CAUSAL DISCOVERY IN STOCHASTIC PROCESSES WITH SIGNATURE KERNEL CONDITIONAL INDEPENDENCE TESTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Inferring the causal structure underlying stochastic dynamical systems from observational data holds great promise in domains ranging from science and health to finance. Such processes can often be accurately modeled via stochastic differential equations (SDEs), which naturally imply causal relationships via 'which variables enter the differential of which other variables'. In this paper, we develop conditional independence (CI) constraints on coordinate processes over selected intervals that are Markov with respect to the acyclic dependence graph (allowing self-loops) induced by a general SDE model. We then provide a sound and complete causal discovery algorithm, capable of handling both fully and partially observed data, and uniquely recovering the underlying or induced ancestral graph by exploiting time directionality assuming a CI oracle. Finally, to make our algorithm practically usable, we also propose a flexible, consistent signature kernel-based CI test to infer these constraints from data. We extensively benchmark the CI test in isolation and as part of our causal discovery algorithms, outperforming existing approaches in SDE models and beyond.

1 INTRODUCTION

Understanding cause-effect relationships from observational data can help identify causal drivers for disease progression in longitudinal data and aid the development of new treatments, act upon the underlying influences of stock prices to support lucrative trading strategies, or speed up scientific discovery by uncovering interactions in complex biological systems such as gene regulatory pathways. Causal discovery (or causal structure learning) has received continued attention from the scientific community for at least two decades (Glymour et al., 2019; Spirtes et al., 2000; Vowels et al., 2022) with a notable uptick in previous years particularly regarding differentiable score-based methods (Zheng et al., 2018; Brouillard et al., 2020; Charpentier et al., 2022; Hägele et al., 2023; Lorch et al., 2021; Annadani et al., 2023; Zheng et al., 2020), building on score matching algorithms (Rolland et al., 2022; Montagna et al., 2023b;a), and other deep-learning based approaches (Chen et al., 2022; Ke et al., 2023; Yu et al., 2019; Ke et al., 2020). Many of these approaches aim at improving the scalability of causal discovery in the number of variables and observations as well as at incorporating uncertainty or efficiently making use of interventional data.

However, causal discovery from time series data has received much less attention and been mostly neglected in these recent advances. At a fundamental level, causal effects in dynamical systems can only 'point into the future', which should make causal discovery in time resolved data intuitively simpler. Nevertheless, except for restricted settings, this promise has not been realized methodologically and causal discovery in time series, especially for systems evolving in continuous time, remains a major challenge (Singer, 1992; Runge et al., 2019; Lawrence et al., 2021; Runge et al., 2023).

Data generating process. We assume data to follow a stochastic process $X := (X^1, \dots, X^d)$, with X^k taking values in $\mathbb{R}^{n_k}$ ($n_k \geq 1$), and X satisfying the following system of stationary, path-dependent SDEs

$$\begin{cases} \mathrm{d}X_t^k = \mu^k(X_{[0,t]})\mathrm{d}t + \sigma^k(X_{[0,t]})\mathrm{d}W_t^k, \\ X_0^k = x_0^k \qquad \text{for } k \in [d] := \{1, \dots, d\}\,. \end{cases} \tag{1}$$

We call this the *SDE model*. The subscript $[0, t]$ at X means that μ^k (the 'drift') and σ^k (the 'diffusion') are functions of the entire solution up to time t, i.e., they are defined on $C([0, +\infty), \mathbb{R}^n)$ (or some suitable subspace thereof), where $n := n_1 + \ldots + n_d$. Each W^k is an m_k-dimensional Brownian motion, and σ^k maps to $n_k \times m_k$-dimensional matrices and is block-diagonal (corresponding to the individual processes) rendering the diffusions orthogonal. The noises W^k together with the (possibly random) initial conditions x_0^k are jointly independent. We refer to Rogers & Williams (2000) for details, Evans (2006) for a concise introduction, and Appendix A.2 for more intuition and discussion of the assumptions made here. Hence, the SDE together with a distribution over initial conditions defines our data generating process. Individual observations are paths, which in practice translate to stochastic, potentially irregularly sampled time series observations for $X^k_{[0,T]}$ with a maximum observation time T, see Figure 1. Appendix A.1 contains an overview table of the notations used.

Induced causal graph. Eq. (1) naturally implies cause-effect relationships: we call $i \in [d]$ a parent of $j \in V := [d]$ ($i \in \mathrm{pa}_j^{\mathcal{G}}$) when either μ^j or σ^j is not constant in the i-th argument. These parental relationships define a directed graph $\mathcal{G} = (V, E)$, which we call *dependence graph* of the SDE model. The **goal of causal discovery** is to infer the graph $\mathcal{G}$ induced by the SDE model from a sample of observed solution paths, depicted in Figure 1. Here, we only consider data generating processes leading to directed graphs without cycles of length greater than one (i.e., only 'loops' $X^k \to X^k$ are allowed). We still call these directed acyclic graphs (DAGs) for simplicity.

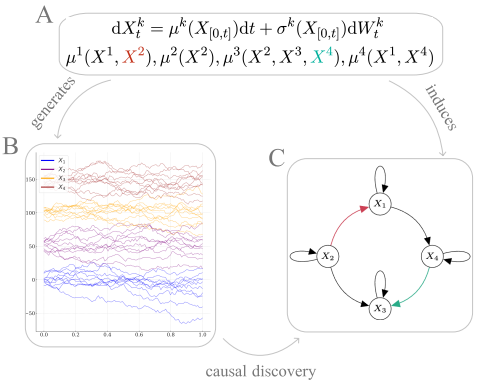


Figure 1: Illustration of causal discovery in the SDE model (A), leveraging conditional independencies in the observed samples (B) to infer the dependence graph of the SDE (C).

Limitations and requirements. The key limitation of our setting is the assumption of acyclicity (except for self-loops). We discuss fundamental issues arising from dropping the acyclicity constraint in Appendix A.3. Further, we focus on stationary SDE models, i.e., the coefficients in eq. (1) are not explicitly time-dependent, motivated by the requirement that causal relationships do not change over time. Despite these limitations, our setting is the first to simultaneously satisfy the following criteria for causal discovery in continuous time systems: (a) It does not rely on the 'discrete-time' assumption, i.e., that all variables are observed at the same, typically homogeneously spaced, time points (neither in the underlying model nor the actual observations). Instead, different variables can be observed at different, irregularly spaced time points in our setting. (b) We can handle partially observed systems with unobserved confounders. (c) Path-dependence (μ^k, σ^k may depend on the entire previous history $X_{[0,t]}$ of paths), including delayed SDEs, is typically not captured in existing approaches (Peters et al., 2022). (d) Beyond 'additive observation noise' it incorporates 'driving noise', where causal dependence may stem from the diffusion, not captured by the means of observed paths.

Contributions. Our main contributions include (a) developing CI constraints on coordinate processes over selected intervals that we prove to be Markov with respect to the acyclic dependence graph (with self-loops) induced by a general SDE model. (b) We then propose an efficient constraint based causal discovery algorithm and prove it to be sound and complete, assuming a CI-oracle, for both fully and partially observed data. Unlike existing constraint based methods in static settings, our algorithm uniquely recovers the full underlying graph or induced ancestral graph by exploiting the direction of time. (c) Finally, we propose a flexible, practical CI test on path space, rendering our method highly practicable. (d) We extensively demonstrate the test's efficacy and achieve superior performance in causal discovery compared to existing methods.

2 Background and Related Work

Causal discovery on time series. While virtually all causal discovery methods on time series exploit that the direction of time constrains edges to point forward, except for Laumann et al. (2023), most existing work assumes observations to be sampled on a regular time grid with a discrete (auto-regressive) law, where past observations $(X_{t-\tau}, \ldots, X_{t-1})$ determine the present $X_t = f(X_{t-\tau}, \ldots, X_{t-1}, \varepsilon_t)$ for some fixed lag τ, which can be seen as a static structural causal model

(SCM) over an 'unrolled graph' with variables at different time steps considered as distinct nodes (Assaad et al., 2022; Hasan et al., 2023; Peters et al., 2017). Based on the seminal work (Granger, 1969), various Granger-type approaches leverage the assumption that past values of a variable can help predict future values to detect lagged causal influences both for linear (Diks & Panchenko, 2006; Granger, 1980) and non-linear (Marinazzo et al., 2008; Shojaie & Fox, 2022; Runge, 2020; Pamfil et al., 2020) functional relationships. While non-parametric constraint-based methods for such discrete-time models (Runge et al., 2019; 2020; 2023) can handle partial observations, they are also fundamentally limited by the 'discrete-time' assumption making them unfit for irregularly sampled observations of continuous time systems, path-dependence, or diffusion dependence. Furthermore, they require numerous tests across all time points and rely on the correct estimation of a fixed 'lookback window' τ. The fundamental problems of the 'discrete-time' assumption have also been discussed in Runge (2018).

Other work focuses on the (static) equilibrium behavior of differential equations (Mooij et al., 2013; Bongers et al., 2018; Bongers & Mooij, 2018), exploiting invariance in mass-action kinetics (Peters et al., 2022), or attempts to directly learn the full dynamical law via non-convex optimization of heavily overparameterized neural network models (Aliee et al., 2021; 2022; Bellot et al., 2022; Wang et al., 2024). The latter are fundamentally limited by the fixed functional modeling structures and therefore also unable to model partially observed settings. Albeit its limited applicability to fully-observed, Markovian SDE models, SCOTCH (Wang et al., 2024), which uses a variational formulation to infer posterior distributions over possible graphs, is the current state-of-the-art baseline.

Constraint-based methods for continuous-time stochastic processes mostly rely on (conditional) local independence, an asymmetric independence relation (Schweder, 1970; Mogensen et al., 2018), which can be used to infer a (partial) causal graph in, e.g., point process models (Didelez, 2008; Meek, 2014; Mogensen et al., 2018; Mogensen & Hansen, 2020). However, besides the limitation to drift-dependencies, there exists no practical test of local independence for diffusions. The only work that actually tests local independence from data is (Christgau et al., 2023), which is heavily restricted to counting processes.

In this work, we overcome the challenges of irregular and partial observations, path- and diffusion dependence, and practical applicability by transforming CI statements of the form

$$X^I_{\mathcal{I}} \perp\!\!\!\perp X^J_{\mathcal{J}} \mid X^K_{\mathcal{K}}, \tag{2}$$

for disjoint $I, J, K \subseteq [d]$ between subsets of coordinate-processes X^I, X^J, X^K over intervals $\mathcal{I}, \mathcal{J}, \mathcal{K} \subseteq [0, T]$ into a practical hypothesis test using the signature kernel and the arrow of time.

Conditional independence tests. As a key building block of graphical models, there is a large literature on conditional independence tests, i.e., testing the null hypothesis $H_0 : X \perp\!\!\!\perp Y \mid Z$. When Z is discrete, CI testing reduces to a series of unconditional tests for each value of Z (Tsamardinos & Borboudakis, 2010). For continuous Z, nonparametric kernel-based methods often either check independence of residuals after kernel (ridge) regressing X and Y on Z (Shah & Peters, 2020; Lundborg et al., 2022; Zhang et al., 2011; Daudin, 1980; Strobl et al., 2019), or measure the distance between joint and marginal kernel mean embeddings (Muandet et al., 2017; Park & Muandet, 2020). For example, KCIPT (Doran et al., 2014) phrases it as a two-sample test of the null $H_0 : P(X, Y, Z) = P(X \mid Z)P(Y \mid Z)P(Z)$ and simulates samples from $P(X \mid Z)P(Y \mid Z)P(Z)$ via a permutation of the data that approximately preserves Z. Lee & Honavar (2017) improve on KCIPT with an unbiased estimate of Maximum Mean Discrepancy (MMD) (Gretton et al., 2006). Other permutation-based approaches require large sample sizes (Sen et al., 2017) or rely on densities (Kim et al., 2022), unfit for our setting.

Laumann et al. (2023) develop a CI test for functional data using the Hilbert-Schmidt conditional independence criterion (HSCIC) (Park & Muandet, 2020) and a permutation test (Berrett et al., 2019), which they use as part of the PC-algorithm (Glymour et al., 2019) for causal discovery. Even though their approach assumes $P_{X|Z}$ to be known (not the case in our setting), it is inapplicable under partial observations, and does not exploit the arrow of time (to identify graphs beyond the Markov equivalence class), we benchmark the work by Laumann et al. (2023) against our method in the experiments. CI testing faces fundamental practical challenges, such as the exponential growth in the number of values to condition on as dimensionality of Z increases, and theoretical limits, as any CI test's power is bounded by its size (Shah & Peters, 2020; Lundborg et al., 2022). This 'no free lunch' statement underscores the need for carefully selecting the right test. Most existing approaches

assume Euclidean spaces and do not generalize to path-valued random variables without densities, a gap we fill by combining kernel-based permutation tests (KCIPT, SDCIT) with the signature kernel for a tailored CI test on path space and proving its consistency.

Signature kernels. Signature kernels (Király & Oberhauser, 2019; Salvi et al., 2021a), a universal class for sequential data, have received attention recently for their efficiency in handling path-dependent problems (Lemercier et al., 2021; Salvi et al., 2021c; Cochrane et al., 2021; Salvi et al., 2021b; Cirone et al., 2023; Issa et al., 2023; Pannier & Salvi, 2024). The definition of the signature kernel requires an initial algebraic setup, which we keep as concise as possible—yet self-contained—here and refer the reader to Cass & Salvi (2024, Chapter 2) for more details. We provide some more informal intuition of signatures in Appendix A.4. Let $\langle \cdot, \cdot \rangle_1$ be the Euclidean inner product on $\mathbb{R}^d$. Denote by $\otimes$ the standard outer product of vector spaces. For any $n \in \mathbb{N}$, we denote by $\langle \cdot, \cdot \rangle_n$ on $(\mathbb{R}^d)^{\otimes n}$ the canonical Hilbert-Schmidt inner product defined for any $a = (a_1, \dots, a_n)$ and $b = (b_1, \dots, b_n)$ in $(\mathbb{R}^d)^{\otimes n}$ as $\langle a, b \rangle_n = \prod_{i=1}^n \langle a_i, b_i \rangle_1$. Define the direct sum of vector spaces $T(\mathbb{R}^d) := \bigoplus_{n=1}^\infty (\mathbb{R}^d)^{\otimes n}$, where it is understood that the direct sum runs over finitely many non-zero levels. The inner product $\langle \cdot, \cdot \rangle_n$ on $(\mathbb{R}^d)^{\otimes n}$ can then be extended by linearity to an inner product $\langle \cdot, \cdot \rangle$ on $T(\mathbb{R}^d)$ defined for any $a = (a_0, a_1, \dots)$ and $b = (b_0, b_1, \dots)$ in $T(\mathbb{R}^d)$ as $\langle a, b \rangle = \sum_{n=0}^\infty \langle a_n, b_n \rangle_n$. Other choices of linear extensions have been studied in Cass et al. (2023). We denote by $\mathcal{T}(\mathbb{R}^d)$ the Hilbert space obtained by completing $T(\mathbb{R}^d)$ with respect to $\langle \cdot, \cdot \rangle$.

The *signature transform* is a classical path-transform from stochastic analysis. For any sub-interval $[s, t] \subset [0, T]$ and any continuous path $X \in C_p([0, T], \mathbb{R}^d)$ of finite p-variation, with $1 \leq p < 3$, it is (canonically) defined as $S(X)_{s,t} := \left(1, S(X)^{(1)}_{s,t}, \dots, S(X)^{(n)}_{s,t}, \dots\right) \in \mathcal{T}(\mathbb{R}^d)$, where $S(X)^{(n)}_{s,t} \in (\mathbb{R}^d)^{\otimes n}$ is the n-fold iterated integral $S(X)^{(n)}_{s,t} = \int_{s<u_1<\dots<u_n<t} \mathrm{d}X_{u_1} \otimes \mathrm{d}X_{u_2} \otimes \dots \otimes \mathrm{d}X_{u_n}$. Given two arbitrary sub-intervals $[a, b], [c, d] \subset [0, T]$, the *signature kernel* $K_S : C_p([s, t], \mathbb{R}^d) \times C_p([s', t'], \mathbb{R}^d) \to \mathbb{R}$ is a positive definite kernel on continuous paths of bounded variation defined as

$$K_S(X, Y) = \langle S(X)_{s,t}, S(Y)_{s',t'} \rangle \ . \tag{3}$$

Salvi et al. (2021a, Thm. 2.5) shows that $K_S(X, Y) = f(t, t')$, where $f : [s, t] \times [s', t'] \to \mathbb{R}$ is the solution of the following path-dependent integral equation:

$$f(t, t') = 1 + \int_s^t \int_{s'}^{t'} f(u, v) \langle dX_u, dY_v \rangle_1 \ , \quad \text{with } f(0, \cdot) = f(\cdot, 0) = 1 \ . \tag{4}$$

This 'kernel trick' allows us to evaluate the signature kernel without explicit computation of the signature transform by solving the partial differential equation (PDE) in eq. (4). We refer to Salvi et al. (2021a) for a numerical approximation scheme to solve this hyperbolic PDE and its error rates and to Appendix A.4 for more mathematical details on the applicability in our setting. In our experiments, we use the JAX library `sigkerax` to efficiently solve eq. (4).

In summary, the signature kernel is a universal kernel that takes (multi-variate) continuous paths—potentially on different intervals—as input, is efficiently computable, and performs well in capturing characteristics of sequential data (Lee & Oberhauser, 2023).

3 Methodology

We start with our constraint-based causal discovery algorithms, where we assume an oracle for CI testing on path space for eq. (2) in Section 3.1. In Section 3.2 we then introduce the signature kernel CI test and state its consistency, which requires a new proof that does not rely on densities. Our test, applicable to arbitrary stochastic processes, time series, or functional data, performs well empirically in our extensive experiments in Section 4 making it a contribution of independent interest.

3.1 Causal Discovery in the Acyclic SDE Model

Throughout this section, we assume access to a CI oracle for eq. (2), where $X^H_{[a,b]}$ denotes the $C([a, b], \mathbb{R}^{|H|})$-valued random variable $\omega \mapsto ([a, b] \ni t \mapsto X^h_t(\omega)^{h \in H})$. Asymptotically, such an oracle can be replaced in practice by a consistent finite-sample CI test for path-valued random variables (see Section 3.2). In Appendix A.3, we prove that when allowing for cycles in the SDE

model, constraint-based causal discovery of the full graph is impossible using arbitrary expressions of the form in eq. (2) despite the flexibility of these conditional independencies. This impossibility result is a key motivation to study the acyclic setting still allowing for loops, which are crucial in dynamic settings as variables should be allowed to depend on themselves infinitesimally into the past.

