INSTRUMENTAL VARIABLES IN CONTINUOUS-TIME STOCHASTIC PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal questions often arise in systems that evolve in continuous time. While instrumental variable (IV) methods provide a principled approach to identify causal effects in static or discrete-time settings, even in the presence of unobserved confounding, they do not yet extend to continuous-time dynamics. In this work, we develop an IV framework for stationary continuous-time stochastic processes parameterized by stochastic differential equations (SDE), where the instrument itself takes the form of a stochastic process. We encode the structural sparsity and causal effects directly at the infinitesimal level of the drift, and contrast these assumptions with the IV assumptions on the discretized version of the same process. Based on these structural assumptions, we derive moment restrictions using path-valued instruments. We show how these restrictions simplify for Ornstein-Uhlenbeck processes and present an identification condition together with a corresponding consistent estimation method. Finally, we extend our estimation framework to nonlinear causal effects by expressing the path-based moment restrictions leveraging the signature kernel. Empirically, our method recovers causal effects in confounded continuous-time systems where maximum-likelihood-based estimation and discrete-time IV methods are systematically biased.

1 INTRODUCTION

Causal effect estimation under unobserved confounding is a central challenge in many scientific domains. In the presence of unobserved confounders, standard estimation techniques fail to distinguish between confounding and the true effect. This distinction is generally impossible from observational data without additional assumptions. The Instrumental Variable (IV) framework forms one of the most prominent settings in which principled strategies for cause-effect identification and estimation exist under certain assumptions. IVs are exogenous variables that affect the covariates, are independent of the hidden confounders, and only affect the outcome through the covariates. IV methods have been thoroughly studied and are an established workhorse in the toolbox of econometricians (Sargan, 1958; Hansen, 1982; Staiger & Stock, 1997) and beyond. They also attracted renewed interest in the context of flexible machine learning techniques (Bennett et al., 2019; Hartford et al., 2017; Ailer et al., 2024).

While a range of methods exists for the static IV setting, many real-world systems in which questions of causation arise are inherently dynamic, including applications in earth system sciences (Runge et al., 2019; Stips et al., 2016), medical settings (Hernan et al., 2001; Lok, 2008) and macroeconomic analyses (Fernández-Villaverde & Rubio-Ramírez, 2007; Sims, 2002). As an example, consider the objective of identifying the glucose-insulin sensitivity of a type 2 diabetes patient, which describes the main mechanism impaired by the disease. Insulin and glucose concentration dynamics may be confounded by unobserved digestive processes, rendering standard effect estimation techniques biased. Using the IV structure of the problem, insulin sensitivity may still be identified using the external insulin administration process as an instrument. We will get back to this example in our empirical validation. Recent work on IV methods for parameter estimation of discrete-time processes Mogensen (2023); Thams et al. (2024) is unsuitable for continuous-time systems. First, observational data is often available only on irregular sampling-intervals, which discrete-time methods fail to handle. Moreover, if the underlying real world system evolves in continuous-time, the structural IV sparsity needs to be modeled on an *infinitesimal scale* rather than over finite time intervals. That is, discrete-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

time models imply structural IV assumptions that do not preserve the actual local IV structure of a continuous-time process, resulting in severe model-misspecification.

Furthermore, our continuous-time formulation has strictly more "identification power" than in the static case. Concretely, in static IV settings identification is fundamentally limited by the variation and number of available instruments. For example, in the linear case, identification requires as many instruments as treatments. The relevant underspecified case, where only few instruments are available compared to treatments, has been acknowledged as an important setting (Ailer et al., 2023), but received little attention. In contrast, in our new continuous-time setting, we show that the temporal structure of the process contains richer identifying information, allowing for identification of multi-dimensional causal effects from a single instrument process.

We summarize our **contributions** as follows:

- We are first to formulate the IV setting in continuous-time stochastic processes parametrized by SDEs, where the IV structure is defined on the infinitesimal level of the process.
- We derive the moment conditions implied by the IV structure in stationary SDE systems as orthogonality between forecasting errors and functionals of the instrumental process path.
- We derive a controllability condition that *identifies* the target parameters in a confounded Ornstein–Uhlenbeck (OU) process, along with a *consistent* estimator.
- We propose an extension for nonlinear objective functions via a kernel-based estimation scheme.
- We demonstrate empirically that our method outperforms possible alternative approaches.

2 Static and Discrete-Time Instrumental Variables

Static IV. In the static setting, the standard IV setup aims to identify the effect of covariates X on outcomes Y under confounding of an unobserved variable U by making use of the instrument variable Z. In general, IV is defined by three assumptions on the underlying data-generating mechanism, *relevance, exclusion and exogeneity* (Hernán & Robins, 2020; Pearl, 2000). Relevance requires that X is not independent of Z ($X \not\perp\!\!\!\perp Z$), the exclusion restriction requires that Z only affects Y through X ($Z \perp\!\!\!\perp Y \mid X, U$) and exogeneity requires that U is independent of Z ($U \perp\!\!\!\perp Z$). The standard "IV graph" encoding these conditions is illustrated in Fig. 4.

Instrumental variable methods and generalized-method-of-moments (GMM) frameworks use the exogeneity of instrument variables together with setting-specific assumptions to formulate effect-identifying moment conditions, and construct associated estimators (Sargan, 1958; Hansen, 1982; Wright, 1928). Albeit the IV setting being restrictive as valid instruments are often hard to find, these methods continue to uncover relevant causal effects, such as the medical cost of air pollution, effects of medical aid programs, among other (Duflo, 2001; Deryugina et al., 2019; Bhuller et al., 2020; Baicker et al., 2013). In recent years, deep learning methods have been applied to both IV- and GMM-setups (Bennett et al., 2019; Saengkyongam et al., 2022; Hartford et al., 2017) and in a different line of work, kernel-based methods are used to approximate causal effects (Ailer et al., 2024; Bennett et al., 2023; Zhang et al., 2023)

IV methods for discrete processes. Recent papers have described instrumental variable techniques in stationary multivariate vector autoregressive processes in discrete time (Mogensen, 2023; Thams et al., 2024). Mogensen (2023) proposes using a time-integrated covariance matrix to discover *normalized parameters*, which can be given a causal interpretation. This also extends to linear Hawkes processes, which describe continuous-time event processes (Mogensen, 2023). Thams et al. (2024) investigate discrete-time instrumental processes by studying graphical properties of the time-unrolled causal graph of the process. They clarify that the previously described static IV conditions are not trivially fulfilled in such a graph, as past instruments violate exclusion by causally affecting both current instrument states and outcomes. They propose two approaches for estimating causal effects in this setting: *conditional IV*, which retrieves the standard IV setting by conditioning on subsets of the process history, and *nuisance IV*, which adds past outcomes to the covariates.

Causal interpretation of SDEs. A growing amount of papers seeks causal interpretations of SDEs, e.g., Hansen & Sokol (2014); Manten et al. (2025); Lorch et al. (2024); Schneider et al. (2026), by defining interventions in SDE systems and performing causal discovery of the local independence graph in SDE systems. We extend this line of work by contributing IV-type estimation for continuous-time processes defined by SDEs.

Instrumental variables in system identification. Adjacent to our setting, IV methods can also be found in the systems engineering and control theory literature, mostly in the context of discrete-time stochastic processes, starting from early papers such as Joseph et al. (1961); Wong & Polak (1967). This line of work often assumes an explicit *control input* which may serve as an instrumental process, alternatively lagged variables are used as instruments (Friedlander et al., 1985; Gilson & Van den Hof, 2005). The goal is most often the identification of a state-space model, instead of the identification of a set of causal parameters.

3 PRELIMINARIES

Setup. We first introduce a general, continuous-time, stochastic data-generating mechanism in which we can encode the structural IV assumptions. Specifically, the data is generated by a stochastic process with states $\mathbf{S}_t = (Y_t, X_t, Z_t, U_t) \in \mathbb{R}^{d_y+d_x+d_z+d_u}$ indexed by time t and driven by a Wiener process $\mathbf{W}_t = (W_t^y, W_t^x, W_t^z, W_t^u)$ with jointly independent components. Process Z acts as an instrument process, X represents the evolving covariates, Y the outcome process, and U the unobserved confounding process. All causal relationships between the variables are encoded in the structure of the drift, which allows for a direct interpretation of the causal effects, in contrast to diffusion coupling, and avoids additional identifiability challenges. Thus, we model the dynamics via the following system of SDEs

$$d\mathbf{S}_t = \underbrace{\begin{pmatrix} f_y(Y_t, X_t) + U_t \\ f_x(Y_t, X_t, Z_t, U_t) \\ f_z(Z_t) \\ f_u(U_t) \end{pmatrix}}_{\mathbf{f}(\mathbf{S}_t)} dt + \underbrace{\begin{pmatrix} \sigma_y(Y_t) & 0 & 0 & 0 \\ 0 & \sigma_x(X_t) & 0 & 0 \\ 0 & 0 & \sigma_z(Z_t) & 0 \\ 0 & 0 & 0 & \sigma_u(U_t) \end{pmatrix}}_{\boldsymbol{\sigma}(\mathbf{S}_t)} \begin{pmatrix} dW_t^y \\ dW_t^x \\ dW_t^z \\ dW_t^u \end{pmatrix} \tag{1}$$

with $\mathbf{S}_0 \sim \pi$, where π is the unique invariant measure of the process. We assume drift $\mathbf{f}$ and diffusion $\boldsymbol{\sigma}$ to be globally Lipschitz and satisfy linear growth conditions, ensuring existence and pathwise uniqueness of a strong solution to Eq. (1) (Pavliotis, 2014), as well as finite first moments $\mathbb{E}_\pi[\|\mathbf{S}\|] < \infty$. The resulting joint process is assumed to be strongly stationary. We assume without loss of generality that the process is zero-mean. All components of $\mathbf{S}_t$ aside from U_t are observed. In Section A, we discuss the assumptions and alternative confounding types based on non-diagonal diffusion matrices, for which our proposed methods remain applicable.

The structural assumptions implied by Eq. (1) directly parallel the static IV assumptions and are illustrated in the sparsity graph of the process, shown in Figure 1. First, we assume *relevance*, meaning that the drift of process X depends non-trivially on process Z. Second, the *exclusion restriction* requires that the instrument process only influences the outcome process Y through the covariate process X, thus there is no direct effect of process Z on Y (Z is not an argument of f_y). Third, *exogeneity* in continuous-time implies that the instrument process Z_t is independent of the Brownian motion increments dW_t^y as well as of the confounding process U_t. The latter is described by the following conditional moment restriction

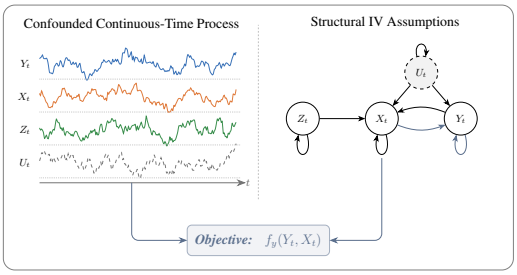


Figure 1: Observed trajectories of processes (Y, X, Z), confounded by the hidden process U, together with the assumed sparse drift graph, are used to identify the objective function $f_y(Y_t, X_t)$.

$$\mathbb{E}\left[U_t \mid \mathcal{F}_s^z\right] = 0, \quad \text{for } s \leq t, \tag{2}$$

where $\mathcal{F}_s^z := \sigma(Z_{[0,s]})$ denotes the filtration of the instrument process up to time s. Analogously to most static IV settings, we assume that U_t enters the drift of the outcome process Y *additively* (Hernán & Robins, 2020). While restrictive, without additivity one typically needs similarly strong alternative assumptions to achieve identifiability (Balke & Pearl, 1997; Imbens & Newey, 2009).

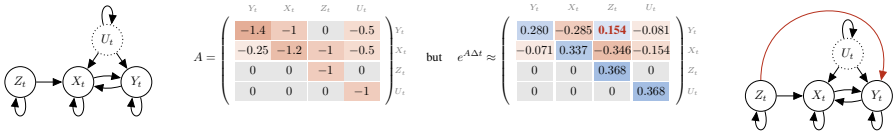
Data. We observe a finite trajectory of the marginal process and record tuples $(Y_{t_i}, X_{t_i}, Z_{t_i}, t_i)$ at n observation times $0 \leq t_1 < t_2 < \ldots < t_n \leq T_{max}$. The states are all measured simultaneously,

however, the spacing between timepoints may be irregular. The confounding process U is latent and never observed. In principle, a single long trajectory suffices for identification and estimation. For practical purposes, in our empirical evaluation we sample multiple shorter trajectories for efficiency.

Objective. The main objective of this paper is the identification of the drift component $f_y(Y_t, X_t)$ describing the causal effect of covariate process X on outcome process Y and the self-dynamics of Y. Although the causal effect of process X on process Y is the primary effect of interest, it is inherently entangled with the self-dynamics of Y and both are therefore jointly identified and estimated.

Bias in MLE. In the given setting, maximum likelihood estimation of f_y is biased due to unobserved confounding. This becomes evident by analyzing the transition densities of the partially observed process. Naive Maximum-Likelihood Estimation (MLE) maximizes $\mathbb{E}\left[\log p_\theta(Y_{t+\Delta t} \mid X_t, Y_t)\right]$ under some approximation of the transition density (Pavliotis, 2014). This transition density depends on the unobserved U_t via $p(Y_{t+\Delta t} \mid X_t, Y_t) = \int p(Y_{t+\Delta t} \mid X_t, Y_t, U_t)\, p(U_t \mid X_t, Y_t)\, dU_t$. Since U_t confounds both X_t and Y_t, a likelihood model omitting U_t generally attributes part of the confounding mechanism to the causal effect f_y. This motivates the use of IV methods that exploit the structure to identify the drift function.

