Author/s:
Wang, Yuanyuan

Title:
On the Identifiability of Linear ODE and SDE Systems for Causal Inference

Date:
2025

On the Identifiability of Linear ODE and SDE Systems for Causal Inference

Yuanyuan Wang

Submitted in total fulfilment of the requirements of the degree of

Doctor of Philosophy

School of Mathematics and Statistics
THE UNIVERSITY OF MELBOURNE

MAY 2025

ORCID: 0009-0007-1395-6959

To my husband and our two little ones – thank you for your love, company, and support.

On the Identifiability of Linear ODE and SDE Systems for Causal Inference

Yuanyuan Wang
Principal Supervisor: Mingming Gong
Co-supervisors: Wei Huang, Xi Geng

Abstract

This thesis investigates the identifiability of linear Ordinary Differential Equation (ODE) and Stochastic Differential Equation (SDE) systems within the context of causal inference. The motivation for this study arises from the widespread use of dynamical systems to model real-world phenomena in various scientific and engineering fields. Reliable identification of model parameters or system generators is crucial for understanding underlying mechanisms, predicting system behavior, and informing policy or interventions. Identifiability for linear ODE systems involves determining when system parameters can be uniquely recovered from observations, while for linear SDE systems, it focuses on identifying when the generator of the system can be uniquely recovered from the observational distribution. Such identifiability is fundamental for meaningful interpretation and reliable causal inference in dynamical systems modeling.

The thesis addresses three critical challenges: identifiability conditions for linear ODE systems from discrete observations, identifiability analysis of linear ODE systems with hidden confounders, and generator identification for linear SDE systems with additive and multiplicative noise. For linear ODE systems from discrete observations, key contributions include deriving explicit identifiability conditions, establishing asymptotic properties such as consistency and asymptotic normality of parameter estimators, and developing methods for causal structure inference. For linear ODE systems with hidden confounders and linear SDE systems, the thesis derives explicit identifiability conditions and validates these results through comprehensive simulations illustrating the correctness and robustness of the derived theoretical results.

Overall, this research advances the theoretical understanding of identifiability in linear dynamical systems, providing essential tools and frameworks for reliable causal inference in scientific and engineering applications.

Declaration

I declare that this thesis and the work presented in it are my own. I confirm that:

- the thesis comprises only my original work towards the Doctor of Philosophy except where indicated in the preface;
- due acknowledgement has been made in the text to all other material used; and
- the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Signed: Yuanyuan Wang

Date: May 2025

Preface

This thesis research was conducted at the School of Mathematics and Statistics, University of Melbourne, Australia. The main contributions are presented in Chapters 3, 4, and 5, and are based on the following publications. In accordance with the University's 2025 thesis submission requirements, I confirm that I made a substantial and leading contribution to each of these works. Specifically, I was responsible for formulating the research problems, developing the theoretical results, proving the theorems, designing and implementing the simulation experiments, and writing the manuscripts. The co-authors provided valuable guidance, feedback, and suggestions throughout the research process.

- **Yuanyuan Wang**, Wei Huang, Mingming Gong, Xi Geng, Tongliang Liu, Kun Zhang, Dacheng Tao
  Identifiability and Asymptotics in Learning Homogeneous Linear ODE Systems from Discrete Observations [1]
  ***JMLR** 2024*

- **Yuanyuan Wang**, Biwei Huang, Wei Huang, Xi Geng, Mingming Gong
  Identifiability Analysis of Linear ODE Systems with Hidden Confounders [2]
  ***NeurIPS** 2024*

- **Yuanyuan Wang**, Xi Geng, Wei Huang, Biwei Huang, Mingming Gong
  Generator Identification for Linear SDEs with Additive and Multiplicative Noise [3]
  ***NeurIPS** 2023*

Funding.

Yuanyuan Wang was supported by the *Australian Government Research Training Program Scholarship* from the University of Melbourne. The support was in terms of both *stipend* and *100% fee remission*. Conference travels were in part funded by the *Faculty of Science* and in part by the *School of Mathematics and Statistics*, both at the University of Melbourne.

Acknowledgements

This thesis marks the end of a long and challenging journey, and I am deeply grateful to the many people who have supported me along the way.

First and foremost, I would like to express my sincere gratitude to my principal supervisor, Dr Mingming Gong. Thank you for your unwavering guidance, encouragement, and belief in my potential — even during the times when I doubted myself. Your sharp insights and patient mentorship have shaped not only this thesis but also my growth as an independent researcher.

To my co-supervisors — Dr Wei Huang, Dr Xi Geng, and Dr Biwei Huang — I offer my heartfelt thanks. Each of you has contributed in a unique way to my academic development. Whether through your insightful feedback, generous time, or thoughtful discussions, I have learned and grown immensely under your guidance.

Beyond academia, this journey would not have been possible without the unconditional love and support of my family. To my husband, Liang Yang — thank you for standing by me through every high and low. Your steady strength, patience, and unshakable belief in me sustained me more than words can express. To my two children — you are my greatest joy and motivation. Thank you for reminding me of what truly matters, and for bringing light and laughter even during the most exhausting days.

To my parents and my parents-in-law — your love, sacrifices, and tireless support, especially in caring for the children, made it possible for me to pursue this work. Thank you for always being there, for never questioning the path I chose, and for giving me the space and trust to follow it through.

Finally, to my dear friends — thank you for your companionship, your listening ears, and the moments of laughter and lightness you brought into my life. Your presence made the difficult days easier and the good ones even better.

Completing this thesis was not a solo effort; it is a collective achievement built on the foundation of many quiet acts of kindness, understanding, and belief. I carry deep gratitude for each of you.

Yuanyuan Wang
Melbourne/Australia, May 2025

Contents

List of Figures

List of Tables

Abbreviations

ODE	Ordinary Differential Equation
ODEs	Ordinary Differential Equations
SDE	Stochastic Differential Equation
SDEs	Stochastic Differential Equations
DAG	Directed Acyclic Graph
CI	Confidence Interval
CR	Confidence Region
SCM	Structural Causal Model
SEM	Structural Equation Model
RDE	Random Differential Equation
NLS	Nonlinear Least Squares
MSE	Mean Squared Error

Symbols

t	Time
t_j	The j-th time point
$\boldsymbol{x}(t), \boldsymbol{z}(t)$	State of observable / latent variables of ODE systems at time t
$\boldsymbol{x}_j$	$\boldsymbol{x}(t_j)$, observable state at time t_j
$\boldsymbol{x}_0, \boldsymbol{z}_0$	Initial conditions of observable / latent variables
d, p	Dimensions of observable / latent variables
$\dot{\boldsymbol{x}}(t)$	Derivative of $\boldsymbol{x}(t)$ w.r.t. time t
X_t	State variable of SDE system at time t
A, B, G	Constant system matrices in ODE/SDE systems
G_k	Diffusion matrix for the k-th noise channel in SDEs
W_t	Standard Brownian motion at time t
$\boldsymbol{\theta}$	System parameters
$\boldsymbol{\theta}^*$	True underlying system parameters
$\hat{\boldsymbol{\theta}}$	Estimated system parameters
$\mathcal{L}$	Generator of the SDE system
$A', \boldsymbol{x}_0', \ldots$	Alternative counterparts of parameters A, $\boldsymbol{x}_0, \ldots$ of ODE systems
$\tilde{A}, \tilde{G}, \ldots$	Alternative counterparts of parmeters $A, G, \ldots$ of SDE systems

Chapter 1

Introduction

1.1 Background and motivations

1.1.1 Motivation and applications of differential equations

Differential equations—including Ordinary Differential Equations (ODEs) and Stochastic Differential Equations (SDEs)—are fundamental tools for modeling dynamical systems across a vast array of scientific and engineering disciplines. These equations describe how system variables evolve over time under deterministic or stochastic laws. Historically rooted in classical mechanics, differential equations have provided a unified mathematical framework for modeling continuous-time evolution in physics, biology, chemistry, economics, and engineering.

In physics, Newtonian mechanics, electromagnetism, and thermodynamics are all governed by ODE systems. In biology, gene regulatory networks, neuronal dynamics, and enzyme kinetics are modeled using ODE or SDE systems. In epidemiology, SIR-type models governed by ODE systems describe the spread of infectious diseases. In economics, macroeconomic indicators, interest rate models, and market dynamics are often formalized using SDE systems to capture uncertainties in policy and behavior. In engineering, both ODE and SDE systems form the backbone of control theory, signal processing, and system identification.

In all these domains, understanding system dynamics is not just a matter of predicting the future, but also of understanding how variables causally influence each other.

This causal interpretation is critical when we want to intervene in the system, simulate alternative policies, or derive counterfactual outcomes. Hence, the identifiability of the underlying ODE or SDE system—whether its structure and parameters/generators can be uniquely inferred from data—is of foundational importance.

To appreciate the ubiquity and importance of identifiability in continuous-time modeling, we consider several representative examples:

Systems Biology: Gene regulatory networks are often modeled using ODE or SDE systems to capture the interactions among genes, proteins, and other molecular components over time [4–6]. For instance, the expression level of one gene may increase in response to the activation of another. These systems guide experimental interventions, such as gene knockouts or targeted drug design. However, if the model is not identifiable, we may infer the wrong regulatory relationship or fail to determine whether a particular gene is a driver or a responder.

Pharmacokinetics: In drug development and clinical pharmacology, ODE systems model how drugs are absorbed, distributed, metabolized, and excreted from the body [7, 8]. Such systems are used to determine safe and effective dosing regimens. If key parameters like absorption rate or clearance cannot be uniquely determined from clinical data (i.e., are unidentifiable), physicians may be misled about the drug's behavior, leading to underdosing, toxicity, or reduced therapeutic efficacy.

Climate Modeling: Simple energy-balance models and more complex components of global climate models use ODE or SDE systems to represent temperature changes, carbon cycle dynamics, and ocean circulation [9–11]. These systems help forecast long-term climate patterns and assess the impact of interventions such as emission reduction policies. Identifiability is crucial here: if two models cannot be distinguished based on observational data, they may suggest contradictory outcomes following policy interventions, making it difficult to reliably guide environmental decision-making.

Econometrics: Economic indicators such as inflation and interest rates are frequently modeled using SDE systems to capture both systematic trends and random economic shocks [12–14]. These systems inform central bank policies and investment decisions. If a model is not identifiable, economic forecasts and risk assessments based on it may be fundamentally flawed, affecting national or global financial stability.

Neural Engineering: In modeling brain activity, ODE or SDE systems are used to describe how neuronal populations interact [15–17]. These systems underpin applications such as seizure prediction, neural prosthetics, and brain-computer interfaces. Identifiability determines whether we can accurately infer connectivity patterns or predict outcomes from interventions like electrical stimulation. Without identifiability, such technologies may misfire or offer misleading diagnostics.