We will use the oracle in the following ways for intervals $[0, s]$ and $[s, s+h]$ with $h > 0$:

- X^i is *symmetrically CI* of X^j given X^K on $[0,T]$ if $X^i_{[0,T]} \perp\!\!\!\perp X^j_{[0,T]} \mid X^K_{[0,T]}$; we then write $X^i \perp\!\!\!\perp_{\text{sym}} X^j \mid X^K$
- X^i is *future-extended h-locally CI* of X^j given X^K at s if $X^i_{[0,s]} \perp\!\!\!\perp X^j_{[s,s+h]} \mid X^j_{[0,s]}, X^K_{[0,s+h]}$;[1] we then write $X^i \perp\!\!\!\perp^+_{s,h} X^j \mid X^K$
- X^i is *conditionally h-locally self-independent* given X^K at s if $X^i_{[0,s]} \perp\!\!\!\perp X^i_{[s,s+h]} \mid X^K_{[0,s+h]}$; we then write $X^i \perp\!\!\!\perp^{\circlearrowleft}_{s,h} \mid X^K$

The key idea to leverage the unidirectional flow of time is to split the observed paths into 'past' and 'future' at some time $t \in [0,T]$. Therefore, we define the *lifted dependence graph* $\tilde{\mathcal{G}} = (\tilde{V}, \tilde{E})$ by setting $\tilde{V} := V_0 \sqcup V_1$, where $\sqcup$ denotes disjoint union and V_0, V_1 are two copies of V, whose elements we subscript with 0 and 1; we include an edge $(i_0 \to i_1) \in \tilde{E}$ if and only if there is a loop $(i \to i) \in E$, and for each edge $(i \to j) \in E$ with $i \neq j$, include edges $(i_0 \to j_0), (i_1 \to j_1), (i_0 \to j_1) \in \tilde{E}$, see Figure 2. Conversely, we say that the graph with edges E is obtained by *collapsing* the graph with edges $\tilde{E}$. Intuitively, V_0 contains all 'past' and V_1 all 'future' variables with the same edges among them (causal relations) with additional edges from 'past' to 'future'. We can now establish the following Markov property.

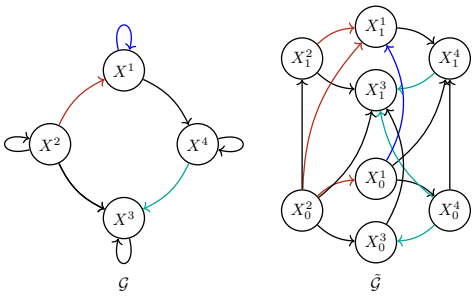


Figure 2: The lifted dependence graph $\tilde{\mathcal{G}}$ (right) for a DAG $\mathcal{G}$ (left) with colors highlighting the correspondence of selected edges.

Proposition 3.1 (Markov property)**.** *Assume the dependence graph $\mathcal{G}$ of the SDE model is acyclic except for loops and associate $X^i_{[0,s]}$ with $i_0 \in V_0$ and $X^i_{[s,s+h]}$ with $i_1 \in V_1$. Then $\tilde{\mathcal{G}}$ is acyclic and the independence relation $\perp\!\!\!\perp^+_{s,h}$ satisfies the global Markov property w.r.t. d-separation in $\tilde{\mathcal{G}}$.*

Our proof (see Appendix A.6) leverages the structure of the SDE-generating mechanism, which factorizes according to the dependence graph, thereby allowing us to formally establish the global Markov property using arguments from stochastic analysis. With an oracle for eq. (2) and a faithfulness type assumption (for more details see Appendix A.8), we now develop new constraint-based causal discovery algorithms that infer the full dependence graph instead of just equivalence classes.

Discovering the full DAG $\mathcal{G}$. To discover the full graph including loops we propose Algorithm 1 that makes use of the time-ordering via our h-local independence models (consider $s, h > 0$ arbitrary but fixed).

Theorem 3.2. *Algorithm 1 is sound and complete for the SDE model, assuming acyclicity except for loops and faithfulness: its output is the true dependence graph $\mathcal{G}$.*

Algorithm 1: Causal discovery for acyclic SDEs.

$\tilde{V} \leftarrow \{k_0, k_1 \mid k \in V\}$
$\tilde{E} \leftarrow \{i_0 \to j_0,\ i_1 \to j_1 \mid i, j \in V,\ i \neq j\}$
$\quad \cup \{i_0 \to j_1 \mid i, j \in V\}$
for $c = 0, \dots, d-2$ **do** ▷ *edge recovery w/o loops*
  for $i, j \in V, i \neq j$ **do**
    for $K \subseteq V \setminus \{i, j\}, |K| = c,$
      s.t. $(k_0 \to j_1) \in \tilde{E}$ for $k \in K$ **do**
      if $X^i \perp\!\!\!\perp^+_{s,h} X^j \mid X^K$ **then**
        $\tilde{E} \leftarrow \tilde{E} \setminus \{i_0 \to j_0, i_1 \to j_1, i_0 \to j_1\}$
$\mathcal{G} = (V, E) \leftarrow \text{collapse}(\tilde{V}, \tilde{E})$
for $k \in V$ **do** ▷ *removing loops*
  if $X^k \perp\!\!\!\perp^{\circlearrowleft}_{s,h} \mid X^{\text{pa}^{\mathcal{G}}_k \setminus \{k\}}$ **then**
    $E \leftarrow E \setminus \{k \to k\}$
return $\mathcal{G}$

The proof (see Appendix A.7) establishes completeness by identifying a separating set and leveraging the global Markov property in Proposition 3.1; soundness is shown via the faithfulness assumption. In particular, our proof does not require the 'strong faithfulness' assumption, but a weaker version, 'parent faithfulness', suffices—see Appendix A.8 for a detailed discussion of faithfulness in our setting.

[1] The '+' denotes conditioning on the future of X^K, not common in the literature (Florens & Fougere, 1996).

Discovering and post-processing the CPDAG. In Algorithm 1, we leverage our Markov property with respect to the lifted graph that also incorporates time directionality to infer the full graph. We now show that we can also use the symmetric criterion $\perp\!\!\!\perp_{\text{sym}}$ to recover the 'Markov equivalence class'—all graphs that are Markov equivalent to $\mathcal{G}$—via the completed partially directed acyclic graph (CPDAG) of $\mathcal{G}$ (Peters et al., 2017, Def. 6.24). Loosely speaking, the CPDAG has the same adjacencies as $\mathcal{G}$, but some edges may remain undirected. Due to space limitations, we present the required global Markov property with respect to $\perp\!\!\!\perp_{\text{sym}}$ in Appendix A.9. This Markov property allows us to apply the sound and complete PC algorithm to infer the CPDAG assuming an CI oracle for $\perp\!\!\!\perp_{\text{sym}}$ and faithfulness (Spirtes et al., 2000). While Algorithm 1 is strictly more informative (returns $\mathcal{G}$ instead of just its CPDAG), the reliability of $\perp\!\!\!\perp^{+}_{s,h}$ in practice might be negatively affected compared to $\perp\!\!\!\perp_{\text{sym}}$ by (a) conditioning on larger sets, (b) shorter time segments (potentially losing information), and (c) the additional choice of parameters s, h. Hence, in real-world applications inferring the CPDAG using $\perp\!\!\!\perp_{\text{sym}}$ may be more robust than inferring $\mathcal{G}$ fully using $\perp\!\!\!\perp^{+}_{s,h}$. Building on these potential benefits, we also develop a post-processing procedure for the CPDAG that again leverages the directionality of time to provably also orient all remaining unoriented edges.

Corollary 3.3 (post-processing)**.** *Let $\mathcal{G} = (V, E)$ be the dependence graph of the acyclic (except for loops) SDE model and $\tilde{\mathcal{G}} = (V, \tilde{E})$ its CPDAG. Then we have $X^j_{[0,T]} \not\perp\!\!\!\perp X^i_0$ but $X^i_{[0,T]} \perp\!\!\!\perp X^j_0$ for all $(i, j) \in E \subset \tilde{E}$ with $(j, i) \in \tilde{E}$ and $i \neq j$.*

The proof in Appendix A.10 relies on the joint independence of Brownian motions $\mathrm{d}W^k_t$ and initial conditions X^k_0. Corollary 3.3 directly motivates an alternative algorithm for causal discovery of $\mathcal{G}$, written out as Algorithm 2 in Appendix A.10, by first constructing the CPDAG (PC algorithm with $\perp\!\!\!\perp_{\text{sym}}$) followed by testing unconditionally the yet unoriented edges as in Corollary 3.3.

Partially observed setting. Finally, we consider the partially observed setting, where only $\{X^{v_k}_t\}_{v_k \in V_{obs}, t \in [0,1]}$ of the process $\{X^{v_i}_t\}_{v_i \in V_{obs} \sqcup V_L, t \in [0,1]}$ from the model in eq. (1) with DAG $\mathcal{G} = (V_{obs} \sqcup V_L, E)$ are observed. Asymmetric local independence based methods for causal discovery are challenged by requiring exponentially many oracle calls in the partially observed setting (Mogensen & Hansen, 2022; Mogensen, 2023). However, since our symmetric conditional independence constraint satisfies the global Markov property with respect to the induced acyclic directed mixed graph (ADMG) $\mathcal{G}_{obs} = (V_{obs}, E')$ (also known as latent projection, with potentially arbitrary unobserved processes) over the observed variables, we can take inspiration from the Fast Causal Inference (FCI) algorithm (Zhang, 2008) by first running its skeleton-discovery-part and then again leveraging the direction of time to uniquely infer the edge-type $\{\leftarrow, \rightarrow, \leftrightarrow\}$ for found adjacencies. This reduces the typical partial ancestral graph output of FCI (representing an equivalence class of ancestral graphs) to a single, maximally informative graph (as informative as an ancestral graph can be). We provide all details in Appendix D.

3.2 Signature Kernel Conditional Independence Test

Our theoretical results in Section 3.1 assume a CI oracle. To make our algorithms usable in practice, we now propose a flexible CI test for path-valued random variables on different time intervals for expressions of the form eq. (2). Recent work using the signature kernel is limited to unconditional hypothesis tests (Chevyrev & Oberhauser, 2022) or only use it as a heuristic measure of conditional independence (not developing a hypothesis test) (Salvi et al., 2021c), and both do not use the time order. Using terms of the signature to detect causality (not specifically with hypothesis testing) was also proposed by Giusti & Lee (2020); Glad & Woolf (2021). Hence, we are the first to combine ideas from kernel-based CI tests and the signature kernel for a practically usable, consistent CI test on path space. While the existing consistency proof for KCIPT (also applying to SDCIT) (Doran et al., 2014) relies on the existence of densities, we prove consistency for testing on path-valued random variables in Appendix A.14 requiring novel arguments. We discuss the sensitivity of constraint-based causal discovery on errors in the CI test in Appendix A.12. Given n samples of path segments $X^{(i)}_{\mathcal{I}}, Y^{(i)}_{\mathcal{J}}, Z^{(i)}_{\mathcal{K}}$ on intervals $\mathcal{I}, \mathcal{J}, \mathcal{K} \subset [0, T]$ for $i \in [n]$, we compute the Gram matrices $k_{XX}, k_{YY}, k_{ZZ} \in \mathbb{R}^{n \times n}$ using the signature kernel K_S and run a kernel-based permutation CI test like KCIPT (Doran et al., 2014) or SDCIT (Lee & Honavar, 2017) to test $X_{\mathcal{I}} \perp\!\!\!\perp Y_{\mathcal{J}} \mid Z_{\mathcal{K}}$.[2] The theoretical foundation combined with the strong empirical performance of our CI test suggest that it may be of independent interest outside of causal discovery applications.

[2] While time observations within a process are not i.i.d., the n process samples are i.i.d. realizations of the SDE model, see Appendix A.13 for details.

Table 1: SHD ($\times 10^2$) comparison of SigKer to the baselines in four bivariate SDE settings: linear, path-dependence, non-linear, and diffusion dependence. We ran SCOTCH for different sparsity parameters λ and numbers of epochs n_e; there is no clear trend in either parameter, but $\lambda = 100, n_e = 2000$ performed best overall. Different learning rates performed worse across the board.

		linear		path-dependence		non-linear		diffusion dependence	
$\times 10^2$	(λ, n_e)	$n = 200$	$n = 400$	$n = 200$	$n = 400$	$n = 200$	$n = 400$	$n = 200$	$n = 400$
CCM		176 ± 7	166 ± 8	186 ± 5	194 ± 3	120 ± 10	106 ± 10	100 ± 10	84 ± 10
Granger		92 ± 5	87 ± 5	85 ± 5	95 ± 6	103 ± 5	103 ± 5	105 ± 5	111 ± 4
PCMCI		89 ± 9	100 ± 10	91 ± 4	125 ± 15	41 ± 8	55 ± 15	98 ± 13	75 ± 23
SCOTCH	100, 1k	74 ± 14	77 ± 17	50 ± 12	46 ± 12	64 ± 15	80 ± 10	75 ± 14	62 ± 10
SCOTCH	50, 2k	79 ± 18	44 ± 17	23 ± 12	73 ± 23	83 ± 11	73 ± 13	$\mathbf{25 \pm 13}$	44 ± 28
SCOTCH	100, 2k	50 ± 17	36 ± 19	64 ± 19	55 ± 15	85 ± 7	91 ± 9	33 ± 16	$\mathbf{9 \pm 8}$
SigKer		$\mathbf{14 \pm 4}$	$\mathbf{7 \pm 3}$	$\mathbf{5 \pm 2}$	$\mathbf{6 \pm 2}$	$\mathbf{28 \pm 5}$	$\mathbf{5 \pm 2}$	72 ± 6	63 ± 5

4 Experiments

In this section, after introducing the baselines and implementation details, we empirically evaluate our CI test in a variety of settings. We then demonstrate strong performance in various causal discovery tasks and also outperform traditional methods in a real-world finance case study.

Baselines. We compare against CCM (Sugihara et al., 2012), PCMCI (Runge et al., 2019), Granger causality (Granger, 1969), a kernel-based approach (Lau) (Laumann et al., 2023), and SCOTCH (Wang et al., 2024), a variational neural SDE method. CCM, Granger, and PCMCI are well-established methods for time series, while Lau and SCOTCH represent the latest advances in causal discovery for functional data and SDE models, respectively. CCM and Granger are limited to bivariate cases, while Lau is only applicable to unconditional tests here. Details are in Appendix C.

Implementation details and metrics. We use `sigkerax` for the signature kernel with an RBF kernel with length scale selected via a median heuristic (see Appendix B.1). For bivariate cases, the causal structure is $X^1 \to X^2$, for $d > 2$, we draw Erdös–Rényi DAGs. Performance is measured by Structural Hamming Distance (SHD, Appendix B.8). For $\perp\!\!\!\perp^+_{s,h}$, $s = 0.1 \cdot T$ (and a fixed $T = 1$) performed best (Table 5). The SDE model is

$$\mathrm{d}X_t = (AX_t + c)\mathrm{d}t + \mathrm{Diag}(BX_t + d)\mathrm{d}W_t, \quad \text{with } A, B \in \mathbb{R}^{d\times d}, c, d \in \mathbb{R}^d\,. \tag{5}$$

except for the 2-dimensional non-linear SDE, which is given by

$$\mathrm{d}\begin{pmatrix} X_t^1 \\ X_t^2 \end{pmatrix} = \begin{pmatrix} -r\omega\sin(\omega t) \\ r\omega\tanh(X_t^2) \end{pmatrix}\mathrm{d}t + \mathrm{Diag}(d^\top)\mathrm{d}W_t\,. \tag{6}$$

Computational complexity. Due to space constraints, we defer the computational complexity analysis of our CI test, the causal discovery algorithm, and the signature kernel evalatuions in Appendix B.9. The overall computation is dominated by finding the permutation that leaves Z (approximately) invariant in the permutation based CI tests (KCIPT, SDCIT).

Choosing (conditional) independence tests. In initial experiments on two- and three-dimensional structures (e.g., forks, colliders, chains) detailed in Appendix B.2, we found bootstrapped SDCIT and HSIC to be the best performing variants of kernel-based (C)I tests (see Figure 6 and Tables 6 to 8 for details) and use them for all subsequent experiments.

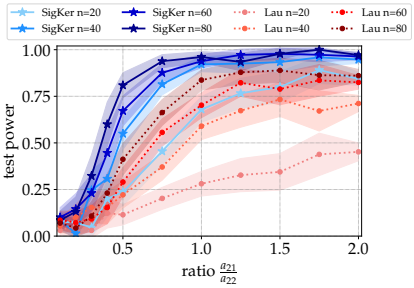


Figure 3: Test power for $X^1_{[0,t]} \perp\!\!\!\perp X^2_{[0,t]}$ over $\frac{a_{21}}{a_{22}}$. Lines (shades) are means (standard errors) over 1000 SDE instances.

Power analysis of the unconditional test. We first demonstrate consistently superior performance of our unconditional independence test (HSIC) SigKer (ours) over the only existing baseline for this setting from Laumann et al. (2023) (Lau) in the linear SDE-setting eq. (5) (with $B \equiv 0$, $d_i = 0.4$), reaching test power near 1 already for $n \geq 40$ when the causal interaction a_{21} (strength of $X^1 \to X^2$) is comparable to X^2's self-dependence (a_{22}). Figure 7 further highlights SigKer's robustness even with *high data missingness*, and Figure 8 demonstrates its effectiveness on *fractional Brownian motions* outside the semi-martingale framework. We are not aware of other CI tests that apply to such settings—establishing it as state-of-the-art CI test for stochastic processes.

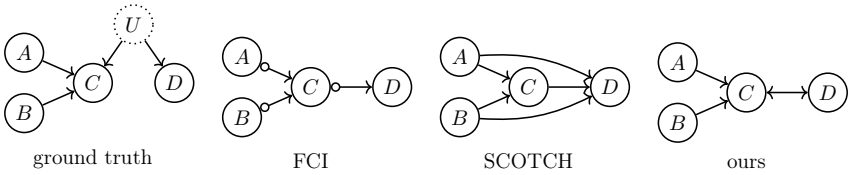


Figure 4: An example graph demonstrating the advantage of our approach under partial observations.

Leveraging the direction of time. Next, we leverage time to orient the edge in the much-studied bivariate causal-discovery setting ($X^1 \to X^2$) via our $\perp\!\!\!\perp^+_{s,h}$ constraint with $K = \emptyset$ for a variety of settings: (i) linear SDEs with dependence in the drift, (ii) diffusion, (iii) or path-dependent, as well as (iv) nonlinear interactions. Table 1 shows that SigKer decisively dominates all baselines in all settings for different sample sizes except diffusion dependence. In this setting, SCOTCH is better *for specific hyperparameter settings*, while performing worse for others. Crucially, one cannot 'cross-validate' or otherwise select such hyperparameters for the causal-discovery task in a data-driven fashion without ground-truth knowledge. The lack of robustness (different optimal hyperparameters for different sample sizes) renders SCOTCH unreliable in practice even for diffusion dependence.

Specifically, we draw the settings for these experiments as follows: for linear drift interactions we sample $a_{21} \sim \mathcal{U}([1, 2.5])$, $a_{11}, a_{22} \sim \mathcal{U}([-0.5, 0.5])$, $a_{12} = 0$, $B \equiv 0$, $d_i \sim \mathcal{U}([0.1, 0.2])$ in a two-dimensional linear SDE model eq. (5). For linear diffusion interactions we draw $a_{11}, a_{22} \sim \mathcal{U}([0.5, 1])$ and $b_{21} \sim \mathcal{U}([1, 4.5])$), the rest set to zero in a two-dimensional linear SDE model eq. (5). For path-dependence, we simulate $\mathrm{d}X^2_t = \mu^2(X^1_{[0,t]})\mathrm{d}t + d_2 \mathrm{d}W^1_t$ via a three-dimensional SDE eq. (5) with $B \equiv 0$, $c \equiv 0$, $a_{23}, a_{31} \sim \mathcal{U}([-3.5, -1] \cup [1, 3.5])$ and $d = (d_1, d_2, 0)^\top$, $d_i \sim \mathcal{U}([0.1, 0.2])$, the rest set to zero. For the nonlinear SDEs we use eq. (6) with $\omega \sim \mathcal{U}([6\pi, 8\pi])$, $r \sim \mathcal{U}([0.5, 1])$, $d_i \sim \mathcal{U}([2, 2.5])$.