Relation to other IV settings. Unlike in the static case, the outcome Y depends on its own past and may interact cyclically with process X leading to substantially different structure of the objective function f_y. Compared to the discrete-time dynamic setting, the local independencies implied by the continuous-time setup in Eq. (1) differ fundamentally from structural assumptions in discrete-time processes described in Mogensen (2023); Thams et al. (2024). In continuous time, the structural sparsity exists only on the infinitesimal level and does not imply an identical sparsity pattern in finite-step transition maps of the corresponding discretely observed process (Mogensen et al., 2018). For example, an equidistantly sampled OU process admits an equivalent representation as a $\mathrm{VAR}(1)$ process. However, when imposing an IV structure on the drift of the OU process, the exclusion restriction is generally not preserved in the corresponding $\mathrm{VAR}(1)$ transition matrix. The key is that sparsity in an infinitesimal generator matrix A does not transfer to the exponential map $e^{A\Delta t}$ for a finite interval (see Section B for detail), e.g., for $\Delta t = 1$:



Hence, when causal mechanisms operate in continuous-time with IV sparsity imposed on the infinitesimal level, *discrete-time IV models are misspecified by construction*.

4 Continuous-Time IV Moment Conditions

We now derive the moment conditions implied by the IV structure using the independence between instrument and confounding processes described in Eq. (2).

Proposition 1 (Continuous-time IV Moment Restriction)**.** *Let $\mathbf{S}_t = (Y_t, X_t, Z_t, U_t)$ be the joint Markov process satisfying Eq.* (1)*, and assume that the IV exogeneity condition defined in Eq.* (2) *holds. Let $R_{t,t+\tau}$ describe the prediction residual at time $t+\tau$*

$$R_{t,t+\tau} = Y_{t+\tau} - Y_t - \int_t^{t+\tau} f_y(Y_s, X_s) ds. \tag{3}$$

Then the following conditional moment restrictions hold

$$\frac{d}{d\tau}\,\mathbb{E}\big[\, Y_{t+\tau} \,\big|\, \mathcal{F}_t^z \,\big] \;=\; \mathbb{E}\big[\, f_y(Y_{t+\tau}, X_{t+\tau}) \,\big|\, \mathcal{F}_t^z \,\big]\,, \tag{4}$$

and, equivalently, for every $t \geq 0$ and $\tau > 0$,

$$\mathbb{E}\big[\, R_{t,t+\tau} \,\big|\, \mathcal{F}_t^z \,\big] \;=\; 0. \tag{5}$$

Proof. We start by following Hansen & Scheinkman (1993) and apply Itô's formula to the test functions $\phi \in C^2(\mathbb{R}^{d_y+d_x+d_z+d_u}; \mathbb{R})$ evaluated on the process state $\mathbf{S}_t$, yielding

$$M_{t+\tau}^{\phi} - M_t^{\phi} \;:=\; \phi(\mathbf{S}_{t+\tau}) - \phi(\mathbf{S}_t) - \int_t^{t+\tau} \mathcal{L}\phi(\mathbf{S}_s)\, ds. \tag{6}$$

Here, the resulting process M_t^ϕ is a continuous local martingale. We choose the linear test function $\phi_j(\mathbf{S}_t) = Y_t$ such that M_t^ϕ becomes a true martingale under the finite first moment and linear growth assumptions. $\mathcal{L}$ denotes the generator of the full process including the unobserved state, as defined in Pavliotis (2014). We take expectations conditioned on the filtration of the instrument process at time t. As $\mathcal{F}_t^z \subset \mathcal{F}_t^{\mathbf{S}}$, the conditional expectation of any future martingale increment remains zero and thus

$$\mathbb{E}[\phi_j(\mathbf{S}_{t+\tau}) \mid \mathcal{F}_t^z] - \mathbb{E}[\phi_j(\mathbf{S}_t) \mid \mathcal{F}_t^z] \;=\; \int_t^{t+\tau} \mathbb{E}[\mathcal{L}\phi_j(\mathbf{S}_s) \mid \mathcal{F}_t^z]\, ds. \tag{7}$$

By expanding the generator for the linear test function as $\mathcal{L}\phi(\mathbf{S}_t) = f_y(Y_t, X_t) + U_t$, Eq. (7) yields

$$\mathbb{E}\left[Y_{t+\tau} - Y_t - \int_t^{t+\tau} \big(f_y(Y_s, X_s) + U_s\big)\, ds \,\middle|\, \mathcal{F}_t^z \right] = 0 \qquad \text{for all } \tau > 0. \tag{8}$$

The conditional expectation of the hidden confounder U_s given $\mathcal{F}_t^z$ vanishes for all $s \in [t, t+\tau]$ by the exogeneity and the linearity of expectation. This directly results in Eq. (5), and under standard regularity conditions we can interchange the integral and expectation to obtain the differential form in Eq. (4), concluding the proof. □

It follows that $\mathbb{E}[\Phi\, R_{t,t+\tau}] = 0$ for all $\mathcal{F}_t^z$ measurable random variables Φ. This implies that we can use the filtration of Z to construct many valid instruments from a single continuous-time instrument process. Choosing $\Phi = g(Z_{[0,t]})$ for any functional $g\colon C([0,t], \mathbb{R}^{d_z}) \to \mathbb{R}$ of the instrument path up to time t, we obtain the following continuum of moments

$$\mathbb{E}\big[g(Z_{[0,t]})\, R_{t,t+\tau}\big] = 0 \qquad \text{for all } g, \tau. \tag{9}$$

Importantly, the conditional expectation with respect to the filtration of the instrument path in Eq. (5) carries much more information than just the past state Z_t. Although the joint process described in Eq. (1) is Markov, this property does not carry over to the partially observed process. Consequently, the entire continuous instrument path up to time t constitutes a valid instrument, giving rise to path-valued moment conditions.

These moment conditions rely on temporal information between observations. When those are not available, for example when only samples from the stationary distribution are accessible, we can still derive stationary moment conditions, which we describe in Section C.

5 The Linear Case: Ornstein-Uhlenbeck Processes

In this section we show that for an OU process the previously described moment equations can achieve point-identification of the targeted causal effect under mild conditions on the joint process. We instantiate the general setup described in Eq. (1) via an OU process with constant diffusion coefficients and a linear drift

$$d\mathbf{S}_t = \underbrace{\begin{pmatrix} \beta_{yy} & \beta_{yx} & 0 & \beta_{yu} \\ \beta_{xy} & \beta_{xx} & \beta_{xz} & \beta_{xu} \\ 0 & 0 & \beta_{zz} & 0 \\ 0 & 0 & 0 & \beta_{uu} \end{pmatrix}}_{=:A} \mathbf{S}_t\, dt + \underbrace{\begin{pmatrix} \sigma_y & 0 & 0 & 0 \\ 0 & \sigma_x & 0 & 0 \\ 0 & 0 & \sigma_z & 0 \\ 0 & 0 & 0 & \sigma_u \end{pmatrix}}_{=:\boldsymbol{\sigma}} \begin{pmatrix} dW_t^y \\ dW_t^x \\ dW_t^z \\ dW_t^u \end{pmatrix}, \tag{10}$$

where each drift coefficient is matrix-valued with dimensions $\beta_{ab} \in \mathbb{R}^{d_a \times d_b}$ for $a, b \in \{y, x, u, z\}$, and the diffusion coefficients are diagonal matrices $\sigma_a \in \mathbb{R}^{d_a \times d_a}$ for $a \in \{y, x, z, u\}$. We assume A to be Hurwitz, so that the process is stable and stationary. Given the system described in Eq. (10), identification of the target function reduces to the identification of the parameters β_{yx} and β_{yy}.

Simplification of moment conditions under joint Gaussianity. For OU processes, Eq. (9) simplifies due to the joint Gaussianity of the process. As $R_{t,t+\tau}$ consists of a sum of Gaussian random variables and linear integrals over Gaussian processes, the distribution of $(R_{t,t+\tau}, Z_{[0,t]})$ is also jointly Gaussian. Thus, the conditional expectation $\mathbb{E}[R_{t,t+\tau} \mid Z_{[0,t]}]$ is linear in the conditioning variables. This has two direct consequences: First, the unconditional formulation Eq. (9) can be restricted to linear functionals g, as any nonlinear transformation g in the moment conditions does not provide any additional identifying information. Second, joint conditioning on multiple past states

of the instrument process does not provide additional information over conditioning on each lagged $Z_{t-\delta}$ individually (Section D). Consequently, in the OU setting, Eq. (9) turns into a continuum of moments $\mathrm{Cov}[R_{t,t+\tau}, Z_{t-\delta}] = 0_{d_y \times d_z}$ for all $\tau > 0, \delta \geq 0$. Due to stationarity and linearity of expectation, varying both τ and δ is redundant. So we can write equivalently

$$\mathrm{Cov}[R_{t,t+\tau}, Z_t] = 0_{d_y \times d_z} \qquad \text{for all } \tau > 0. \tag{11}$$

Identification. In an OU process, U_t also enters the drift of X additively, such that the confounder's contribution to the cross-covariance dynamics between X and Z vanishes by exogeneity as well. Thus, differentiating Eq. (11) with respect to τ together with the equivalent equation for the X and Z cross-covariance yields a linear ODE in the covariance states $C(\tau) = [C_{YZ}(\tau), C_{XZ}(\tau)] = [\mathbb{E}[Y_{t+\tau} Z_t^\top], \mathbb{E}[X_{t+\tau} Z_t^\top]] \in \mathbb{R}^{(d_x+d_y)\times d_z}$, in which the confounding no longer appears explicitly. The resulting system contains an additional forcing term $C_{ZZ}(\tau) \in \mathbb{R}^{d_z \times d_z}$, which can be interpreted as a control input.

$$\frac{d}{d\tau}\begin{pmatrix} C_{YZ}(\tau) \\ C_{XZ}(\tau) \end{pmatrix} = \underbrace{\begin{pmatrix} \beta_{yy} & \beta_{yx} \\ \beta_{xy} & \beta_{xx} \end{pmatrix}}_{=:A_c} \begin{pmatrix} C_{YZ}(\tau) \\ C_{XZ}(\tau) \end{pmatrix} + \underbrace{\begin{pmatrix} 0 \\ \beta_{xz} \end{pmatrix}}_{=:b} C_{ZZ}(\tau). \tag{12}$$

Consequently, in the case of a confounded OU process, the IV moment conditions therefore amount to matching the lagged covariances dynamics between instrument and outcome described by Eq. (12) to those estimated from data. With these prerequisites, we obtain the following identifiability statement.

Proposition 2 (Identification in the OU setting)**.** *Consider an OU as in Eq.* (10) *with Hurwitz drift matrix A and non-degenerate instrument diffusion $\sigma_z \sigma_z^\top \succ 0$ and let $a := d_x + d_y$. Further, define the controllability matrix*

$$\mathcal{C} = \begin{pmatrix} b & A_c b & \cdots & A_c^{a-1} b \end{pmatrix} \in \mathbb{R}^{a \times a d_z}. \tag{13}$$

For timepoints $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_a) \in (0, \tau_{max}]^a$ drawn i.i.d. from an absolutely continuous distribution over $(0, \tau_{max}]$, if $\mathrm{rank}(\mathcal{C}) = a$, the linear system obtained by integrating the first row of Eq. (12)

$$C_{YZ}(\tau_i) - C_{YZ}(0) = \beta_{yy} \int_0^{\tau_i} C_{YZ}(s)\, ds + \beta_{yx} \int_0^{\tau_i} C_{XZ}(s)\, ds, \qquad i \in \{1, \ldots, a\}, \tag{14}$$

almost surely has a unique solution, identifying the causal effect β_{yx} and the self-effect β_{yy}.

Proof sketch. The proof (presented in detail in Section E) relies on three steps. Given Eq. (14), the proposition can directly be viewed as a rank condition on a matrix of a observed timepoints of the trajectory $C(\tau)$. In LTI systems, the controllability condition Eq. (13) ensures that the trajectory can span the whole state-space (Bellman & Åström, 1970). However, for a fixed control input, there may still exist initial conditions for which some modes of the system remain unobserved. Using the stationary Lyapunov equation, the proof shows that such initial conditions do not occur under a non-degenerate instrument diffusion $\sigma_z \sigma_z^\top \succ 0$ almost surely. Overall, the proposition thus states that given controllability, the causal parameters β_{yx} and β_{yy} are identified by fitting the covariance ODE in Eq. (12) at a finite number of lags τ.

Estimation. We estimate the drift coefficients through a two-stage procedure based on the integrated covariance identities in Eq. (14). For fixed $L > a$ lags sampled as in Proposition 2, define

$$B_\ell := C_{YZ}(\tau_\ell) - C_{YZ}(0), \qquad G_\ell := \int_0^{\tau_\ell} C(s)\, ds, \qquad \ell = 1, \ldots, L. \tag{15}$$

The first stage estimates these moments from the observed trajectory. Specifically, we estimate $C_{YZ}(\tau_\ell)$ by Nadaraya–Watson kernel regression (Walk, 2010) of the observed cross-products $Y_{t_j} Z_{t_i}^\top$ on their time lags $t_j - t_i$, using a bandwidth that shrinks with sample size. Together with the empirical zero-lag covariance $\widehat{C}_{YZ,n}(0) = n^{-1} \sum_{i=1}^n Y_{t_i} Z_{t_i}^\top$, this yields $\widehat{B}_{\ell,n} := \widehat{C}_{YZ,n}(\tau_\ell) - \widehat{C}_{YZ,n}(0)$. We estimate G_ℓ directly by importance-weighted averages of lagged cross-products involving both Y and X, correcting for the possibly nonuniform lag density induced by the known observation-gap distribution.

The second stage fits the empirical moment equations by least squares. Stacking $\widehat{B}_n = (\widehat{B}_{1,n}, \ldots, \widehat{B}_{L,n})$ and $\widehat{G}_n = (\widehat{G}_{1,n}, \ldots, \widehat{G}_{L,n})$, the estimator admits the closed-form expression

$$(\widehat{\beta}_{yy,n}, \widehat{\beta}_{yx,n}) := \widehat{B}_n \widehat{G}_n^\dagger \in \operatorname*{arg\,min}_{\Theta \in \mathbb{R}^{d_y \times a}} \left\| \widehat{B}_n - \Theta \widehat{G}_n \right\|_F^2, \tag{16}$$

where † denotes the Moore–Penrose pseudoinverse, selecting the minimum-Frobenius-norm solution when the minimizer is nonunique. This combines nonparametric covariance estimation with a single linear least-squares solve for the causal drift coefficients. All details of the construction, a differential version of the estimator that performs similarly in experiments, and the proof of the following consistency statement are in Section F.

