi(\tau))^{-1}e^{-jw\tau}}_{\text{the inverse Fourier at } w=0}, \frac{1}{N}\sum_{k=0}^{N-1} S_{Z_t}(w_k))\right]\right\} \\
&= \mathbb{E}_{z\sim q(Z_t|O_t)}\left\{\sum_{\mathcal{F}_0^+, Z_t, N\in(N_0,T)}^{\mathcal{F}_T} \log p\left[\mathcal{N}(\hat{U}\mathrm{PSD}_{Z_t}(H(0)), \frac{1}{N}\sum_{k=0}^{N-1} S_{Z_t}(w_k))\right]\right\} \qquad \text{(A3)}
\end{aligned}$$

D.3 Explicit control for convolution prior

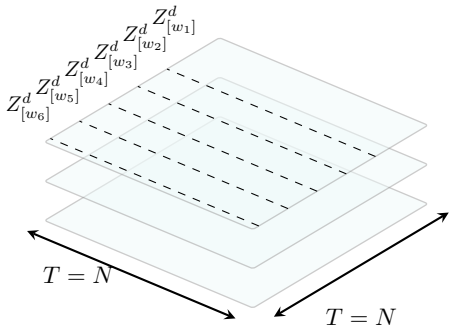

Figure A2: Visually Time-adaptive PSD Computation

Encoder-PSD flow As shown in Eq. (A3), a key component of our module is to efficiently compute the decomposition of PSD matrix. However, under milder regularity conditions, the PSD decomposition is not unique, thus only be recovered up to the minimal-phase. By the Theorem, the energy of the time domain and frequency domain is equivalent. Therefore, the encoded distribution is not sufficient to decompose the PSD matrix for which a reparameterization is needed. An encoder receives a T-length sequence O_t and returns the latent variable vector. Fast Fourier Transformation converts the latent sequence to a vector of equal length up to t:

$$[Z_{\mathcal{F}_0}, Z_{\mathcal{F}_t}, \cdots, Z_T] \Rightarrow \{[Z[f_0], Z[f_k], \cdots, Z[K]] | K = 0, 1, 2, \cdots, T\}$$

And the flow method is enforced by solving the following Wilson Factorization optimization problem for each $[Z[f_0], Z[f_k], \cdots, Z[K]]$, finding the transfer matrix

$$H^\dagger = \arg\min_{\Sigma_t = \sigma^2 I} PSD(Z_t) - H^\dagger \Sigma_t H^{\dagger H}$$

That is then sent to evaluate the true prior distribution, supporting the joint optimization of all loss components.

The summation of kernel products and integrated noise variables is guaranteed to converge to the true time-adaptive process under $\mathcal{F}_t$, provided that the time discretization is sufficiently dense. The latent variable Z_t^Δ is sampled from the encoder distribution q_ϕ and passed to the PSD decomposition module to compute the frequency-domain representation of the full kernel matrix $F_w[1 - \Phi_t]$ and the power spectral density $S_{\hat{R}_t}$. We further remark that the key step, spectrum decomposition, is completed for the entire encoded trajectory $\hat{Z}_{t_0:T}$, and the prior structure is ensured by segmenting filtration. This features the major difference in prior work that recursively constructs an equal-length sliding window for each latent. Filtration segmentation can work with causal masks that a more expressive encoder leverages. Note that transformer modules are not a required component for shorter sequences, i.e. $T < 100$. However, when the sequence is extremely long, as simulated in the conventional class of stochastic point processes, a transformer can be used in place of a common MLP encoder to learn much more expressive latent embeddings by utilizing the filtration attention from arbitrarily long past events.

Overall training loss. To encourage sparsity in transferring kernels, we follow the widely used penalty to jointly optimize:

$$\mathcal{L}_{Total} = \mathcal{L}_{\text{Recon}} - \beta \mathcal{L}_{\text{KLD}} - \gamma |\Phi| - \omega \mathcal{L}_{\text{PSD}} \tag{A4}$$

This training objective ensures the learned latent process is driven by a family of generalized white processes, as, in the Encoder-PSD flow, the decomposition is enforced by the prescribed isotropic noise, which omits any discriminator module as used in [31]. The coefficients in sparsity loss and PSD accuracy are registered as tunable hyperparameters.

Table A2: Reporting the best performance for each baseline

Method	Metric	Kernel Ave.	Exp	Power.	Rect.	Nonlin.	Nonpar.
TDRL	MCC	0.657	0.629	0.653	0.773	0.584	**0.644**
	$\mathcal{L}_{vae}$	0.449	0.308	0.302	0.302	0.871	**0.461**
BetaVAE	MCC	0.419	0.395	0.414	0.420	0.433	0.433
	$\mathcal{L}_{vae}$	9.480	8.538	7.533	8.424	11.683	11.220
SlowVAE	MCC	0.410	0.384	0.405	0.420	0.425	0.412
	$\mathcal{L}_{vae}$	362.890	395.107	448.105	452.472	238.520	280.247
PCL	MCC	0.440	0.469	0.379	0.430	0.474	0.449
	$\mathcal{L}_{vae}$(train)	0.693	0.693	0.694	0.693	**0.693**	0.693
MUTATE	MCC	**0.811**	**0.922**	**0.784**	**0.964**	**0.885**	0.501
	$\mathcal{L}_{vae}$	**0.670**	**0.448**	**0.508**	**0.253**	0.942	1.201