Causal discovery. For causal discovery, we sample 40 linear SDEs from eq. (5) (with $a_{ij} \sim \mathcal{U}([-2, -1] \cup [1, 2])$ for $j \neq i$ and $a_{ii} \sim \mathcal{U}([-0.5, 0.5])$) with random DAG adjacency structures for $d \in \{3, 5, 10, 20, 50\}$ and use $n = 200$ sample paths from each SDE as input for the algorithms. Table 2 shows that our Algorithm 1 (denoted by $\perp\!\!\!\perp^+_{s,h}$) and Algorithm 2 ($\perp\!\!\!\perp_{\mathrm{sym}}$+pp and $\perp\!\!\!\perp_{\mathrm{sym}}$ the result before post-processing) clearly outperform PCMCI and SCOTCH up to $d = 10$. SCOTCH's heavy dependence on hyperparameter choices for λ (graph sparsity) and n_e (the number of epochs), can sometimes render it superior when charitably picking the best setting. However, since good values particularly for λ cannot be known up front nor selected in a data-driven fashion, one must interpret SCOTCH's performance more conservatively—arguably in terms of its worst case performance over a set of reasonable hyperparameter settings. Hence, SigKer—free of any hyperparameter choices—broadly outperforms the state-of-the-art, such as SCOTCH even in the setting SCOTCH was specifically tailored to (SDE models). Unlike such methods, our approach continues to work well for other data generating mechanisms as our CI test handles stochastic processes beyond the SDE model.

Table 2: SHD ($\times 10^2$) comparison of SigKer to PCMCI and SCOTCH (different λ and n_e) in causal discovery. Means and standard errors are over 40 SDE instances.

	$d = 3$	$d = 5$	$d = 10$	$d = 20$	$d = 50$
PCMCI	29 ± 16	243 ± 37	793 ± 84	3530 ± 159	$19.6k \pm 857$
SCOTCH 100, 2k	188 ± 28	417 ± 86	250 ± 61	1525 ± 1160	10275 ± 6176
SCOTCH 200, 2k	110 ± 21	270 ± 48	530 ± 223	$\mathbf{370 \pm 174}$	$\mathbf{538 \pm 70}$
SCOTCH 200, 1k	400 ± 17	1370 ± 29	6425 ± 80	705 ± 77	7863 ± 2670
$\perp\!\!\!\perp^+_{s,h}$	26 ± 5	80 ± 8	284 ± 19	1026 ± 40	4946 ± 113
$\perp\!\!\!\perp_{\mathrm{sym}}$+p.p.	$\mathbf{13 \pm 4}$	$\mathbf{38 \pm 7}$	$\mathbf{157 \pm 15}$	725 ± 439	4593 ± 93
$\perp\!\!\!\perp_{\mathrm{sym}}$ only	57 ± 84	147 ± 11	436 ± 19	1294 ± 48	6005 ± 98

Summary and discussion of results. Our CI test for $\perp\!\!\!\perp^+_{s,h}$ reliably detects the direction of time and outperforms strong baselines (CCM, Granger, PCMCI), consistently improving with larger sample sizes (Table 1). SigKer also dominates PCMCI in causal discovery across dimensions and consistently beats most hyperparameter settings of SCOTCH. While isolated hyperparameter settings for SCOTCH perform better than SigKer in diffusion dependence and high-dimensional causal discovery, its strong dependence on hyperparameters, particularly the sparsity parameter λ, renders it unreliable in practice. Finally, SCOTCH and PCMCI are tailored to Markovian SDEs (or SDEs with a fixed lag). Instead, our CI test is broadly applicable and effective across data modalities which we further demonstrate empirically in a functional data example where both PCMCI and SCOTCH fail to detect path-dependence, see Table 10.

The partially observed setting. One of the main benefits of constraint-based causal discovery (especially with a non-parametric test) is the ability to handle partial observations. Score-based methods like SCOTCH are fundamentally challenged in this setting, since there could in principle be infinitely many unobserved variables, impossible to model, e.g., with neural-network based approaches. We showcase this advantage on the challenging example graph in Figure 4, where SCOTCH falsely infers an adjacency $A - D$ in 88 out of 100 runs, whereas our algorithm correctly handles the unobserved confounder and only (falsely) predicts this adjacency in 8 instances.

Real-world pairs trading example. To demonstrate the applicability of our developed methods on real-data, we evaluate pairs trading strategies on ten stocks from the VBR Small-Cap ETF over a three-year period (2010/01/01–2012/12/31). We only provide a concise summary of our results here and refer to Appendix B.7 for more details on this proof-of-concept study. We assess our method's effectiveness by quantifying the profit-and-loss (P&L) profile of the generated pairs trading strategy as a substitution for ground truth. Pairs were selected based on pairwise p-values from different hypothesis tests: 1) cointegration via the Augmented Dickey-Fuller (ADF) test, 2) Granger causality, and 3) our method. Table 3 shows that our strategy's P&L substantially outperforms the baselines in total return, APR, Sharpe ratio, and maxDDD while being almost on par with ADF & Granger in the fifth relevant metric maxDD. This case-study highlights the broad applicability and potential downstream impact of the findings in this work.

Table 3: SigKer outperforms baselines in most pairs trading performance metrics. ↑ / ↓ indicates 'higher/lower is better'.

	return ↑	**APR ↑**	**Sharpe ↑**	**maxDD ↓**	**maxDDD ↓**
ADF	0.004	0.004	0.090	0.087	230
Granger	-0.010	-0.011	-0.230	0.056	219
ADF & Granger	0.008	0.008	0.242	**0.022**	153
SigKer	**0.076**	**0.077**	**1.500**	0.027	**21**

5 Conclusion and Future Work

We introduce constraint based causal discovery algorithms for both fully and partially observed data in the form of stochastic processes, proving them to be sound and complete assuming a CI oracle. These are based on novel Markov properties corresponding to the developed CI constraints with respect to the induced dependence graph. Our algorithms critically leverage the directionality of time to uniquely identifying the full underlying graph or induced ancestral graph, critically improving over existing constraint-based algorithms in the static case that only output equivalence classes. Our framework also efficiently captures path- or diffusion-dependence. Finally, we propose a practical and consistent kernel-based CI test on path-space leveraging recent advances in signature kernels that empirically outperforms existing alternatives across a wide range of settings also beyond the SDE model such as functional data and fractional Brownian motions.

Due to the identified fundamental limitations in the cyclic setting, we limit ourselves to acyclic (except for self-loops) dependence graphs and assume causal relationships to not change over time. Relaxing both of these assumptions and assessing how much of the causal structure can in principle be learned in the cyclic settings are immediate interesting directions for future work. We highlight that the field of causality represents just one of the many potential applications of our conditional independence test for path-valued random variables. Exploring applications in different domains is a worthwhile direction for future work that could also catalyze new methodological developments.

Reproducibility Statement

Significant effort was made to ensure reproducibility, both for the theoretical and experimental results. The complete proofs of Proposition 3.1, Theorem 3.2, Corollary 3.3 can be found in Appendix A.6 Appendix A.7, Appendix A.10, respectively. Additionally, we discuss the limitations and requirements in Section 1, which outlines the assumptions made throughout the paper. Details regarding implementation and metrics can be found in Section 4, along with further information on the data-generating mechanism and results evaluation throughout Section 4. The implementation of baselines is described in Appendix C, and details on the real-world pairs trading example can be found in Appendix A.11. We will also make all code used to produce the results in this paper openly available.

REFERENCES

Hananeh Aliee, Fabian J Theis, and Niki Kilbertus. Beyond predictions in neural odes: Identification and interventions. *arXiv preprint arXiv:2106.12430*, 2021. 3

Hananeh Aliee, Till Richter, Mikhail Solonin, Ignacio Ibarra, Fabian Theis, and Niki Kilbertus. Sparsity in continuous-depth neural networks. *Advances in Neural Information Processing Systems*, 35:901–914, 2022. 3

Holly Andersen. When to expect violations of causal faithfulness and why it matters. *Philosophy of Science*, 80(5):672–683, 2013. doi: 10.1086/673937. 23

Yashas Annadani, Nick Pawlowski, Joel Jennings, Stefan Bauer, Cheng Zhang, and Wenbo Gong. Bayesdag: Gradient-based posterior sampling for causal discovery. *arXiv preprint arXiv:2307.13917*, 2023. 1

Charles K Assaad, Emilie Devijver, and Eric Gaussier. Discovery of extended summary graphs in time series. In *Uncertainty in Artificial Intelligence*, pp. 96–106. PMLR, 2022. 3

Alexis Bellot, Kim Branson, and Mihaela van der Schaar. Neural graphical modelling in continuous-time: consistency guarantees and algorithms. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=SsHBkfeRF9L. 3

Thomas Berrett, Yi Wang, Rina Barber, and Richard Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 10 2019. doi: 10.1111/rssb.12340. 3

Stephan Bongers and Joris M Mooij. From random differential equations to structural causal models: The stochastic case. *arXiv preprint arXiv:1803.08784*, 2018. 3

Stephan Bongers, Tineke Blom, and Joris M Mooij. Causal modeling of dynamical systems. *arXiv preprint arXiv:1803.08784*, 2018. 3

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885 – 2915, 2021. doi: 10.1214/21-AOS2064. URL https://doi.org/10.1214/21-AOS2064. 19

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020. 1

T Cass, T Lyons, and X Xu. Weighted signature kernels. *Annals of Applied Probability*, 2023. 4, 20, 28

Thomas Cass and Cristopher Salvi. Lecture notes on rough paths and applications to machine learning. *arXiv preprint arXiv:2404.06583*, 2024. 4

Thomas Cass and William F Turner. Topologies on unparameterised path space. *Journal of Functional Analysis*, 286(4):110261, 2024. 20

Bertrand Charpentier, Simon Kibler, and Stephan Günnemann. Differentiable DAG sampling. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9wOQOgNe-w. 1

Hang Chen, Keqing Du, Xinyu Yang, and Chenguang Li. A review and roadmap of deep learning causal discovery in different variable paradigms. *arXiv preprint arXiv:2209.06367*, 2022. 1

Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning, 2016. URL https://arxiv.org/abs/1603.03788. 21

Ilya Chevyrev and Harald Oberhauser. Signature moments to characterize laws of stochastic processes. *J. Mach. Learn. Res.*, 23(1), jan 2022. ISSN 1532-4435. 6, 20

Alexander Mangulad Christgau, Lasse Petersen, and Niels Richard Hansen. Nonparametric conditional local independence testing. *The Annals of Statistics*, 51(5):2116 – 2144, 2023. doi: 10.1214/23-AOS2323. URL https://doi.org/10.1214/23-AOS2323. 3, 19

Nicola Muça Cirone, Maud Lemercier, and Cristopher Salvi. Neural signature kernels as infinite-width-depth-limits of controlled resnets. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023. 4

Thomas Cochrane, Peter Foster, Varun Chhabra, Maud Lemercier, Terry Lyons, and Cristopher Salvi. Sk-tree: a systematic malware detection algorithm on streaming trees via the signature kernel. In *2021 IEEE international conference on cyber security and resilience (CSR)*, pp. 35–40. IEEE, 2021. 4

Christa Cuchiero, Philipp Schmocker, and Josef Teichmann. Global universal approximation of functional input maps on weighted spaces. *arXiv preprint arXiv:2306.03303*, 2023. 20

JJ Daudin. Partial association measures and an application to qualitative regression. *Biometrika*, 67 (3):581–590, 1980. 3

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, 2008. 3

Cees Diks and Valentyn Panchenko. A new statistic and practical guidelines for nonparametric granger causality testing. *Journal of Economic Dynamics and Control*, 30(9-10):1647–1669, 2006. 3

Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pp. 132–141, 2014. 3, 6, 28, 29, 30

Lawrence C Evans. An introduction to stochastic differential equations version 1.2. *Lecture Notes, UC Berkeley*, 2006. 2

Jean-Pierre Florens and Denis Fougere. Noncausality in continuous time. *Econometrica: Journal of the Econometric Society*, pp. 1195–1212, 1996. 5, 19

Anne Gégout-Petit and Daniel Commenges. A general definition of influence between stochastic processes. *Lifetime data analysis*, 16(1):33–44, 2010. 20

Chad Giusti and Darrick Lee. Iterated integrals and population time series analysis. In Nils A. Baas, Gunnar E. Carlsson, Gereon Quick, Markus Szymik, and Marius Thaule (eds.), *Topological Data Analysis*, pp. 219–246, Cham, 2020. Springer International Publishing. ISBN 978-3-030-43408-3. 6

Will Glad and Tom Woolf. Path signature area-based causal discovery in coupled time series. In Sisi Ma and Erich Kummerfeld (eds.), *Proceedings of The 2021 Causal Analysis Workshop Series*, volume 160 of *Proceedings of Machine Learning Research*, pp. 21–38. PMLR, 16 Jul 2021. URL https://proceedings.mlr.press/v160/glad21a.html. 6

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019. 1, 3

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969. 3, 7

Clive WJ Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980. 3

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006. 3

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007. 30

Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath, Bernhard Schölkopf, and Andreas Krause. Bacadi: Bayesian causal discovery with unknown interventions. In *International Conference on Artificial Intelligence and Statistics*, pp. 1411–1436. PMLR, 2023. 1

Uzma Hasan, Emam Hossain, and Md Osman Gani. A survey on causal discovery methods for iid and time series data. *Transactions on Machine Learning Research*, 2023. 3

Zacharia Issa, Blanka Horvath, Maud Lemercier, and Cristopher Salvi. Non-adversarial training of neural sdes with signature kernel scores. *Advances in Neural Information Processing Systems*, 2023. 4

Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Quantifying causal influences. *The Annals of Statistics*, pp. 2324–2358, 2013. 29

Prince Joseph Erneszer Javier. causal-ccm a Python implementation of Convergent Cross Mapping, 6 2021. 37

Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007. 27

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stephan Bauer, Hugol Larochelle, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions, 2020. URL https://openreview.net/forum?id=H1gN6kSFwS. 1

Nan Rosemary Ke, Silvia Chiappa, Jane X Wang, Jorg Bornschein, Anirudh Goyal, Melanie Rey, Theophane Weber, Matthew Botvinick, Michael Curtis Mozer, and Danilo Jimenez Rezende. Learning to induce causal structure. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=hp_RwhKDJ5. 1

Ilmun Kim, Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Local permutation tests for conditional independence. *The Annals of Statistics*, 50(6):3388–3414, 2022. 3

Franz J Király and Harald Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20, 2019. 4

Felix Laumann, Julius Von Kügelgen, Junhyung Park, Bernhard Schölkopf, and Mauricio Barahona. Kernel-based independence tests for causal structure learning on functional data. *Entropy*, 25(12): 1597, 2023. 2, 3, 7, 29, 31, 34, 37

Steffen Lauritzen and Kayvan Sadeghi. Unifying Markov properties for graphical models. *The Annals of Statistics*, 46(5):2251 – 2278, 2018. doi: 10.1214/17-AOS1618. URL https://doi.org/10.1214/17-AOS1618. 24

Steffen L Lauritzen, A Philip Dawid, Birgitte N Larsen, and H-G Leimer. Independence properties of directed markov fields. *Networks*, 20(5):491–505, 1990. 25

Andrew R Lawrence, Marcus Kaiser, Rui Sampaio, and Maksim Sipos. Data generating process to evaluate causal discovery techniques for time series data. *arXiv preprint arXiv:2104.08043*, 2021. 1

Darrick Lee and Harald Oberhauser. The signature kernel. *arXiv preprint arXiv:2305.04625*, 2023. 4, 21

Sanghack Lee and Vasant G Honavar. Self-discrepancy conditional independence test. In *Uncertainty in artificial intelligence*, volume 33, 2017. 3, 6, 30

Maud Lemercier and Terry Lyons. A high order solver for signature kernels. *arXiv preprint arXiv:2404.02926*, 2024. 20

Maud Lemercier, Cristopher Salvi, Theodoros Damoulas, Edwin Bonilla, and Terry Lyons. Distribution regression for sequential data. In *International Conference on Artificial Intelligence and Statistics*, pp. 3754–3762. PMLR, 2021. 4

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. Dibs: Differentiable bayesian structure learning. *Advances in Neural Information Processing Systems*, 34:24111–24123, 2021. 1

Anton Rask Lundborg, Rajen D Shah, and Jonas Peters. Conditional independence testing in hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1821–1850, 2022. 3

Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. Kernel method for nonlinear granger causality. *Physical review letters*, 100(14):144103, 2008. 3

Christopher Meek. Toward learning graphical and causal process models. In *CI@ UAI*, pp. 43–48, 2014. 3

Søren W. Mogensen. Faithful graphical representations of local independence. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2989–2997. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/mogensen24a.html. 23

Søren Wengel Mogensen. Weak equivalence of local independence graphs. *arXiv preprint arXiv:2302.12541*, 2023. 6

Søren Wengel Mogensen and Niels Richard Hansen. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1):539 – 559, 2020. 3

Søren Wengel Mogensen and Niels Richard Hansen. Graphical modeling of stochastic processes driven by correlated noise. *Bernoulli*, 28(4):3023 – 3050, 2022. 6

Søren Wengel Mogensen, Daniel Malinsky, and Niels Richard Hansen. Causal learning for partially observed stochastic dynamical systems. In *Conference on Uncertainty in Artificial Intelligence,*, pp. 350–360, 2018. 3

Francesco Montagna, Atalanti A. Mastakouri, Elias Eulig, Nicoletta Noceti, Lorenzo Rosasco, Dominik Janzing, Bryon Aragam, and Francesco Locatello. Assumption violations in causal discovery and the robustness of score matching. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=IyTArtpuCK. 1

Francesco Montagna, Nicoletta Noceti, Lorenzo Rosasco, Kun Zhang, and Francesco Locatello. Scalable causal discovery with score matching. In Mihaela van der Schaar, Cheng Zhang, and Dominik Janzing (eds.), *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pp. 752–771. PMLR, 11–14 Apr 2023b. URL https://proceedings.mlr.press/v213/montagna23b.html. 1

Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, pp. 440–448, Arlington, Virginia, USA, 2013. AUAI Press. 3

James Morrill, Adeline Fermanian, Patrick Kidger, and Terry Lyons. A generalised signature method for multivariate time series feature extraction. *arXiv preprint arXiv:2006.00873*, 2020. 21

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017. 3, 20

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*, pp. 1595–1605. PMLR, 2020. 3

Alexandre Pannier and Cristopher Salvi. A path-dependent pde solver based on signature kernels. *arXiv preprint arXiv:2403.11738*, 2024. 4

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in neural information processing systems*, 33:21247–21259, 2020. 3, 28, 29

Judea Pearl. *Causality*. Cambridge university press, 2009. 25

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. 3, 6, 21, 22

Jonas Peters, Stefan Bauer, and Niklas Pfister. Causal models for dynamical systems. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 671–690. Association for Computing Machinery, 2022. 2, 3, 19

Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30 (4):962 – 1030, 2002. doi: 10.1214/aos/1031689015. URL https://doi.org/10.1214/aos/1031689015. 37

L. C. G. Rogers and David Williams. *Diffusions, Markov processes, and martingales. Vol. 2*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000. ISBN 0-521-77593-0. doi: 10.1017/CBO9781107590120. URL https://doi.org/10.1017/CBO9781107590120. Itô calculus, Reprint of the second (1994) edition. 2, 18, 21, 23

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pp. 18741–18753. PMLR, 2022. 1

Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, 2018. 3

Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1388–1397. PMLR, 2020. 3

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5 (11):eaau4996, 2019. 1, 3, 7, 37