1.1.2 Causal reasoning in dynamical systems

Recent years have witnessed increasing interest in leveraging ODE and SDE systems for causal reasoning [18–21]. Causal relationships describe how variables actively influence one another and how the system responds to deliberate interventions—in contrast to purely statistical associations. This distinction is critical in domains like biology, epidemiology, and economics, where interventions such as administering a drug, knocking out a gene, or changing a monetary policy are common and consequential.

ODE and SDE systems provide a natural framework for capturing time-continuous causal relationships [22]. In these systems, the structure of the equations reflects how variables influence each other over time, and interventions can be represented by modifying specific components of the system. For example, in an ODE system, setting a state variable to follow an externally imposed trajectory corresponds to a hypothetical intervention, enabling one to model counterfactual behaviors. In SDE systems, the generator—which determines how the stochastic process evolves—plays a central role in characterizing both observational and interventional distributions.

When interpreted causally, the parameters of an ODE system encode direct causal effects. For instance, replacing the differential equation governing $x_i(t)$ with a user-defined function $u_i(t)$ corresponds to an intervention $\mathtt{do}(x_i(t) = u_i(t))$ in the structural causal model framework. Identifiability of the system ensures that such inferences are grounded in a uniquely recoverable model. Without identifiability, causal interpretation and prediction under intervention become ambiguous.

The role of identifiability. Identifiability is a fundamental prerequisite for meaningful inference and decision-making in dynamical systems. If a system is not identifiable, then multiple models can explain the same observed behavior, leading to ambiguity in parameter estimation, uncertainty in causal interpretation, and potentially invalid predictions under intervention.

For ODE systems, we typically ask whether the parameters of the model (e.g., rate constants or interaction strengths) can be uniquely determined from observational data. For SDE systems, we instead focus on whether the generator is uniquely determined by the observed data distribution. This is because different combinations of drift and diffusion parameters may produce indistinguishable behavior in the observed data, making the generator a more fundamental and identifiable representation of the system's dynamics.

Without identifiability, any attempt to interpret causal relationships, estimate intervention effects, or simulate future behavior becomes unreliable, even when the data are ideal. Identifiability is thus a necessary condition for trustworthy modeling, particularly in high-stakes settings such as medicine, climate science, or economic policy.

1.1.3 Problem statement

This thesis addresses the fundamental problem of **identifiability** in time-continuous causal models: under what conditions can we uniquely recover the parameters of a linear ODE system or the infinitesimal generator of a linear SDE system from observational data?

We focus on three central questions:

1. When can the matrix governing a linear ODE system be identified from discrete, noise-free observations, possibly without knowledge of the initial condition?
2. In linear ODE systems with latent variables, what assumptions ensure that partial or full causal structure can be recovered from observations?
3. For linear SDE systems, under what conditions can the generator be uniquely determined from the observational distribution, and does this guarantee identifiability of post-intervention distributions?

These questions are crucial for using ODE and SDE systems in causal inference, where identifying direct influences and predicting the effects of interventions depend critically on the model being uniquely determined by the data.

Identifiability ensures that model estimation and causal interpretation are meaningful. Without it, even perfect observational data may be consistent with multiple, indistinguishable models—especially in continuous-time systems where feedback and indirect effects obscure causal structure. Thus, identifiability is a necessary condition for valid causal discovery and intervention modeling.

This thesis investigates two complementary notions of identifiability. For linear ODE systems, we derive conditions under which the **parameters** can be uniquely recovered from some amount of observations. For linear SDE systems, we establish when the **generator** is identifiable from the observational distribution—ensuring that the effects of interventions are likewise identifiable. These two cases reflect the different mathematical properties of deterministic versus stochastic systems.

While identifiability has been studied extensively in control theory and systems biology, several key challenges remain:

- **Discrete measurements:** Classical identifiability results often assume access to full, continuous trajectories and known initial conditions. In practice, however, data are typically sampled at discrete time points, and sometimes degraded through aggregation or rescaling. Moreover, the initial conditions are often unknown.
- **Hidden confounders:** Many real-world systems involve latent variables that affect the dynamics of observed variables. These hidden confounders can induce spurious correlations and hinder causal inference.
- **Stochastic dynamics:** In SDE systems, the generator determines both trajectory distributions and post-intervention behavior. Yet identifiability results have primarily focused on additive noise under strong assumptions like ergodicity, leaving more general cases largely unexplored.

This thesis addresses these challenges by developing identifiability theory for more realistic settings and providing statistical tools that support practical causal inference.

1.1.4 Related work

1.1.4.1 Identifiability and estimation of linear ODE systems

Identifiability of linear ODE systems. Identifiability of parameters in linear ODE systems has been a long-standing focus in control theory, where key foundational results were established under full-state observability and known initial conditions [23–27].

Stanhope et al. [28] conducted a systematic identifiability analysis based on a **single** continuous trajectory under the assumption of known initial conditions. Our work extends this framework to more realistic scenarios with discrete observations and unknown initial conditions. Specifically, our approach allows for explicit computation of system parameters from equally spaced, noise-free observations sampled from a single trajectory. More importantly, we develop identifiability conditions built entirely on the system parameters, without requiring assumptions such as linear independence among observed vectors. These advances also enable the derivation of asymptotic properties of the Nonlinear Least Squares (NLS) estimator, including an analytic form of its asymptotic normality covariance matrix. This facilitates a causal structure inference method based on finite noisy observations.

Qiu et al. [29] explored practical aspects of identifiability analysis of linear ODE systems with unknown initial conditions from a single trajectory. They proposed several quantitative scores for assessing the identifiability of linear ODEs in empirical settings. While their work is diagnostic in nature, our contributions offer a theoretical basis for inference in single-trajectory scenarios.

Chapter 3 formalizes these results, establishing when and how the system matrix and initial condition can be identified from a discretely observed single trajectory. It also proves that under mild conditions, the NLS estimator is consistent and asymptotically normal. These results enable valid confidence intervals and hypothesis tests for causal structure inference.

Parameter estimation in ODE systems. Various methods have been developed for parameter estimation in ODE systems, including the NLS approach [30–32], two-stage smoothing [5, 33, 34], and principal differential analysis [35, 36]. Bayesian frameworks [37] and neural network-based methods [38, 39] have also emerged in recent years. However, most neural approaches lack identifiability guarantees and require multiple trajectories. The NLS method remains a widely used technique, though few works have rigorously established its asymptotic properties under realistic sampling assumptions. Our work provides such guarantees with minimal assumptions and demonstrates the theoretical soundness of the NLS estimator in discrete settings.

ODE systems with hidden confounders. Latent variables present a significant challenge to causal inference in ODE systems, yet systematic identifiability analysis in this context is limited. Chapter 4 fills this gap by investigating linear ODE systems under two general forms of hidden confounding: (i) independent latent inputs with known functional forms, and (ii) dynamically evolving latent states structured by a Directed Acyclic Graph (DAG). Sufficient identifiability conditions are derived for each case, supporting causal discovery in scenarios where full observability is not attainable.

1.1.4.2 Identifiability of linear SDE systems

Identifiability of SDE systems is typically studied in the context of Gaussian diffusions and continuous trajectories [40–42]. These works often assume strong assumptions such as ergodicity or stability, and typically focus on parameter estimation from a single long trajectory. An alternative line of work [43–45] investigates discrete-time analogues by mapping SDEs to vector autoregressive models. However, these approaches usually impose restrictive spectral conditions on the drift matrix, limiting their applicability to more general settings.

In the context of causal inference, Hansen and Sokol [21] established a foundational framework in which the infinitesimal generator of an SDE plays the role of the system's causal mechanism. Specifically, the generator uniquely determines post-intervention distributions under Lipschitz interventions, thereby linking observational and interventional dynamics. However, an open and fundamental question remains: under what conditions can this generator be identified from observational data alone? Chapter 5 of this thesis addresses this challenge by developing a rigorous identifiability analysis for linear SDE systems. It establishes necessary and sufficient conditions for identifying the generator in the additive-noise setting and introduces generic sufficient conditions for systems with multiplicative noise. These results are accompanied by geometric interpretations that provide insight into the structural mechanisms underlying identifiability. Importantly, the chapter demonstrates that generator identifiability ensures the recovery of all Lipschitz-continuous post-intervention distributions, thereby offering a robust foundation for causal reasoning in stochastic dynamical systems.

1.1.4.3 Differential equations and causal reasoning

Differential equations—both deterministic and stochastic—provide a natural formalism for representing causal relationships in continuous-time systems. A growing body of work [18–20] has developed formal links between ODEs, Random Differential Equations (RDEs), and Structural Causal Models (SCMs), showing how differential equations can encode stable causal mechanisms that generalize to interventions.

In stochastic settings, Hansen and Sokol [21] propose that the infinitesimal generator of an SDE system uniquely characterizes both observational and interventional behavior, positioning it as the causal mechanism of the system. This generator-based perspective establishes a strong theoretical foundation for causal inference in continuous-time stochastic systems.

On the algorithmic front, causal discovery approaches leveraging neural differential equations have gained traction. In particular, neural ODE frameworks with sparsity constraints [46, 47] and neural SDE models [48] offer flexible tools for uncovering causal structure from data. However, these methods often prioritize predictive accuracy and model flexibility over identifiability, leaving a gap in understanding when the recovered structure reflects true causal relationships.

This thesis directly addresses this gap by focusing on the fundamental question of identifiability. We establish conditions under which the causal mechanisms of a dynamical system—whether encoded through ODE parameters or the generator of an SDE—can be uniquely recovered from observational data. Our analysis spans a variety of realistic observational regimes, including discrete or continuous-time sampling, finite data, partially-observability, and measurement noise. In certain regimes, we further develop statistical procedures for consistent estimation and inference. These results provide a principled foundation for causal reasoning in dynamical systems, complementing algorithmic advances with rigorous identifiability guarantees.

This thesis advances the field by providing a rigorous identifiability analysis for both linear ODE and SDE systems with causal interpretation. Chapter 3 develops causal structure inference methods based on noisy discrete data, Chapter 4 addresses the role of hidden confounders in dynamical systems, and Chapter 5 formalizes the identifiability of SDE generators from observational distributions and their link to post-intervention behavior. These contributions unify and extend the theoretical foundation connecting differential equations to causal inference.

1.2 Overview and contributions

1.2.1 Main contributions

1. **Identifiability of linear ODE systems from discrete observations.** Chapter 3 establishes conditions under which a homogeneous linear ODE system is identifiable from a sequence of discrete observations sampled from a single trajectory, even when the initial condition is unknown. Under these conditions, the Nonlinear Least Squares (NLS) estimator is shown to be consistent and asymptotically normal. Confidence intervals and hypothesis tests are derived to support causal structure inference.

2. **Identifiability of linear ODE systems with hidden confounders.** Chapter 4 analyzes linear ODE systems affected by hidden confounders. It provides identifiability conditions for two cases: (i) latent inputs with known functional forms, and (ii) dynamically interacting latent variables following a Directed Acyclic Graph (DAG) structure. The conditions apply to various data regimes, including continuous and discrete observations from single or multiple trajectories.