Theorem 3 (Consistent estimator in the OU setting)**.** *In the setting of Proposition 2, the estimator in Eq.* (63) *is consistent, i.e.,* $\widehat{\beta}_{yy,n} \xrightarrow{p} \beta_{yy}$ *and* $\widehat{\beta}_{yx,n} \xrightarrow{p} \beta_{yx}$ *as* $n \to \infty$.

In summary, we have established identifiability and developed a consistent estimator for the linear (OU) continuous-time IV setting. Before demonstrating the practical efficacy of our estimator empirically, we also develop a tractable estimation procedure for the generically unidentifiable and complex nonlinear case.

6 The Nonlinear Case

In the OU setting, identification is governed by a controllability condition, which can ultimately be tested for from observed data. In the nonlinear case, identification and practical estimation become much more challenging, as linear functionals g do no longer exhaust the identifying information contained in the moment equation Eq. (4) and furthermore, the target parameters generally cannot be moved out of the expectation, as in Eq. (12).

Identification. It is known from the static case, that nonparametric instrumental variable estimation, which seeks to recover a nonlinear function from conditional moment conditions, is an ill-posed problem (Bennett et al., 2023; Zhang et al., 2023; Shen et al., 2026; van der Laan et al., 2026). In the nonlinear setting, the goal is to identify the parameters $\theta \in \mathbb{R}^p$ parameterizing the target drift function $f_{y;\theta} \in \mathcal{H}$ in the general SDE model described in Eq. (1). Generally, θ is unidentifiable for flexible function classes $\mathcal{H}$, i.e., the conditional moment restrictions in Eq. (5) are satisfied for multiple parameter settings. The standard remedy from the literature (Bennett et al., 2023) is to regularize the problem to obtain one compatible candidate θ in the identified set, for instance, aiming at the least norm solution. We follow this approach in our continuous-time setting.

Estimation. We start from the unconditional form of the previously derived moment condition $\mathbb{E}\big[g(Z_{[0,t]})\, R_{t,t+\tau;\theta}\big] = 0$ for all g, τ, where $R_{t,t+\tau;\theta}$ now depends on the drift function $f_{y;\theta}$ via Eq. (3). We evaluate the residual at a fixed time horizon $\tau = h$, where h is a hyperparameter. For a moderate choice of h, the signal to noise ratio of the drift contribution increases. At the same time, the correlation to the past instrument path typically decreases as h increases. When the sampling frequency is sufficiently high, the integral terms in $R_{t,t+\tau}$ can be approximated using standard quadrature schemes. Otherwise, simulation-based estimation methods are required.

We can then formulate the nonlinear estimation problem as a minimax problem of the risk $\mathcal{R}_z$ associated with the moment conditions Eq. (5) (Bennett et al., 2023; Ailer et al., 2024). Concretely, we minimize the parameters of the target function with respect to the worst case functional g, where we assume that g lies in the reproducing kernel Hilbert space (RKHS) $\mathcal{G}$ induced by some characteristic kernel $k_z : C([s,t],\mathbb{R}^d) \times C([s',t'],\mathbb{R}^d) \to \mathbb{R}$. Together with the regularization term $\Omega : \mathbb{R}^p \to \mathbb{R}$ this gives the following dual optimization problem

$$\theta^\star = \arg\min_\theta \mathcal{R}_z = \arg\min_\theta \left[\sup_{g\in\mathcal{G},\ \|g\|_\mathcal{G}\le 1} \big(\mathbb{E}\left[R_{t,t+h;\theta}\, g(Z_{[0,t]})\right]\big)^2 + \lambda\Omega(\theta) \right]. \tag{17}$$

Following Zhang et al. (2023), we assume that g lies within a unit ball of its RKHS $\mathcal{G}$. This allows us to avoid a computationally expensive explicit dual optimization, as the inner supremum admits an analytical solution, so that we obtain the risk

$$\mathcal{R}_z(\theta) = \mathbb{E}_{i,j}\left[R^i_{t,t+h;\theta}\, R^j_{t,t+h;\theta}\, k_z\big(Z^i_{[0,t]}, Z^j_{[0,t]}\big) \right] + \lambda\Omega(\theta). \tag{18}$$

Using the V-statistic to empirically estimate the risk, we finally get

$$\widehat{\mathcal{R}}_z(\theta) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n R^i_{t,t+h;\theta}\, R^j_{t,t+h;\theta}\, k_z\big(Z^i_{[0,t]}, Z^j_{[0,t]}\big) + \lambda\Omega(\theta) = \frac{1}{n^2}\mathbf{r}^\top K_z\, \mathbf{r} + \lambda\Omega(\theta), \tag{19}$$

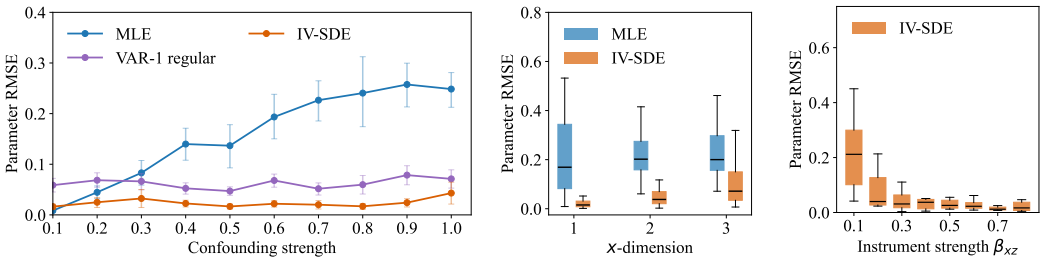


Figure 2: RMSE of β_{yy} and β_{yx} for a confounded OU process (one instrumental and one outcome process throughout, except in (b)). **Left:** Estimation error under varying confounding strength. **Center:** Recovery of multidimensional causal effects from a single instrument process. **Right:** Estimation error as a function of instrument–covariate coupling strength β_{xz}. Error bars indicate 95% confidence intervals, boxplots show medians and interquartile ranges. Each evaluation is based on 20 systems (left) and 10 systems (middle/right).

where $K_z \in \mathbb{R}^{n\times n}$ describes the kernel Gram matrix, $\mathbf{r}$ the residual vector, and n the number of sampled path sections.

Signature kernel. The risk $\mathcal{R}_z$ assumes access to a characteristic kernel on path spaces to express the functionals of the instrument paths in the RKHS $\mathcal{G}$. We propose using the signature kernel (Manten et al., 2025; Salvi et al., 2021; Lee & Oberhauser, 2025), which has gained increasing popularity in recent years for various types of sequential data. More details are found in Section H.

7 Experiments

OU setup and baselines. We evaluate our method on simulated OU process trajectories across different systems. For each system, we randomly sample drift and diffusion matrices with sparsity according to Eq. (10), and all nonzero parameters bounded away from zero. The observation time-points are irregularly sampled with uniformly distributed observation intervals, detail are described in Section I.

We compare our method to parameter estimates obtained from naive Maximum Likelihood estimates (Pavliotis, 2014) ignoring the unobserved state U. Due to a lack of continuous-time IV baselines, we also compare to the VAR(1) estimator presented in Thams et al. (2024) working with the same number of regularly sampled data from the same trajectory. Additionally, in Section I, we compare to the discrete-time estimator for the normalized parameters in (Mogensen, 2023), where we construct the ratio estimates from both individually estimated variables β_{yx} and β_{yy}.

IVSDE identifies causal effects under confounding. Figure 2(left) shows the parameter estimation error over both β_{yy} and β_{yx} in a confounded OU process, with increasing confounding strength through parameters β_{xu} and β_{yu}. The MLE estimate of the parameters is clearly increasingly biased under growing confounding, reflecting that MLE cannot distinguish between the confounding effect and the causal influence of process X on Y. By contrast, the continuous-time IV estimate remains stable across confounding levels. Given the same number of samples on a regular grid, discrete IV methods for time series remain biased as they assume the IV sparsity in the finite-time transition maps instead of on the infinitesimal level, as explained in Section B.

The proposed method identifies multi-dimensional causal effects with a single instrument. Unlike static IV methods, which require at least as many instruments as causal effects for identification in the linear setting, a single instrumental process in continuous-time can identify multi-dimensional causal effects. Figure 2 (middle) shows the parameter estimation error depending on the dimensionality of the covariate process given a single instrument process. The total number of estimated parameters is always $d_x + 1$, accounting for both β_{yx} and β_{yy}. The results reveal that the proposed IV method may even be used to estimate four-dimensional causal effects using a single instrumental process and still obtains a lower average error than MLE. However, the variance of the IV estimator increases with growing dimensionality of the targeted causal effect.

Influence of instrument strength. The proposed IV method relies on the coupling between instrument and covariate process, which is also captured by the relevance assumption. Figure 2

(right) shows that the mean estimation error and variance decrease with stronger coupling parameter β_{xz}. This is consistent with the intuition that larger coupling parameters induce more exogenously generated variation in the covariate process, thus providing a stronger instrument.

Kernel-based estimation of confounded nonlinear effects. In the nonlinear setting, inspired by Saengkyongam et al. (2022), we set up the drift f_y as $f_y(Y_t, X_t) = \alpha Y_t + \beta X_t + \sum_{i=1}^{p} w_i e^{-(X_t - c_i)^2/\sigma^2_{\text{rbf}}} - \sum_{i=1}^{p} w_i e^{-(X_t + c_i)^2/\sigma^2_{\text{rbf}}}$, where $w \in \mathbb{R}^p$ describes a randomly sampled parameter vector of radial-basis function weights and $c \in \mathbb{R}^p$ their centers. We parameterize the target function $f_{y;\theta}$ with a basis function expansion containing the same radial basis function centers and variances, as well as two linear terms for X and Y. Figure 3 (left) clearly shows that the IV-based estimator outperforms the MLE estimation for larger confounding values, while the MLE performs better for small and no confounding. This behavior of these estimation methods can be explained by two observations. First, IV methods are inherently less efficient than MLE, since they only exploit the variation explained by the instrument instead of the full likelihood. Furthermore, the signature kernel-based optimization introduces a more complex optimization landscape, which might contribute to the lower performance in settings where MLE is only slightly biased. With higher confounding, however, this is outweighed by the bias of the MLE estimate.

Table 1: Parameter RMSE on the diabetes dataset.

Method	RMSE	
IV-SDE	**0.0157**	(0.0035)
MLE	0.0998	(0.0003)

Neural parametrization. We can also use a small MLP to parameterize the class of the target drift function f_y. In this setting, we assume the ground truth drift function to be linear in Y and use a sigmoid function to describe the dependence on the covariates X. Figure 3 (right) shows that we recover the qualitative shape of the target function.

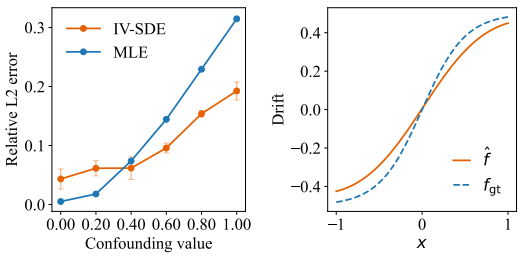


Figure 3: Results for a nonlinear drift parametrization. (left) Relative drift RMSE over the X_t, Y_t distribution. (right) Ground truth sigmoidal drift f_{gt} and learned drift $\hat{f}$ parameterized with an MLP.

Diabetes dataset. We setup a semi-synthetic experiment motivated by pharmacokinetic/pharmacodynamic models for type 2 diabetes (Subramanian et al., 2024; Naik et al., 2026). We aim to identify the patient's insulin sensitivity, which describes how their glucose level responds to insulin and is the main physiological factor impaired by the disease. Both glucose and insulin levels are confounded by unobserved gut glucose levels, and external insulin dosing schedule is used as an instrument. The resulting drift is modeled as $g_y(\mathbf{S}) := f_y(Y_t, X_t) + c_{yu}U_t = c_{yu}U_t - a_y Y_t - \gamma_{xy} X_t Y_t$. Table 1 shows that compared to an MLE-based parameter fit, our nonlinear estimator retrieves all function parameters with lower RMSE across five runs. More runs including parameter ablations are in Section I.

8 Conclusion

We introduced a framework for instrumental variable estimation in continuous-time stochastic processes governed by SDEs by deriving path-based moment conditions resulting from the structural IV assumptions. For OU processes, we established a controllability-based identification condition together with a consistent estimator. For nonlinear effects, we proposed a minimax estimation scheme based on path-valued moment conditions. We have shown the empirical performance on synthetic data and proposed SDE simulators for diabetes from the literature. Our results also verify that a single instrument process indeed carries enough information to recover multi-dimensional effects.

We believe that with the currently growing interest in learning continuous-time dynamics (Kostic et al., 2024; Lorch et al., 2024) the causal interpretation of these models is becoming increasingly relevant. Our results also highlight once more that the transition from discrete time to continuous time is often not trivial as local structure does often not transfer to finite-step transitions. Across our simulated systems, including the glucose–insulin example, respecting this structure improves estimation. Interesting direction for future work include a better understanding of identifiability in the nonlinear case and uncertainty quantification that remain reliable with weak instruments. Also, theoretical extensions beyond stationarity would be interesting, albeit stationarity does not appear to be mandatory for practical estimation.

AI USE STATEMENT

In this work, we used generative AI tools to help generate synthetic data sets, support formulating mathematical claims, assist in the writing of proofs, propose and refine hypotheses, design or provide feedback on research methodology and experiments, support with the implementation of methods and support qualitative and thematic data analysis. We have not used generative AI tools to provide critical ingredients for proving mathematical claims, help develop conceptual frameworks, interpret results and assist with translation. Cleaning and reformatting datasets is not applicable to this work.

While we used AI in the research process, the final content of this paper was manually constructed, checked and verified with care. We have manually reviewed all LLM written code for functionality and correctness. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We prove all main mathematical statements in the appendix and the main paper itself and precisely refer to previous methods, when necessary. Further, we provide a thorough discussion of the method and its main assumptions in the appendix. We supply full descriptions of all synthetically generated datasets and used methods in the appendix. We provide the main methods in an anonymized form as source code at time of submission, the repository with full reproducible result plots will be available upon acceptance.