Jakob Runge, Xavier-Andoni Tibau, Matthias Bruhns, Jordi Muñoz-Marí, and Gustau Camps-Valls. The causality for climate competition. In *NeurIPS 2019 Competition and Demonstration Track*, pp. 110–120. PMLR, 2020. 3

Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023. 1, 3

Cristopher Salvi, Thomas Cass, James Foster, Terry Lyons, and W. Y. The signature kernel is the solution of a Goursat PDE. *SIAM Journal on Mathematics of Data Science*, 3(3):873–899, 2021a. 4, 20

Cristopher Salvi, Maud Lemercier, Thomas Cass, Edwin V Bonilla, Theodoros Damoulas, and Terry J Lyons. Siggpde: Scaling sparse gaussian processes on sequential data. In *International Conference on Machine Learning*, pp. 6233–6242. PMLR, 2021b. 4

Cristopher Salvi, Maud Lemercier, Chong Liu, Blanka Horvath, Theodoros Damoulas, and Terry Lyons. Higher order kernel mean embeddings to capture filtrations of stochastic processes. *Advances in Neural Information Processing Systems*, 34:16635–16647, 2021c. 4, 6

Tore Schweder. Composable markov processes. *Journal of applied probability*, 7(2):400–410, 1970. 3, 20

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010. 37

Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/02f039058bd48307e6f653a2005c9dd2-Paper.pdf. 3

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020. 3

Ali Shojaie and Emily B Fox. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9:289–319, 2022. 3

Carl-Johann Simon-Gabriel and Bernhard Schölkopf. Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions. *The Journal of Machine Learning Research*, 19(1):1708–1736, 2018. 20

Hermann Singer. Dynamic structural equations in discrete and continuous time. In *Economic Evolution and Demographic Change: Formal Models in Social Sciences*, pp. 306–320. Springer, 1992. 1

Alexander Sokol and Niels Hansen. Causal interpretation of stochastic differential equations. *Electronic Journal of Probability*, 19, 03 2013. doi: 10.1214/EJP.v19-2891. 20

Peter Spirtes and Christopher Meek. Learning bayesian networks with discrete variables from data. In *KDD*, volume 1, pp. 294–299, 1995. 27

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000. 1, 6, 23, 27

Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1), 2019. 3

George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *science*, 338(6106):496–500, 2012. 7

Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *Ann. Statist. 35 (6) 2769 - 2794*, 2007. 37

Ioannis Tsamardinos and Giorgos Borboudakis. Permutation testing improves bayesian network learning. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 322–337. Springer, 2010. 3

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2), April 2013. ISSN 0090-5364. doi: 10.1214/12-aos1080. URL http://dx.doi.org/10.1214/12-AOS1080. 23

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, pp. 255–270, USA, 1990. Elsevier Science Inc. ISBN 0444892648. 21, 22

Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022. 1

Benjie Wang, Joel Jennings, and Wenbo Gong. Neural structure learning with stochastic differential equations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=V1GM9xDvIY. 3, 7, 37

Naftali Weinberger. Faithfulness, coordination and causal coincidences. *Erkenntnis*, 83(2):113–133, April 2018. ISSN 1572-8420. doi: 10.1007/s10670-017-9882-6. URL https://doi.org/10.1007/s10670-017-9882-6. 23

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pp. 7154–7163. PMLR, 2019. 1

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16):1873–1896, 2008. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2008.08.001. URL https://www.sciencedirect.com/science/article/pii/S0004370208001008. 6, 37

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pp. 804–813, Arlington, Virginia, USA, 2011. AUAI Press. ISBN 9780974903972. 3, 30

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. 1

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425. PMLR, 2020. 1

APPENDIX

TABLE OF CONTENTS

A Proofs and Theoretical Digressions **18**

A.1 Summary of Notations 18

A.2 Intuition and Details for the SDE Model 18

A.3 Counterexample for Cyclic Causal Discovery in the SDE Model 19

A.4 Signature Kernel Details 20

A.5 Testing on Time Points 21

A.6 Proof of Proposition 3.1 21

A.7 Proof of Theorem 3.2 22

A.8 Faithfulness and Causal Minimality 23

A.9 Global Markov Property for the Symmetric Independence Criterion 24

A.10 Proof of Corollary 3.3 and Algorithm for Robust Causal Discovery 25

A.11 Post-Processing Without Jointly Independent Initial Values 25

A.12 Sensitivity of Constraint-Based Methods to Errors in Conditional Independence Testing 27

A.13 I.i.d. Assumption of the Conditional Independence Test 28

A.14 Consistency of the Conditional Independence Test 28

B Experiments and Evaluation **29**

B.1 Kernel Heuristic and Interval Choice 29

B.2 Empirical Comparison of Different Tests 30

B.3 Additional Experimental Results for the Bivariate Case 30

B.4 Building Blocks for Causal Discovery 31

B.5 Causal Discovery 34

B.6 Conditional Independence Testing Beyond the SDE Model 34

B.7 Real-World Pairs Trading Example 35

B.8 Evaluation Metrics 36

B.9 Computational Complexity 36

C Baselines **37**

D Details on the Partially Observed Setting **37**

A Proofs and Theoretical Digressions

A.1 Summary of Notations

A summary of notation is provided in Table 4.

Table 4: Summary of notations

Symbol	Meaning
$[m]$	short for $\{1, \ldots, m\}$, $m \in \mathbb{N}$
X	stochastic process $X = (X^1, \ldots, X^d) \in C([0,T], \mathbb{R}^{n_1+\ldots+n_d})$
d	number of coordinate processes $X = (X^1, \ldots, X^d)$ or nodes in graph
X^k	k-th coordinate process of $X = (X^1, \ldots, X^d)$, $X_t^k \in \mathbb{R}^{n_k}$
n_k	dimension of the k-th coordinate process X^k
W^k	k-th Brownian motion, $W_t^k \in \mathbb{R}^{m_k}$
m_k	dimension of the k-th Brownian motion W^k
n	total dimension of the stochastic process X, $n = n_1 + \ldots + n_d$
μ^k	k-th drift component $\mu^k : \mathbb{R}^n \to \mathbb{R}^{n_k}$
σ^k	k-th diffusion component $\sigma^k : \mathbb{R}^n \to \mathbb{R}^{n_k \times m_k}$
$T \in \mathbb{R}_{>0}$	maximum observation time
$t, s \in [0,T]$	time points within the observation time
$\mathcal{G} = (V, E)$	dependence graph with nodes V and edges E
$V \cong [d]$	set of vertices in the dependence graph
$E \subseteq V \times V$	set of directed edges between nodes V
$\mathrm{pa}_v^{\mathcal{G}}$	parents of node v in graph $\mathcal{G}$ (omitted if clear from context)
$\tilde{\mathcal{G}} = (\tilde{V}, \tilde{E})$	lifted dependence graph of graph $\mathcal{G} = (V, E)$
V_0, V_1	disjoint copies of V, with $0/1$ indicating past/future
$\tilde{V} := V_0 \sqcup V_1$	node set in lifted dependence graph
$\tilde{E} \subseteq \tilde{V} \times \tilde{V}$	edge set in lifted dependence graph
$S(X)_{s,t} \in \mathcal{T}(\mathbb{R}^d)$	signature transform of path X over interval $[s,t]$
$K_S(X, Y)$	signature kernel of paths X, Y
$T(\mathbb{R}^d)$	the direct sum $\bigoplus_{n=1}^{\infty} (\mathbb{R}^d)^{\otimes n} := \{a = (a_n)_{n\in\mathbb{N}} = \prod_{n\in\mathbb{N}} (\mathbb{R}^d)^{\otimes n};\ \max\{n : a_n \neq 0\} < \infty\}$
$\mathcal{T}(\mathbb{R}^d)$	the completion of $T(\mathbb{R}^d)$ w.r.t. $\langle a, b\rangle = \sum_{n=1}^{\infty} a_n b_n$ (well-defined)
$\mathcal{O}(\cdot)$	Landau asymptotic notation ("Big-O notation")

A.2 Intuition and Details for the SDE Model

Making the actual dependence of μ^k and σ^k on their arguments in eq. (1) explicit, we can rewrite it as

$$\begin{cases} \mathrm{d}X_t^k = \mu^k(X_{[0,t]}^{\mathrm{pa}_k})\mathrm{d}t + \sigma^k(X_{[0,t]}^{\mathrm{pa}_k})\mathrm{d}W_t^k, \\ X_0^k = x_0^k \qquad k \in [d]. \end{cases} \tag{7}$$

Then μ^k, σ^k are functions defined on $C([0,+\infty), \mathbb{R}^{\dim(\mathrm{pa}_k)})$ (or some suitable subspace thereof): Lipschitz conditions on the coefficients that guarantee existence and uniqueness for this type of SDE can be found in Rogers & Williams (2000), which we assume to hold throughout. Each W^k is an m_k-dimensional Brownian motion (a collection of m_k independent 1-dimensional Brownian motions, that is), σ^k maps to the space of $n_k \times m_k$-dimensional matrices, and the noises W^k together with the (possibly random) initial conditions x_0^k are jointly independent. In other words, the system can be written as an $n := n_1 + \ldots + n_d$-dimensional SDE driven by an $m := m_1 + \ldots + m_d$-dimensional Brownian motion, with a block-diagonal diffusion coefficient σ (since the noise is unobserved, this structure has to be imposed if we wish not to deal with unobserved confounding). The superscript pa_k refers to the parents of the kth node in $\mathcal{G}$, which may (and most often does) contain k itself. Intuitively, all of this means that, for all times t and small increments Δt, the increment of the solution $X_{t+\Delta t}^k - X_t^k$ is random with distribution that is well-approximated by a multivariate normal with mean $\mu^k(X_{[0,t]}^{\mathrm{pa}_k})\Delta t$ and covariance function $\sigma^k(X_{[0,t]}^{\mathrm{pa}_k})\sigma^k(X_{[0,t]}^{\mathrm{pa}_k})^\intercal \Delta t$, and independent of the history

of the system up to time t. This interpretation can actually be made precise by showing that these piecewise constant paths converge in law to the true solution (usually known as the weak solution, when viewed in this way).

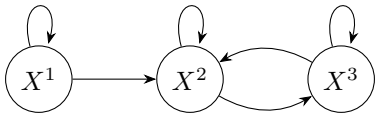
We will refer to $\mathcal{G}$ as the *dependence graph* of the eq. (1), which we refer to as the *SDE model* for brevity. Compared to that of Peters et al. (2022), this model is slightly more general in that (i) it allows for path-dependence and (ii) for each node to represent a multidimensional process; the special case of state-dependent SDE—i.e., in which μ and σ only depend on X_t, the value of X at time t—continues to be an important special case, although our model also accommodates delayed SDEs (the coefficients depend on the value of X at a prior instant in time, e.g., $X_{t-\tau}$ for fixed or possibly random/time-dependent $\tau > 0$), and SDEs that depend on quantities involving the whole past of X, such as the average $t^{-1}\int_0^t X_s \mathrm{d}s$. We note that the choice not to make the coefficients explicitly time-dependent is deliberate and motivated by the requirement that the system be *causally stationary* (the causal relations between the variables do not change over time).

Given that we are considering the dynamic setting, it is generally natural to allow for $\mathcal{G}$ to have cycles. This comes at no additional requirement of consistency constraints as it does in the static case (see for example Bongers et al. (2021)), since the causal arrows in the model eq. (1) should be thought of as 'pointing towards the infinitesimal future, with infinitesimal magnitude', integrated over the whole time interval considered. Indeed, the system of SDEs does not require any global consistency to be well-posed, other than the regularity and growth conditions that guarantee global-in-time existence and uniqueness. On the other hand, we will be interested in the potential for constraint-based causal discovery of such systems, qualified by the following assumption on the types of conditional independencies that we allow to be tested in continuous time:

A.3 Counterexample for Cyclic Causal Discovery in the SDE Model

The following example demonstrates that constraint-based causal discovery of the full graph is impossible using expressions of the form eq. (2) when allowing for cycles.

Example A.1. An oracle for eq. (2) is not powerful enough to rule out a directed edge $X^1 \to X^3$ for an SDE model with the following dependence graph:



We now go through this example and describe a different type of oracle that would be required for causal discovery in cyclic SDE models. Concretely, we claim that an oracle for eq. (2) is not powerful enough to rule out a directed edge $X^1 \to X^3$ for an SDE model with the graph in Example A.1. Testing $X^1_{[0,s]} \perp\!\!\!\perp X^3_{[s,s+h]} \mid X^{2,3}_{[0,s]}$, will not remove the edge, due to the open path

$$X^1_{[0,s]} \to X^1_{[s,s+h]} \to X^2_{[s,s+h]} \to X^3_{[s,s+h]}.$$

Testing $X^1_{[0,s]} \perp\!\!\!\perp X^3_{[s,s+h]} \mid X^3_{[0,s]}, X^2_{[0,s+h]}$, on the other hand, runs into the collider

$$X^1_{[0,s]} \to X^1_{[s,s+h]} \to X^2_{[s,s+h]} \leftarrow X^3_{[s,s+h]}.$$

There is no other way of splitting up the time interval (even allowing for more than 2 subintervals) that overcomes both these problems: when testing $X^1_{[0,s]} \perp\!\!\!\perp X^3_{[s,s+h]}$ (as is necessary in order to rely on time to obtain direction of the arrow), if X^2 is not conditioned on over $[s, s+h]$ it will be a mediator, and if it is, it will be a collider.

What would be needed to detect the edge in the above example (and to perform causal discovery more generally in the cyclic case), is the availability of tests of *strong instantaneous non-causality in the Granger sense* as defined in Florens & Fougere (1996). This reflects the fact that the SDE model is 'acyclic at infinitesimal scales'. Unfortunately, such a property—which has to do with the Doob-Meyer decompositions of functions of the process w.r.t. two different filtrations—is much more difficult to test for, as it is infinitesimal in nature. Recent progress in this direction was made in Christgau et al. (2023) for the case of SDEs driven by jump processes; this local independence

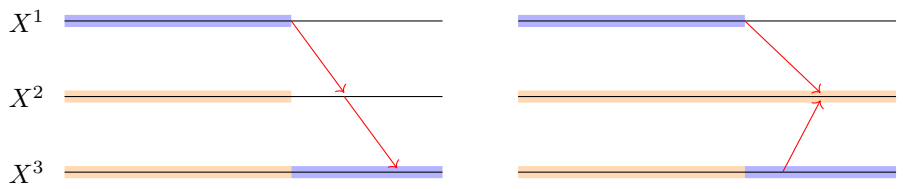


Figure 5: The blue highlighter is for intervals over which the path is being tested for independence, while orange is for conditioning. The first figure illustrates the path opened by the unconditioned-on mediator $X^2_{[s,s+h]}$, while the second illustrates the path opened by the same variable acting as a conditioned-on collider. While not quite possible to draw extra/figures this way, it is helpful to think of the arrows as pointing from values to increments $X_t \to (Y_{t+\mathrm{d}t} - Y_t)$, infinitesimally, with the total causal effect accrued over time.

criterion, which goes back to Schweder (1970), is however *weak* in the sense that it only takes into account the Doob-Meyer decomposition of X and not functions of it, and is not therefore able, for example, to detect dependence in the diffusion coefficient (Gégout-Petit & Commenges, 2010). Here we take the view that a continuously-indexed path, queried over an interval, can be considered as an observational distribution (cf. the discussion in Sokol & Hansen (2013) on whether the infinitesimal generator can be considered 'observational').

A.4 Signature Kernel Details

Additional mathematical details. For any $p \in [1, 2)$, let $\mathcal{K} \subset C_p([0, T], \mathbb{R}^d)$ be a compact subset such that the signature transform is a continuous injection from $\mathcal{K}$ to $\mathcal{T}(\mathbb{R}^d)$. An example of compact set $\mathcal{K}$ satisfying such conditions is the set of time-augmented paths started at a common origin, where $C_p([0, T], \mathbb{R}^d)$ is endowed with the p-variation topology. Further examples of compact sets and choices of topologies on path-space are discussed by Cass & Turner (2024). Then, the signature kernel is *universal* over $\mathcal{K}$ in the sense that the associated RKHS, whose elements are restricted to act on $\mathcal{K}$, is dense in $C(\mathcal{K})$, the space of continuous real-valued functions on $\mathcal{K}$, endowed in the topology of uniform convergence. See Cass et al. (2023, Proposition 3.3) for a proof. The compactness assumption can be relaxed by either renormalizing the signature transform so that its range is always guaranteed to be bounded (Chevyrev & Oberhauser, 2022) or by operating on weighted path-spaces with a different choice of topology Cuchiero et al. (2023).

The signature kernel is well-defined and can be shown, under similar compactness conditions, to be universal also when $p \geq 2$ (Salvi et al., 2021a; Lemercier & Lyons, 2024). In the specific setting of stochastic differential equations (SDEs) considered in this paper, i.e. when $p \in [2, 3)$, one can use a classical limiting procedure from rough analysis to show that the integral equation in eq. (4) takes the Stratonovich form $f(t, t') = 1 + \int_s^t \int_{s'}^{t'} f(u, v) \langle \circ dX_u, \circ dY_v \rangle_1$. We refer the interested reader to Salvi et al. (2021a, Section 4.3) for further details.

Another important property that a kernel should possess in order to ensure theoretical guarantees of most kernel methods, including tests for conditional independence (Muandet et al., 2017), is that of being *characteristic*. Loosely speaking, a kernel k is said characteristic to a topological space if for any (Borel) probability measure μ on that space, the μ-expectation of the feature map $k(x, \cdot)$ uniquely characterises μ. An important result from the theory of kernels relates to the equivalence between the notions of universality and characteristicness. In particular, these three notions can be shown to be equivalent for kernels defined over general locally convex topological vector spaces (Simon-Gabriel & Schölkopf, 2018, Theorem 6).

Intuition. The signature transform of a (potentially multidimensional) continuous path (or time series) is a sequence of iterated integrals, static features, which form an expressive representation of the path. An analogy is often drawn to monomials in (multiple) indeterminants, with which one can express continuous functions. Another useful analogy is the Fourier transform. While Fourier transforms (or wavelet transforms) capture the frequency content of different modes (over different times), the signature captures information about order, area, and so forth (due to the *iterated* integrals). For example, the depth 1 (or order 1) iterated integral describes the changes of each

variable throughout the time interval; depth 2 describes the signed area that is swept by the curve (an illustration of depth 1 and 2 of a path can be found in Morrill et al. (2020)); higher depths capture properties of the volume, and so on, but those can be hard to grasp visually. We refer the reader to Chevyrev & Kormilitzin (2016); Lee & Oberhauser (2023) as introductory texts on the signature (kernel) method that contain more visualizations and examples for building intuitions about these mathematical objects.

A.5 Testing on Time Points

It is also possible to replace any of the path-valued random variables in eq. (2) by a random variable at a single time-point (as for example done in the post-processing by testing on X_0) using a regular rbf-kernel. Note, that in addition, if we knew in advance that there is no path-dependence, the unconditional independence $X^i \perp\!\!\!\perp^+_{s,h} X^j$ is equivalent to the conditional independence $X^i_s \perp\!\!\!\perp X^j_{s+h} - X^j_s \mid X^j_s$ (at fixed times $s, s+h$): this is because without path-dependence X^j_u for $u \in [s,t]$ only depends on $X^j_{[s,t]}$ via X^j_s. Testing on entire (segments of) paths is always required when allowing for path-dependence.