3. **Generator identification for linear SDE systems.** Chapter 5 studies identifiability of the generator for linear SDE systems. For additive noise, a necessary and sufficient condition is established. For multiplicative noise, a sufficient condition is introduced. Both conditions are proved to be generic and interpreted geometrically. The chapter further shows that generator identifiability ensures the identifiability of all Lipschitz-continuous post-intervention distributions.

Together, these results provide a unified theory of identifiability for linear ODE and SDE systems, accommodating practical data limitations and enabling reliable causal inference from time-series data.

1.2.2 Structure of the thesis

Chapter 2 presents the necessary mathematical preliminaries, including system formulations, observational regimes, and causal semantics. Chapters 3 through 5 contain the main theoretical results and simulation-based validations. Chapter 6 concludes the thesis and outlines potential directions for future research.

Chapter 2

Preliminaries

This chapter lays the mathematical and conceptual groundwork for the identifiability analysis of linear dynamical systems investigated in this thesis. We begin with formal definitions of linear ODE and SDE systems and their causal semantics, including settings with hidden confounders. Then we review various observational regimes that affect identifiability and conclude with a detailed definition of identifiability in both deterministic and stochastic contexts. These preliminaries set the stage for the identifiability theorems, estimation results, and simulation-based verifications presented in the subsequent chapters.

2.1 Linear ODE systems

2.1.1 Time-homogeneous formulation

We consider homogeneous linear ODE systems of the form:

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= A\boldsymbol{x}(t)\,, \\ \boldsymbol{x}(0) &= \boldsymbol{x}_0\,, \end{aligned} \tag{2.1}$$

where $t \in [0, \infty)$, $\boldsymbol{x}(t) \in \mathbb{R}^d$ denotes the d-dimensional state vector, $\dot{\boldsymbol{x}}(t)$ is its time derivative, $A \in \mathbb{R}^{d\times d}$ is the system matrix, and $\boldsymbol{x}_0 \in \mathbb{R}^d$ is the initial state.

The system admits a unique solution given by:

$$\boldsymbol{x}(t; \boldsymbol{x}_0, A) = e^{At}\boldsymbol{x}_0,$$

where e^{At} denotes the matrix exponential. This describes a single trajectory in $\mathbb{R}^d$ initialized at $\boldsymbol{x}_0$ and evolving under the linear flow induced by A.

2.1.2 Systems with hidden confounders

We also study linear ODE systems in which some variables are latent or unobserved. We consider two types of hidden confounding:

1. **Independent latent confounders:** Latent variables exhibit no causal relationships among themselves, leading to the following linear ODE system:

$$\begin{bmatrix} \dot{\boldsymbol{x}}(t) \\ \dot{\boldsymbol{z}}(t) \end{bmatrix} = \begin{bmatrix} A & B \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(t) \\ \boldsymbol{z}(t) \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \boldsymbol{f}(t) \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{x}(0) \\ \boldsymbol{z}(0) \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{z}_0 \end{bmatrix}. \tag{2.2}$$

2. **Causally related latent confounders:** Latent variables exhibit causal relationships among themselves, specifically, they follow a DAG structure, represented as:

$$\begin{bmatrix} \dot{\boldsymbol{x}}(t) \\ \dot{\boldsymbol{z}}(t) \end{bmatrix} = \begin{bmatrix} A & B \\ \mathbf{0} & G \end{bmatrix} \begin{bmatrix} \boldsymbol{x}(t) \\ \boldsymbol{z}(t) \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{x}(0) \\ \boldsymbol{z}(0) \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{z}_0 \end{bmatrix}. \tag{2.3}$$

In both systems, $\boldsymbol{x}(t) \in \mathbb{R}^d$ denotes the state of observable variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$, while $\boldsymbol{z}(t) \in \mathbb{R}^p$ denotes the state of latent variables $\boldsymbol{z} = (z_1, z_2, \ldots, z_p)$. The matrix $B \in \mathbb{R}^{d\times p}$ encodes the influence of latent confounders on the observed variables. The matrix $G \in \mathbb{R}^{p\times p}$ in (2.3) models the latent causal structure and is assumed to be strictly upper triangular, reflecting the DAG assumption. $\boldsymbol{f}(t)$ in (2.2) is a general function of time t. Inference of the parameters A, B, and G from observations of $\boldsymbol{x}(t)$ alone poses significant challenges due to the unobserved dynamics of $\boldsymbol{z}(t)$.

2.1.3 Causal structure and semantics

When an ODE system captures the underlying causal mechanisms of a dynamical process, it offers a natural and principled framework for modeling time-continuous causal relationships among variables. The causal structure encoded in such systems can often be read directly from the differential equations themselves [18, 22].

Let the system variables be denoted by $x_1, \ldots, x_d$. If the time derivative $\dot{x}_i$ depends on x_j, this indicates a direct causal influence from x_j to x_i, represented as a directed edge $x_j \to x_i$ in the associated causal graph. For instance, in the linear ODE system (2.1), the matrix A determines the causal structure among variables: $A_{ij} \neq 0$ implies a direct causal link from x_j to x_i. Since A is time-invariant, the system corresponds to an autonomous structural causal model, where causal relations do not change over time.

In systems with latent confounding, the full causal structure includes both observed and unobserved variables. Figure 2.1 illustrates example causal graphs corresponding to the ODE systems in (2.2) and (2.3). Notably, cycles and self-loops among observable variables may arise due to feedback dynamics encoded in A.

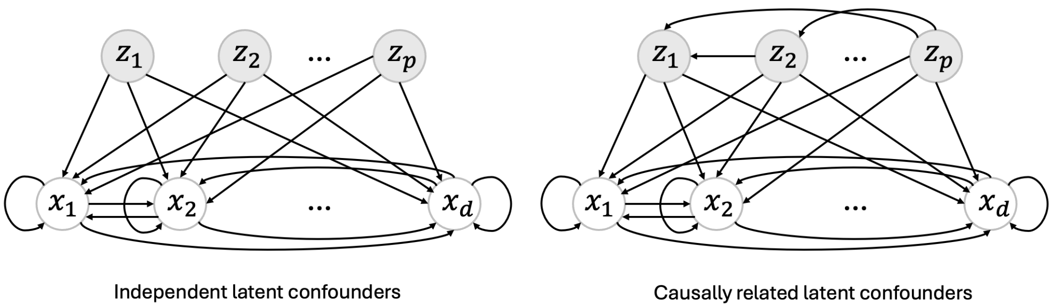

FIGURE 2.1: Example causal structures of the ODE system (2.2) and (2.3).

2.2 Linear SDE systems

We study multidimensional linear SDEs with both additive and multiplicative noise. Let $W := \{W_t = [W_{1,t}, \ldots, W_{m,t}]^\top : 0 \leqslant t < \infty\}$ denote an m-dimensional standard Brownian motion defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\})$. Let $X_t \in \mathbb{R}^d$ denote the state of the system at time t, and let $\boldsymbol{x}_0 \in \mathbb{R}^d$ be the initial condition. We consider two classes of linear SDEs.

2.2.1 Additive noise

The first class involves additive noise and takes the form:

$$
\begin{aligned}
dX_t &= AX_t dt + G dW_t \,, \\
X_0 &= \boldsymbol{x}_0 \,,
\end{aligned}
\tag{2.4}
$$

where $0 \leqslant t < \infty$, $A \in \mathbb{R}^{d\times d}$ and $G \in \mathbb{R}^{d\times m}$ are some constant matrices. This SDE admits an explicit strong solution (cf. [49]):

$$
X_t := X(t; \boldsymbol{x}_0, A, G) = e^{At}\boldsymbol{x}_0 + \int_0^t e^{A(t-s)} G dW_s \,. \tag{2.5}
$$

Throughout this work, we refer to the strong solution, in the sense of [50], unless stated otherwise.

2.2.2 Multiplicative noise

We also consider systems with state-dependent (multiplicative) noise, described by the following SDE:

$$
dX_t = AX_t dt + \textstyle\sum_{k=1}^m G_k X_t dW_{k,t} \,, \quad X_0 = \boldsymbol{x}_0 \,, \tag{2.6}
$$

where $0 \leqslant t < \infty$, $A, G_k \in \mathbb{R}^{d\times d}$ for $k = 1, \ldots, m$ are some constant matrices. This formulation captures state-dependent volatility and introduces greater modeling flexibility but also increases the complexity of theoretical analysis, particularly for identifiability.

In general, an explicit closed-form solution for (2.6) is not available. However, when the matrices $A, G_1, \ldots, G_m$ commute pairwise—that is, if

$$
AG_k = G_k A \quad \text{and} \quad G_k G_l = G_l G_k
$$

holds for all $k, l = 1, \ldots, m$, then an explicit solution can be obtained (cf. [51]):

$$
X_t := X(t; \boldsymbol{x}_0, A, \{G_k\}_{k=1}^m) = \exp\left\{ \left(A - \frac{1}{2}\sum_{k=1}^m G_k^2 \right) t + \sum_{k=1}^m G_k W_{k,t} \right\} \boldsymbol{x}_0 \,.
$$

2.2.3 Generator and distributional semantics

The generator of a stochastic process provides an infinitesimal characterization of its evolution and plays a central role in connecting sample path behavior with distributional properties. For a sufficiently regular function f, the generator $\mathcal{L}$ of a stochastic process X_t is defined as

$$(\mathcal{L}f)(\boldsymbol{x}) = \lim_{s \to 0} \frac{\mathbb{E}[f(X_{t+s}) - f(X_t) | X_t = \boldsymbol{x}]}{s},$$

which captures the instantaneous rate of change in the expected value of $f(X_t)$, conditioned on the current state $X_t = \boldsymbol{x}$.

Both additive and multiplicative linear SDEs considered above conform to the general Itô process:

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t\,, \quad X_0 = \boldsymbol{x}_0\,. \tag{2.7}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ is the drift term and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times m}$ is the diffusion term, both assumed to be locally Lipschitz continuous. The generator $\mathcal{L}$ of the SDE (2.7) can be explicitly computed by utilizing Itô's formula (cf. [49]).

Proposition 2.2.1. Let X be a stochastic process defined by the SDE (2.7). The generator $\mathcal{L}$ of X on $C_b^2(\mathbb{R}^d)$ is given by

$$(\mathcal{L}f)(\boldsymbol{x}) := \sum_{i=1}^{d} b_i(\boldsymbol{x}) \frac{\partial f(\boldsymbol{x})}{\partial x_i} + \frac{1}{2} \sum_{i,j=1}^{d} c_{ij}(\boldsymbol{x}) \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j} \tag{2.8}$$

for $f \in C_b^2(\mathbb{R}^d)$ and $\boldsymbol{x} \in \mathbb{R}^d$, where $c(\boldsymbol{x}) = \sigma(\boldsymbol{x}) \cdot \sigma(\boldsymbol{x})^\top$ is a $d \times d$ matrix, and $C_b^2(\mathbb{R}^d)$ denotes the space of continuous functions on $\mathbb{R}^d$ that have bounded derivatives up to order two.