REFERENCES

Elisabeth Ailer, Jason Hartford, and Niki Kilbertus. Sequential underspecified instrument selection for cause-effect estimation. In *International Conference on Machine Learning*, pp. 408–420. PMLR, 2023. URL https://arxiv.org/abs/2302.05684.

Elisabeth Ailer, Niclas Dern, Jason Hartford, and Niki Kilbertus. Targeted sequential indirect experiment design. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=U3Rgdb4li9.

Katherine Baicker, Sarah L. Taubman, Heidi L. Allen, Mira Bernstein, Jonathan H. Gruber, Joseph P. Newhouse, Eric C. Schneider, Bill J. Wright, Alan M. Zaslavsky, and Amy N. Finkelstein. The oregon experiment-effects of medicaid on clinical outcomes. *New England Journal of Medicine*, 368(18):1713–1722, 2013.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997. doi: 10.1080/01621459.1997.10474074.

Richard Bellman and Karl Johan Åström. On structural identifiability. *Mathematical Biosciences*, 7 (3–4):329–339, 1970. doi: 10.1016/0025-5564(70)90132-X.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32, 2019. URL https://arxiv.org/abs/1905.12495.

Andrew Bennett, Nathan Kallus, Xiaojie Mao, Whitney Newey, Vasilis Syrgkanis, and Masatoshi Uehara. Minimax instrumental variable regression and l_2 convergence guarantees without identification or closedness. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2291–2318. PMLR, 2023. URL https://arxiv.org/abs/2302.05404.

A.R. Bergstrom. *Continuous time stochastic models and issues of aggregation over time*, volume 2 of *Handbook of Econometrics*. Elsevier, 1 edition, None 1984. doi: None. URL https://ideas.repec.org/h/eee/ecochp/2-20.html.

Manudeep Bhuller, Gordon Dahl, Katrine Løken, and Magne Mogstad. Incarceration, recidivism, and employment. *Journal of Political Economy*, 2020.

Andrew Chesher. Identification in nonseparable models. *Econometrica*, 71(5):1405–1441, 2003. doi: https://doi.org/10.1111/1468-0262.00454. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00454.

Tatyana Deryugina, Garth Heutel, Nolan H. Miller, David Molitor, and Julian Reif. The mortality and medical costs of air pollution: Evidence from changes in wind direction. *American Economic Review*, 109(12):4178–4219, December 2019. doi: 10.1257/aer.20180279. URL https://www.aeaweb.org/articles?id=10.1257/aer.20180279.

Esther Duflo. Schooling and labor market consequences of school construction in indonesia: Evidence from an unusual policy experiment. *American Economic Review*, 91(4):795–813, September 2001. doi: 10.1257/aer.91.4.795. URL https://www.aeaweb.org/articles?id=10.1257/aer.91.4.795.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5 edition, 2019. doi: 10.1017/9781108591034.

Jesús Fernández-Villaverde and Juan F Rubio-Ramírez. Estimating macroeconomic models: A likelihood approach. *The Review of Economic Studies*, 74(4):1059–1087, 2007.

B Friedlander, P Stoica, and T Söderström. Instrumental variable methods for arma parameter estimation. *IFAC Proceedings Volumes*, 18(5):29–36, 1985.

Marion Gilson and Paul Van den Hof. Instrumental variable methods for closed-loop system identification. *Automatica*, 41(2):241–249, 2005.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912775.

Lars Peter Hansen and Jose Alexandre Scheinkman. Back to the future: Generating moment implications for continuous-time markov processes. Working Paper 141, National Bureau of Economic Research, September 1993. URL http://www.nber.org/papers/t0141.

Niels Hansen and Alexander Sokol. Causal interpretation of stochastic differential equations. *Electronic Journal of Probability*, 19(none), January 2014. ISSN 1083-6489. doi: 10.1214/ejp.v19-2891. URL http://dx.doi.org/10.1214/EJP.v19-2891.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1414–1423. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/hartford17a.html.

M. Hernan, B. Brumback, and JM. Robins. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96(None):440–448, June 2001. doi: None. URL https://ideas.repec.org/a/bes/jnlasa/v96y2001mjunep440-448.html.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020. Available online at https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

Guido W. Imbens and Whitney K. Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512, 2009. doi: https://doi.org/10.3982/ECTA7108. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA7108.

P Joseph, J Lewis, and J Tou. Plant identification in the presence of disturbances and application to digital adaptive systems. *Transactions of the American Institute of Electrical Engineers, Part II: Applications and Industry*, 80(1):18–24, 1961.

Vladimir R Kostic, Hélène Halconruy, Timothée Devergne, Karim Lounici, and Massimiliano Pontil. Learning the infinitesimal generator of stochastic diffusion processes. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=H7SaaqfCUi.

Darrick Lee and Harald Oberhauser. The signature kernel. In *Signature Methods in Finance: An Introduction with Computational Applications*, pp. 85–124. Springer, 2025.

Judith J. Lok. Statistical modeling of causal effects in continuous time. *The Annals of Statistics*, 36 (3), 6 2008. ISSN 0090-5364. doi: 10.1214/009053607000000820. URL http://dx.doi.org/10.1214/009053607000000820.

Lars Lorch, Andreas Krause, and Bernhard Schölkopf. Causal modeling with stationary diffusions. In *International Conference on Artificial Intelligence and Statistics*, pp. 1927–1935. PMLR, 2024. URL https://arxiv.org/abs/2310.17405.

Georg Manten, Cecilia Casolo, Emilio Ferrucci, Søren Wengel Mogensen, Cristopher Salvi, and Niki Kilbertus. Signature kernel conditional independence tests in causal discovery for stochastic processes. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Nx4PMtJ1ER.

Matthew A. Masten and Alexandre Poirier. Salvaging falsified instrumental variable models. *Econometrica*, 89(3):1449–1469, 2021. doi: https://doi.org/10.3982/ECTA17969. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA17969.

Søren Wengel Mogensen. Instrumental processes using integrated covariances. In *Conference on Causal Learning and Reasoning*, pp. 620–641. PMLR, 2023.

Søren Wengel Mogensen, Daniel Malinsky, and Niels Richard Hansen. Causal learning for partially observed stochastic dynamical systems. In *Conference on Uncertainty in Artificial Intelligence*, 2018. URL https://api.semanticscholar.org/CorpusID:215768132.

Parvaiz Ahmad Naik, Bijal M. Yeolekar, Kolade M. Owolabi, and Mahesh Yeolekar. Modeling and analyzing glucose-insulin interactions during diabetes through fractional dynamics in presence of glucagon. *Journal of Advanced Research*, 2026. ISSN 2090-1232. doi: https://doi.org/10.1016/j.jare.2026.01.020. URL https://www.sciencedirect.com/science/article/pii/S2090123226000457.

Grigorios A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations*, volume 60 of *Texts in Applied Mathematics*. Springer, New York, NY, 2014. ISBN 978-1-4939-1322-0. doi: 10.1007/978-1-4939-1323-7.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000. ISBN 0-521-77362-8.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D. Mahecha, Jordi Muñoz-Marí, Egbert H. van Nes, Jonas Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Schölkopf, Peter Spirtes, George Sugihara, Jie Sun, Kun Zhang, and Jakob Zscheischler. Inferring causation from time series in earth system sciences. *Nature Communications*, 10(1):2553, Jun 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-10105-3. URL https://doi.org/10.1038/s41467-019-10105-3.

Sorawit Saengkyongam, Leonard Henckel, Niklas Pfister, and Jonas Peters. Exploiting independent instruments: Identification and distribution generalization. In *International Conference on Machine Learning*, pp. 18935–18958. PMLR, 2022.

Cristopher Salvi, Thomas Cass, James Foster, Terry Lyons, and Weixin Yang. The signature kernel is the solution of a goursat pde. *SIAM Journal on Mathematics of Data Science*, 3(3):873–899, January 2021. ISSN 2577-0187. doi: 10.1137/20m1366794. URL http://dx.doi.org/10.1137/20M1366794.

J. D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415, 1958. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1907619.

Nora Schneider, Georg Manten, and Niki Kilbertus. Conservative continuous-time treatment optimization. *arXiv preprint arXiv:2603.16789*, 2026.

Zikai Shen, Dimitri Meunier, Houssam Zenati, Arthur Gretton, Nathan Kallus, and Aurélien Bibaut. Instrumental variable analysis without structural equations, 2026. URL https://arxiv.org/abs/2604.24660.

Christopher A. Sims. Solving linear rational expectations models. *Computational Economics*, 20 (1):1–20, 2002. doi: 10.1023/A:1020517101123. URL https://doi.org/10.1023/A:1020517101123.

Douglas Staiger and James H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, 1997. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2171753.

Adolf Stips, Diego Macias, Clare Coughlan, Elisa Garcia-Gorriz, and X. San Liang. On the causal structure between co2 and global temperature. *Scientific Reports*, 6(1):21691, Feb 2016. ISSN 2045-2322. doi: 10.1038/srep21691. URL https://doi.org/10.1038/srep21691.

Vijaya Subramanian, Jonatan I. Bagger, Vinayak Harihar, Jens J. Holst, Filip K. Knop, and Tina Villsbøll. An extended minimal model of ogtt: estimation of α- and β-cell dysfunction, insulin resistance, and the incretin effect. *American Journal of Physiology-Endocrinology and Metabolism*, 326(2):E182–E205, 2024. doi: 10.1152/ajpendo.00278.2023. URL https://doi.org/10.1152/ajpendo.00278.2023. PMID: 38088864.

Nikolaj Thams, Rikke Søndergaard, Sebastian Weichwald, and Jonas Peters. Identifying causal effects using instrumental time series: Nuisance iv and correcting for the past. *Journal of Machine Learning Research*, 25(302):1–51, 2024.

Lars van der Laan, Nathan Kallus, and Aurélien Bibaut. Nonparametric instrumental variable inference with many weak instruments, 2026. URL https://arxiv.org/abs/2505.07729.

Harro Walk. Strong laws of large numbers and nonparametric estimation. In Luc Devroye, Bülent Karasözen, Michael Kohler, and Ralf Korn (eds.), *Recent Developments in Applied Probability and Statistics: Dedicated to the Memory of Jürgen Lehn*, pp. 183–214. Physica-Verlag HD, Heidelberg, 2010. ISBN 978-3-7908-2598-5. doi: 10.1007/978-3-7908-2598-5_8. URL https://doi.org/10.1007/978-3-7908-2598-5_8.

Kwan Wong and Elijah Polak. Identification of linear discrete time systems using the instrumental variable method. *IEEE Transactions on Automatic Control*, 12(6):707–718, 1967.

P.G. Wright. *The Tariff on Animal and Vegetable Oils*. Investigations in international commercial policies. Macmillan, 1928. URL https://books.google.de/books?id=hig5AAAAMAAJ.

Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. Instrumental variable regression via kernel maximum moment loss. *Journal of Causal Inference*, 11(1):20220073, 2023.

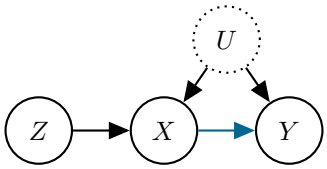


Figure 4: Standard instrumental-variable causal graph. The instrument Z affects the outcome Y only through the treatment X, while the latent variable U confounds the relationship between X and Y. The causal effect of interest is the direct dependence of Y on X highlighted in blue.

A Extended Setup Discussion

Our setup relies on the standard instrumental variable assumptions, and it is worth discussing them in more detail and disentangling them from other assumption made throughout this paper.

A.1 Static IV assumptions

The instrumental variable setting itself is defined by the three conditions *relevance, exclusion and exogeneity* (Hernán & Robins, 2020; Pearl, 2000). Relevance requires that X is not independent of Z, the exclusion restriction requires that Z only affects Y through X and exogeneity requires that U is independent of Z. A simple *structural causal model* fulfilling these instrumental variable conditions is found in Fig. 4. Relaxing any of these conditions is known to yield impossibility statements without strong replacements such as sign- or subset validity restrictions. For such very specific, IV-related settings, one can provably only obtain bounds or sensitivity regions and no point-identification (Masten & Poirier, 2021; Balke & Pearl, 1997). While the literature contains ample discussion of the IV assumptions, they have not impaired the practical relevance and usage of IV estimation.

In addition to these three defining conditions, IV methods require a fourth assumption for identification Hernán & Robins (2020), which usually takes the form of the additivity assumption. This assumption implies that the confounder only enters the outcome variable in an additive way. This assumption is substantive, but it is just as well standard and provably necessary: If this additive separability is not assumed the effect is unidentifiable unless other strong assumptions are introduced. Such alternatives include adding monotone control functions or shape restrictions (Imbens & Newey, 2009; Chesher, 2003) or give up on identifiability and analyze partial identification.

A.2 Transfer to the SDE setting

We transfer the previously described defining assumptions of the IV setting together with the necessary additivity assumption to the continuous-time dynamic setting. We make the standard assumption of a a diagonal diffusion matrix with coordinate-wise state-dependent entries. Consequently, we assume all causal effects between the variables to be encoded in the drift functions. We discuss a relaxation of this in a later section.

We encode the structural IV assumptions in the sparse interaction graph of the SDE. Thus, the **relevance** condition requires the drift of X to depend on the instrument Z. Similarly, the **exclusion** restriction requires that Z does not enter the drift of outcome process Y and only influences Y through process X. **Exogeneity** is best expressed through the path-filtration condition $\mathbb{E}(U_s|\mathcal{F}_t^z) = 0, \quad s \geq t$. The **Additivity** assumption can be directly transferred to the dynamic setting by requiring additivity of the unobserved state U_t in the drift function of process Y.

Without the assumption of a diagonal, component-wise state-dependent diffusion, other definitions of the IV setup would also be possible in continuous-time, for which our proposed moment conditions would remain valid.