A.6 Proof of Proposition 3.1

Proof of Proposition 3.1. Since $\mathcal{G}$ is acyclic, a directed cycle in $\tilde{\mathcal{G}}$ that is not a loop must contain nodes both in V_0, V_1. But there can be no such directed cycles, since edges only travel in the direction $V_0 \to V_1$, by construction of $\tilde{\mathcal{G}}$. Since $\tilde{\mathcal{G}}$ also free of loops, it is a DAG. We now show that there exists a structural causal model (SCM) over $\tilde{\mathcal{G}}$ with the path-valued random variables of the statement: Markovianity will then follow from Peters et al. (2017, Proposition 6.31), original to Verma & Pearl (1990). In other words, we must show there exist Borel-measurable functions (setting $t := s+h$)

$$\begin{aligned} F^k_0 \colon \mathbb{R}^{\dim(\mathrm{pa}_k)} \times C([0,s],\mathbb{R}^{m_k}) \times C([0,s],\mathbb{R}^{\dim(\mathrm{pa}_k\setminus\{k\})}) &\rightharpoonup C([0,s],\mathbb{R}^{n_k}) \\ (x_0^{\mathrm{pa}(k)}, W^k_{[0,s]}; X^{\mathrm{pa}_k\setminus\{k\}}_{[0,s]}) &\mapsto X^k_{[0,s]} \end{aligned}$$

and

$$\begin{aligned} F^k_1 \colon C([s,t],\mathbb{R}^{m_k}) \times C([0,s],\mathbb{R}^{\dim(\mathrm{pa}_k)}) \times C([s,t],\mathbb{R}^{\dim(\mathrm{pa}_k\setminus\{k\})}) &\rightharpoonup C([0,s],\mathbb{R}^{n_k}) \\ (W^k_{[s,t]}; X^{\mathrm{pa}_k}_{[0,s]}, X^{\mathrm{pa}_k\setminus\{k\}}_{[s,t]}) &\mapsto X^k_{[0,s]} \end{aligned}$$

with the initial conditions x^0 and Brownian path segments $W^k_{[0,s]}, W^k_{[s,t]}$ collectively independent. This independence follows from the definition of the SDE model and from independence of Brownian increments over disjoint or consecutive intervals. These functions are partial in that they are not defined on the whole space of continuous functions on the interval: we only need to show them to be defined on a measurable set of paths that contains all solutions of SDEs on the required interval. F^k_1 is defined as follows, with care to make explicit the dependence on the solution on the two intervals via the operation of path-concatenation $*$:

$$\begin{aligned} F^k_1(X^k_{[s,t]}) = \text{solution of} \quad X^k_u = X^k_s + \int_0^u \mu^k(X^{\mathrm{pa}_k\setminus\{k\}}_{[0,s]} * X^{\mathrm{pa}_k\setminus\{k\}}_{[s,u]}, X^k_{[0,s]} * X^k_{[s,u]})\mathrm{d}t \\ + \sigma^k(X^{\mathrm{pa}_k\setminus\{k\}}_{[0,s]} * X^{\mathrm{pa}_k\setminus\{k\}}_{[s,u]}, X^k_{[0,s]} * X^k_{[s,u]})\mathrm{d}W^k_t \end{aligned}$$

We have also separated out the k-th component of the solution from the rest of the arguments on which σ^k, μ^k are dependent, which we consider part of the measurable adapted dependence on $W^{\mathrm{pa}\setminus\{k\}}$ needed in the existence and uniqueness theorem. F^k_0 is defined similarly (with no dependence on past path-segments and with x^k_0 replacing X^k_s). Such functions are well-known to be well-defined and measurable (Rogers & Williams, 2000, Theorem 10.4). □

Remark A.2. In testing, using the symmetric criterion $\perp\!\!\!\perp_{\mathrm{sym}}$ has proven to be more reliable than the asymmetric one used in the test above. A procedure, which is a little redundant, but which makes good use of the superior performance of $\perp\!\!\!\perp_{\mathrm{sym}}$ would work as follows:

1. Run the PC algorithm to discover the CPDAG corresponding to the variables $X^k_{[0,T]}$.

2. Pass this CPDAG as input to Algorithm 1: by this we mean that $\tilde{V}$ is the same and the input $\tilde{E}$ is given by

$$\{i_0 \to j_0,\ i_1 \to j_1, i_0 \to j_1 \mid i, j \in V,\ (i \to j) \text{ or } (i - j) \in E'\} \cup \{k_0 \to k_1 \mid k \in V\}$$

   where $(i \to j)$ denotes an undirected arrow and E' is the CPDAG output by the classical PC algorithm.

3. The loop-recovery step is identical as before.

The algorithm would run very similarly as before, but performing fewer CI tests in the first phase, since it can make use of the information contained in the partially-directed skeleton.

Remark A.3. If we knew in advance that there is no path-dependence, the test $X^i \perp\!\!\!\perp^+_{s,h} X^j$ (with no additional conditioning set) is equivalent to the test on values of the process $X^i_s \perp\!\!\!\perp X^j_{s+h} \mid X^j_s$: this is because X^j_u, for $u \in [s,t]$ only depends on $X^j_{[s,t]}$ via X^j_s. However, this kind of statement no longer works in the conditional case (by arguments similar to Example A.1), and testing on whole paths is strictly necessary when allowing for path-dependence.

Remark A.4. CI testing of SDEs is conceptually similar to static CI on discretely sampled data, when the sampling rate is lower than the actual frequency at which causal effects propagate. This is because, in continuous time and for any $h > 0$, there are going to be causal effects that occur at time scales less than h. Whenever we just discretely sample time series with no instantaneous effects, Peters et al. (2017, Thm. 10.3) provides full causal discovery, even in the presence of cycles in the summary graph, by performing d^2 tests $X^i_s \perp\!\!\!\perp X^j_{s+\Delta s} \mid X^{[d]\setminus\{i,j\}}_{[0,s]}$. Potentially absent loops could then be removed as in Algorithm 1. Ignoring that this might suffer from the conditioning set being too large, our signature method can be of help in this setting too, if there is path-dependence: the conditioning variable can be taken to be $S(X^{[d]\setminus\{i,j\}})_{0,s}$.

A.7 Proof of Theorem 3.2

Proof of Theorem 3.2. We have to show that $E^{Alg} = E$ (where E^{Alg} the edges detected by Algorithm 1).

"$\subseteq$:" This direction can be shown by finding an appropriate separation set in $\tilde{\mathcal{G}}$ by exploiting the global Markov property Proposition 3.1. We begin by proving correctness of the recovery of the skeleton modulo loops. This will follow if we show that for $(i \to j) \notin E$ for $i, j \in V$ distinct

$$i_0 \perp\!\!\!\perp^{\tilde{\mathcal{G}}}_{d-sep} j_1 \mid \{j_0\} \cup \{k_0, k_1 \mid k \in \mathrm{pa}^{\mathcal{G}}_j \setminus \{j\}\},$$

where $\perp\!\!\!\perp^{\tilde{\mathcal{G}}}_{d-sep}$ denotes d-separation in the graph $\tilde{\mathcal{G}}$. Indeed, eventually, the set K in the algorithm will take the value $\{k_0, k_1 \mid k \in \mathrm{pa}^{\mathcal{G}}_j \setminus \{j\}\}$, at which point the three edges (i_0, j_0), (i_1, j_1), (i_0, j_1) (which are either all present or all absent, inductively, by construction of $\tilde{E}$) are deleted. (Of course it may be that these edges are deleted before this, if a smaller or different d-separating set is found, but note that edges are never added.) All undirected paths between i_0 and j_1 factor as $i_0 \cdots h_0 \to l_1 \cdots j_1$, where the dots stand for a possibly empty undirected path. All such paths with $l_1 = j$ are blocked by $\{j_0\} \cup \{k_0 \mid k \in \mathrm{pa}^{\mathcal{G}}_j \setminus \{j\}\}$, so we focus our attention on the case in which $l_1 \cdots j_1$ is non-empty. Here we follow a similar argument to that of Verma & Pearl (1990, Lemma 1). If $l_1 \cdots j_1$ is of the form $l^1_1 \cdots l^r_1 \to j_1$ (with $l^1_1 = l_1$) it is blocked by $\{k_1 \mid k \in \mathrm{pa}^{\mathcal{G}}_j \setminus \{j\}\}$. Assume instead it is of the form $l^1_1 \cdots l^r_1 \leftarrow j_1$ and let $1 \le q \le r$ be such that l^q_1 is the first collider on the entire path $i_0 \cdots j_1$, starting from j_1 and travelling back: certainly this exists (and belongs to V_1), thanks to the arrows $h_0 \to l^1_1$ and $l^r_1 \leftarrow j_1$. For the path $i_0 \cdots j_1$ to be open given the conditioning set, it must be the case that l^p_1 must be an ancestor (in $\tilde{\mathcal{G}}$) or member of $\{k_1 \mid k \in \mathrm{pa}^{\mathcal{G}}_j \setminus \{j\}\}$ (l^p_1 cannot be an ancestor of nodes in V_0): this would yield a directed cycle, which is impossible.

We argue similarly for the loop-removal phase. If $(i \in i) \notin E \iff (i_0, i_1) \notin \tilde{E}$, we must show

$$i_0 \perp\!\!\!\perp^{\tilde{\mathcal{G}}}_{d-sep} i_1 \mid \{k_0, k_1 \mid k \in \mathrm{pa}^{\mathcal{G}}_i \setminus \{i\}\}$$

Consider a path $i_0 \cdots h_0 \to l_1 \cdots i_1$ in $\tilde{\mathcal{G}}$. If the segment $l_1 \cdots i_1$ is non-empty we conclude that the path is blocked by the same argument as above (with i_1 replacing j_1). Assume that the path is of the

form $i_0 \cdots h_0 \to i_1$ (with $i_0 \neq h_0$ since $(i_0, i_1) \notin \tilde{E}$): in this case $h \in \mathrm{pa}_i^{\mathcal{G}} \setminus \{i\}$ and the path is again blocked. Thus all paths are blocked and the implication follows.
$\supseteq$: Assume $(i,j) \in E$, thus $(i_0, j_1) \in \tilde{E}$. Hence as $\tilde{\mathcal{G}}$ a DAG by construction, $\nexists\, S \subseteq V \setminus \{i,j\}$ s.t. $\{i_0\} \perp\!\!\!\perp_d^{\tilde{\mathcal{G}}} \{j_1\} \mid S_0 \cup \{j_0\} \cup S_1$. By the faithfulness assumption Definition A.5 one therefore has $X^i_{[0,s]} \not\perp\!\!\!\perp X^j_{[s,s+h]} \mid X^S_{[0,s]}, X^j_{[0,s]}, X^S_{[s,s+h]}\ \forall\, S \subseteq V \setminus \{i,j\}$. Hence $(i,j) \in E^{Alg}$. □

A.8 Faithfulness and Causal Minimality

In the above proof, we have used the following notion of *faithfulness*:

Definition A.5 ((Global) Faithfulness of $\perp\!\!\!\perp^+_{s,h}$ in $\tilde{\mathcal{G}}$)**.** Let $\{X^i_t\}_{i \in [d]}$ be the coordinate processes of a solution of eq. (1) for $t \in [0,1]$, $\mathcal{G} = (V \cong [d], E)$ the dependence graph defined by eq. (1). The independence relation $\perp\!\!\!\perp^+_{s,h}$ is called *(globally) faithful* w.r.t. the DAG $\tilde{\mathcal{G}}$, if

$$X^A \perp\!\!\!\perp^+_{s,h} X^B \mid X^C \;\Rightarrow\; A_0 \perp\!\!\!\perp_d^{\tilde{\mathcal{G}}} B_1 \mid B_0, C_0, C_1 \tag{8}$$

for all $A, B, C \subseteq V$ pairwise disjoint.

Faithfulness is a standard assumption in the causal discovery literature. Some type of faithfulness-like assumption is strictly necessary to do causal discovery since, otherwise, it would be impossible to make any inference about adjacencies in the underlying graph from the observed distribution. While we used the strong faithfulness assumption Definition A.5 for the proof above, various weaker faithfulness-like assumptions in the case of DAG-based causal discovery could be used, e.g., the strictly weaker assumption known as 'parent faithfulness' (Mogensen (2024) gives a definition of parent faithfulness in the context of stochastic processes). This assumption suffices in our algorithm without making any changes to it (or the respective proofs) in the acyclic case.

The strength of the faithfulness assumption in real-world data is heavily debated. Besides Weinberger (2018), who states that 'proposed counterexamples to the causal faithfulness condition are not genuine', Spirtes et al. (2000) argues that for parametric models, 'the measure of the set of free parameter values for any DAG that lead to such cancellations is zero for any smooth prior probability density, such as the Gaussian or exponential one, over the free parameters'. What is meant in our case is, that when allowing the diffusion coefficients σ in eq. (1) to range in some 'reasonable' finite-dimensional family (such as linear or affine functions) and drawing the parameters of such a family from a distribution which has positive density w.r.t. the Lebesgue measure (and independently of the noise), the resulting distribution on path-space, split over $[0,s]$ and $[s,t]$, will almost surely be faithful for any $0 < s < t$ w.r.t. the lifted dependence graph. We do not attempt a proof of this statement here, as it may require considerable effort and technique. On the other side, Andersen (2013) or Uhler et al. (2013) (among others) caution against making the strong faithfulness assumption easily. Many of their concerns are ameliorated by only assuming the weaker parent faithfulness.

Causal minimality, on the other hand, is generally easier to understand and can be expected to hold for the arrow $X^i_{[0,s]} \to X^j_{[s,t]}$ whenever the following condition is satisfied: the conditional distribution $X^j_u \mid X^{\mathrm{pa}_j^{\mathcal{G}} \setminus \{i\}} = x^j$ admits a positive density on some submanifold M^x_u of $\mathbb{R}^n$ and for all u there exist $x^i_1, x^i_2 \in M^x_u$, $x^i_1 \neq x^i_2$, s.t. $(\mu^j, \sigma^j)(x, x^i_1) \neq (\mu^j, \sigma^j)(x, x^i_2)$.

Example A.6 (Causal minimality)**.** The fact that we are allowing for variables to have more than one dimension means the causal minimality condition is not a trivial one, as the following examples demonstrate. Let $X^1 = (W, 0)$ where W is a 1-dimensional Brownian motion (so that $n_1 = 2$), and let μ^1 and/or σ^1 depend on X^1 only through its second coordinate. Then, even though this dependence may be non-trivial, causal minimality does not hold. This may lead one to believe that one can generically only expect causal minimality to hold if $m = n$, but this is not the case. Take, for example, $d = 4 = n$ and assume $\sigma^3 \equiv 0$, i.e., X^3 has no driving Brownian motion, so that m is only 3. Assume, furthermore, that $\mathrm{pa}_4 = \{4\}$. Then, by Hörmander's Lie bracket-generating condition (see for example §V.38 in Rogers & Williams (2000)), (X^1, X^2, X^3) has a density in $\mathbb{R}^3$ for any choice of an initial condition, if $(\sigma^1(x), 0, 0), (0, \sigma^2(x), 0)$ span $\mathbb{R}^2$ for all $x \in \mathbb{R}^3$ and $[(\sigma^1(x), 0, 0), \mu] = \sigma^1 \partial_1 \mu - \mu^i \partial_i \sigma^1$ or $[(\sigma^2(x), 0, 0), \mu]$ span the third direction. Arrows going from the first three nodes to the fourth will be necessary for Markovianity as long as the coefficients of X^4 depend on the respective variables on the support of such density. The point is that even though (X^1, X^2, X^3) is only driven by a two-dimensional Brownian motion, the coefficients of the SDE (at

the initial condition) impart a 'twist' to the solution, making causal minimality a trivial condition on (σ^4, μ^4) which is verified as soon as there is any dependence.

A.9 Global Markov Property for the Symmetric Independence Criterion

In this section, we provide a proof for the global Markov property of the symmetric independence criterion in order for the PC-algorithm to be applicable. It is based on the notion of independence models:

Definition A.7. Let $V \cong [d]$ be a set. An *independence model* $\mathcal{J}(V)$ over V is a ternary relation over disjoint subsets of V,

$$\mathcal{J}(V) \subset \{(A, B, C) \mid A, B, C \subset V \text{ disjoint}\}$$

When (A, B, C) is a triple in $\mathcal{J}(V)$, $(A, B, C) \in \mathcal{J}(V)$, we also use $\langle A, B \mid C\rangle$ to denote (A, B, C). This notation highlights the fact that C is a conditioning set. An independence model $\mathcal{J}(V)$ is called a *semigraphoid* if 1.-4. hold for all disjoint $A, B, C \subset V$.

1. (symmetry) $\langle A, B \mid C\rangle \in \mathcal{J}(V) \Rightarrow \langle B, A \mid C\rangle \in \mathcal{J}(V)$
2. (decomposition) $\langle A, B \mid C\rangle \in \mathcal{J}(V)$ and $D \subset B \Rightarrow \langle A, D \mid C\rangle \in \mathcal{J}(V)$
3. (weak union) $\langle A, B \cup D \mid C\rangle \in \mathcal{J}(V) \implies \langle A, B \mid C \cup D\rangle \in \mathcal{J}(V)$
4. (contraction) $\langle A, B \mid C\rangle \in \mathcal{J}(V)$ and $\langle A, D \mid B \cup C\rangle \in \mathcal{J}(V) \implies \langle A, B \cup D \mid C\rangle \in \mathcal{J}(V)$

A semigraphoid $\mathcal{J}(V)$ is called *graphoid* if 5. holds.

5. (intersection) $\langle A, B \mid C \cup D\rangle \in \mathcal{J}(V)$ and $\langle A, D \mid B \cup C\rangle \in \mathcal{J}(V) \implies \langle A, B \cup D \mid C\rangle \in \mathcal{J}(V)$

An independence model $\mathcal{J}(V)$ called *compositional* if 6. holds.

6. (composition) $\langle A, B \mid C\rangle \in \mathcal{J}(V)$ and $\langle A, D \mid C\rangle \in \mathcal{J}(V) \implies \langle A, B \cup D \mid C\rangle \in \mathcal{J}(V)$

The graphical criterion of d-separation defines a independence model on a DAG $\mathcal{G} = (V, \mathcal{E})$ by

$$\mathcal{J}_{d-sep}(\mathcal{G}) = \{(A, B, C) \mid A, B, C \subset V \text{ disjoint}, A \perp\!\!\!\perp_{d-sep} B \mid C\} \tag{9}$$

Theorem A.8 (Lauritzen & Sadeghi (2018))**.** *For any DAG $\mathcal{G} = (V, \mathcal{E})$, the independence model $\mathcal{J}_{d-sep}(\mathcal{G})$ is a compositional graphoid.*

Let $V = [d]$. Given random variables $X^i : (\Omega, \mathcal{A}, P) \to (\mathcal{X}^i, \mathcal{A}_i)$, $i \in V$, and $A, B, C \subset V$ disjoint, conditional independence

$$X^A \perp\!\!\!\perp X^B \mid X^C :\Leftrightarrow \sigma(X^A) \perp\!\!\!\perp \sigma(X^B) \mid \sigma(X^C)$$

defines the *probabilistic independence model*

$$\mathcal{J}(P) = \{(A, B, C) \mid A, B, C \subset V \text{ disjoint}, X_A \perp\!\!\!\perp X_B \mid X_C\}\,. \tag{10}$$

We will use this as a *symmetric conditional independence* relation. Note that X^i may take values in a path-space in which case it is a stochastic process. The independence model $\mathcal{J}(P)$ is a semigraphoid which is an immediate consequence of sub-σ-algebra properties.

For $v \in V$, we let $\mathrm{nd}_v^{\mathcal{G}}$ denote the set of *nondescendants* of v, i.e., the set of nodes i such that there is no directed path from v to i.