2.2.4 Causal interpretation and interventions

In the setting of SDE systems, by applying causal reasoning of the system we focus on deriving conditions under which the post-intervention distributions can be determined from the observational distribution. Here, for an observational SDE system, an intervention fixes the l-th coordinate to a Lipschitz control $\zeta\big(X_t^{(-l)}\big)$, where $X^{(-l)}$ denotes the $(d-1)$-dimensional vector that results from the removal of the l-th coordinate of $X \in \mathbb{R}^d$, and produces an post-intervention SDE.

Let $\mathcal{L}$ denote the generator of the observational SDE, we know both drift and diffusion coefficients enter $\mathcal{L}$. Crucially, if two observational SDEs share the same generator, under mild conditions, their post-intervention distributions coincide (Lemma 5.3 of [21]). Hence identifying $\mathcal{L}$ from the observational distribution is **sufficient** to determine every post-intervention law, even though the underlying parameter tuple (A, G) or $(A, \{G_k\})$ may itself be non-unique. Chapter 5 derives the concrete conditions that guarantee such generator-identifiability.

2.3 Observational regimes

Identifiability depends as much on how the system is observed as on its intrinsic dynamics. In this thesis, we formalize five observational settings that serve as a unifying framework. All experiments presented in Chapters 3–5 fall within these settings: some use a single setting, while others combine multiple settings (e.g., partial measurements together with discrete time points). Not all settings appear in every chapter, but collectively they cover the full range of experiments in this thesis. They fall into two broad classes.

2.3.1 Direct state observations

i) **Full trajectories.** Every component of the state $\boldsymbol{x}(t) \in \mathbb{R}^d$ is recorded continuously on a time–interval $[0, T]$ or $[0, \infty)$. Each trajectory starting from a given initial state $\boldsymbol{x}_0$ is a single trajectory.

ii) **Discrete time points.** Only a finite sequence $\{\boldsymbol{x}(t_k)\}_{k=0}^n$ is available, the spacing $t_{k+1} - t_k = \Delta t$ could be the same or different among different steps, representing equally-spaced or randomly-spaced discrete observations.

iii) **Partial measurements.** In addition to observable $\boldsymbol{x}(t)$, there are also unobservable $\boldsymbol{z}(t)$ interact with the system, making the system only partially observable.

2.3.2 Degraded observations

iv) **Temporal aggregation.** The observations are the average values of k consecutive, non-overlapping measurements, with $k \geq 2$.

v) **Time scaling.** Experiments sometimes normalise time via $\tau = k\,t$, so the recorded path is $\tilde{x}(\tau) = x(k^{-1}\tau)$.

The remainder of the thesis analyses identifiability of linear ODE and SDE systems under these observational regimes, always under the idealisation of noise-free measurements so as to isolate structural from statistical issues.

2.4 Identifiability

Let Θ denote the parameter space of a model family and

$$\Theta \ni \theta \;\longmapsto\; \mathcal{D}(\theta)$$

be the data map, i.e. the ideal, noise-free observational object produced under a fixed regime from Section 2.3. We now make two target-specific notions of identifiability explicitly.

2.4.1 Parameter-identifiability for linear ODE systems

Let Θ_{ODE} be the parameter space corresponding to the linear ODE systems we consider.

Definition 2.1 (Parameter-identifiability)**.** The linear ODE system is identifiable under a given observational regime if

$$\mathcal{D}(\theta_1) = \mathcal{D}(\theta_2) \implies \theta_1 = \theta_2, \qquad \forall\, \theta_1, \theta_2 \in \Theta_{\mathrm{ODE}}.$$

Here $\mathcal{D}(\theta)$ is a set of exact state values (full trajectory, discrete samples, aggregated measurements, etc.). Chapters 3 and 4 establish conditions under which Definition 2.1 holds in each observational regime.

2.4.2 Generator-identifiability for linear SDE systems

For linear SDE systems (2.4) and (2.6), many distinct parameter tuples (A, G) or $(A, \{G_k\})$ can generate the same stochastic law. The causally relevant object is therefore the infinitesimal generator. Let $\mathcal{L}$ be the corresponding mapping from parameter to generator, let Θ_{SDE} be the parameter space corresponding to the SDE systems we consider.

Definition 2.2 (Generator-identifiability)**.** The generator of a linear SDE system is identifiable under a specified observational regime if

$$\mathcal{D}(\theta_1) = \mathcal{D}(\theta_2) \implies \mathcal{L}(\theta_1) = \mathcal{L}(\theta_2), \qquad \forall\, \theta_1, \theta_2 \in \Theta_{\mathrm{SDE}}.$$

Distinct (A, G) pairs sharing the same $\mathcal{L}$ produce identical observational and interventional behaviour; identifying $\mathcal{L}$ is both necessary and sufficient for causal prediction, while identifying each parameter separately is unnecessary and often impossible [21]. Chapter 5 establishes conditions under which Definition 2.2 holds for additive noise and for multiplicative noise.

Once either system parameters for ODE systems or generator for SDE systems are uniquely identified, corresponding continuous-time interventions can be safely made, providing a reliable causal inference of the dynamical systems.

Chapter 3

Identifiability and Asymptotics in Learning Linear ODE Systems from Discrete Observations

Ordinary Differential Equations (ODEs) have recently gained a lot of attention in machine learning. However, the theoretical aspects, for example, identifiability and asymptotic properties of statistical estimation are still obscure. This chapter derives a sufficient condition for the identifiability of homogeneous linear ODE systems from a sequence of equally-spaced error-free observations sampled from a single trajectory. When observations are disturbed by measurement noise, we prove that under mild conditions, the parameter estimator based on the Nonlinear Least Squares (NLS) method is consistent and asymptotic normal with $n^{-1/2}$ *convergence rate. Based on the asymptotic normality property, we construct confidence sets for the unknown system parameters and propose a new method to infer the causal structure of the ODE system, that is, inferring whether there is a causal link between system variables. Furthermore, we extend the results to degraded observations, including aggregated and time-scaled ones. We also construct simulations with various system dimensions to illustrate the established theoretical results.*

This chapter is derived from the following publication: Identifiability and Asymptotics in Learning Homogeneous Linear ODE Systems from Discrete Observations [1]

3.1 Introduction

Ordinary Differential Equations (ODEs) have been widely used to model dynamic systems in various scientific fields such as physics [52–54], biology [55–58], and economics [59–61]. In recent years, ODEs are also attracting increasing attention in the machine learning community. For instance, ODEs have been used to build new families of deep neural networks [62–64] and the connection between ODEs and structural causal models has been established [18, 19].

Existing works mostly focus on the parameter estimation of ODEs [32, 33, 35, 55, 65–68]. However, before estimating the unknown parameters, it is essential to perform an identifiability analysis of an ODE system; that is, uncovering the mathematical conditions under which the parameters can be uniquely determined from noise-free observations. If a system is not identifiable, the estimation procedure may produce erroneous and misleading parameter estimates [29]. This is detrimental in many applications; for example, the estimated parameter can easily lead to wrong causal structures of non-identifiable ODE systems.

The contribution of this chapter are summarized as follows:

Derive identifiability condition for linear ODEs from discrete observations. We derive the condition for the identifiability of homogeneous linear ODE systems from discrete observations (collected at discrete time points). Specifically, we consider the setting where the observations are sampled from a single trajectory generated from one initial condition. This setting is prevalent in applications that can only access a single trajectory due to the unrepeatable process of the measurements collection. Our identifiability analysis is built upon the work of Stanhope et al. [28], which constructs a systematic study of the identifiability of homogeneous linear ODE systems from a continuous trajectory with known initial conditions. Our research extends this framework to more practical scenarios where we only have discrete observations and do not know the initial conditions.

Derive asymptotic properties for NLS estimator. Based on our identifiability results, we study the asymptotic properties of parameter estimation from data with measurement noises. We focus on the estimator obtained by the Nonlinear Least Squares (NLS) method, which is simple and widely used in dynamical systems [69–73]. However, the asymptotic properties of NLS based estimators for ODE systems have not been systematically studied due to the lack of identifiability conditions. We prove that under mild conditions, the NLS estimator is consistent and asymptotic normal with $n^{-1/2}$ convergence rate. In addition, based on the established asymptotic normality theory, we construct the confidence sets of unknown parameters and propose a new method to infer the causal structure of ODE systems, that is, inferring whether there is a causal link between system variables.

Extend theoretical results to degraded observations. We extend the consistency and asymptotic normality results to the observations with degraded quality, including aggregated and time-scaled observations. The aggregated observations are usually caused by time aggregation in the data collection process [74]. The time-scaled observations result from data preprocessing to fit the ODE model. We prove that the ODE model generating the original observations is identifiable from the degraded observations. The asymptotic properties can be naturally extended given the identifiability results. Simulations with various system dimensions are constructed to verify the developed theoretical results.

3.2 Identifiability condition of linear ODE systems

Linear ODE systems hold significant importance in modelling and comprehending the dynamics of various real-world phenomena across diverse disciplines. For instance, well-established examples such as the Spring-Mass-Damper system [75], the simple population growth model [76], and the heat cooling problem model [77] all belong to the category of linear ODE systems. These systems often admit closed-form analytical solutions, thereby facilitating precise predictions of system behaviour and enabling detailed analysis. Moreover, linear ODE systems serve as a foundation for approximating nonlinear systems through linearization techniques, providing valuable insights into the more intricate nonlinear system behaviour. In this chapter, we focus on an important special case of linear ODE systems: the homogeneous linear ODE system.

A homogeneous linear ODE system can be defined as:

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= A\boldsymbol{x}(t)\,, \\ \boldsymbol{x}(0) &= \boldsymbol{x}_0\,, \end{aligned} \tag{3.1}$$

where $t \in [0, \infty)$ denotes the independent variable time, $\boldsymbol{x}(t) \in \mathbb{R}^d$ denotes the state of the ODE system at time t, $\dot{\boldsymbol{x}}(t)$ denotes the first derivative of $\boldsymbol{x}(t)$ with respect to time t, and we refer to both the parameter matrix $A \in \mathbb{R}^{d\times d}$ and the initial condition $\boldsymbol{x}_0 \in \mathbb{R}^d$ as the system parameters. In this chapter we focus on ODE systems with complete observation, that is, all state variables are observable. Therefore, the measurement model can be described as:

$$\boldsymbol{y}(t) = \boldsymbol{x}(t)\,. \tag{3.2}$$

The solution of ODE system (3.1) can be explicitly expressed as:

$$\boldsymbol{x}(t; \boldsymbol{x}_0, A) = e^{At}\boldsymbol{x}_0\,, \tag{3.3}$$

which is also called a trajectory. The symbol e in Equation (3.3) denotes the matrix exponential function. In this chapter, we focus on identifiability analysis of the ODEs from observations sampled from a **single** d-dimensional trajectory generated with an initial condition $\boldsymbol{x}_0$. Under the setting of our work, the term identifiability means that given the error-free observation from a single trajectory of the ODE system, whether the system parameters A and $\boldsymbol{x}_0$ can be uniquely determined.