D.4 Extended Results

E Related Work

Causal disentanglement and learning time series. Although estimating and predicting time series is a classical problem in both traditional statistics and modern machine learning, representation learning has opened new avenues for leveraging latent information to better characterize time series data [36, 37]. Recently, learning causal representations in time series has become a foundational approach for enabling new scientific discoveries. This line of research primarily focuses on establishing identifiability of causal latent variables by exploiting nonstationary data [31, 4] and modular distribution shifts [5, 38] with sparsity constraints [39, 40] on the latent transition. Those works solve the identifiability problem of latent causal models by leveraging sufficient variability that can come from proper interventions or passive distribution shifts. Another line of research focuses on learning the underlying causal graph among latent variables

Learning Causality in Stochastic Processes. While learning causality remains a considerably more challenging task than causal discovery or representation learning, several efforts have been made to bridge these areas. Here we review existing approaches that link causal learning with stochastic modeling. Our scope is not limited to causal representation learning with stochastic processes, but extends to a broader set of problems that are closely related to either domain.

One representative direction in causal learning for dynamical systems is the study of Granger causality—a broader and looser notion compared to strictly structured causal models [41]. It is widely acknowledged that full causal recovery in such systems is impossible. Consequently, even the most recent work on stochastic processes can only determine whether a point process a is Granger-causal or non-causal with respect to another process b, typically formalized through *local independence* and the δ-separation rule [42]. Another active line of work concerns identifiability in dynamical systems [43]. However, to the best of our knowledge, none of these models provides provable guarantees for highly dynamical systems such as self-exciting or more general stochastic processes.

Connections between causal representation and dynamical systems have also been explored through ordinary differential equations (ODEs) [44]. Technically, these approaches recover only a set of parameters that are difficult to interpret as causal in the latent space, or at best allow stochastic dynamics in the observed variables. More recently, causal diffusion models have been proposed [45, 18], yet they largely treat diffusion as a standard denoising process and thus do not permit a well-structured stochastic latent causal representation.

Another important line of research investigates interventions on stochastic processes and the corresponding post-intervention distributions, which serve as the basis for causal inference [46, 47, 48, 49, 18]. The first attempt to introduce a causal interpretation into stochastic differential equations (SDEs) was made by the authors of [50], where interventions are defined as the removal of single variables in SDEs. They showed that causal principles in SDEs can be formalized as interventions, with the resulting post-interventional distribution identifiable via the infinitesimal generator. However, such interventions are too restrictive to capture more complex dynamical scenarios. Following this initial

line of work, [47] further develops methods for estimating stationary causal models by minimizing the deviation of stationarity of diffusion.

Stochastic representation in biological science. Prior work on the dynamics of cancer genomics has investigated modeling the underlying stochastic processes, typically under the assumption that all mutations can be identified through observable changes in protein binding and synthesis. A seminal line of studies focuses on methods and conditions under which mutation rates are treated as fixed for each driver mutation during tumor progression. Under these assumptions, the evolutionary dynamics can be effectively modeled using a linear Moran process with fixed population size [51, 52]. One branch of this literature aims to identify driver mutations that are directly manifested in observations. However, given the limited prior knowledge, mutation interactions are unlikely to be strictly linear or fixed. As a result, due to the inherent stochasticity and dynamical nature of cancer development, most driver mutations remain latent and their patterns are not readily discernible in protein sequences [53, 54, 55]. More recently, [55] reformulated this problem by introducing a framework to distinguish between weak and strong driver mutations to precisely characterize cancer progression. For example, in several cell cycles, a normal cell must accumulate multiple mutations in tumor-susceptibility genes to trigger oncogenesis. This process is inherently stochastic, and many mutational events may be dependent, self-exciting, or regulated by other processes. Identifying latent processes underlying disease-specific mutations and recovering their causal relationships is therefore crucial for computational biology and the planning of sequential cancer treatment regimens.

F Conclusion and Limitation

Our paper makes extensions of causal representation learning framework to stochastic causal dynamics (i.e., multivariate Hawkes Processes), a topic not yet covered in current CRL literature. We propose a new perspective that a branch of stochastic processes can be viewed as the corresponding equivalent class through INAR representation and weak convergence. Under those conditions, we show that the latent stochastic process can be identified up to a component-wise transformation and a scaling permutation matrix. Our theoretical result bridges the gap between stochastic modeling and causal representation. We also propose a novel framework to learn the time-adaptive transition dynamics to accurately estimate the latent processes. However, our work avoids the most complicated case for a fully nonparametric kernel, which, most of the time, can be replaced with a simpler kernel. Future direction may include solving this condition and causal representation learning for stochastic differential processes that manifest in rich scientific questions.