Definition A.9 (Directed global and local Markov properties)**.** Let $\mathcal{G} = (V, \mathcal{E})$ be a DAG and $\mathcal{J}(V)$ be an independence model over V. The independence model $\mathcal{J}(V)$ satisfies the *global Markov property w.r.t.* $\mathcal{G}$ $:\Leftrightarrow$

$$\langle A, B \mid C\rangle \in \mathcal{J}_{d-sep}(\mathcal{G}) \quad \implies \quad \langle A, B \mid C\rangle \in \mathcal{J}(V) \qquad \forall A, B, C \subset V \text{ disjoint} \tag{11}$$

The independence model $\mathcal{J}(V)$ satisfies the *directed local Markov property w.r.t.* $\mathcal{G}$ $:\Leftrightarrow$

$$\langle \{v\}, \left(\mathrm{nd}_v^{\mathcal{G}} \backslash \mathrm{pa}_v^{\mathcal{G}}\right) \mid \mathrm{pa}_v^{\mathcal{G}}\rangle \in \mathcal{J}(V) \text{ for all } v \in V \tag{12}$$

Theorem A.10 (Lauritzen et al. (1990))**.** *Let $\mathcal{G} = (V, \mathcal{E})$ be a DAG, let and $\mathcal{J}(V)$ a semigraphoid over V. The directed global and local Markov properties are equivalent, that is,*

$$\mathcal{J}(V) \quad \textit{satisfies eq. (11) w.r.t. } \mathcal{G} \;\Leftrightarrow\; \mathcal{J}(V) \quad \textit{satisfies eq. (12) w.r.t. } \mathcal{G} \tag{13}$$

We therefore only have to establish the local Markov property for the symmetric independence model $\perp\!\!\!\perp_{\text{sym}}$:

Proposition A.11. *Let X be a set of variables induces by the model in eq. (7) with a constant and diagonal diffusion matrix such that it induces the DAG $\mathcal{G} = (V, \mathcal{E})$. Then the system satisfies the* directed local Markov property *w.r.t. $\mathcal{G}$,*

$$X^i \perp\!\!\!\perp_{\text{sym}} \text{nd}^{\mathcal{G}}_{X^i} \setminus \text{pa}^{\mathcal{G}}_{X^i} \mid \text{pa}^{\mathcal{G}}_{X^i} \qquad \forall i \in [d] \tag{14}$$

Proof. The idea of the proof is that all information about node X^i is contained in its parents $pa_i := \text{pa}^{\mathcal{G}}_{X^i}$, the Brownian motion W^i, and the initial condition X^i_0. A similar idea can be used to show the result in the case of a SCM Pearl (2009). We let $\tilde{\mathcal{J}}(P)$ denote the semigraphoid induced by $\perp\!\!\!\perp_{\text{sym}}$ on the variable set $\tilde{V} = \{X^1, \dots, X^n\} \cup \{W^1, \dots, W^n\} \cup \{X^1_0, \dots, X^n_0\}$, and we let $\mathcal{J}(P)$ denote the semigraphoid induced by $\perp\!\!\!\perp_{\text{sym}}$ on the variable set $V = \{X^1, \dots, X^n\}$. We define the DAG

$$\tilde{\mathcal{G}} = \left(\tilde{V} = \{X^1, \dots, X^n\} \cup \{W^1, \dots, W^n\} \cup \{X^1_0, \dots, X^n_0\},\ \tilde{E} = E \cup \{X^j \leftarrow W^j\} \cup \{X^j \leftarrow X^j_0\}\right). \tag{15}$$

By construction, $\mathcal{F}_t(X^i) \subset \mathcal{F}_t(W^i) \vee \mathcal{F}_t(X^{\text{pa}^{\mathcal{G}}_i}) \vee \sigma(X^i_0)$, and therefore

$$\mathcal{F}_t(X^i) \perp\!\!\!\perp \mathcal{F}_t(X^{\text{nd}^{\mathcal{G}}_i \setminus \text{pa}_i} \vee W^{-i} \vee X^{-i}_0) \mid \mathcal{F}_t(X^{\text{pa}_i}) \vee \mathcal{F}_t(W^i) \vee \sigma(X^i_0)$$

where the parent and nondescendant sets are to be read in the graph $\mathcal{G}$, $W^{-i} = \{W_1, \dots, W_n\} \setminus \{W_i\}$, and $X^{-i}_0 = \{X^1_0, \dots, X^n_0\} \setminus \{X^i_0\}$. The above statement corresponds to eq. (12) when $v \in \{X_1, \dots, X_n\}$. When $v = W_i$ or $v = X^i_0$, v has no parents in $\tilde{\mathcal{G}}$ and eq. (12) also holds. Therefore, the independence model $\tilde{\mathcal{J}}(P)$ satisfies the local directed Markov property w.r.t. to $\tilde{\mathcal{G}}$.

The independence model $\tilde{\mathcal{J}}(P)$ is a semigraphoid, and by Theorem A.10 it satisfies eq. (11) w.r.t. $\tilde{\mathcal{G}}$. We have $X^i \perp\!\!\!\perp_{d-sep} X^{\text{nd}_i \setminus \text{pa}_i} \mid X^{\text{pa}_i}$ in $\tilde{\mathcal{G}}$, and one therefore has

$$\mathcal{F}_t(X^i) \perp\!\!\!\perp \mathcal{F}_t(X^{\text{nd}_i \setminus \text{pa}_i}) \mid \mathcal{F}_t(X^{\text{pa}^i})$$

establishing eq. (12) for $\mathcal{J}(P)$ w.r.t. $\mathcal{G}$. □

A.10 Proof of Corollary 3.3 and Algorithm for Robust Causal Discovery

Proof. This is clear from the data generating mechanism in eq. (1) where all Brownian motions $\mathrm{d}W^k_t$ and initial conditions X^k_0 are jointly independent. Since $X^j_{[0,T]}$ is fully determined by $\{X^{\text{an}_j}_0, \mathrm{d}W^{\text{an}_j}_t\}$ where an_j are the ancestors of j in $\mathcal{G}$, it follows that $i \in \text{an}_j$ if and only if $X^j_{[0,T]} \not\perp\!\!\!\perp X^i_0$. Since we already know from the CPDAG that i and j are adjacent and the graph is acyclic, $i \in \text{an}_j$ implies $i \to j$. □

The full written out algorithm then looks like Algorithm 2 and is sound and complete by Corollary 3.3.

A.11 Post-Processing Without Jointly Independent Initial Values

In real-world settings where there is no natural 'starting point' of the processes, we can typically not assume joint independence of initial values X^i_0 as 'the process also happened before'. In this subsection we extend Algorithm 2 to a version that is applicable in scenarios where we cannot assume

Algorithm 2 Robust Causal discovery for acyclic SDE models leveraging the initial values

$V_{pre} \leftarrow \{k \in V\}$
$E_{pre} \leftarrow \{i \to j,\ j \to i \mid i, j \in V,\ i \neq j\}$
for $c = 0, \dots, d-2$ **do** ▷ *Adjacency detection*
  for $i, j \in V, i \neq j$ **do**
    for $K \subseteq V \setminus \{i, j\}, |K| = c,$
      s.t. $(k - j) \in E_{pre}$ for $k \in K$ **do**
      if $X^i \perp\!\!\!\perp X^j \mid X^K$ **then**
        $E_{pre} \leftarrow E_{pre} \setminus \{i \to j, j \to i\}$
        $S_{ij} \leftarrow V$
for each triple $i, j, k \in V$ with $i - j - k$ and $i \not- k$ **do** ▷ *Collider Orientation*
  if $j \notin S_{ik}$ **then**
    $E_{pre} \leftarrow E_{pre} \setminus \{(j, i), (j, k)\}$
Apply the Meek-Rules
for $i, j \in V, i \neq j$ s.t. $(i, j), (j, i) \in E$ **do**
  if $X_0^i \perp\!\!\!\perp X_{[0,1]}^j$ **then**
    $E \leftarrow \setminus \{i \to j\}$
  else
    $E \leftarrow \setminus \{j \to i\}$
for $k \in V$ **do** ▷ *removing loops*
  if $X^k \perp\!\!\!\perp^{\circlearrowleft}_{s,h} \mid X^{\mathrm{pa}_k^{\mathcal{G}} \setminus \{k\}}$ **then**
    $E \leftarrow E \setminus \{k \to k\}$
return $\mathcal{G}$

Algorithm 3 Robust causal discovery for acyclic SDE models.

$V_{pre} \leftarrow \{k \in V\}$
$E_{pre} \leftarrow \{i \to j,\ j \to i \mid i, j \in V,\ i \neq j\}$
for $c = 0, \dots, d-2$ **do** ▷ *Adjacency detection*
  for $i, j \in V, i \neq j$ **do**
    for $K \subseteq V \setminus \{i, j\}, |K| = c,$
      s.t. $(k - j) \in E_{pre}$ for $k \in K$ **do**
      if $X^i \perp\!\!\!\perp X^j \mid X^K$ **then**
        $E_{pre} \leftarrow E_{pre} \setminus \{i \to j, j \to i\}$
        $S_{ij} \leftarrow V$
for each triple $i, j, k \in V$ with $i - j - k$ and $i \not- k$ **do** ▷ *Collider Orientation*
  if $j \notin S_{ik}$ **then**
    $E_{pre} \leftarrow E_{pre} \setminus \{(j, i), (j, k)\}$
Apply the Meek-Rules
$\tilde{V} \leftarrow V$
$\tilde{E}_{post} \leftarrow \{i_0 \to j_0,\ i_1 \to j_1, i_0 \to j_1 \mid (i, j) \in E_{pre}\}$
  $\cup \{i_0 \to i_1 \mid i \in V\}$ ▷ *Lifted Graph*
for $i, j \in V, i \neq j$ s.t. $(i_0, j_1), (j_0, i_1) \in E$ **do**
  for $K \in \mathcal{K}_{i \to j}$ **do**
    if $X^i \perp\!\!\!\perp^{+}_{s,h} X^j \mid X^K$ **then**
      $\tilde{E} \leftarrow \tilde{E} \setminus \{i_0 \to j_0, i_1 \to j_1, i_0 \to j_1\}$
      Break the loop
$\mathcal{G} = (V, E) \leftarrow \mathrm{collapse}(\tilde{V}, \tilde{E})$
for $k \in V$ **do** ▷ *removing loops*
  if $X^k \perp\!\!\!\perp^{\circlearrowleft}_{s,h} \mid X^{\mathrm{pa}_k^{\mathcal{G}} \setminus \{k\}}$ **then**
    $E \leftarrow E \setminus \{k \to k\}$
return $\mathcal{G}$

the initial values $(X_0^i)_{i\in[d]}$ to be jointly independent. The algorithm works similar as to Algorithm 2, except for the fact that for each pair $i, j \in V$ with an unoriented edge, we have to condition on the parents of j when testing for the existence of edge $i \to j$ to prevent information flowing from i to j. As it can happen that j is connected to $k \neq i$ with another undirected edge, the set $\mathcal{K}_{i\to j}$ is required:

$$\mathcal{K}_{i\to j} := \mathrm{pa}_{j,known}^{\mathcal{G}_{post}} \times \mathcal{P}(V_{j,un}),$$

where $\mathrm{pa}_{j,known}^{\mathcal{G}_{post}} := \{k \in V\backslash\{i,j\} : (k_0, j_1) \in \tilde{E}_{post}, (j_0, k_1) \notin \tilde{E}_{post}\}$, $V_{j,un} := \{k \in V\backslash\{i\} : (k_0, j_1) \in \tilde{E}_{post}, (j_0, k_1) \in \tilde{E}_{post}\}$. $\mathcal{K}_{i\to j}$ is required to implement the procedure of conditioning on all known parents of j ($\mathrm{pa}_{j,known}^{\mathcal{G}_{post}}$) and in case there are unspecified other neighbors $V_{j,un}$ of j, loops through the options of either condition on them or don't condition on them. In case j has no other undirected adjacencies, this set only contains one element, the parents of j except i.

Proposition A.12. *Algorithm 3 is sound and complete assuming faithfulness, meaning if we denote the output graph of Algorithm 3 by $\mathcal{G}_{ext} = (V, E_{ext})$, the output edges E_{ext} coincide with the edges of the true DAG $\mathcal{G} = (V, E)$.*

Proof.

"$\supset$": We have to show $(i,j) \in E \implies (i,j) \in E_{ext}$. This direction is clear but will still be stated here. Since the original PC algorithm (see e.g. Spirtes et al. (2000)) is sound, $(i,j) \in \tilde{E}_{post}$ Since $(i,j) \in E$, i_0 and j_1 cannot be separated in the lifted graph $\forall K \subset V\backslash\{i,j\}$. Thus by the faithfulness assumption $X^i \not\perp\!\!\!\perp_{s,h}^{+} X^j \mid X^K \; \forall K \subset V\backslash\{i,j\}$ which holds especially for $K \in \mathcal{K}_{i\to j}$, hence $(i,j) \in E_{ext}$.

"$\subset$": *Proof by contraposition:* $(i,j) \in E_{ext} \implies (i,j) \in E$ is equivalent to $(i,j) \notin E \implies (i,j) \notin E_{ext}$. Since the PC result $\mathcal{G}_{pre} = (V, E_{pre})$ has the correct skeleton and v-structures, the only option would be that wlog $i \neq j$ s.t. $j \to i \in \mathcal{G}$, $(i_0, j_1), (j_0, i_1) \in \tilde{E}_{post}$ meaning the original PC was to detect an adjacency but unable to discard (i,j). But then j_0 is separated from i_1 by the parents $\mathrm{pa}_i^{\mathcal{G}}$ of i, which do not contain j. Therefore, they are contained in $\mathcal{K}_{j\to i}$ as a single element and by the global Markov property, we obtain $X^j \perp\!\!\!\perp_{s,h}^{+} X^i \mid X^{\mathrm{pa}_i^{\mathcal{G}}}$, thus $(i,j) \notin E_{ext}$.

□

A.12 Sensitivity of Constraint-Based Methods to Errors in Conditional Independence Testing

Constraint-based causal discovery methods such as PC or FCI heavily rely on conditional independence tests to infer the causal structure. However, the adaptive nature of these tests, where the results of subsequent tests are contingent upon the outcomes of earlier ones, complicates the error analysis of these approaches. While this is a widely acknowledged challenge faced by all constraint-based methods, how the final graph depends on type I and II errors remains an important open problem in the field. Already early work by Spirtes & Meek (1995) presented a simulation study showing the sensitivity of different methods to the different types of errors, underscoring a variable trade-off between sensitivity and specificity across different sample sizes in skeleton and edge orientation detection depending on the size of the graph and number of samples. Instead, Kalisch & Bühlman (2007) provide theoretical results showing that the PC algorithm is uniformly consistent for high-dimensional settings under a mild sparsity assumption on the DAGs, where the number of nodes can grow quickly with the sample size. However, they critically rely on the restrictive assumption that the observational distribution is a multi-variate Gaussian for their finite sample results, where a strong theory for testing CI with partial correlations is available.

The performance of conditional independence tests is critical for the robustness of constraint-based methods and consistency (as demonstrated for our test in Appendix A.14) is a critical first step. In practice, extensive simulation studies are another useful approach to benchmark the actual performance of the final causal discovery method.

A.13 I.I.D. ASSUMPTION OF THE CONDITIONAL INDEPENDENCE TEST

KCIPT and SDCIT both require i.i.d. samples as input. In our analysis, a solution path of all variables in X is treated as a single example originating from the SDE model. By generating n solution paths as independent realizations of a stochastic process (e.g., independent initial conditions and independent Brownian motions in the SDE), we obtain n i.i.d. samples. It is important to note that while multiple observations from each path at different time points are considered part of a single sample, these within-path observations are not assumed to be i.i.d. However, as each path is an independent realization, the collection of different paths meets the i.i.d. requirement necessary for the effective application of KCIPT and SDCIT in our setting.

A.14 CONSISTENCY OF THE CONDITIONAL INDEPENDENCE TEST

In this section, we argue, that under certain assumptions on our system, Theorem 1 of Doran et al. (2014) still holds even without assuming the existence of a density. At first, it is to be noted that the signature kernel eq. (3) is bounded on compact sets $C \subset \mathrm{BV}(C(I, \mathbb{R}^{d_x}))$ of BV paths, an assumption that can be made as we evaluate our SDEs over finite, discrete time steps. Satisfying this boundedness, it defines an RKHS over these compact sets. In addition, the signature kernel is characteristic on these bounded sets (see Proposition 3.3 Cass et al. (2023)). We use the notation $X : (\Omega, \mathcal{A}, P) \to (\mathcal{X}, \mathcal{A}_\mathcal{X})$, $Y : (\Omega, \mathcal{A}, P) \to (\mathcal{Y}, \mathcal{A}_\mathcal{Y})$, $Z : (\Omega, \mathcal{A}, P) \to (\mathcal{Z}, \mathcal{A}_\mathcal{Z})$ with $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ being compact subsets in $\mathrm{BV}(C(I, \mathbb{R}^{n_x}))$, $BV(C(I, \mathbb{R}^{n_y}))$, $BV(C(I, \mathbb{R}^{n_z}))$ and denote the RKHS's defined by the signature kernel as $(\mathcal{H}_\mathcal{X}, k_\mathcal{X}), (\mathcal{H}_\mathcal{Y}, k_\mathcal{Y}), (\mathcal{H}_\mathcal{Z}, k_\mathcal{Z})$.

Our measure-theoretical argumentation now follows the definitions in Park & Muandet (2020) and the original proof in Doran et al. (2014). Let H be a Banach space, $\mathcal{A}'_H \subset \mathcal{A}_H$ a sub-σ-algebra of the Borel-σ-algebra.