In its most general form, the moment condition presented in Proposition 1 can be written as

$$\mathbb{E}[R_{t,t+h}|\mathcal{F}_t^Z] = \int_t^{t+h} \mathbb{E}[U_s|\mathcal{F}_t^Z]ds + \mathbb{E}[M_{t+h}^Y - M_t^Y|\mathcal{F}_t^Z]. \tag{20}$$

$$M_t^Y = \int_0^t \sigma_{yy}(S_s) dW_s^Y + \int_0^t \sigma_{yx}(S_s) dW_s^X + \int_0^t \sigma_{yz}(S_s) dW_s^Z + \int_0^t \sigma_{yu}(S_s) dW_s^U. \quad (21)$$

Hence, the moment restrictions remain valid as long as M_t is a martingale and the exogeneity of the instrument holds. In case of an arbitrary non-diagonal and state-dependent diffusion matrix, M_t remains a martingale. However, a non-diagonal diffusion matrix may break the validity of the moments through a violation of the exogeneity condition. This may occur in two ways:

First, exogeneity may be violated if Z and U are directly mixed by sharing a mutual innovation, specifically $\left(\sigma(S_t)\sigma(S_t)^\top\right)_{zu} \neq 0$. This introduces shared information between the history of Z and the history of U, which through persistence in the latent process also allows for a prediction of future U. Second, exogeneity may also not hold, if the diffusion coefficient of either process depends on the state of the other described by $\sigma_u = \sigma_u(U_t, Z_t)$ or $\sigma_z = \sigma_z(Z_t, U_t)$.

We note that, as long as exogeneity is fulfilled, diffusion mixing as well as state dependent diffusions may introduce proper confounding bias in the naive MLE estimator even without drift confounding, while the moment conditions in Proposition 1 and thus the proposed estimation remain valid. Consequently, we may relax the assumption on a diagonal diffusion matrix by only requiring that exogenity holds, and could also define a more general continuous-time IV setting than described by the SDE system in Eq. (1).

We decide to present this paper with the more restrictive drift confounding as in Eq. (1), because we argue that proper causal confounding is intuitively conveyed via the drift instead of shared diffusion noise or state-dependent volatility. We discuss the influence of a non-diagonal diffusion matrix on our proposed methods in Section G.

A.3 Discussion of further assumptions.

Beyond the previously discussed conditions defining the IV setting, we make the following further assumptions on the process:

Zero-mean. We assume the joint process to have zero-mean. This holds without loss of generality, as the observed parts of the process can be centered with the observed mean. The moment conditions in both the linear and nonlinear setup retrieve the function $f_y(Y_t, X_t)$ up to translation, as we can only identify the sum of the intercept of $f_y(Y_t, X_t)$ and the unobserved $\mu_u = \mathbb{E}(U_t)$. However, for prediction of process Y under intervention on process X this translation cancels and the total outcome drift of process Y remains unchanged. Hence, we can assume $\mathbb{E}[U_t] = 0$ without loss of generality. As presented, the current model setup requires $d_u = d_y$. We interpret U as the effective latent contribution instead of necessarily the full underlying state. Its dynamics may also depend on additional unobserved variables, the proposed moment conditions will remain valid as long as exogeneity is fulfilled.

Stationarity. We assume the joint process to be stationary, the non-stationary setting is only adjacent to our setup. We make use of stationarity for identification and estimation from a single curve. Together with the assumption of ergodicity, we can estimate the moments using time averages from a single trajectory. In the nonstationary setting, the moments would generally depend on the absolute observation time. Thus, they could not be pooled along single trajectories without introducing additional assumptions. Estimation would instead require e.g. multiple independent trajectories or assumptions on the local stationarity of the process. We leave this extension to future work.

B Distinction from Discrete-Time Models

In the following section we demonstrate the difference between structural assumptions in discrete-time models and continuous-time models as also described in Mogensen et al. (2018). We specifically show that local IV structure on an infinitesimal level does not imply IV sparsity in finite interval transition maps in discrete-time models.

Consider as an example a general OU process with states $\mathbf{S}_t$ with drift matrix A and diagonal diffusion matrix $\boldsymbol{\sigma}$. When all components of $\mathbf{S}_t$ are observed on an equidistant grid with sampling interval $h > 0$, Bergstrom (1984) showed that the discrete observations satisfy the $\mathrm{VAR}(1)$ process with

transition matrix Φ:

$$\mathbf{S}_t \;=\; e^{Ah}\,\mathbf{S}_{t-h} \;+\; \eta_t^h = \Phi \mathbf{S}_{t-h} + \eta_t^h, \qquad t = h, 2h, \ldots, \tag{22}$$

Here, the noise follows:

$$\eta_t^{(h)} \;=\; \int_{t-h}^{t} e^{A(t-s)}\,\boldsymbol{\sigma}\, dW_s, \tag{23}$$

thus satisfying

$$\mathbb{E}[\eta_t^{(h)}] = 0, \qquad \mathbb{E}\Big[\eta_t^{(h)}(\eta_t^{(h)})^\top\Big] \;=\; \int_0^h e^{As}\,\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\, e^{A^\top s}\, ds \tag{24}$$

and $\mathbb{E}[\eta_t^{(h)}(\eta_{t-r}^{(h)})^\top] = 0$ for all $r \geq h$.

The discrete-time transition matrix $\Phi = e^{Ah}$, however, generally does not maintain the sparsity pattern of drift matrix A. Therefore, the local independencies in A are typically not visible in the discrete $\mathrm{VAR}(1)$ representation Mogensen et al. (2018).

Specifically, for the OU process with IV structure described in Eq. (10) with states $\mathbf{S}_t = (Y_t, X_t, Z_t, U_t)$, the IV sparsity encoded in the drift (Eq. (10)) does not carry over to the transition matrix in the $\mathrm{VAR}(1)$ model of the process. While the block-diagonal structure of A ensures that the exogeneity of the instrument and confounding process are preserved in the finite-time transition matrix, the exclusion restriction does not hold. Specifically, $A_{yz} = 0$ does not imply $\Phi_{yz} = 0$. For a concrete example, consider

$$A = \begin{pmatrix} -1.4 & -1 & 0 & -0.5 \\ -0.25 & -1.2 & -1 & -0.5 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \tag{25}$$

but for $h = 1$

$$\Phi = e^{Ah} \approx \begin{pmatrix} 0.2803 & -0.2845 & 0.1541 & -0.0806 \\ -0.0711 & 0.3372 & -0.3462 & -0.1538 \\ 0 & 0 & 0.3679 & 0 \\ 0 & 0 & 0 & 0.3679 \end{pmatrix}, \tag{26}$$

where $\Phi_{yz} \neq 0$ despite $A_{yz} = 0$.

Consequently, imposing a local IV structure in a continuous-time model implies a different assumption about the underlying mechanism than enforcing IV sparsity in discrete-time models.

C Stationarity moment conditions

In the setting where only observations from the stationary distribution are accessible, we can still setup moment conditions: since $\mathbb{E}[\mathcal{L}\phi(\mathbf{S}_t)] = 0$ (presented by Hansen & Scheinkman, 1993), we can directly make use of the stationary covariance between any function of the instrument state Z_t and the outcome Y_t. For this, we choose the test function as $\phi(\mathbf{S}_t) = g(Z_t)Y_t$, where $g(\cdot)$ is a projection of the instrument state. Expanding the generator expression and using the exogeneity of the instrument, this yields

$$\mathbb{E}\Big[g(Z_t)(f_y(Y_t, X_t))\Big] = -\mathbb{E}\left[Y_t\left(f_z(Z_t)g'(Z_t) + \frac{1}{2}\sigma_z^2(Z_t)g''(Z_t)\right)\right]. \tag{27}$$

While this moment condition can be used for estimation and may even identify the target function by using different projections g under a sufficiently rich instrument process, it also assumes knowledge of the instrument dynamics in the form of the instrument drift and diffusion. We therefore only focus on the previously introduced moment conditions in the presentation of this paper.

D Simplification of Moment Conditions under Joint Gaussianity

The following section briefly shows that the moment restrictions conditioned on the full instrument path reduce to moment restrictions conditioned on the individual instrument states in an OU system.

Since the OU process is jointly Gaussian, the vector $(R_{t,t+\tau}, Z_{s_1}, \ldots, Z_{s_k})$ is jointly Gaussian with zero mean for any $s_1, \ldots, s_k \leq t$. The conditional expectation of a jointly Gaussian vector is linear in the conditioning variables, so:

$$\mathbb{E}\big[R_{t,t+\tau} \mid Z_{s_1}, \ldots, Z_{s_k}\big] = \mathrm{Cov}[R_{t,t+\tau}, \mathbf{Z}]\,\mathrm{Cov}[\mathbf{Z}, \mathbf{Z}]^{-1}\mathbf{Z}, \tag{28}$$

where $\mathbf{Z} = (Z_{s_1}, \ldots, Z_{s_k})$. Since $\mathrm{Cov}[R_{t,t+\tau}, Z_s] = 0$ for all $s \leq t$, the path-valued moment conditions reduce to moment conditions based on all individual variables Z_s.

We further show that the continuum of moments over δ and τ may be reduced to the single-parameter family in Eq. (11). Starting from just the continuum over τ

$$\mathrm{Cov}\big[R_{t,t+\tau}, Z_t\big] = 0 \qquad \text{for all } \tau > 0. \tag{29}$$

we define $\Delta = \delta + \tau$ and write, using linearity of expectation

$$\mathrm{Cov}\big[R_{t,t+\Delta}, Z_t\big] = \mathrm{Cov}\big[R_{t,t+\delta}, Z_t\big] + \mathrm{Cov}\big[R_{t+\delta,t+\delta+\tau}, Z_t\big] = 0 \qquad \text{for all } \delta, \tau > 0. \tag{30}$$

Evaluating the first term to zero and using stationarity yields:

$$\mathrm{Cov}\big[R_{t,t+\tau}, Z_{t-\delta}\big] = 0 \qquad \text{for all } \tau, \delta > 0 \tag{31}$$

which is equivalent to the continuum of moments over both τ and δ.

E Proof for Proposition 2

The proof for Proposition 2 consists of three steps.

E.1 Step 1: reduction to rank condition

Given is the ODE system with states $C(\tau) = (C_{YZ}(\tau), C_{XZ}(\tau)) \in \mathbb{R}^{(d_y+d_x)\times d_z}$ and forcing term $C_{ZZ}(\tau) \in \mathbb{R}^{d_z \times d_z}$:

$$\frac{d}{d\tau}\begin{pmatrix} C_{YZ}(\tau) \\ C_{XZ}(\tau) \end{pmatrix} = \underbrace{\begin{pmatrix} \beta_{yy} & \beta_{yx} \\ \beta_{xy} & \beta_{xx} \end{pmatrix}}_{=:A_c} \begin{pmatrix} C_{YZ}(\tau) \\ C_{XZ}(\tau) \end{pmatrix} + \underbrace{\begin{pmatrix} 0 \\ \beta_{xz} \end{pmatrix}}_{=:b} C_{ZZ}(\tau), \quad C(0) = c_0 \tag{32}$$

Furthermore, sampling $a := d_x + d_y$ random lags $\tau \geq 0$, integrating the first row of the covariance ODE, and stacking the equations results in the linear system described in Eq. (14)

$$K\,G(\tau) \;=\; B(\tau). \tag{33}$$

where

$$G(\boldsymbol{\tau}) := \left(\int_0^{\tau_1} C(s)\,ds,\, \ldots,\, \int_0^{\tau_a} C(s)\,ds\right) \in \mathbb{R}^{a \times a d_Z},$$

$$B(\boldsymbol{\tau}) := (C_{YZ}(\tau_1) - C_{YZ}(0),\, \ldots,\, C_{YZ}(\tau_a) - C_{YZ}(0)) \in \mathbb{R}^{d_Y \times a d_Z},$$

and $K = (\beta_{yy}, \beta_{yx}) \in \mathbb{R}^{d_Y \times a}$. The parameter matrix K is uniquely determined iff $\mathrm{rank}(G(\tau)) = a$.

E.2 Step 2: Trajectory spans $\mathbb{R}^a$

For the multi-input LTI system, we can determine the Kalman controllability of the pair (A_c, b) via the rank condition of the controllability matrix $\mathcal{C}$, defined as:

$$\mathcal{C} = \begin{pmatrix} b & A_c b & \cdots & A_c^{a-1} b \end{pmatrix} \in \mathbb{R}^{a \times a d_Z}. \tag{34}$$

This condition guarantees that every eigenmode of A_c can be excited through some combination of input and delays. This equivalently means that the controllable subspace equals $\mathbb{R}^a$.

When (A_c, b) is fully controllable and the diffusion of Z is not degenerate, implying $\sigma_z \sigma_z^\top \succ 0$, the trajectory could still fail to span $\mathbb{R}^a$, only if the initial condition is "unluckily" aligned with the system dynamics and control input.

In standard control problems, the set of such initial conditions has measure zero when x_0 is randomly sampled, and thus usually this setting only obtains limited consideration. In our case however, c_0 is not chosen freely, but fixed by the full system dynamics via the stationary Lyapunov equation. Therefore, it remains to be shown that given the control input in form of the autocorrelation of process Z, the initial condition c_0 does not lie in such a subspace, in which the eigenmodes of A_c are not fully visible in the trajectory. The following proof shows that given a non-degenerate instrument variation and controllability, the space covered by the $C(\tau)$ is guaranteed to stretch $\mathbb{R}^a$.