3.2.1 Identifiability condition from a whole trajectory

Given a fixed initial condition $\boldsymbol{x}_0$, Stanhope et al. [28] derived a necessary and sufficient condition for identifying the ODE system (3.1) from a whole trajectory, $\{e^{At}\boldsymbol{x}_0\}_{t\in[0,\infty)}$, that is, $\{\boldsymbol{x}_0, A\boldsymbol{x}_0, \ldots, A^{d-1}\boldsymbol{x}_0\}$ are linearly independent. The notation A^k denotes the kth power of the matrix A, represented as the product of k copies of matrix A, that is, $A^k = A \times A \times \ldots \times A$. However, in practice, we cannot usually observe the initial condition $\boldsymbol{x}_0$. Under this practical circumstance, we need to treat the initial condition also as a system parameter and identify it from the data. In the following, we extend the identifiability definition and condition in [28] to the case where both parameter matrix A and initial condition $\boldsymbol{x}_0$ are system parameters.

Definition 3.1. Let (M^0, Ω) be given parameter spaces, with $M^0 \subset \mathbb{R}^d$ and $\Omega \subset \mathbb{R}^{d\times d}$. The ODE system (3.1) is said to be identifiable in (M^0, Ω), if for all $\boldsymbol{x}_0, \boldsymbol{x}_0' \in M^0$ and all $A, A' \in \Omega$, with $(\boldsymbol{x}_0, A) \neq (\boldsymbol{x}_0', A')$, it holds that $\boldsymbol{x}(\cdot; \boldsymbol{x}_0, A) \neq \boldsymbol{x}(\cdot; \boldsymbol{x}_0', A')$.

Here $\boldsymbol{x}(\cdot;\boldsymbol{x}_0,A) \neq \boldsymbol{x}(\cdot;\boldsymbol{x}_0',A')$ means that there exists at least one $t \geq 0$ such that $\boldsymbol{x}(t;\boldsymbol{x}_0,A) \neq \boldsymbol{x}(t;\boldsymbol{x}_0',A')$. Then we establish the condition for identifiability of the ODE system based on Definition 3.1.

Lemma 3.2. *Suppose that $M^0 \subset \mathbb{R}^d$ and $\Omega \subset \mathbb{R}^{d\times d}$ are both open subsets. Then the ODE system (3.1) is identifiable in (M^0,Ω) if and only if $\{\boldsymbol{x}_0, A\boldsymbol{x}_0, \ldots, A^{d-1}\boldsymbol{x}_0\}$ are linearly independent for all $\boldsymbol{x}_0 \in M^0$ and all $A \in \Omega$.*

The proof of Lemma 3.2 is a straightforward extension of the proof of Theorem 2.5 in [28] and can be found in Appendix A.1.1. From the Lemma, we can see that the condition for identifying the system parameters $(A,\boldsymbol{x}_0)$ is the same as that for only identifying A, except that the linear independence condition needs to hold for all possible $\boldsymbol{x}_0$ in M^0.

3.2.2 Identifiability condition from discrete observations

In practice, typically, we can only access a sequence of discrete observations sampled from a trajectory instead of knowing the whole trajectory. Thus, from now on, we focus on the case where only discrete observations from a trajectory are available. In particular, We extend the identifiability definition of the ODE system (3.1) as follows.

Definition 3.3 ($(\boldsymbol{x}_0,A)$-identifiability)**.** For $\boldsymbol{x}_0 \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$, for any $n_0 \geq 1$, let $t_j, j = 1,\ldots,n_0$ be any n_0 time points and $\boldsymbol{x}_j := \boldsymbol{x}(t_j;\boldsymbol{x}_0,A)$ be the error-free observation of the trajectory $\boldsymbol{x}(t;\boldsymbol{x}_0,A)$ at time t_j. We say the ODE system (3.1) is $(\boldsymbol{x}_0,A)$ identifiable from $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n_0}$, if for all $\boldsymbol{x}_0' \in \mathbb{R}^d$ and all $A' \in \mathbb{R}^{d\times d}$, with $(\boldsymbol{x}_0',A') \neq (\boldsymbol{x}_0,A)$, it holds that $\exists j$ for $j = 1,\ldots,n_0$, such that $\boldsymbol{x}(t_j;\boldsymbol{x}_0',A') \neq \boldsymbol{x}(t_j;\boldsymbol{x}_0,A)$.

This definition is inspired by [29, Definition 1.6], and it is not a simple extension of Definition 3.1 to discrete observations when $M^0 := \mathbb{R}^d$ and $\Omega := \mathbb{R}^{d\times d}$. The reason is that in Definition 3.3, the initial condition $\boldsymbol{x}_0$ and parameter matrix A are fixed, and $\boldsymbol{x}_0'$ and A' are an arbitrary vector and an arbitrary matrix in M^0 and Ω respectively. However, in Definition 3.1, both $\boldsymbol{x}_0$ and $\boldsymbol{x}_0'$ are arbitrary vectors in M^0 and both A and A' are arbitrary matrices in Ω. In other words, Definition 3.3 describes an intrinsic property of a single system instead of a collective property of a set of systems. In dealing with the identifiability problem of an ODE system, we aim to check whether the true underlying system parameter $(\boldsymbol{x}_0,A)$ is uniquely determined by error-free observations.

Therefore, $(\boldsymbol{x}_0, A)$-identifiability described in Definition 3.3 is a more natural way to define the identifiability of the ODE system from the practical perspective, and all of the other relevant definitions and theorems in the rest of this chapter are derived based on Definition 3.3.

To derive the identifiability condition from discrete observations, we focus on the equally-spaced observations. Specifically, data are collected on an equally-spaced time grid, that is, $t_{j+1} - t_j = \Delta t$ for a constant $\Delta t > 0$ and $j = 1, 2, \ldots$. The motivation is that a time series is most commonly collected equally spaced in practice, which follows the standard rules of collecting data either from a scientific experiment or a natural phenomenon. Then, based on Definition 3.3, we derive a sufficient condition for the identifiability of the ODE system from discrete observations.

Theorem 3.4. *For $\boldsymbol{x}_0 \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$, the ODE system (3.1) is $(\boldsymbol{x}_0, A)$ identifiable from* ***any*** *$d+1$ equally-spaced error-free observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{d+1}$, if the following two conditions are satisfied.*

A1 $\{\boldsymbol{x}_0, A\boldsymbol{x}_0, \ldots, A^{d-1}\boldsymbol{x}_0\}$ *are linearly independent.*

A2 *Parameter matrix A has d distinct real eigenvalues.*

The proof of Theorem 3.4 can be found in Appendix A.1.2. Here, **any** $d+1$ equally-spaced error-free observations means that the time interval between two consecutive observations: Δt can take any positive value. In other words, the identifiability of the ODE system will not be influenced by the time-space between consecutive observations.

Now, in addition to the identifiability condition from a whole trajectory (condition A1), when only discrete observations are available, we further require that A has d distinct real eigenvalues (condition A2). This condition seems to be restrictive, however, due to the limited observations (a set of equally-spaced observations sampled from a single trajectory), the condition cannot be relaxed. The reasons are as follows: (1) **Distinct eigenvalues**: as discussed in [29], almost every $A \in \mathbb{R}^{d\times d}$ (with respect to the Lebesgue measure on $\mathbb{R}^{d\times d}$) has d distinct eigenvalues based on random matrix theory [78, 79]. Therefore, A has d distinct eigenvalues is a natural and reasonable assumption. In addition, to guarantee the ODE system (3.1) is $(\boldsymbol{x}_0, A)$ identifiable from any $d+1$ equally-spaced observations sampled from a single trajectory, we need d consecutive observations of them to be linearly independent. To ensure any d equally-spaced observations sampled from a single trajectory are linearly independent, matrix A has d distinct eigenvalues is a necessary condition. Moreover, in Section 3.3.2, to derive the explicit formula of the asymptotic covariance matrix of the NLS estimator's asymptotic normal distribution, we require the parameter matrix has d distinct eigenvalues. (2) **Real eigenvalues**: according to [28, Corollary 6.4], a matrix with complex eigenvalues is not identifiable from any set of equally-spaced observations sampled from a single trajectory. Therefore, we require the eigenvalues to be real.

Our identifiability condition here is sufficient but not necessary. However, it allows us to derive explicit expressions of A and $\boldsymbol{x}_0$ in terms of the observations. To see this, we set $\Phi(t) := e^{At}$, and let $\boldsymbol{X}_j$ denote the matrix $(\boldsymbol{x}_j, \boldsymbol{x}_{j+1}, \dots, \boldsymbol{x}_{j+d-1}) \in \mathbb{R}^{d\times d}$ for $j = 1, 2$. Then $\boldsymbol{X}_2 = \Phi(\Delta t)\boldsymbol{X}_1$. We show in the proof that $\boldsymbol{X}_1$ is nonsingular if A has d distinct eigenvalues, and thus, $\Phi(\Delta t) = \boldsymbol{X}_2\boldsymbol{X}_1^{-1}$. Finally, we can obtain a unique real A by taking logarithm of $e^{A\Delta t} = \boldsymbol{X}_2\boldsymbol{X}_1^{-1}$ if A has d distinct real eigenvalues [28, Theorem 6.3]. The initial condition $\boldsymbol{x}_0$ can then be calculated by $e^{-At_1}\boldsymbol{x}_1$. Please refer to Appendix A.1.2 for the detail.

Worth mentioning that we can not check whether the proposed identifiability conditions are satisfied in practice since we do not have access to the real underlying system parameters in practical applications. Nevertheless, studying the identifiability conditions of the dynamical system provides us with a better understanding of the system. Moreover, in practice, we can use models satisfying the conditions (through constrained parameter estimation) to learn the real-world data to ensure the learned model's identifiability.

Remark If the ODE system is identifiable from a set of discrete observations sampled from a single trajectory, the system can also be identifiable from the whole corresponding trajectory.

3.2.3 Identifiability condition from degraded observations

In practice, there are cases where we have no records of the original observations of the ODE system and can only access the degraded observations instead. Is the ODE system still identifiable? Furthermore, will the parameter change? We are interested in addressing the problem of identifiability under degraded observations. In particular, we are interested in the aggregated and time-scaled observations.

According to Theorem 3.4, under assumptions A1 and A2, the ODE system (3.1) is $(\boldsymbol{x}_0, A)$ identifiable from n equally-spaced error-free observations $\boldsymbol{X} := \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$, for any $n > d$. Based on these n observations, in the following, we define the aggregated and time-scaled observations.