Definition A.13. Let $F : (\Omega, \mathcal{A}, P) \to (H, \mathcal{A}_H)$ a Bochner P-integrable random variable. The *conditional expectation of* H w.r.t. E is a Bochner P-integrable RV $F : (\Omega, \mathcal{A}, P) \to (H, \mathcal{A}'_H)$ s.t.

$$\forall A \in \mathcal{A}'_H \quad \int_A F \mathrm{d}P = \int_A F' \mathrm{d}P \,.$$

Due to the existence and (almost sure) uniqueness, such an F' is denoted $\mathbb{E}[F \mid \mathcal{A}'_H]$ with the notation $\mathbb{E}[F \mid Z] := \mathbb{E}[F \mid \sigma(Z)]$ for a general $Z : (\Omega, \mathcal{A}, P) \to (\mathcal{Z}, \mathcal{A}_\mathcal{Z})$ which in turn can be used to define the *conditional expectation* $P(A \mid \mathcal{A}'_H) := \mathbb{E}[\mathbb{1}_A \mid \mathcal{A}'_H]$ and can be used to define the *conditional kernel mean embedding*

$$\mu_{X|Z} := \mu_{P_{X|Z}} := \mathbb{E}_{X|Z}[k_\mathcal{X}(X, \,\cdot\,) \mid Z]$$

As proven in Theorem 5.8 of Park & Muandet (2020) under the assumption, that $P_{X|Z} := \mathbb{E}[\,\cdot, \mid Z]$ admits a regular version (meaning its can be written as a transition kernel, see Park & Muandet (2020)) and $k_\mathcal{X} \otimes k_\mathcal{Y}$ is characteristic, it holds that

$$X \perp\!\!\!\perp Y \mid Z \;\Leftrightarrow\; \|\mu_{XY|Z} - \mu_{X|Z} \otimes \mu_{Y|Z}\|_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}} = 0 \;\Leftrightarrow\; \|\mu_{XY|Z} \otimes \mu_Z - \mu_{X|Z} \otimes \mu_{Y|Z} \otimes \mu_Z\|_{\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y} \otimes \mathcal{H}_\mathcal{Z}} = 0$$

where the second equivalence is trivial. Let now $\sigma \in \mathcal{S}_n$ be a permutation such that its permutation matrix $M_\sigma = (m_{ij})$

$$m_{ij} = \begin{cases} 1 & \text{if } j = \sigma(i) \,, \\ 0 & \text{if } j \neq \sigma(i) \,. \end{cases}$$

satisfies $\mathrm{Tr}(M_\sigma) = 0$. The empirical estimates for the given sample $\{(x_i, y_i, z_i)\}_{i \in [n]}$ and its permuted version under σ

$$\hat{\mu}_{P_{X,Y,Z}} := \frac{1}{n} \sum_{i=1}^{n} k_\mathcal{X}(x_i, \cdot) \otimes k_\mathcal{Y}(y_i, \cdot) \otimes k_\mathcal{Z}(z_i, \cdot), \tag{16}$$

$$\hat{\mu}_{P'_{X,Y,Z}} := \frac{1}{n} \sum_{i=1}^{n} k_\mathcal{X}(x_i, \cdot) \otimes k_\mathcal{Y}(y_{\sigma(i)}, \cdot) \otimes k_\mathcal{Z}(z_i, \cdot). \tag{17}$$

Similar to Laumann et al. (2023), can now write the distance between test statistic and the approximated test statistic in terms of the sum of the distance between the empirical estimates and its respective counterparts by using the inverse and the regular triangle inequality

$$\begin{aligned}
&\left|\|\mu_{XY|Z}\otimes\mu_Z-\mu_{X|Z}\otimes\mu_{Y|Z}\otimes\mu_Z\|_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Z}}-\|\hat{\mu}_{P_{X,Y,Z}}-\hat{\mu}_{P'_{X,Y,Z}}\|_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Z}}\right| \\
\leq &\left|\|\mu_{XY|Z}\otimes\mu_Z-\mu_{X|Z}\otimes\mu_{Y|Z}\otimes\mu_Z-\hat{\mu}_{P_{X,Y,Z}}+\hat{\mu}_{P'_{X,Y,Z}}\|_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Z}}\right| \\
\leq &\|\mu_{XY|Z}\otimes\mu_Z-\hat{\mu}_{P_{X,Y,Z}}\|_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Z}}+\|\mu_{X|Z}\otimes\mu_{Y|Z}\otimes\mu_Z-\hat{\mu}_{P'_{X,Y,Z}}\|_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Z}}.
\end{aligned}$$

By Park & Muandet (2020), the first term tends to zero for $n\to\infty$. With a similar line of argumentation as in Doran et al. (2014) on can establish a majorant for the estimated HSCIC, namely

$$\begin{aligned}
&\|\hat{\mu}_{P_{X,Y,Z}}-\hat{\mu}_{P'_{X,Y,Z}}\|_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Z}} \\
=&\left\|\frac{1}{n}\sum_{i=1}^n k_\mathcal{X}(x_i,\cdot)\otimes k_\mathcal{Y}(y_i,\cdot)\otimes k_\mathcal{Z}(z_i,\cdot)-\frac{1}{n}\sum_{i=1}^n k_\mathcal{X}(x_i,\cdot)\otimes k_\mathcal{Y}(y_{\sigma(i)},\cdot)\otimes k_\mathcal{Z}(z_i,\cdot)\right\|_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Z}} \\
=&\left\|\frac{1}{n}\sum_{i=1}^n k_\mathcal{X}(x_i,\cdot)\otimes k_\mathcal{Y}(y_i,\cdot)\otimes k_\mathcal{Z}(z_i,\cdot)-\frac{1}{n}\sum_{i=\sigma(j),j=1}^n k_\mathcal{X}(x_{\sigma^{-1}(i)},\cdot)\otimes k_\mathcal{Y}(y_i,\cdot)\otimes k_\mathcal{Z}(z_{\sigma^{-1}(i)},\cdot)\right\|_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Z}} \\
\leq&\frac{1}{n}\sum_{i=1}^n\left\|\left(k_\mathcal{X}(x_i,\cdot)-k_\mathcal{X}(x_{\sigma^{-1}(i)},\cdot)\right)\otimes k_\mathcal{Y}(y_i,\cdot)\otimes\left(k_\mathcal{Z}(z_i,\cdot)-k_\mathcal{Z}(z_{\sigma^{-1}(i)},\cdot)\right)\right\|_{\mathcal{H}_\mathcal{X}\otimes\mathcal{H}_\mathcal{Y}\otimes\mathcal{H}_\mathcal{Z}} \\
\leq&\frac{1}{n}\sum_{i=1}^n\underbrace{\left\|k_\mathcal{X}(x_i,\cdot)-k_\mathcal{X}(x_{\sigma^{-1}(i)},\cdot)\right\|_{\mathcal{H}_\mathcal{X}}}_{\leq 2M_x}\underbrace{\|k_\mathcal{Y}(y_i,\cdot)\|_{\mathcal{H}_\mathcal{Y}}}_{\leq M_y}\left\|\left(k_\mathcal{Z}(z_i,\cdot)-k_\mathcal{Z}(z_{\sigma^{-1}(i)},\cdot)\right)\right\|_{\mathcal{H}_\mathcal{Z}} \\
\leq&\, 2M_xM_y\left(\frac{1}{n}\operatorname{Tr}(PD^{RKHS})\right)
\end{aligned}$$

where the assumptions can be made due to the boundedness of the kernels. Hence, under the assumption that

$$\frac{1}{n}\sum_{i=1}^n\delta_{(x_i,y_{\sigma(i)},z_i)}\to P_{X|Z}\otimes P_{Y|Z}\otimes P_Z \tag{18}$$

a statement that holds true in the case of densities (see supplementary material of Janzing et al. (2013)), since $P_{|Z}$ is regular, $\hat{\mu}_{P'_{X,Y,Z}}\to\mu_{X|Z}\otimes\mu_{Y|Z}\otimes\mu_Z$.

Hence under the assumptions that $P_{\cdot|Z}$ admits a regular version (an assumption also made by Laumann et al. (2023)), the considered random variables operate on compact sets of BV-paths and eq. (18) holds, the test is consistent if and only if $\frac{1}{n}\operatorname{Tr}(PD^{RKHS})\to 0$. This is the exact same statement as in Theorem 1 of Doran et al. (2014).

B Experiments and Evaluation

This section provides details about the experimental framework, elaborating on the data generation process, the parameters and architectures employed, and presents additional results.

B.1 Kernel Heuristic and Interval Choice

In kernel methods, besides selecting the right kernel function, the correct choice of its parameters is vital. A reasonable choice for radially symmetric kernels $k(x_i,x_j)=f(\|x_i-x_j\|/\sigma)$, $f:\mathbb{R}_+\to\mathbb{R}_+$ is to choose σ to lie in the same order of magnitude as the pairwise distances $(\|x_i-x_j\|)_{i,j}$. Choosing the bandwidth for the RBF kernel in our setting with samples $(x_i)_{i\in[n]}$, $x_i\in\mathcal{C}^0(I,\mathbb{R}^d)$ is not straight forward, as the RBF kernel ultimately acts on signatures and there is no immediate way in obtaining a

Table 5: Performance of *future-extended h-locally conditionally independence criterion* in the bivariate setting for the two heuristic σ_1 and σ_2 for different sample sizes. The SHD is measured for different values of s in $\perp\!\!\!\perp^{+}_{s,t}$ with $t = 1$, defining different lengths of intervals in the past and future. Overall the performance is best of shorter values of the past intervals compared to the future ($s < \frac{t}{2}$).

	σ_1		σ_2	
$(\times 10^2)$	$n = 400$	$n = 600$	$n = 400$	$n = 600$
$s = 0.1$	$\mathbf{76.2 \pm 2.5}$	$\mathbf{58.4 \pm 2.7}$	14.9 ± 2.0	15.8 ± 2.0
$s = 0.2$	77.3 ± 2.7	72.2 ± 2.8	$\mathbf{12.9 \pm 1.7}$	$\mathbf{10.9 \pm 1.7}$
$s = 0.3$	80.2 ± 2.3	76.2 ± 2.6	21.8 ± 2.2	17.8 ± 2.0
$s = 0.4$	89.1 ± 2.0	92.0 ± 2.1	31.7 ± 2.4	25.7 ± 2.2
$s = 0.5$	89.1 ± 2.0	98.1 ± 2.0	38.6 ± 2.5	27.7 ± 2.2
$s = 0.6$	82.4 ± 1.9	96.0 ± 2.1	60.0 ± 2.7	48.5 ± 2.9
$s = 0.7$	102.0 ± 1.2	104.0 ± 2.0	80.2 ± 2.2	61.4 ± 2.7
$s = 0.8$	104.0 ± 1.4	95.0 ± 1.5	81.1 ± 2.0	87.1 ± 2.4
$s = 0.9$	95.0 ± 1.6	104.4 ± 1.8	102.0 ± 1.2	101.0 ± 2.1

'typical scale' of pairwise distances of signatures. To still set the bandwidth purely based on observed data, there are two natural ways to implement a median heuristic based on two different collections of pairwise distances. The first option is to choose

$$\sigma_1 = \text{Median}\left\{ \|x_i - x_j\|_{L^2} \,\middle|\, i, j \in [n] \right\},$$

where $\|\cdot\|_{L^2}$ is the is the L^2 norm of entire paths, i.e., integrated over the time interval I. Given that in practice we observe the paths x_i still at fixed time points $t_1 < \ldots < t_m$ within the interval I, we can also consider the heuristic

$$\sigma_2 = \text{Median}\left\{ \|x_{i,t_l} - x_{j,t_k}\|_2 \,\middle|\, i, j \in [n], l, k \in [m] \right\},$$

where $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^d$ and we consider pairwise distances of all observations across all paths and time points. In early benchmarks, we have empirically found σ_2 to yield a better heuristic by comparing the performance of both heuristics by evaluation of the SHD in a causal discovery task in the linear bivariate setting $X^1 \to X^2$ without diffusion interaction with $a_{21} \sim \mathcal{U}\left((-2.5, -1] \cup [1, 2.5)\right)$, $a_{11}, a_{22} \sim \mathcal{U}\left((-0.5, 0.5)\right)$ and $\sigma_i \sim \mathcal{U}\left([0, 0.5)\right)$. The experiment is performed for different choices of intervals $[0, s]$, $[s, s{+}h]$ in the future-extended h-local conditional independence. As can be seen from Table 5, σ_2 and the intervals $[0, 0.1]$, $[0.1, 1]$ seem to perform best on average and are therefore used in the rest of the experiments. In all the experiments of the paper, in the implementation of the signature kernel we use a depth parameter of 4, the RBF kernel and we add time as an extra dimension to the signature kernel.

B.2 Empirical Comparison of Different Tests

Different kernel-based conditional and unconditional tests are compared. HSIC with bootstrap (HSIC_b) and with γ-distribution approximation (HSIC_γ) (Gretton et al., 2007) were implemented for the unconditional setting, while KCIPT (Doran et al., 2014), KCIT with bootstrap (KCIT_b), KCIT with γ approximation (KCIT_γ) (Zhang et al., 2011) and SDCIT for the conditional setting (Lee & Honavar, 2017). Figure 6 shows that SDCIT and HSIC_b are overall the best performing ones in terms of type I and type II error. In this experiment, we sampled 100 different linear SDEs and we sample 400 different trajectories within time interval [0,1]. The number of bootstrap samples over permutation test is set to 100, the number of permutations for a single permutation test to 20000, the number of null samples via Monte Carlo from all values in permutation test to 20000 and 1000 for HSIC, the number of null samples for KCIT-bootstrap to 20000 and the number of null samples for SDCIT is set to 1000. These are the tests parameters also used in the rest of the experiments.

B.3 Additional Experimental Results for the Bivariate Case

Figure 7 shows that in the simple unconditional bivariate setting of the power analysis in Section 4 our test maintains high power up to substantial levels of missingness and only starts degrading once

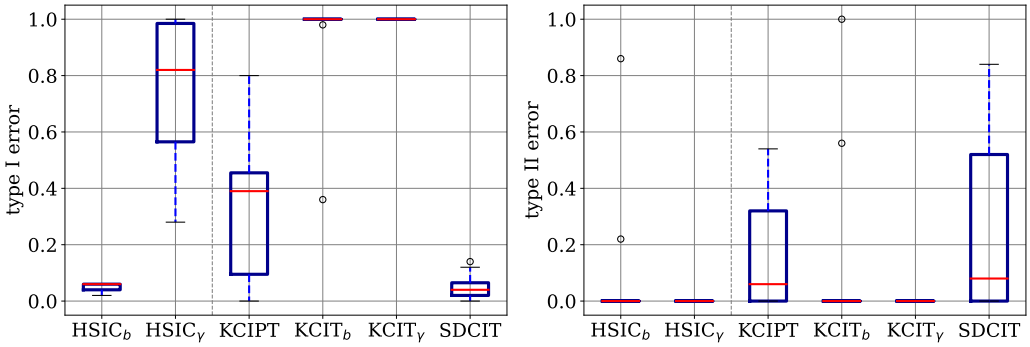


Figure 6: Performance in terms of type I and type II errors of different unconditional (HSIC$_b$ and HSIC$_\gamma$) and conditional tests (KCIPT, KCIT$_b$, KCIT$_\gamma$, SDCIT). Among the conditional tests, SDCIT outperforms KCIPT in terms of type I error, while performig in a similar range in terms of type II error. Among the unconditional tests, HSIC$_b$ outperforms HSIC$_\gamma$

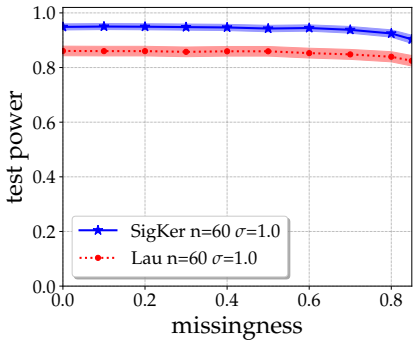


Figure 7: Test power over the missingness level (percentage of observations that is being dropped at random). While both our method as well as Laumann et al. (2023) maintain their power up to high levels of missingness, our test still consistently outperforms the baseline. Lines show means over 1000 settings in the power analysis in Section 4 and shaded regions show standard errors.

more than 80% of observations are dropped at random starting from 64 time steps over the interval $[0, 1]$ for 1000 different SDEs.

B.4 Building Blocks for Causal Discovery

The performance of the conditional and unconditional independence tests was evaluated using widely recognized causal structures that are at the foundation of causal discovery. Specifically, we examined scenarios involving two variables, both connected and unconnected, and three-variable configurations comprising chain, fork, and collider structures. The analysis of type I and type II errors, as shown in Table 6, indicates robust performance for both conditional SDCIT and unconditional HSIC tests. These results refer to the symmetric conditional independence $\perp\!\!\!\perp_{sym}$, while future-extended h-local conditional independence are presented in Table 8. The tables show that the test is able to reconstruct the presence and the directionality of the edges overall all building blocks of causal structure learning. We also perform the same experiments for the setting in which the causal dependence only comes from the diffusion. Referring to the linear SDE in eq. (5), $a_{ij} = a_{ji} = 0$, $a_{11}, a_{22} \sim \mathcal{U}[-0.5, 0.5]$, $\sigma_{12}, \sigma_{21} \sim \mathcal{U}[(-2.5, -1.0) \cup (1.0, 2.5)]$ and $\sigma_{11}, \sigma_{22} \sim \mathcal{U}[-0.5, 0.5]$. b_i and d_i are sampled from $\mathcal{U}[-0.1, 0.1]$ and $\mathcal{U}[-0.2, 0.2]$. Table 7 includes the results for all the different causal structures.

Table 6: Test performance on the key building blocks for causal discovery. Here, $\perp\!\!\!\perp$ refers to $\perp\!\!\!\perp_{\text{sym}}$ except for $X_p^1 \perp\!\!\!\perp X_f^1$ which means $X_{[0,t/2]}^1 \perp\!\!\!\perp X_{[t/2,t]}^1$. We use SDCIT for CI tests and bootstrapped HSIC unconditional independence. We simulate 40 samples of 512 trajectories for 10 different SDEs (400 tests per column) with 25% of the 128 original observations dropped uniformly at random. The **error** represent type I or type II error when the null hypothesis H_0 should be rejected (✗) or accepted (✓), respectively.

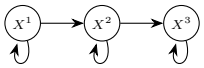
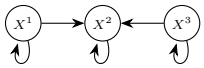
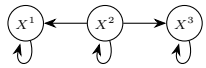
graph							
H_0	$X^1 \perp\!\!\!\perp X^2 \mid X^3$	$X^1 \perp\!\!\!\perp X^3 \mid X^2$	$X^1 \perp\!\!\!\perp X^2 \mid X^3$	$X^1 \perp\!\!\!\perp X^3 \mid X^2$	$X^1 \perp\!\!\!\perp X^3 \mid X^2$	$X_p^1 \perp\!\!\!\perp X_f^1$	$X^1 \perp\!\!\!\perp X^2$	$X^1 \perp\!\!\!\perp X^2$
should reject	✓	✗	✓	✓	✗	✓	✗	✓
error	0.0000	0.0075	0.0000	0.0000	0.9600	0.0000	0.0450	0.0000

Table 7: Test performance on the key building blocks for causal discovery. Here, $\perp\!\!\!\perp$ refers to $\perp\!\!\!\perp_{\text{sym}}$ except for $X_p^1 \perp\!\!\!\perp X_f^1$ which means $X_{[0,t/2]}^1 \perp\!\!\!\perp X_{[t/2,t]}^1$ in a setting where the dependence only comes from the diffusion. We use SDCIT for CI tests and bootstrapped HSIC unconditional independence. We simulate 40 samples of 512 trajectories for 10 different SDEs (400 tests per column) with 25% of the 128 original observations dropped uniformly at random. The **error** represent type I or type II error when the null hypothesis H_0 should be rejected (✗) or accepted (✓), respectively.

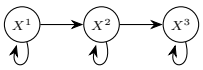
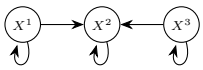
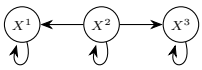
graph							
H_0	$X^1 \perp\!\!\!\perp X^2 \mid X^3$	$X^1 \perp\!\!\!\perp X^3 \mid X^2$	$X^1 \perp\!\!\!\perp X^2 \mid X^3$	$X^1 \perp\!\!\!\perp X^3 \mid X^2$	$X^1 \perp\!\!\!\perp X^3 \mid X^2$	$X_p^1 \perp\!\!\!\perp X_f^1$	$X^1 \perp\!\!\!\perp X^2$	$X^1 \perp\!\!\!\perp X^2$
should reject	✓	✗	✓	✓	✗	✓	✗	✓
error	0.0000	0.0072	0.0000	0.0000	0.9600	0.00	0.07	0.01

Table 8: Test performance on the chain, fork and collider causal structures for the *future-extended h-locally conditionally independence criterion*. Here, $\perp\!\!\!\perp$ refers to $\perp\!\!\!\perp^{+}_{s,h}$ and we use SDCIT. We simulate 10 samples of 512 trajectories for 10 different SDEs (100 tests per column). The **error** represent type I or type II error when the null hypothesis H_0 should be rejected (✗) or accepted (✓), respectively.