Lyapunov condition. We first study the stationary covariance matrix Σ, which satisfies the Lyapunov equation

$$A\Sigma + \Sigma A^\top = -\sigma\sigma^\top, \tag{35}$$

where

$$A = \begin{pmatrix} \beta_{yy} & \beta_{yx} & 0 & \beta_{yu} \\ \beta_{xy} & \beta_{xx} & \beta_{xz} & \beta_{xu} \\ 0 & 0 & \beta_{zz} & 0 \\ 0 & 0 & 0 & \beta_{uu} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} & \Sigma_{YZ} & \Sigma_{YU} \\ \Sigma_{YX}^\top & \Sigma_{XX} & \Sigma_{XZ} & \Sigma_{XU} \\ \Sigma_{YZ}^\top & \Sigma_{XZ}^\top & \Sigma_{ZZ} & 0 \\ \Sigma_{YU}^\top & \Sigma_{XU}^\top & 0 & \Sigma_{UU} \end{pmatrix}, \tag{36}$$

and the block-diagonal diffusion matrix is given by

$$\sigma\sigma^\top = \operatorname{diag}\left(\sigma_y\sigma_y^\top, \sigma_x\sigma_x^\top, \sigma_z\sigma_z^\top, \sigma_u\sigma_u^\top\right). \tag{37}$$

Evaluating sub-blocks of Eq. (35) for $C(0) = c_0 = (\Sigma_{YZ}, \Sigma_{XZ})$ and $C_{ZZ}(0) = \Sigma_{ZZ}$ results in:

$$\beta_{zz}\,\Sigma_{ZZ} + \Sigma_{ZZ}\,\beta_{zz}^\top \;=\; -\,\sigma_z\sigma_z^\top. \tag{38}$$

and

$$A_c\,c_0 \;+\; c_0\,\beta_{zz}^\top \;+\; b\,\Sigma_{ZZ} \;=\; 0. \tag{39}$$

Span analysis. Next, we assume the existence of a direction p that defines a subspace not covered by the trajectory, such that

$$p^\top C(\tau) = 0 \quad \text{for all } \tau \geq 0. \tag{40}$$

Now, we define $f(\tau) := p^\top C(\tau)$. Since $C(\tau)$ is real-analytic in τ, so is f and the derivatives of f exist everywhere on $[0, \infty)$. Equation (40) implies that every derivative of f is zero, as the direction p is never observed along the entire trajectory.

$$f^{(k)}(0) = 0 \quad \text{for all } k \geq 0.$$

We will show that this condition cannot be fulfilled under controllability by using the evaluations of $f^{(k)}(0)$ and the stationary Lyapunov conditions derived earlier, Eq. (38) and Eq. (39).

We now investigate the trajectories of the covariances C and the instrument's lagged covariances jointly.

$$\dot{C} = A_c C(\tau) + b C_{ZZ}(\tau), \qquad C(0) = c_0, \qquad C_{ZZ}(\tau) = e^{\beta_{zz}\tau}\Sigma_{ZZ}. \tag{41}$$

Starting with $f(0) = 0$ we obtain the condition $p^\top c_0 = 0$. From the first derivative, we obtain:

$$p^\top \dot{C}(0) = 0 = p^\top A_c c_0 + p^\top b \Sigma_{ZZ} = p^\top (A_c c_0 + b\Sigma_{ZZ}) \tag{42}$$

For the second derivative, we obtain after inserting the first derivative:

$$p^\top \ddot{C}(0) = 0 = p^\top A_c (A_c c_0 + b\Sigma_{ZZ}) + p^\top b \beta_{zz} \Sigma_{ZZ} \tag{43}$$

Separately, we now multiply Eq. (39) with $p^\top A_c$:

$$p^\top (A_c^2 c_0 + A_c c_0 \beta_{zz}^T + A_c b \Sigma_{ZZ}) = 0 \tag{44}$$

and now use Eq. (42) to obtain:

$$p^\top (A_c^2 c_0 - b\Sigma_{ZZ}\beta_{zz}^T + A_c b \Sigma_{ZZ}) = 0 \tag{45}$$

Subtracting this equation from Eq. (43):

$$p^\top (b\beta_{zz}\Sigma_{ZZ} + b\Sigma_{ZZ}\beta_{zz}^\top) = 0 \tag{46}$$

Now we can insert Eq. (38):

$$p^\top b\sigma_z\sigma_z^T = 0 \tag{47}$$

Since we assume a non-degenerate diffusion for Z, we conclude: $p^\top b = 0$.

Higher-order derivatives. We can now similarly cascade through the expressions for the higher derivatives. The general form for a higher derivative is given by:

$$C^{(m)}(0) = A_c^m c_0 + \sum_{j=0}^{m-1} A_c^{m-j-1} b\beta_{zz}^j \Sigma_{ZZ} \tag{48}$$

We now proceed by induction and assume as induction hypothesis that $p^\top A_c^i b = 0$ for all $i \in \{0, \dots, k-1\}$. For arbitrary $k+1, k+2$, we obtain first observe

$$p^\top C^{(k+2)}(0) = 0 = p^\top A_c^{k+2} c_0 + \sum_{j=0}^{k+1} p^\top A_c^{k+1-j} b\beta_{zz}^j \Sigma_{ZZ} \tag{49}$$

We now use the induction hypothesis to kill all terms of j for $j \geq 2$, so that only the following remains:

$$0 = p^\top A_c^{k+2} c_0 + p^\top A_c^{k+1} b\Sigma_{ZZ} + p^\top A_c^k b\beta_{zz}\Sigma_{ZZ} \tag{50}$$

As before, we separately multiply Eq. (39) by A_c^{k+1} and $p^\top$:

$$p^\top A_c^{k+2} c_0 + p^\top A_c^{k+1} c_0 \beta_{zz}^\top + p^\top A_c^{k+1} b\Sigma_{ZZ} = 0 \tag{51}$$

To obtain the value for $p^\top A_c^{k+1} c_0$, we multiply the general derivative at $m = k+1$ and use the inductive hypothesis for all terms with $j \geq 1$:

$$p^\top C^{(k+1)}(0) = 0 = p^\top A_c^{k+1} c_0 + p^\top A_c^k b\Sigma_{ZZ} \tag{52}$$

Substituting this yields:

$$p^\top A_c^{k+2} c_0 - p^\top A_c^k b\Sigma_{ZZ}\beta_{zz}^T + p^\top A_c^{k+1} b\Sigma_{ZZ} = 0 \tag{53}$$

We now subtract Eq. (50) from Eq. (53) and obtain:

$$p^\top A_c^k b(\beta_{zz}\Sigma_{ZZ} + \Sigma_{ZZ}\beta_{zz}^\top) = 0 \tag{54}$$

Again using Eq. (38):

$$p^\top A_c^k b\sigma_z\sigma_z^\top = 0 \tag{55}$$

which as before under a non-degenerate instrument diffusion implies $p^\top A_c^k b = 0$ and thus completes the induction.

By Cayley-Hamilton, we can express A_c^a and higher order polynomials through the lower order polynomials, thus we do not obtain further conditions from higher derivatives and only need to consider all derivatives up to $a+1$.

$$p^\top \mathcal{C} = p^\top \begin{pmatrix} b & A_c b & \cdots & A_c^{a-1} b \end{pmatrix} = 0 \tag{56}$$

with $\mathcal{C} \in \mathbb{R}^{a \times ad_Z}$.

If $\mathcal{C}$ has full row rank, it has a trivial left nullspace, and no such direction $p^\top$ exists.

Consequently, the full rank of $\mathcal{C}$ guarantees that the trajectory of $C(\tau)$ spans $\mathbb{R}^a$, and the degeneracy of a badly chosen initial condition cannot occur.

E.3 Step 3: Sampling points $\boldsymbol{\tau}$

If the trajectory of $C(s)$ spans $\mathbb{R}^a$, the same is true for the integrated covariance curve, as any direction orthogonal to its integrals would also need to be orthogonal to the covariance curve itself. Hence, tuples of a timepoints, where $\operatorname{rank}(G) = a$ exist. Degenerate choices of sampling points can still occur. Intuitively, this happens, when the sampled times repeatedly hit an eigenfrequency of the system, such that the columns of G are linearly dependent. However, such sampling points lie on a lower-dimensional subset of the sampling space. Thus, as the timepoints are drawn from a continuous distribution over positive sampling times, this set has Lebesgue measure zero.

Consequently a randomly drawn lags almost surely give $\operatorname{rank}(G(\boldsymbol{\tau})) = a$ and identify K.

Note: Sufficiency of less timepoints. Each instrument process Z contributes one moment condition at every sampled timepoint τ, giving a total of ad_z moment conditions for a sampled timepoints. Compared to the a parameters that are identified, this generally already puts the linear equation system Eq. (14) in the overidentified setting, if $d_z > 1$. The reason for this is that the correlation with different instrument processes might excite different and partially overlapping eigenmodes of matrix A_c. Under the controllability criterion, each instrument excites at most a eigenmodes, and all instruments jointly excite exact a eigenmodes. Thus, given controllability, the number of timepoints required for the just identified setting lies between a/d_Z and a. Sampling from all instrument processes at a times thus gives a conservative identification strategy.

F Estimator Consistency in the Linear Case

We present two estimators based on either the integral weak form formulation of the covariance ODE or the differential formulation. We show consistency for the integral form of the estimator.

Integral linear IV estimator. The fixed number of lags $0 < \tau_1 < \cdots < \tau_L$, with $L > a$, is drawn from a continuous distribution on $(0, \tau_{max}]$, as defined in Proposition 2. Further, denote

$$V_t := \begin{pmatrix} Y_t \\ X_t \end{pmatrix}, \qquad C(s) := \mathbb{E}[V_{t+s} Z_t^\top] = \begin{pmatrix} C_{YZ}(s) \\ C_{XZ}(s) \end{pmatrix}. \tag{57}$$

We define

$$B_\ell := C_{YZ}(\tau_\ell) - C_{YZ}(0), \qquad G_\ell := \int_0^{\tau_\ell} C(s)\, ds, \tag{58}$$

and stack the blocks as $B = (B_1, \ldots, B_L)$ and $G = (G_1, \ldots, G_L)$. Integrating the covariance ODE gives

$$B = KG, \qquad \text{with } K = (\beta_{yy}, \beta_{yx}). \tag{59}$$

We assume (Y, X, Z) are observed at times $t_1 = 0, t_2, \ldots, t_n$, with positive, i.i.d. gaps $\Delta_i = t_{i+1} - t_i$ having known density q, which is bounded, positive on $(0, T]$, continuous and independent of the process. We may then define k-timestep gaps as $\Delta_{i,k} := t_{i+k} - t_i$ and, considering the next m samples, compute the expected number of successive samples at each time gap s as $\rho(s) := \sum_{k \geq 1}^{m} q^{*k}(s)$. We choose $mT > \tau_{max}$, such that our sampling interval covers the maximum lag to be evaluated. Each integrated block is then estimated by

$$\widehat{G}_{\ell,n} := \frac{1}{n-m} \sum_{i=1}^{n-m} \sum_{k=1}^{m} \frac{\mathbb{1}\{\Delta_{i,k} \leq \tau_\ell\}}{\rho(\Delta_{i,k})} V_{t_{i+k}} Z_{t_i}^\top. \tag{60}$$

Following Walk (2010), we define a kernel $\mathcal{K}$ and bandwidth $h_n > 0$, with $\mathcal{K}_h(v) := h^{-1}\mathcal{K}(v/h)$. The bandwidths $h_n > 0$ are chosen as $h_n = c_m n^{-\gamma}$, with constants $c_m > 0$ and $0 < \gamma < 1$. We use this kernel to define a Nadaraya-Watson estimator to nonparametrically regress the covariance function at the chosen lags τ_ℓ. The covariance function at the lag does not admit an estimator in the form of a simple sample average, as we almost surely do not observe a sample at exactly τ_ℓ, thus, we use this nonparametric first stage. The regression function takes the form of $g(\Delta) = \mathbb{E}(Y_{t+\Delta}\, Z_t^T | \Delta)$. Further, we define $J_\ell := \{k \in \{1, \ldots, m\} : q^{*k}(\tau_\ell) > 0\}$ which selects the subsequent point indices,

whose lag distributions are positive at the target lag τ_ℓ. We then estimate the covariance function at lag zero and each sampled lag τ_ℓ as follows:

$$\widehat{C}_{YZ,n}(\tau_\ell) := \frac{\sum_{i=1}^{n-m}\sum_{k\in J_\ell}\mathcal{K}_{h_n}(\Delta_{i,k}-\tau_\ell)Y_{t_{i+k}}Z_{t_i}^\top}{\sum_{i=1}^{n-m}\sum_{k\in J_\ell}\mathcal{K}_{h_n}(\Delta_{i,k}-\tau_\ell)}, \qquad \widehat{C}_{YZ,n}(0) := \frac{1}{n}\sum_{i=1}^{n} Y_{t_i}Z_{t_i}^\top. \tag{61}$$

Further, we denote

$$\widehat{B}_{\ell,n} := \widehat{C}_{YZ,n}(\tau_\ell) - \widehat{C}_{YZ,n}(0), \qquad \widehat{B}_n := (\widehat{B}_{1,n},\dots,\widehat{B}_{L,n}), \tag{62}$$

and $\widehat{G}_n := (\widehat{G}_{1,n},\dots,\widehat{G}_{L,n})$. We fit the sample moments by least squares. The analytic minimizer of Eq. (59) can be obtained as

$$\widehat{K}_n := \widehat{B}_n\widehat{G}_n^\top(\widehat{G}_n\widehat{G}_n^\top)^{-1}, \tag{63}$$

which gives the final parameter estimates for the integral linear estimator.

Theorem 3 (Consistent estimator in the OU setting)**.** *In the setting of Proposition 2, the estimator in Eq.* (63) *is consistent, i.e.,* $\widehat{\beta}_{yy,n} \xrightarrow{p} \beta_{yy}$ *and* $\widehat{\beta}_{yx,n} \xrightarrow{p} \beta_{yx}$ *as* $n \to \infty$.

Proof. We start by showing the consistency of the integrated blocks. Stationarity and ergodicity, together with finite second moments of the process imply that $\widehat{C}_{YZ,n}(0) \xrightarrow{p} C_{YZ}(0)$.