Definition 3.5 (aggregated observations)**.** Let $k \geq 2$ be an integer and $\tilde{n} = \lfloor n/k \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor function. For each $j = 1, 2, \dots, \tilde{n}$, we define $\tilde{\boldsymbol{x}}_j := \big(\boldsymbol{x}_{(j-1)k+1} + \boldsymbol{x}_{(j-1)k+2} + \dots + \boldsymbol{x}_{jk}\big)/k$, the average values of k consecutive, non-overlapping observations in $\boldsymbol{X}$, starting from time $\tilde{t}_j := t_{(j-1)k+1}$. We call $\tilde{\boldsymbol{X}} := \{\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \dots, \tilde{\boldsymbol{x}}_{\tilde{n}}\}$ a set of aggregated observations from $\boldsymbol{X}$.

For notational simplicity, we use the same notation for aggregated and time-scaled observations from here on.

Definition 3.6 (time-scaled observations)**.** Let $k > 0$ be a constant and $\tilde{n} = n$. For each $j = 1, 2, \dots, \tilde{n}$, defining $\tilde{t}_j := kt_j$ be the scaled time, the time-scaled observation at time $\tilde{t}_j$, denoted by $\tilde{\boldsymbol{x}}_j$, equals $\boldsymbol{x}_j$. We call $\tilde{\boldsymbol{X}} := \{\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \dots, \tilde{\boldsymbol{x}}_{\tilde{n}}\}$ the set of time-scaled observations from $\boldsymbol{X}$ at the scaled time grid $\{\tilde{t}_1, \dots, \tilde{t}_{\tilde{n}}\}$.

Remark Aggregated/time-scaled observations $\tilde{\boldsymbol{X}}$ follow new ODE systems as (3.1) but with different initial condition and parameter matrix, denoted respectively by $\tilde{\boldsymbol{x}}_0$ and $\tilde{A}$, and we call the new ODE systems aggregated/time-scaled ODE systems.

Having defined the aggregated/time-scaled observations, we now derive the conditions for identifying the original ODE system with parameters $\boldsymbol{x}_0$ and A from them. In addition to conditions A1 and A2, a common identifiability condition from aggregated/time-scaled observations is

A3 The sample size of the aggregated/time-scaled observations $\tilde{n} > d$.

3.2.3.1 Identifiability condition from aggregated observations

Though the identifiability of vector auto-regressive model from aggregated observations has been studied [74, 80], the identifiability condition of ODEs from aggregated observations remains unknown. Based on Theorem 3.4, we derive the following corollary.

Corollary 3.7 (aggregated observations)**.** *If conditions* A1-A3 *are satisfied, where* $\boldsymbol{x}_0 \in \mathbb{R}^d$ *and* $A \in \mathbb{R}^{d\times d}$*, then the aggregated ODE system parameters* $\tilde{\boldsymbol{x}}_0$ *and* $\tilde{A}$ *can be uniquely determined by the aggregated observations* $\tilde{\boldsymbol{X}}$*, and the original ODE system* (3.1) *(with parameters* $\boldsymbol{x}_0$ *and* A*) is* $(\boldsymbol{x}_0, A)$ *identifiable from the aggregated observations* $\tilde{\boldsymbol{X}}$*, with* $\boldsymbol{x}_0 = k(I + e^{A\Delta_t} + \cdots + e^{A(k-1)\Delta_t})^{-1}\tilde{\boldsymbol{x}}_0$ *and* $A = \tilde{A}$*.*

The proof of Corollary 3.7 can be found in Appendix A.1.3. This corollary implies that the ODE system (3.1) is still identifiable from the aggregated observations under mild conditions. Moreover, the new parameter matrix corresponding to the aggregated ODE system is the same as that of the actual model, that is, $\tilde{A} = A$.

3.2.3.2 Identifiability condition from time-scaled observations

Time plays a critical role in ODE systems. However, in practical applications, using the actual time of the system directly may cause inconvenience. A common practice is to use the method defined in Definition 3.6 to scale the actual timeline into a fixed one, such as [0,1], to simplify the calculation [38, 77]. How will the time scaling affect the causal relationship between variables, that is, parameter matrix A? This motivates us to derive the following corollary.

Corollary 3.8 (time-scaled observations)**.** *If conditions* A1-A3 *are satisfied, where* $\boldsymbol{x}_0 \in \mathbb{R}^d$ *and* $A \in \mathbb{R}^{d\times d}$*, then the time-scaled ODE system parameters* $\tilde{\boldsymbol{x}}_0$ *and* $\tilde{A}$ *can be uniquely determined by the time-scaled observations* $\tilde{\boldsymbol{X}}$*, and the original ODE system* (3.1) *(with parameters* $\boldsymbol{x}_0$ *and* A*) is* $(\boldsymbol{x}_0, A)$ *identifiable from the time-scaled observations* $\tilde{\boldsymbol{X}}$*, with* $\boldsymbol{x}_0 = \tilde{\boldsymbol{x}}_0$ *and* $A = k\tilde{A}$*.*

The proof of Corollary 3.8 can be found in Appendix A.1.4. This corollary implies that the ODE system (3.1) is identifiable from the time-scaled observations, and the new parameter matrix corresponding to the time-scaled ODE system is reduced by a factor of k from the parameter matrix corresponding to the actual model, that is, $\tilde{A} = A/k$. This corollary provides theoretical support for time scaling and implies that one can safely scale the actual timeline into a fixed one for the ODE system (3.1).

3.3 Asymptotic properties of the NLS estimator

Now that we have established the sufficient condition for the identifiability of the ODE system from discrete error-free observations. However, in practical applications, the observations are typically disturbed by measurement noise. In this case, one can not calculate the true unknown parameters explicitly from $e^{A\Delta t} = \boldsymbol{X}_2\boldsymbol{X}_1^{-1}$ and $\boldsymbol{x}_0 = e^{-At_1}\boldsymbol{x}_1$. Instead, we resort to parameter estimation procedures to estimate the unknown parameters. In this section, we will investigate the asymptotic properties of the parameter estimator based on the Nonlinear Least Squares (NLS) method. First, we introduce the measurement model and the NLS method.

Measurement model. Suppose the system state $\boldsymbol{x}(t)$ in ODE system (3.1) is measured with noise at time points $t_1, \dots, t_n$, with $t_i \in [0, T]$ for all $i = 1, \dots, n$, and $0 < T < +\infty$. Abusing notation a bit, from now on, we use $\boldsymbol{\theta} := (\boldsymbol{x}_0, A) \in \mathbb{R}^{d+d^2}$ to vectorize the parameters $\boldsymbol{x}_0, A$. We further let $\Theta := (M^0, \Omega)$ to denote the parameter space, where $M^0 \subset \mathbb{R}^d$ and $\Omega \subset \mathbb{R}^{d\times d}$. The true parameter is denoted as $\boldsymbol{\theta}^* := (\boldsymbol{x}_0^*, A^*)$. Then the measurement model can be described as:

$$\boldsymbol{y}_i = \boldsymbol{x}(t_i; \boldsymbol{\theta}^*) + \boldsymbol{\epsilon}_i = e^{A^* t_i}\boldsymbol{x}_0^* + \boldsymbol{\epsilon}_i\,, \tag{3.4}$$

for all $i = 1, \dots, n$, where $\boldsymbol{y}_i \in \mathbb{R}^d$ denotes the noisy observation at time t_i and $\boldsymbol{\epsilon}_i \in \mathbb{R}^d$ is the measurement error at time t_i.

Nonlinear least squares (NLS) method is well used for parameter estimation in nonlinear regression models including ODEs [32, 81]. In the following, based on the identifiability condition we build in Section 3.2, with mild assumptions, we show the consistency and asymptotic normality of the NLS estimator.

Suppose the ODE system (3.1) is $(\boldsymbol{x}_0^*, A^*)$ identifiable from a set of equally-spaced error-free observations sampled from a single trajectory: $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, with $\boldsymbol{x}_i = \boldsymbol{x}(t_i; \boldsymbol{\theta}^*) = e^{A^* t_i}\boldsymbol{x}_0^*$ and $t_i \in [0, T]$ for all $i = 1, \dots, n$, then the true parameter $\boldsymbol{\theta}^* = (\boldsymbol{x}_0^*, A^*)$ uniquely minimizes:

$$M(\boldsymbol{\theta}) = \frac{1}{T}\int_0^T \| \, e^{A^* t}\boldsymbol{x}_0^* - e^{At}\boldsymbol{x}_0 \, \|_2^2 \, dt\,, \tag{3.5}$$

where $\| \cdot \|_2$ denotes the Euclidean norm. The proof is straightforward, since the ODE system (3.1) is $(\boldsymbol{x}_0^*, A^*)$ identifiable from $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, the ODE system (3.1) is also $(\boldsymbol{x}_0^*, A^*)$

identifiable from the corresponding trajectory at time $[0, T]$, which implies that $M(\boldsymbol{\theta})$ attains its unique global minimum at $\boldsymbol{\theta}^*$. In practical applications, typically, one can only access the noisy observation $\boldsymbol{y}_i$'s as described in (3.4). Therefore, we propose to estimate $\boldsymbol{\theta}^*$ by minimizing the empirical version of $M(\boldsymbol{\theta})$, which is

$$M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \| \boldsymbol{y}_i - e^{At_i}\boldsymbol{x}_0 \|_2^2 \ . \tag{3.6}$$

That is, the NLS estimator of $\boldsymbol{\theta}^*$ is defined as

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta})\,.$$

Assumptions. Now we investigate the asymptotic properties of the NLS estimator $\hat{\boldsymbol{\theta}}_n$. We first list all the required assumptions:

A4 Parameter space Θ is a compact subset of $\mathbb{R}^{d+d^2}$.

A5 Error terms $\{\boldsymbol{\epsilon}_i\}$ for $i = 1, \ldots, n$ are independent and identically distributed random vectors with mean zero and covariance matrix $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, where $0 < \sigma_j^2 < \infty$ for all $j = 1, \ldots, d$.

A6 We have n equally-spaced observations $\boldsymbol{Y} := \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$, where $\boldsymbol{y}_i$ is defined by measurement model (3.4). Without loss of generality, we assume observation time starts with $t_1 = 0$, ends with $t_n = T$, and thus the equal time space $\Delta t = T/(n-1)$.

A7 $\boldsymbol{\theta}^*$ is an interior point of the parameter space Θ.

In addition to the aforementioned assumptions A4-A7, assumptions A1 and A2 stated in Theorem 3.4 are required with respect to the true parameter $\boldsymbol{\theta}^* = \{\boldsymbol{x}_0^*, A^*\}$. These two assumptions guarantee the ODE system (3.1) is $(\boldsymbol{x}_0^*, A^*)$ identifiable from any $d+1$ equally-spaced error-free observations sampled from the trajectory $\boldsymbol{x}(\cdot; \boldsymbol{x}_0^*, A^*)$. A1 and A2 are needed because the identifiability of the system is a prerequisite for obtaining a consistent parameter estimator. A4 is commonly used in deriving consistency for parameter estimators, as demonstrated in references such as [82, Thm 2.1, Thm 2.5, Thm 2.6]. While alternative conditions for consistency exist without the compactness assumption, they typically require the objective function to be convex [82]. Given that our objective function $M_n(\boldsymbol{\theta})$ is not convex, the compactness assumption remains indispensable. The compactness assumption implicitly requires having known bounds on the true parameter values, which can be challenging to check in real-world situations since the true parameters of the ODE systems are unknown. However, since our derived confidence sets do not depend on the bounds of the parameter values, we may safely assume enormously large parameter boundaries. Consequently, in practical applications, there is no need to verify this assumption, and the derived statistical inferences will remain unaffected. A5 is a common way to define measurement noise, note that we do not require the error terms follow a normal distribution. It is worth mentioning that the error terms are not necessarily identically distributed. In other words, one can easily generate our theoretical results for cases that only require the error terms being independently distributed. A6 ensures the observations are collected at equally-spaced time points, following the rules of observations collection required by Theorem 3.4. Assumption A7 is a standard condition for proving asymptotic normality [32, 83, 84].