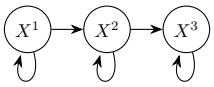
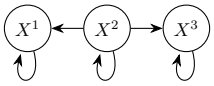
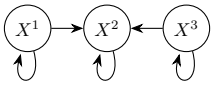
graph						
H_0	$X^1 \perp\!\!\!\perp X^2 \mid X^3$	$X^2 \perp\!\!\!\perp X^1 \mid X^3$	$X^1 \perp\!\!\!\perp X^3 \mid X^2$	$X^3 \perp\!\!\!\perp X^1 \mid X^2$	$X^2 \perp\!\!\!\perp X^3 \mid X^1$	$X^3 \perp\!\!\!\perp X^2 \mid X^1$
should reject	✗	✓	✗	✗	✗	✓
error	0.07	0.33	0.04	0.04	0.03	0.00
graph						
H_0	$X^1 \perp\!\!\!\perp X^2 \mid X^3$	$X^2 \perp\!\!\!\perp X^1 \mid X^3$	$X^1 \perp\!\!\!\perp X^3 \mid X^2$	$X^3 \perp\!\!\!\perp X^1 \mid X^2$	$X^2 \perp\!\!\!\perp X^3 \mid X^1$	$X^3 \perp\!\!\!\perp X^2 \mid X^1$
should reject	✓	✗	✗	✗	✗	✓
error	0.12	0.02	0.04	0.02	0.04	0.06
graph						
H_0	$X^1 \perp\!\!\!\perp X^2 \mid X^3$	$X^2 \perp\!\!\!\perp X^1 \mid X^3$	$X^1 \perp\!\!\!\perp X^3 \mid X^2$	$X^3 \perp\!\!\!\perp X^1 \mid X^2$	$X^2 \perp\!\!\!\perp X^3 \mid X^1$	$X^3 \perp\!\!\!\perp X^2 \mid X^1$
should reject	✗	✓	✓	✓	✓	✗
error	0.08	0.4	0.56	0.46	0.36	0.16

Table 9: SHD ($\times 10^2$) of discovered graphs, for 200 samples and d nodes in the SDE model. The mean and standard error of SCOTCH for different values of λ and n_e for 20 different SDEs is presented. The best performing setting for higher dimensional graphs is $\lambda = 200$ and $n_e = 2k$.

$\times 10^2$	(λ, n_e)	$d = 3$	$d = 5$	$d = 10$	$d = 20$	$d = 50$
SCOTCH	100, 2k	188 ± 28	417 ± 86	250 ± 61	1525 ± 1160	10275 ± 6176
SCOTCH	200, 2k	110 ± 21	270 ± 48	530 ± 223	$\mathbf{370 \pm 174}$	$\mathbf{538 \pm 70}$
SCOTCH	1, 1k	$\mathbf{20 \pm 11}$	$\mathbf{85 \pm 20}$	$\mathbf{375 \pm 28}$	1455 ± 103	6095 ± 153
SCOTCH	50, 1k	330 ± 27	1255 ± 54	5895 ± 107	1340 ± 243	2994 ± 941
SCOTCH	200, 1k	400 ± 17	1370 ± 29	6425 ± 80	705 ± 77	7863 ± 2670

B.5 Causal Discovery

We tested SCOTCH using various sparsity parameters and epochs to identify the optimal configuration. Table 9 confirms that the configuration with $\lambda = 200$ and $n_e = 2000$ outperforms others in the tested settings, and hence is chosen for the comparison with our method in Table 2. The table also reveals that SCOTCH's performance is not robust to changes in λ, showing substantial variations under the same number of epochs, optimizer, and learning rate with different λ values. These results demonstrate that the SCOTCH's performance can significantly deteriorate with changes in the parameter λ, while keeping the number of epochs constant. For example, when λ is adjusted from 200 to 100, the Structural Hamming Distance (SHD, scaled by 10^2) increases from an average of 538 to 10,275 for $n_e = 2000$.

B.6 Conditional Independence Testing Beyond the SDE Model

In this section, we aim to demonstrate that our proposed non-parametric conditional independence (CI) test extends beyond the setting of SDEs and even beyond semi-martingales, rendering it more broadly applicable than most existing methods, which usually assume independent (additive) noise.

Fractional Brownian Motion. To illustrate this, we start by applying our test to a stochastic process driven by fractional Brownian motions (fBMs). Fractional Brownian motion is a generalization of standard Brownian motion, a continuous time Gaussian process $B_H(t)$ with zero mean and $\mathbb{E}[B_H(t)B_H(s)] = \frac{1}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right)$ that incorporates a parameter $H \in (0, 1)$, the so called *Hurst parameter*, which governs long-range dependencies. For

- $H = 0.5$ it reduces to the regular Brownian Motion.
- $H > 0.5$ increments are positively correlated.
- $H < 0.5$ increments are negatively correlated.

Specifically, we consider the following system of stochastic differential equations driven by fractional Brownian motions

$$\begin{pmatrix} dX_t \\ dY_t \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ a_{21} & 0 \end{pmatrix} \begin{pmatrix} X_t \\ Y_t \end{pmatrix} dt + \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} \begin{pmatrix} dW_t^1 \\ dW_t^2 \end{pmatrix}, \tag{19}$$

with W_t^1, W_t^2 independent fBMs with Hurst parameter H and $a_{21}, d_1, d_2 \sim \mathcal{U}([-2, 2])$. To quantify the effectiveness of our test in this context, we measure the test power for different strengths of interaction for different Hurst parameters. Figure 8 shows this power analysis for 500 random settings of eq. (19). The power of our test steadily approaches 1 as the interaction strength increases to 1 regardless of H. We note that since the above example does not have independent increments, Algorithm 1 is not applicable anymore for causal discovery in such settings. Instead, the fractional Brownian motion example merely highlights the usefulness of our CI test independently from its application in causal discovery in the SDE model.

Functional data. For further corroboration of our claims, we turn to the following example from functional data analysis (FDA), taken from Laumann et al. (2023), to test causal discovery in a non SDE setting. After drawing a DAG $\mathcal{G} = (V, E)$, processes for source vertices, i.e., vertices $i = v_i$

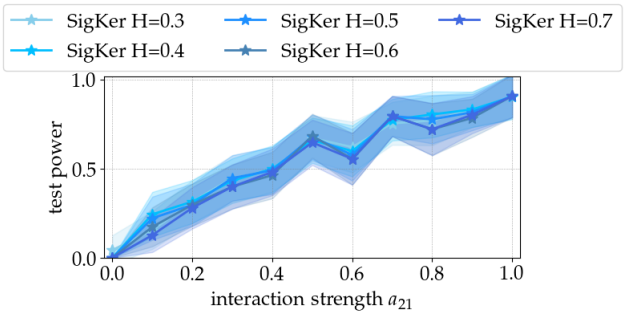


Figure 8: Test power over for different fBM-driven processes eq. (19) for different H. The CI test picks up dependence for as little as 100 samples given a certain amount of interaction-strength. Lines show means over 500 settings and shaded regions show standard errors.

Table 10: SHD ($\times 10^2$) comparison of SigKer to the baselines PCMCI and SCOTCH (with $\lambda = 200$ and $n_e = 2000$) in the functional data setting. We report mean and standard errors over 40 instances.

$\times 10^2$	$d = 3$	$d = 5$	$d = 10$	$d = 20$	$d = 50$
$\perp\!\!\!\perp_{\text{sym}}$	40 ± 8	22 ± 10	509 ± 25	1542 ± 60	6786 ± 223
$\perp\!\!\!\perp_{\text{sym}}$+p.p.	42 ± 8	141 ± 13	489 ± 28	1498 ± 56	6529 ± 250
PCMCI	110 ± 30	320 ± 42	1150 ± 121	6518 ± 293	7799 ± 250
SCOTCH	558 ± 14	1861 ± 43	8567 ± 63	35467 ± 833	105165 ± 10934

without parents in $\mathcal{G}$, are generated according to

$$X^i(t) = \sum_{m=1}^{M} c_m^i \phi_m(t) + \epsilon_t^i$$

with Fourier basis functions $\phi_1(t) = 1$, $\phi_2(t) = \sqrt{2}\sin(2\pi t)$, $\phi_3(t) = \sqrt{2}\cos(2\pi t)$ and so on, the weights $c_m^i \sim \mathcal{N}(0,1)$, and the additive noise $\epsilon_t^i \sim \mathcal{N}(0,1)$. For all vertices i with at least one parent, we set

$$X^i(t) = a \sum_{k \in \text{pa}_i^{\mathcal{G}}} \int_0^t X^k(s)\beta^k(s,t)ds + \epsilon_t^i$$

with $\beta^k(s,t) = 8(s - c_1^k)^2 - 8(t - c_2^k))^2$, $c_1^k, c_2^k \sim \mathcal{U}_{[0,1]}$ and $a = 1$.

We draw 40 DAGs with 40 distinct FD generating mechanisms for each dimension $d \in \{3, 5, 10, 20, 50\}$. Table 10 shows that SCOTCH cannot capture the underlying dependencies, as it is tailored specifically to the SDE model. It is thus misspecified for such a function generating mechanism. Under such (arguably mild) misspecification, SCOTCH is even outperformed by PCMCI, whereas our method still performs much better than both baselines.

B.7 Real-World Pairs Trading Example

In the pairs trading experiment, stock price data is downloaded from Yahoo Finance for a predefined list of stocks over a specific period, divided into training (1st January 2010 to 31st December 2011) and trading intervals (1st January 2012 to 31st December 2012). The chose stocks are Trinity Industries (TRN), Brandywine Realty Trust (BDN), Commercial Metals Company (CMC), The New York Times Company (NYT), New York Community Bancorp (NYCB), The Wendy's Company (WEN), CNX Resources Corporation (CNX). Logarithms of the stock prices are computed to stabilize variance and normalize the prices. During the training period, pairs of stocks are selected based on various statistical tests: Cointegration Test (Engle-Granger) to determine if a pair of stocks is cointegrated, suggesting a long-term equilibrium relationship; Augmented Dickey-Fuller (ADF) Test to assess the stationarity of the spread (ratio) of a pair's prices; and Granger Causality Test to check if

the price of one stock in the pair can predict the price of the other. Pairs are selected based on the p-values from these tests, with a threshold of 0.05 determining significance. In the trading phase, a rolling window is used to continuously recalculate the mean and standard deviation of the spread between the selected pairs. Trades are initiated based on the z-score of the spread, where a high z-score indicates a short position and a low z-score indicates a long position. Positions are managed and closed when the spread returns to a predefined z-score threshold. The strategy's performance is evaluated based on total return, annual percentage rate (APR), Sharpe ratio, maximum drawdown (maxDD), and maximum drawdown duration (maxDDD).

B.8 Evaluation Metrics

Following common conventions in the causal discovery literature, we evaluate our algorithms using the following metrics.

Structural Hamming Distance (SHD). Given two directed graphs $\mathcal{G}_1 = (V, E_1)$, $\mathcal{G}_2 = (V, E_2)$ over the same nodes with different sets of edges and let $A_1, A_2 \in \{0,1\}^{n\times n}$ be their adjacency matrices. Then we define the structural hamming distance (SHD) as

$$\text{SHD} = \sum_{i,j\in[n]} |(A_1)_{ij} - (A_2)_{ij}|$$

Normalized SHD. In order to compare the recovery performance of our algorithms for different types of graph-sizes, the normalized SHD is defined as

$$\text{SHD}_{\text{norm}} = \frac{\text{SHD}}{d(d-1)}$$

B.9 Computational Complexity

Conditional Independence Test. For KCIPT, we denote by B the number of outer bootstraps, by b the number of inner bootstraps, by n the number of i.i.d. samples, and by n_{MC} the number of Monte-Carlo samples. We spell out the computational complexity of all steps in performing a single conditional independence test.

- computing B permutations of the n-samples: $\mathcal{O}(n)$
- the loop over the B outer bootstrap contains in each iteration:
  - permuting matrices: $\approx \mathcal{O}(n^{2.4})$ (assuming fast matrix multiplication algorithms, otherwise $\mathcal{O}(n^3)$)
  - splitting matrices using indexing: $\mathcal{O}(n^2)$
  - computing RKHS distances: $\mathcal{O}(n^2)$
  - solving the linear program to find the best permutation (this depends on the solver, but for a primal-dual interior point methods it is $\mathcal{O}(\sqrt{n}\log(1/\epsilon)(mn+m^2))$, where m is the number of constraints, n the number of variables, and ϵ the accuracy tolerance for the solution. Overall, in our case with $m = n(n-1)$ constraints, we obtain: $\mathcal{O}(\sqrt{n}\log(1/\epsilon)(n^2(n-1)+n^2\cdot(n-1)^2))$ which becomes $\mathcal{O}(n^{4.5}\log(1/\epsilon))$
  - permuting matrices: $\mathcal{O}(n^{2.4})$
  - computing the original statistic: $\mathcal{O}(n)$
  - loop over b inner permutations
    - permuting matrices: $\mathcal{O}(n^{2.4})$
    - computing the permuted statistic: $\mathcal{O}(n^2)$
  - computing the p-value: $\mathcal{O}(b)$
- computing the original statistic: $\mathcal{O}(B)$
- for n_{MC} Monte Carlo samples computing the null sample: $\mathcal{O}(B\cdot b)$
- computing the final p-value: $\mathcal{O}(n_{MC})$

Causal discovery. The computational complexity of our causal discovery algorithm is in the worst case $\mathcal{O}(d^{d_{\max}}\cdot d^2)$, where d is the number of nodes and $d_{\max}$ is the maximal degree in the graph, meaning the maximal number of adjacencies of any node. For sparse graphs (e.g., graphs with a low maximal degree), the complexity can reduce to being polynomial in the number of nodes.

Signature kernel. The computational complexity of evaluating the signature kernel scales quadratically in the number of time points within a process using the `sigkerax` Python package with an RBF lifting kernel and again quadratically in the number of samples when computing the full Gram matrix. The is another package, `iisignature`, that allows for linear scaling in the number of time points when using a linear lifting kernel. Both implementation support highly parallelized execution on GPU accelerators.

C BASELINES

In Section 4, we benchmark our method against other baselines, including Granger causality, CCM, PCMCI, and Laumann (Laumann et al., 2023). For the latter, we used their implementation provided in the `causal-fda` package. For the Granger-implementation for two variables ($d = 2$), we used Seabold & Perktold (2010), for CCM we used Javier (2021), and for PCMCI we used the `tigramite` package (Runge et al., 2019). In PCMCI, tests for edges are conducted by applying distance correlation-based independence tests (Székely et al., 2007) between the variables' residuals after regressing out other nodes using Gaussian processes. For SCOTCH implementation (Wang et al., 2024), we use the package `causica`. For SCOTCH, we always use a learning rate of 0.001 and keep the same default parameters for the learning algorithm.

D DETAILS ON THE PARTIALLY OBSERVED SETTING

Since in the partially observed setting there could in principle be infinitely many unobserved variables, special tools are required. Usually, the graphical framework of *maximal ancestral graphs* (MAGs) (loosely speaking DAGs with also bidirected edges) is used, which encodes ancestral relations in the DAG $\mathcal{G}$ and aims at graphically representing conditional independencies implied by $\mathcal{G}$ involving only the (marginal of the) observed variables. The unique MAG $M^{\mathcal{G}}$ corresponding to the partially observed DAG $\mathcal{G}$ can be constructed via Zhang (2008) the following rules.

- For $v_1, v_2 \in V_{obs}$, $v_1 - v_2$ in $M^{\mathcal{G}}$ if and only if there exists an inducing path relative to V_L between v_1, v_2 in $\mathcal{G}$.
- For each pair of adjacent vertices $v_1 - v_2$ in $M^{\mathcal{G}}$ we orient the edge between them as follows:
  - $v_1 \to v_2$ if $v_1 \in \mathrm{an}^{\mathcal{G}}_{v_2}$ and $v_2 \notin \mathrm{an}^{\mathcal{G}}_{v_1}$
  - $v_1 \leftarrow v_2$ if $v_2 \in \mathrm{an}^{\mathcal{G}}_{v_1}$ and $v_1 \notin \mathrm{an}^{\mathcal{G}}_{v_2}$
  - $v_1 \leftrightarrow v_2$ if $v_2 \notin \mathrm{an}^{\mathcal{G}}_{v_1}$ and $v_1 \notin \mathrm{an}^{\mathcal{G}}_{v_2}$

An *inducing path* is a path with all colliders on the path being in $\mathrm{an}^{\mathcal{G}}_{\{v_1,v_2\}} \cap V_{obs}$ and all other nodes on the path in V_L. The obtained MAG $M^{\mathcal{G}}$ therefore preserves causal ancestral relations with respect to $\mathcal{G}$. The *Fast Causal Inference* (FCI) algorithm (Zhang, 2008) run on the observed marginal outputs an equivalence class of MAGs, known as a *Partial Ancestral Graph* (PAG), which captures the same adjacencies but leaves some endpoints unoriented as there are multiple MAGs with the same conditional independence relations. For example, in Figure 4 FCI cannot rule out an additional latent confounder between A and C such that the induced MAG has the same m-separation properties. Unlike FCI, by leveraging the direction of time, we can recover the full underlying MAG $M^{\mathcal{G}}$, which is substantially more informative than FCI's PAG. First, we establish adjacencies analogous to the first part of Algorithm 2 (or simply by running FCI with our symmetric criterion) and then follow the steps in the construction of the MAG from a DAG to orient these adjacent edges $v_i - v_j$, $v_i \neq v_j \in V_{obs}$:

- $v_i \to v_j$ if $X_0^i \not\perp\!\!\!\perp_{\mathrm{sym}} X^j_{[0,T]}$ and $X_0^j \perp\!\!\!\perp_{\mathrm{sym}} X^i_{[0,T]}$
- $v_i \leftarrow v_j$ if $X_0^i \perp\!\!\!\perp_{\mathrm{sym}} X^j_{[0,T]}$ and $X_0^j \not\perp\!\!\!\perp_{\mathrm{sym}} X^i_{[0,T]}$
- $v_i \leftrightarrow v_j$ if $X_0^i \not\perp\!\!\!\perp_{\mathrm{sym}} X^j_{[0,T]}$ and $X_0^j \not\perp\!\!\!\perp_{\mathrm{sym}} X^i_{[0,T]}$

We note that these criteria precisely encode the (non-)ancestral relationships in the MAG construction rules. Uniqueness of the obtained result follows from Richardson & Spirtes (2002, Corollary 3.10) as two ancestral graphs $\mathcal{G}_1, \mathcal{G}_2$ are equal, if they share the same adjacencies and ancestral relations.

We have thus constructed a sound and complete algorithm (assuming access to a CI oracle) to recover the unique MAG $M^{\mathcal{G}}$ for the ground truth DAG $\mathcal{G}$ in the partially observed SDE

model. To illustrate our causal discovery algorithm in practice, we conduct experiments on an SDE model with a graph $\mathcal{G}$ as depicted in Figure 4 where $V_{obs} = \{A, B, C, D\}$, $V_L = \{U\}$, $E = \{(A, C), (B, C), (U, C), (U, D)\}$. While FCI run on data coming from a Markov model with this DAG is able to detect the adjacency structure, it is unable to decide whether A is truly an ancestor of C. It will correctly infer the arrow-head into C, but cannot rule out an arrow-head at A, see Figure 4. Neural network approaches like SCOTCH that directly infer the functional relationships by model fitting are not able to deal with such partially observed settings and are typically expected to predict edges that do not exists in the ground truth graph, since they can not model potentially infinitely many unobserved processes. To demonstrate this concretely, we ran SCOTCH as well as our FCI-inspired algorithm on 100 SDEs with the adjacency from our example graph $\mathcal{G}$ and compare how often A and D are (falsely) adjacent in the output. SCOTCH predicted an edge between A and D in 88% and ours algorithm in only 8% of all settings, see also Figure 4.