To show consistent estimation of $\widehat{G}$, we first notice that $\mathbb{E}[V_{t_{i+k}}Z_{t_i}^\top \mid \Delta_{i,k}=s] = C(s)$ and write

$$\begin{aligned}
G_\ell &= \int_0^{\tau_\ell} C(s)\,ds \\
&= \sum_{k=1}^{m}\int_0^{\tau_\ell}\frac{\mathbb{E}[V_{t+s}Z_t^\top]}{\rho(s)}q^{*k}(s)\,ds \\
&= \sum_{k=1}^{m}\mathbb{E}\left[\frac{\mathbb{1}\{\Delta_{i,k}\le\tau_\ell\}}{\rho(\Delta_{i,k})}\mathbb{E}\left[V_{t_{i+k}}Z_{t_i}^\top \mid \Delta_{i,k}\right]\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{m}\frac{\mathbb{1}\{\Delta_{i,k}\le\tau_\ell\}}{\rho(\Delta_{i,k})}V_{t_{i+k}}Z_{t_i}^\top\right] = \mathbb{E}[W_{i,\ell}].
\end{aligned} \tag{64}$$

We then estimate the $\widehat{G}_{\ell,n}$ via the sample average as described in Eq. (60). Using the independence of the observation times from the process implies that the process $W_{i,\ell}$ is stationary and ergodic. Further, we can show integrability by using Cauchy Schwarz and stationarity

$$\mathbb{E}|W_{i,\ell}| \le \int_0^{\tau_\ell}\mathbb{E}|V_{t+s}Z_t^T|ds \le \tau_\ell\sqrt{\mathbb{E}|V_t|^2\mathbb{E}|Z_0|^2} < \infty, \tag{65}$$

where the first inequality comes from using the triangle inequality over the sum up to m timepoints. This thus allows to invoke Birkhoff's ergodicity theorem (Durrett, 2019) and yields $\widehat{G}_{\ell,n} \xrightarrow{p} G_\ell$.

Following (Walk, 2010, Thm. 3b), we show the consistency of the kernel regression of $C_{YZ}(\tau_\ell)$ with either a uniform or Gaussian kernel. First notice that the first equation in Eq. (61) can be written as a weighed sum of Nadaraya-Watson estimators for each $k \in J_\ell$.

$$\widehat{C}_{YZ,n}(\tau_\ell) = \sum_{k\in J_\ell} w_{k,\ell}\widehat{C}^{(k)}_{YZ,n}(\tau_\ell), \qquad w_{k,\ell} := \frac{\sum_{i=1}^{n-m}\mathcal{K}_{h_n}(\Delta_{i,k}-\tau_\ell)}{\sum_{j\in J_\ell}\sum_{i=1}^{n-m}\mathcal{K}_{h_n}(\Delta_{i,j}-\tau_\ell)}, \tag{66}$$

with

$$\widehat{C}^{(k)}_{YZ,n}(\tau_\ell) := \frac{\displaystyle\sum_{i=1}^{n-m}\mathcal{K}_{h_n}(\Delta_{i,k}-\tau_\ell)Y_{t_{i+k}}Z_{t_i}^\top}{\displaystyle\sum_{i=1}^{n-m}\mathcal{K}_{h_n}(\Delta_{i,k}-\tau_\ell)}. \tag{67}$$

For each k the sequence of samples $(\Delta_{i,k}, Y_{t_{i+k}}Z_{t_i}^\top)$ is geometrically mixing, as seen from the OU transition matrix, which takes the form of a matrix exponential. Geometric mixing is stronger than

the polynomial mixing required by by Walk (2010). Thus, we may choose $h_n = c_m(n-m)^{-\gamma}$ with any $0 < \gamma < 1$ and $c_m > 0$. Consequently, each of the estimators $\widehat{C}^{(k)}_{YZ,n}(\tau_\ell)$ fulfills the conditions presented in Walk (2010) and is consistent. As the weights are nonnegative and sum up to one, the average of these estimators and thus the final estimate for the endpoint covariance are consistent as well. This yields $\widehat{C}_{YZ,n}(\tau_\ell) \xrightarrow{p} C_{YZ}(\tau_\ell)$, as well as $\widehat{B}_{\ell,n} \xrightarrow{p} B_\ell$ through the continuous mapping theorem, and for the stacked matrices $\widehat{G}_n \xrightarrow{p} G$ and $\widehat{B}_n \xrightarrow{p} B$.

If the identification condition defined in Eq. (13) holds, matrix G has full rank. Hence, GG^T is almost surely invertible. We use the least-squares solution for the final estimate of the parameter matrix. Following the continuous-mapping theorem, this function is continuous around an open set of the true matrix GG^T. Thus, given the consistent estimation of $\widehat{B}$ and $\widehat{G}$, the estimate of K is consistent.

$$\widehat{K}_n \xrightarrow{p} BG^\top (GG^\top)^{-1} = K. \tag{68}$$

□

On a simulation grid. The nonparametric first stage using the kernel-based estimation of $C_{YZ}(\tau_\ell)$ can be replaced by a simple sample average in case of working on a simulation grid with random dropouts defined by a simulation interval δ_{sim} and a uniformly random dropout percentage per point p_{drop}, as in such a setup each τ_ℓ has a non-zero chance of being sampled. In this case, we can fix the kernel bandwidth h to $0 < h < \delta_{sim}$, which retrieves the standard sample average for the covariance estimation at the given lag τ_ℓ. We note that this discrete version does not keep the previously presented consistency result due to numeric approximation errors.

Differential linear IV estimator. As a second estimator, we propose the differential estimator. We define the differential IV estimator by binning the covariances across lags to estimate the covariance curves on a fine grid in the first stage. In the second stage we then differentiate the curves at the lags $0 < \tau_1 < \cdots < \tau_L$, which are sampled as previously described, and then solve an equation system based on the differential form of the covariance ODE via least squares. The preceding consistency argument for the integral estimator does not extend directly to this differential form. Establishing consistency would require a separate analysis of the derivative estimator given the specific smoothing and sampling schemes.

G Influence of a non-diagonal diffusion matrix.

In this section, we discuss the effects of a non-diagonal diffusion matrix for the linear case specifically. As mentioned in Section A, the moment conditions themselves remain valid assuming a non-diagonal diffusion matrix, as long as exogeneity of the instrument is fulfilled. Specifically, in case of an OU process this requires $\left(\sigma(S_t)\sigma(S_t)^\top\right)_{zu} = 0$.

However, we note that the identification argument presented in Proposition 2 might not hold in case of diffusion mixing of Y and Z. In this case, even under the required controllability condition, the initial condition derived via the Lyapunov equation, as done in Section E, can be unluckily chosen such that the resulting covariance curves do not cover the required state space of $\mathbb{R}^a$. Consequently, in this case the moment equations do not necessarily identify the drift function anymore.

Nevertheless, we empirically investigated the effect of two special cases of non-diagonal diffusion matrices in OU-processes on our proposed estimation, as shown in Table 2.

In the first case, we mix the diffusions between X and Y with increasing strength up to the level of the diagonal entry, thus $\sigma_{xy,k} = k\sigma_x$ and $\sigma_{yx,k} = k\sigma_y$. The results show that with increasing mixing strength the MSE of the estimation increases. Stronger mixing increases the variance of the outcome and thereby increases the variance of the moment conditions, resulting in lower estimation efficiency.

In a second experiment, we mix the diffusions between Z and X. Interestingly, moderate mixing may decrease the MSE of the estimation. The increased coupling between process X and Z increases the effective instrument strength and thus allows for a better effect estimation. Beyond a certain mixing level, however, the increase in the overall estimation variance dominates this gain in relevance, causing the estimation error to increase again. Further, the experiments confirm that a dense covariance matrix violates the assumptions, resulting in a higher estimation error.

Table 2: Total MSE of the IV-SDE estimator under different diffusion configurations and mixing strengths.

Configuration	IV-SDE MSE (total)
$x - y,\ k = 0.0$	0.0036
$x - y,\ k = 0.3$	0.0068
$x - y,\ k = 0.5$	0.0151
$x - y,\ k = 0.9$	0.0162
$z - x,\ k = 0.3$	0.0031
$z - x,\ k = 0.5$	0.0009
$z - x,\ k = 0.9$	0.0011
Violation (dense, $k = 0.6$)	0.2306

H THE SIGNATURE KERNEL

The signature kernel (Lee & Oberhauser, 2025; Manten et al., 2025) relies on the signature features $\mathbf{S}(X) \in \mathcal{T}(\mathbb{R}^d)$ of continuous bounded variation paths $X \in C([0,T], \mathbb{R}^d)$, where $\mathcal{T}$ describes the tensor algebra over $\mathbb{R}^d$. These features can be understood as a sequence of n-fold iterated integrals of the path X against itself, which together give a universal feature map of continuous paths. The signature kernel $k_z : (C([0,T], \mathbb{R}^d) \times C([0,T], \mathbb{R}^d) \to \mathbb{R}$ between paths X, Y is the tensor-algebra inner product of their signatures, $k_z(X,Y) = \langle \mathbf{S}(X), \mathbf{S}(Y) \rangle$. The kernel can be efficiently evaluated without explicitly computing the signature features by using the kernel trick, which then requires solving an associated hyperbolic PDE (Salvi et al., 2021).

I EXPERIMENTAL DETAILS

The linear experiments were performed on CPU. The nonlinear experiments were run on a single A100 GPU in under two hours for all evaluations.

I.1 LINEAR (OU) EXPERIMENTS

The linear experiments were all simulated with the following SDE system:

$$d\mathbf{S}_t = \underbrace{\begin{pmatrix} \beta_{yy} & \beta_{yx} & 0 & \beta_{yu} \\ \beta_{xy} & \beta_{xx} & \beta_{xz} & \beta_{xu} \\ 0 & 0 & \beta_{zz} & 0 \\ 0 & 0 & 0 & \beta_{uu} \end{pmatrix}}_{A} \mathbf{S}_t\, dt + \underbrace{\begin{pmatrix} \sigma_y & 0 & 0 & 0 \\ 0 & \sigma_x & 0 & 0 \\ 0 & 0 & \sigma_z & 0 \\ 0 & 0 & 0 & \sigma_u \end{pmatrix}}_{\boldsymbol{\sigma}} \begin{pmatrix} dW_t^y \\ dW_t^x \\ dW_t^z \\ dW_t^u \end{pmatrix}. \tag{69}$$

where $\beta_{ab} \in \mathbb{R}^{d_a \times d_b}$ for $a, b \in \{y, x, z, u\}$. Unless stated otherwise, we use $d_y = d_x = d_z = d_u = 1$.

Parameter sampling. All non-zero drift coefficients are sampled uniformly from $[-1, -0.2] \cup [0.2, 1]$, ensuring existing edges of the sparse connectivity graph are bounded away from zero. Diffusion coefficients are sampled uniformly from $[0.2, 0.5]$. All sampled systems are filtered for stability (negative real parts of all eigenvalues of A) and controllability (full rank of the controllability matrix $\mathcal{C}$ defined in Eq. (13)).

Simulation. For the left plot in Figure 2, trajectories are simulated by sampling from the analytic transition densities at irregularly sampled observation times. The irregular observation density q is set to be uniform on an interval of $(0, 0.4]$. We evaluate each method on 8000 sample paths of length $T_{max} = 20\,\mathrm{s}$, initialized from the analytically determined stationary distribution. For the plots in Figure 2 in the middle and right, we sampled the trajectories on a simulation grid of $0.01s$, kept a random small subset of points and used the IV-estimator on a simulation grid, as described in Section F.

Experiment-specific settings. In the confounding strength experiment, we simultaneously vary β_{xu} and β_{yu} over the range $[0.1, 1.0]$, sampling 200 random systems across all confounding values.

For the covariate dimension experiment, we sample 30 systems in total. For the instrument strength experiment, we sample 80 systems in total, varying β_{xz} over $[0.1, 0.8]$.

Baselines. We note that the only available baselines, besides the naive MLE estimator, for causal IV methods in time are presented in Mogensen (2023) and Thams et al. (2024), both of which are designed for discrete time methods. As both were not designed to handle irregularly spaced timepoints in continuous-time systems, we cannot directly compare them fairly on the same data. Thus, besides applying the $\mathrm{VAR}(1)$ estimator naively on irregular samples, we also sample the same number of regularly spaced timepoints from the simulated trajectories, and apply the discrete-time estimators on these.

To construct the MLE estimator, we assume the full system to be Markov with respect to just the observed states (Y, X, Z) and use the analytically available and exact Gaussian transition densities to estimate the drift parameters.

We applied the discrete $\mathrm{VAR}(1)$ estimator on regularly sampled timepoints with spacing of $0.2s$ and match the number of continuous-time timepoints. This yields an estimate for the discrete-time transition matrix. We estimate the continuous-time target parameters from the transition matrix by applying the matrix logarithm and dividing by the sampling interval. We also naively apply the discrete time estimator on the irregular timepoints.

We further compare to the discrete-time IV method presented Mogensen (2023) which estimates the *normalized parameters* using integrated covariance matrices. Here, the *normalized parameters* correspond to the ratio β_{yx}/β_{yy}. The method does not identify the target parameters themselves, thus we need to compare our results to their ratio estimate. We also apply this method on the regularly spaced timepoints.

Extended results discussion. Table 3 shows the mean coefficient error across all 200 simulated systems, with 20 sampled systems per confounding strength and irregular spacing. The results clearly demonstrate smaller estimation errors of the IVSDE estimators under stronger confounding. Both IV estimators, the differential IVSDE estimatior (IVSDE-diff), which bins the data across 0.05s bins, and the continuous-time integral IVSDE estimator (IVSDE-int) presented in Section F perform similarly well on the data. In case of minimal or no confounding, the MLE yields slightly lower estimation errors, as it uses the full transition densities and is thus more efficient. The naively applied $\mathrm{VAR}(1)$ method fails on irregularly spaced data.

The results for the direct ratio estimation presented in Table 4 suffer from the same discrete sampling bias as the discrete $\mathrm{VAR}(1)$ method. As a consequence, using the proposed IVSDE method to first individually estimate the parameters and then construct ratio estimates still outperforms this direct ratio estimation, despite being a more complex two-step estimation.

Table 3: Mean per-system parameter RMSE over 20 systems each with sample standard deviations reflecting variability across systems of varying difficulty. (reg.) denotes that the regularly sampled timepoints were used.