3.3.1 Consistency

In this subsection, we study the consistency of the NLS estimator. An estimator is said to be consistent, if it converges in probability to the true value of the parameter.

Theorem 3.9. *Suppose assumptions* A1, A2 *are satisfied with respect to* $\boldsymbol{\theta}^*$ *and assumptions* A4-A6 *hold, the NLS estimator* $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}^*$*, as* $n \to \infty$*.*

The notation $\xrightarrow{p}$ stands for convergence in probability. The proof of Theorem 3.9 can be found in Appendix A.1.5. This theorem shows the consistency of our NLS estimator $\hat{\boldsymbol{\theta}}_n$ to the true $\boldsymbol{\theta}^*$. That is, as n goes to infinity, the NLS estimator $\hat{\boldsymbol{\theta}}_n$ converges to the true system parameters $\boldsymbol{\theta}^*$ with probability approaching 1. Note that the consistency of the estimator is a necessary condition for the estimator's asymptotic normality.

In the case where we only observe degraded data, we let $\tilde{\boldsymbol{Y}} := (\tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2, \ldots, \tilde{\boldsymbol{y}}_{\tilde{n}})$ denote the noisy aggregated/time-scaled data, which are collected from the original observations $\boldsymbol{Y}$ in the same way as the corresponding error-free observations $\tilde{\boldsymbol{X}}$ from $\boldsymbol{X}$ defined in Definition 3.5/3.6. Then we can estimate the corresponding parameters $\tilde{\boldsymbol{\theta}}^* := (\tilde{\boldsymbol{x}}_0^*, \tilde{A}^*)$ by minimizing the NLS objective function in (3.6), with the data replaced by $\tilde{\boldsymbol{Y}}$. Let such an estimator be $\hat{\tilde{\boldsymbol{\theta}}} := (\hat{\tilde{\boldsymbol{x}}}_0, \hat{\tilde{A}})$. Using the relationship between $\tilde{\boldsymbol{\theta}}^*$ and $\boldsymbol{\theta}^*$ found in Corollary 3.7/3.8, we can define a mapping $g : \mathbb{R}^{d+d^2} \to \mathbb{R}^{d+d^2}$ from $\tilde{\boldsymbol{\theta}}^*$ to $\boldsymbol{\theta}^*$, and obtain an estimator of $\boldsymbol{\theta}^*$ by $\hat{\boldsymbol{\theta}}_{\tilde{n}} := g(\hat{\tilde{\boldsymbol{\theta}}})$. In the following, we present the expressions of g and the consistency of the NLS estimators by using the aggregated and time-scaled observations, respectively.

Corollary 3.10 (aggregated observations)**.** *Suppose assumptions* A1, A2 *are satisfied with respect to* $\boldsymbol{\theta}^*$ *and assumptions* A3-A6 *hold, the NLS estimator* $\hat{\boldsymbol{\theta}}_{\tilde{n}} \xrightarrow{p} \boldsymbol{\theta}^*$*, as* $\tilde{n} \to \infty$*, where* $\hat{\boldsymbol{\theta}}_{\tilde{n}} := g(\hat{\tilde{\boldsymbol{\theta}}}) := \big(k(I + e^{\hat{\tilde{A}}\Delta_t} + \cdots + e^{\hat{\tilde{A}}(k-1)\Delta_t})^{-1}\hat{\tilde{\boldsymbol{x}}}_0, \hat{\tilde{A}}\big)$.

Corollary 3.11 (time-scaled observations)**.** *Suppose assumptions* A1, A2 *are satisfied with respect to* $\boldsymbol{\theta}^*$ *and assumptions* A3-A6 *hold, the NLS estimator* $\hat{\boldsymbol{\theta}}_{\tilde{n}} \xrightarrow{p} \boldsymbol{\theta}^*$*, as* $\tilde{n} \to \infty$*, where* $\hat{\boldsymbol{\theta}}_{\tilde{n}} := g(\hat{\tilde{\boldsymbol{\theta}}}) := (\hat{\tilde{\boldsymbol{x}}}_0, k\hat{\tilde{A}})$.

The proofs of Corollary 3.10 and Corollary 3.11 can be found in Appendix A.1.6 and Appendix A.1.7, respectively. Now in addition to the assumptions mentioned in Theorem 3.9, assumption A3 is also required for the consistency from degraded observations. These two corollaries show that the NLS estimators obtained from degraded observations are consistent to the true parameters of the original ODE (3.1). Since we have established the identifiability conditions for the original system parameters of the ODE (3.1) from the degraded observations in Corollary 3.7/3.8 in Section 3.2, the consistency of their NLS estimators is a natural result from Theorem 3.9. To see this, we first show that the NLS estimator from aggregated/time-scaled observations converges to the true system parameter corresponding to the new ODE system, that is, $\hat{\tilde{\boldsymbol{\theta}}} \xrightarrow{p} \tilde{\boldsymbol{\theta}}^*$, as $\tilde{n} \to \infty$. Since we have derived the mapping g such that $g(\tilde{\boldsymbol{\theta}}^*) = \boldsymbol{\theta}^*$ in Corollary 3.7/3.8, then by multivariate continuous mapping theorem, one takes the function $g(\cdot)$ with respect to $\hat{\tilde{\boldsymbol{\theta}}}$ and $\tilde{\boldsymbol{\theta}}^*$, respectively, one can reach the conclusion $g(\hat{\tilde{\boldsymbol{\theta}}}) \xrightarrow{p} g(\tilde{\boldsymbol{\theta}}^*)$, as $\tilde{n} \to \infty$. That is, $\hat{\boldsymbol{\theta}}_{\tilde{n}} \xrightarrow{p} \boldsymbol{\theta}^*$, as $\tilde{n} \to \infty$.

3.3.2 Asymptotic normality

After establishing the consistency of our NLS estimators, we can study their asymptotic distributions.

Theorem 3.12. *Suppose assumptions* A1, A2 *are satisfied with respect to* $\boldsymbol{\theta}^*$ *and assumptions* A4-A7 *hold, we have* $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{d} N(\mathbf{0}, H^{-1}VH^{-1})$*, as* $n \to \infty$*, where* $H = \nabla^2_{\boldsymbol{\theta}} M(\boldsymbol{\theta}^*)$*, is the Hessian matrix of* $M(\boldsymbol{\theta})$ *at* $\boldsymbol{\theta}^*$ *and* $V = \lim_{n\to\infty} var(\sqrt{n}\nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}^*))$, *with* $\nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}^*)$ *the gradient of* $M_n(\boldsymbol{\theta})$ *at* $\boldsymbol{\theta}^*$*.* $M(\boldsymbol{\theta})$ *and* $M_n(\boldsymbol{\theta})$ *are defined in Equations* (3.5) *and* (3.6)*, respectively.*

The notation $\xrightarrow{d}$ stands for convergence in distribution. The proof of Theorem 3.12 can be found in Appendix A.1.8. From the theorem, we see that the NLS estimator is asymptotically normal, and the convergence rate is $n^{-1/2}$. Here, the rate meets the one of the standard parametric NLS estimator [71, 81]. This result is reasonable because our model in (3.4) is a parametric one.

Now, with the asymptotic normality result, if we can estimate the asymptotic covariance matrix $\Sigma^* := H^{-1}VH^{-1}$, we can perform statistical inference for the unknown system parameters. In particular, we derive the explicit expressions of matrices H and V in (A.47) and (A.42). According to their formulae, they are functions of the true system parameter $\boldsymbol{\theta}^*$, which are unknown in practice. Therefore, we approximate H and V by substituting $\boldsymbol{\theta}^*$ with the NLS parameter estimate $\hat{\boldsymbol{\theta}}_n$ in their formulae. Then the inference can be performed based on the approximated covariance matrix, denoted by $\hat{\Sigma}_n$. In the following two subsections, we introduce the details of the inference. For those who are not familiar with statistical inference, please refer to [85, 86] for relevant concepts and methods.

3.3.2.1 Confidence sets for unknown parameters

Based on Theorem 3.12, we can derive the approximate confidence sets for the unknown true parameters $\boldsymbol{\theta}^*$. In our work, we employ the term "confidence set" as a general descriptor for a range of values within which we have a certain level of confidence that the true parameters reside. This umbrella term encompasses two specific types of confidence sets: confidence intervals (CIs) and confidence regions (CRs). When the parameter under consideration is one-dimensional, we refer to the confidence set as a confidence interval (CI). When the parameter is multidimensional, we use the term confidence region (CR) to describe the confidence set. For the detailed theoretical definition, please refer to [85].

We first construct the simultaneous confidence region (CR) for all the $d+d^2$ parameters $\boldsymbol{\theta}_i^*$, where $\boldsymbol{\theta}_i^*$ denotes the ith entry of $\boldsymbol{\theta}^*$, and $i = 1, \ldots, d+d^2$. According to Theorem 3.12, we have

$$n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^\top \{\Sigma^*\}^{-1}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{d} \chi^2_{d+d^2}\,, \text{ as } n \to \infty\,.$$

Therefore, we approximate the $1-\alpha$ CR by the set

$$\left\{\boldsymbol{\theta}: \quad n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})^\top \hat{\Sigma}_n^{-1}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \leq \chi^2_{d+d^2}(1-\alpha)\right\}, \tag{3.7}$$

where $\chi^2_m(1-\alpha)$ denotes the upper-tail critical value of χ^2 distribution with m degrees of freedom at significance level α. Please note that in statistical hypothesis testing, the significance level α represents the probability of rejecting the null hypothesis when it is actually true. Typically set to 5% or lower depending on the field of study, α is chosen by the experimenter. In the meantime, $1-\alpha$ denotes the corresponding confidence level, where CI/CR contains the true parameter value $(1-\alpha)\%$ of the time.