Confounding strength	MLE full	VAR-1	VAR-1 (reg.)	IVSDE-diff	IVSDE-int
0.1	**0.0069** $\pm$ 0.01	0.6590 $\pm$ 0.73	0.0516 $\pm$ 0.03	0.0104 $\pm$ 0.01	0.0123 $\pm$ 0.01
0.2	0.0379 $\pm$ 0.02	1.2766 $\pm$ 1.33	0.0625 $\pm$ 0.03	**0.0148** $\pm$ 0.01	0.0192 $\pm$ 0.02
0.3	0.0691 $\pm$ 0.05	1.5407 $\pm$ 1.64	0.0576 $\pm$ 0.03	**0.0182** $\pm$ 0.02	0.0202 $\pm$ 0.03
0.4	0.1244 $\pm$ 0.07	1.1334 $\pm$ 1.04	0.0464 $\pm$ 0.03	0.0244 $\pm$ 0.02	**0.0176** $\pm$ 0.01
0.5	0.1146 $\pm$ 0.08	0.8761 $\pm$ 0.73	0.0431 $\pm$ 0.02	**0.0097** $\pm$ 0.01	0.0137 $\pm$ 0.01
0.6	0.1746 $\pm$ 0.09	0.8853 $\pm$ 0.85	0.0613 $\pm$ 0.03	0.0214 $\pm$ 0.01	**0.0184** $\pm$ 0.01
0.7	0.2065 $\pm$ 0.10	0.9870 $\pm$ 0.81	0.0456 $\pm$ 0.02	**0.0112** $\pm$ 0.01	0.0136 $\pm$ 0.02
0.8	0.2112 $\pm$ 0.12	0.6470 $\pm$ 0.30	0.0506 $\pm$ 0.03	0.0225 $\pm$ 0.03	**0.0126** $\pm$ 0.01
0.9	0.2397 $\pm$ 0.10	1.0516 $\pm$ 0.82	0.0675 $\pm$ 0.04	0.0255 $\pm$ 0.04	**0.0185** $\pm$ 0.02
1.0	0.2359 $\pm$ 0.08	0.9859 $\pm$ 1.15	0.0616 $\pm$ 0.04	0.0309 $\pm$ 0.03	**0.0290** $\pm$ 0.03

Table 4: Ratio RMSE. Ratio RMSE across varying confounding strength. Each value represents the mean over 20 sampled systems.

Confounding strength	MLE full	VAR-1	VAR-1 (reg.)	IVSDE-diff	IVSDE-int	Direct ratio (reg.)
0.1	0.0472	0.8777	0.3202	**0.0166**	0.0174	0.1005
0.2	0.1711	1.9936	0.1894	**0.0504**	0.0526	0.0613
0.3	1.2192	2.2831	0.1939	0.0772	**0.0649**	0.3011
0.4	0.3261	2.1911	0.2962	**0.0218**	0.0233	0.1542
0.5	0.5827	1.6280	0.2167	**0.0338**	0.0348	0.1127
0.6	1.1755	1.2828	0.2407	**0.0344**	0.0442	0.1520
0.7	2.6015	3.6445	0.2048	0.0305	**0.0261**	0.2258
0.8	15.2417	3.0491	0.5627	0.0615	**0.0566**	0.2535
0.9	84.2825	2.4190	0.3090	0.0916	**0.0794**	0.1910
1.0	1.7591	1.4126	0.2655	0.0462	**0.0372**	0.2251

I.2 Kernel-based estimation of a basis-function expansion

Simulation. We simulate paths from the following nonlinear SDE system, where all processes are one-dimensional except the instrument process Z_t, which is three-dimensional ($d_z = 3$):

$$dY_t = \left(\alpha Y_t + \beta X_t + \sum_{i=1}^{p} w_i e^{-(X_t - c_i)^2/\sigma_{\text{rbf}}^2} - \sum_{i=1}^{p} w_i e^{-(X_t + c_i)^2/\sigma_{\text{rbf}}^2} + \beta_{yu} U_t \right) dt + \sigma_y \, dW_t^y \tag{70}$$

$$dX_t = (\beta_{xz} Z_t + \beta_{xx} X_t + \beta_{xu} U_t) \, dt + \sigma_x \, dW_t^x \tag{71}$$

$$dZ_t = \beta_{zz} Z_t \, dt + \sigma_z \, dW_t^z \tag{72}$$

$$dU_t = -a_u U_t \left(U_t^2 - 0.5^2 \right) dt + \sigma_u \, dW_t^u \tag{73}$$

where the confounding process U_t follows a nonlinear mean-reverting dynamics, resulting in a non-Gaussian distribution of X. The drift of Y is parameterized using two linear expressions in X and Y, and pairs of radial basis functions symmetric about the origin, which preserves the zero mean of process Y. As described in the main text, we assume access to densely sampled observations and therefore use the time series directly on the simulation grid. In settings with coarser sampling, simulation-based estimation methods would be required to reliably approximate the integral terms in the residual $R_{t,t+h;\theta}$.

Basis function parametrization. We parametrize the target drift $f_{y;\theta}$ using a basis function expansion consisting of $p = 4$ radial basis functions (RBFs) with centers $\{c_i\}_{i=1}^4$ placed uniformly over $[0.2, 2]$, covering approximately 96% of the marginal distribution of X_t under stationarity, together with two linear terms in X_t and Y_t. The bandwidth of each RBF is set to the ground truth $\sigma_{\text{rbf}} = 0.35$. The resulting parametrization is:

$$f_{y;\theta}(Y_t, X_t) = \alpha Y_t + \beta X_t + \sum_{i=1}^{4} w_i e^{-(X_t - c_i)^2/\sigma_{\text{rbf}}^2} - \sum_{i=1}^{4} w_i e^{-(X_t + c_i)^2/\sigma_{\text{rbf}}^2} \tag{74}$$

where $\theta = (\alpha, \beta, w_1, \ldots, w_4) \in \mathbb{R}^6$ are the parameters to be estimated.

Closed-form estimation. Given the parametrization $f_{y;\theta}(Y_t, X_t) = B_t^\top \theta$, which is linear in parameters, where $B_t \in \mathbb{R}^p$ collects the basis function evaluations at time t, the residual $R_{t,t+h;\theta}$ is linear in θ:

$$R_{t,t+h;\theta} = Y_{t+h} - Y_t - \int_t^{t+h} B_s^\top \theta \, ds = Y_{t+h} - Y_t - \bar{B}_{t,h}^\top \theta \tag{75}$$

where $\bar{B}_{t,h} = \int_t^{t+h} B_s \, ds \in \mathbb{R}^p$ is the time-integrated basis function vector over the horizon h, approximated via numerical quadrature on the dense simulation grid, and $\mathbf{e}$ describes the difference $Y_{t+h} - Y_t$. As a result, the V-statistic risk $\widehat{\mathcal{R}}_z(\theta)$ in Eq. (19) reduces to a quadratic objective in θ, and the minimization problem admits the closed-form kernel ridge regression solution:

$$\hat{\theta} = \left(\bar{\mathbf{B}}^\top K_z \bar{\mathbf{B}} + \lambda I \right)^{-1} \bar{\mathbf{B}}^\top K_z \mathbf{e} \tag{76}$$

where $\bar{\mathbf{B}} \in \mathbb{R}^{n\times p}$ stacks the integrated basis vectors $\bar{B}_{t_i,h}$ across all n sampled path sections, $K_z \in \mathbb{R}^{n\times n}$ is the signature kernel Gram matrix evaluated on the instrument paths $\{Z^i_{[0,t]}\}_{i=1}^n$, $\mathbf{e} \in \mathbb{R}^n$ collects the observed outcome increments $Y^i_{t+h} - Y^i_t$, and $\lambda > 0$ is the regularization parameter. This allows the full estimation to be performed in a single linear solve, without iterative optimization.

Sampling scheme. We sample 5000 path segments, each consisting of the forward prediction horizon h and the lookback time ℓ over Z, which is used as the instrument path. We use the following hyperparameters: prediction horizon $h = 0.3\,\mathrm{s}$, stride between consecutive sections $= 2\,\mathrm{s}$, and lookback window for the instrument path $\ell = 1.5\,\mathrm{s}$ and $\lambda = 0.001$.

I.3 Nonlinear drift function with a neural network

Furthermore, we show that the kernel-based estimation scheme also allows for a neural parameterization of the drift function. We keep the same setup for the instruments and covariates as described in the previous section. The drift of Y now depends on X via a sigmoid nonlinearity S, resulting in the following:

$$dY_t = (\alpha Y_t + S(kX_t) + U_t)\,dt + \sigma_y\,dW^y_t \tag{77}$$

where we choose k=4. We parameterize f_θ with a small MLP with two inputs (X_t, Y_t), two hidden layers of dimension 16, and tanh activation functions.

We sample $n = 10000$ path segments and optimize the empirical risk in Eq. (19) with the following hyperparameters using the Adam optimizer: integration horizon $h = 0.3\,\mathrm{s}$, lookback $\ell = 1.5\,\mathrm{s}$, weight decay $\lambda = 1 \times 10^{-4}$, learning rate $\eta = 10^{-3}$, and $N = 1000$ gradient steps.

I.4 Diabetes dataset.

Setup. We setup a semi-synthetic experiment to demonstrate the applicability of our proposed method on problems of real-world complexity and relevance. We motivate our setting from pharmacokinetics / pharmacodynamics models of type 2 diabetes. These models describe the dynamics and interactions of insulin and glucose concentrations in the blood. In type 2 diabetes specifically, one specific mechanism of the disease is found in the insulin resistance build up by patients, as muscle and fat cells take up glucose less efficiently. Consequently, identifying the insulin sensitivity in such setting allows to judge a patient's personal disease progression and guide future insulin dosing. Both insulin production as well as blood glucose concentration can be confounded by unobserved gut glucose. Here, the insulin production is stimulated by incretin hormones, such as GLP-1, which are secreted by enteroendocrine cells in the gut, while intestinal glucose absorption provides a direct additive contribution to the blood glucose concentration.

The setup thus fulfills the requirements of our setting, with scheduled external insulin dosages as instruments, the insulin concentration as the covariate process, glucose concentration as the outcome process, and the insulin sensitivity as the estimand. It illustrates our method in a system combining nonlinear interactions, feedback, and an unobserved common driver.

We follow the structural setup of the model presented in Naik et al. (2026); Subramanian et al. (2024), select the parameters for the simulation setting such that we obtain a stable SDE and set the instrument process to be a two-dimensional OU process, resulting in the following SDE system

$$\begin{aligned}
dZ_{j,t} &= m_{Z,j}\big(\mu_Z - Z_{j,t}\big)\,dt + \sigma_{Z,j}\,dW^{Z_j}_t, \qquad j = 1,\dots,k,\\
dX_t &= \left(\sum_{j=1}^{k} a_{Z,j} Z_{j,t} + c_{ux}U_t + c_{yx}Y_t - a_x X_t\right)\,dt + \sigma_x\,dW^x_t,\\
dY_t &= (c_{yu}U_t - a_y Y_t - \gamma_{xy}X_tY_t)\;dt + \sigma_y\,dW^y_t,\\
dG_t &= (\mu_g - a_g G_t)\;dt + \sigma_g\,dW^g_t,\\
dU_t &= (g_{gu}G_t - a_u U_t)\;dt + \sigma_u\,dW^u_t.
\end{aligned} \tag{78}$$

Table 5: Parameter values for the two-instrument diabetes dataset configuration.

Parameter	Description	Value
k	Number of instruments	2
μ_Z	Common instrument equilibrium mean	0
$(m_{Z,1}, m_{Z,2})$	Instrument mean-reversion rates	$(0.4,\ 0.8)$
$(a_{Z,1}, a_{Z,2})$	Instrument effects on X	$(0.35,\ 0.35)$
$(\sigma_{Z,1}, \sigma_{Z,2})$	Instrument noise amplitudes	$(0.5,\ 0.5)$
c_{ux}	Effect of U on X	0.60
c_{yx}	Effect of Y on X	0.10
a_x	Decay rate of X	0.30
σ_x	Noise amplitude of X	0.02
c_{yu}	Effect of U on Y	2.10
$\mathbf{a_y}$	Decay rate of Y	0.10
$\boldsymbol{\gamma_{xy}}$	**Insulin sensitivity**	0.15
σ_y	Noise amplitude of Y	0.02
μ_g	Constant inflow into G	0.05
a_g	Decay rate of G	0.05
σ_g	Noise amplitude of G	0.005
g_{gu}	Effect of G on U	0.05
a_u	Decay rate of U	0.05
σ_u	Noise amplitude of U	0.02

Table 6: Hyperparameter sensitivity in the diabetes model. in terms of RMSE. The standard setting across all runs is a Lookback of $0.7s$, a horizon of 0.2 and a RBF bandwidth of 0.5.

Study	Setting	RMSE		
		a_y	γ_{xy}	Overall
Lookback	0.2	0.0615	0.0126	0.0444
	0.5	0.0113	0.0117	**0.0115**
	0.7	0.0371	0.0154	0.0284
	0.9	0.0477	0.0217	0.0370
Horizon	0.1	0.0239	0.0094	0.0182
	0.2	0.0398	0.0105	0.0291
	0.3	0.0236	0.0081	0.0177
RBF bandwidth	0.3	0.0429	0.0271	0.0359
	0.5	0.0398	0.0105	0.0291
	0.8	0.0211	0.0065	0.0157
MLE baseline	—	0.1188	0.0763	0.0998

We first center the SDE system, then apply our nonlinear estimator using a basis function expansion and the signature kernel to estimate the target parameters (Eq. (76)), and then transform the estimated parameters back to the original coordinate system. We compute the final estimation error over the parameters after transforming back to the original coordinates.

Result and ablation. Furthermore, we run a small ablation over the parameters of the proposed nonlinear estimation method, specifically, the lookback, the residual integration horizon and the RBF bandwidth, as shown in Table 6. Overall, the results clearly show that the proposed IV-SDE method can estimate the causal effect with a lower error compared to the biased MLE estimate. Furthermore, the results remain stable across all settings, demonstrating the method's robustness across reasonable parameter ranges.

Although this SDE system does not fulfill the linear growth assumptions of our theoretical setup defined in Section 3, the simulated trajectories remain numerically stable under the chosen parameter settings, and our proposed estimator accurately retrieves the objective function. This shows that the estimator may also be applicable beyond the stated regularity conditions.

Alternative examples. We note that further examples for confounded dynamic settings fulfilling the main IV assumptions can be found e.g. a thermodynamic model of two bodies, where the thermal conductivity between them is the estimand, a known heat source drives one body, and the ambient temperature acts as a common confounder. Another motivating example concerns the interaction between two brain regions, where the stimulation of one region by an e.g. optogenic external impulse acts as the instrument process, while signal from a third region acts as a common confounder.