We then construct a pointwise confidence interval (CI) for each $\boldsymbol{\theta}^*_i$, $i = 1, \ldots, d+d^2$. Based on Theorem 3.12, we can derive that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{ni} - \boldsymbol{\theta}^*_i) \xrightarrow{d} N(0, D_i(\Sigma^*)), \text{ as } n \to \infty,$$

where $\hat{\boldsymbol{\theta}}_{ni}$ denotes the ith entry of $\hat{\boldsymbol{\theta}}_n$ and $D_i(M)$ denotes the ith diagonal entry of matrix M. Then, the $1-\alpha$ CI for each $\boldsymbol{\theta}^*_i$ can be estimated by

$$\left[\hat{\boldsymbol{\theta}}_{ni} - z_{\alpha/2}\sqrt{D_i(\hat{\Sigma}_n)/n}, \quad \hat{\boldsymbol{\theta}}_{ni} + z_{\alpha/2}\sqrt{D_i(\hat{\Sigma}_n)/n}\right], \tag{3.8}$$

where $z_{\alpha/2}$ is the two-tailed critical value of the standard normal distribution at significance level α. That is, if $Z \sim N(0,1)$, then $\mathbb{P}(Z < -z_{\alpha/2}) + \mathbb{P}(Z > z_{\alpha/2}) = \alpha$.

3.3.2.2 Infer the causal structure of the ODE system

Another application of Theorem 3.12 is to infer the causal structure among variables within the ODE system, specifically by testing the hypothesis $a^*_{jk} = 0$, where a^*_{jk} denotes the jk-th entry of the true parameter matrix A^*, with $j, k = 1, \ldots, d$. When $a^*_{jk} \neq 0$, as delineated by the expression of the ODE system (3.1), the derivative of $x_j(t)$ is influenced by $x_k(t)$, implying a causal link from variable x_k to variable x_j, as referenced in Schölkopf et al. [22]. Here, x_j denotes the j-th variable of the ODE system, and $x_j(t)$ denotes the state of the j-th variable at time t. Note that the ODE system is fully observable, that is, there are no latent variables interacting with the system.

Then based on the derived $1-\alpha$ CI for each parameter in (3.8), we propose to conduct a hypothesis test:

$$H_0 : a^*_{jk} = 0 \;\text{ vs }\; H_1 : a^*_{jk} \neq 0 \tag{3.9}$$

by assessing the following inequality:

$$|\hat{a}_{jk}| > z_{\alpha/2}\sqrt{D_{d+(j-1)d+k}(\hat{\Sigma}_n)/n}\,, \tag{3.10}$$

where $\hat{a}_{jk}$ denotes the estimator of a^*_{jk}, which is the test statistic and equals the $d + (j-1)d + k$-th entry of $\hat{\boldsymbol{\theta}}_n$. If (3.10) holds, we have significant evidence to reject the null hypothesis at the significance level α, and conclude that $a^*_{jk} \neq 0$, thereby affirming a causal link from variable x_k to variable x_j.

3.3.2.3 Asymptotic normality of NLS estimators from degraded observations

Here we present the asymptotic normality of NLS estimators from degraded observations. Recalling the transformation rules from $\tilde{\boldsymbol{\theta}}^*$ to $\boldsymbol{\theta}^*$, g, defined in Corollary 3.10/3.11, we denote its gradient at $\tilde{\boldsymbol{\theta}}^*$ by $G := \nabla g(\tilde{\boldsymbol{\theta}}^*)$. We have derived the explicit formulae of $H(T, \boldsymbol{\theta}^*)$ and $V(T, \boldsymbol{\theta}^*)$ as matrix functions of T and $\boldsymbol{\theta}^*$ in (A.47)/(A.42). Then we establish the following corollaries.

Corollary 3.13 (aggregated observations)**.** *Suppose assumptions* A1, A2 *are satisfied with respect to* $\boldsymbol{\theta}^*$*, assumptions* A3-A6 *hold and assumption* A7 *is satisfied with respect to* $\tilde{\boldsymbol{\theta}}^*$*, we have* $\sqrt{\tilde{n}}(\hat{\boldsymbol{\theta}}_{\tilde{n}} - \boldsymbol{\theta}^*) \xrightarrow{d} N(\mathbf{0}, G\tilde{H}^{-1}\tilde{V}\tilde{H}^{-1}G^\top)$*, as* $\tilde{n} \to \infty$*, where* $\hat{\boldsymbol{\theta}}_{\tilde{n}}$ *is defined in Corollary 3.10,* $\tilde{H} = H(\tilde{T}, \tilde{\boldsymbol{\theta}}^*)$ *and* $\tilde{V} = V(\tilde{T}, \tilde{\boldsymbol{\theta}}^*)/k$*, with* $\tilde{T} = (\lfloor n/k \rfloor - 1)kT/(n-1)$ *and* $\tilde{\boldsymbol{\theta}}^* = (\tilde{\boldsymbol{x}}^*_0, \tilde{A}^*) = \big((I + e^{A^*\Delta_t} + \cdots + e^{A^*(k-1)\Delta_t})\boldsymbol{x}^*_0/k, A^*\big)$*.*

Corollary 3.14 (time-scaled observations)**.** *Suppose assumptions* A1, A2 *are satisfied with respect to* $\boldsymbol{\theta}^*$*, assumptions* A3-A6 *hold and assumption* A7 *is satisfied with respect to* $\tilde{\boldsymbol{\theta}}^*$*, we have* $\sqrt{\tilde{n}}(\hat{\boldsymbol{\theta}}_{\tilde{n}} - \boldsymbol{\theta}^*) \xrightarrow{d} N(\mathbf{0}, G\tilde{H}^{-1}\tilde{V}\tilde{H}^{-1}G^\top)$*, as* $\tilde{n} \to \infty$*, where* $\hat{\boldsymbol{\theta}}_{\tilde{n}}$ *is defined in Corollary 3.11,* $\tilde{H} = H(kT, \tilde{\boldsymbol{\theta}}^*)$ *and* $\tilde{V} = V(kT, \tilde{\boldsymbol{\theta}}^*)$*, with* $\tilde{\boldsymbol{\theta}}^* = (\tilde{\boldsymbol{x}}^*_0, \tilde{A}^*) = (\boldsymbol{x}^*_0, A^*/k)$*.*

The proofs of Corollary 3.13 and Corollary 3.14 can be found in Appendix A.1.9 and Appendix A.1.10, respectively. With the consistency property of the NLS estimators, these two corollaries can be directly derived from Theorem 3.12 by using multivariate

Delta method. The explicit formulae for matrices G, $\tilde{H}$ and $\tilde{V}$ for aggregated/time-scaled observations are derived in the proofs.

Worth to be noted that matrix $\tilde{V}$ in Corollary 3.13 is not $V(T, \boldsymbol{\theta}^*)$ by substituting T and $\boldsymbol{\theta}^*$ with $\tilde{T}$ and $\tilde{\boldsymbol{\theta}}^*$, but also reduced by a factor of k. The reason is that, the equation of V includes the variance matrix Σ of the error term $\boldsymbol{\epsilon}_i$ in (A.42). By the generation rules of aggregated observations defined in Definition 3.5, the variance of the aggregated noise term $\tilde{\boldsymbol{\epsilon}}_i$ becomes k times smaller than that of the original one, that is $\tilde{\Sigma} = \Sigma/k$. And thanks to the reduced variance, we will show that the parameter estimates of aggregated observations can reach the same level of accuracy as that of the original observations with a much smaller sample size in the simulation results in subsection 3.4.4.1.

Now that we have derived the asymptotic normality results from aggregated/time-scaled observations. We can perform statistical inference for the unknown original system parameters $\boldsymbol{\theta}^*$ using the same way introduced in subsection 3.3.2.1 and 3.3.2.2.

3.4 Simulations

The simulations in this section are designed to illustrate the asymptotic results established in Section 3.3. The identifiability conditions in Section 3.2 are deterministic and established by theoretical proofs, and therefore do not require simulation-based validation. In contrast, the asymptotic properties of the estimator are statistical in nature and thus benefit from numerical verification through simulations.

3.4.1 Data simulation

For each $d = 2, 3, 4$, we first randomly generate a $d \times d$ parameter matrix A_d^* and a $d \times 1$ initial condition $\boldsymbol{x}_{0d}^*$ as the true system parameters for each d-dimensional ODE system (3.1). Moreover, to test whether $a_{jk}^* = 0$, we randomly set several entries to zero in each A_d^*. Without loss of generality, we set $T = 1$. Then n equally-spaced noisy observations are generated based on Equation (3.4) in $[0, 1]$ time interval with error term $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathrm{diag}(0.05^2, \ldots, 0.05^2))$. We tested various sample sizes for each d-dimensional ODE system. For each configuration, we run 200 random replications. The A_d^* and $\boldsymbol{x}_{0d}^*$

are shown below.

$$A_2^* = \begin{bmatrix} 1.76 & -0.1 \\ 0.98 & 0 \end{bmatrix}, \; A_3^* = \begin{bmatrix} 1.76 & 0 & 0.98 \\ 2.24 & 0 & -0.98 \\ 0.95 & 0 & -0.1 \end{bmatrix}, \; A_4^* = \begin{bmatrix} 1.76 & 0.9 & 0 & 2.24 \\ 1.87 & -0.98 & 0 & -1.15 \\ -1.1 & 0 & 0.64 & 0 \\ 1.26 & 0.12 & 0.94 & 0 \end{bmatrix},$$

and $\boldsymbol{x}_{02}^* = [1.87, -0.98]^\top$, $\boldsymbol{x}_{03}^* = [0.41, 0.14, 1.45]^\top$, $\boldsymbol{x}_{04}^* = [-0.42, 1.01, 1.97, -0.38]^\top$.

Note that since the NLS loss function (3.6) is a non-convex function, in practical application, one may require a global optimization technique to obtain the NLS estimates. However, in this chapter, we focus on the theoretical statistical properties analysis of the NLS estimator. Theoretically, the non-convexity of the NLS loss function does not influence our derived theoretical results. Therefore, for the purpose of illustrating our theoretical results, we do not apply the global optimization technique in our simulation due to its high computational cost. Instead, we use a bound-constrained minimization technique [87] to obtain the NLS estimates. In order to get the global minimum NLS estimate (or a local minimum that is close enough to the global minimum), we apply two tricks when implementing the optimization method. Firstly, we initialize the parameter with a value close to the true parameter (for example, $\boldsymbol{\theta}^* - 0.001$). Secondly, we constrain the bounds of the parameter within a reasonable neighbourhood of the true parameter (for example, $[\boldsymbol{\theta}^* - 0.5, \boldsymbol{\theta}^* + 0.5]$). According to the simulation results below, the attained NLS estimates are precise enough to illustrate the correctness of our theoretical results. In other words, suppose we apply a global optimization technique to obtain the NLS estimates. In that case, the simulation results will be more supportive of the correctness of our theoretical results with the help of potentially more accurate NLS estimates.

3.4.2 Metrics

Mean Squared Error (MSE) is introduced to check the consistency of the parameter estimator. It is defined as:

$$\text{MSE} = \frac{1}{N} \sum_{j=1}^{N} \| \hat{\boldsymbol{\theta}}_n^{(j)} - \boldsymbol{\theta}^* \|_2^2 \,,$$

where $\hat{\boldsymbol{\theta}}_n^{(j)}$ denotes the estimated parameter of the jth replication, and N is the number of replications.

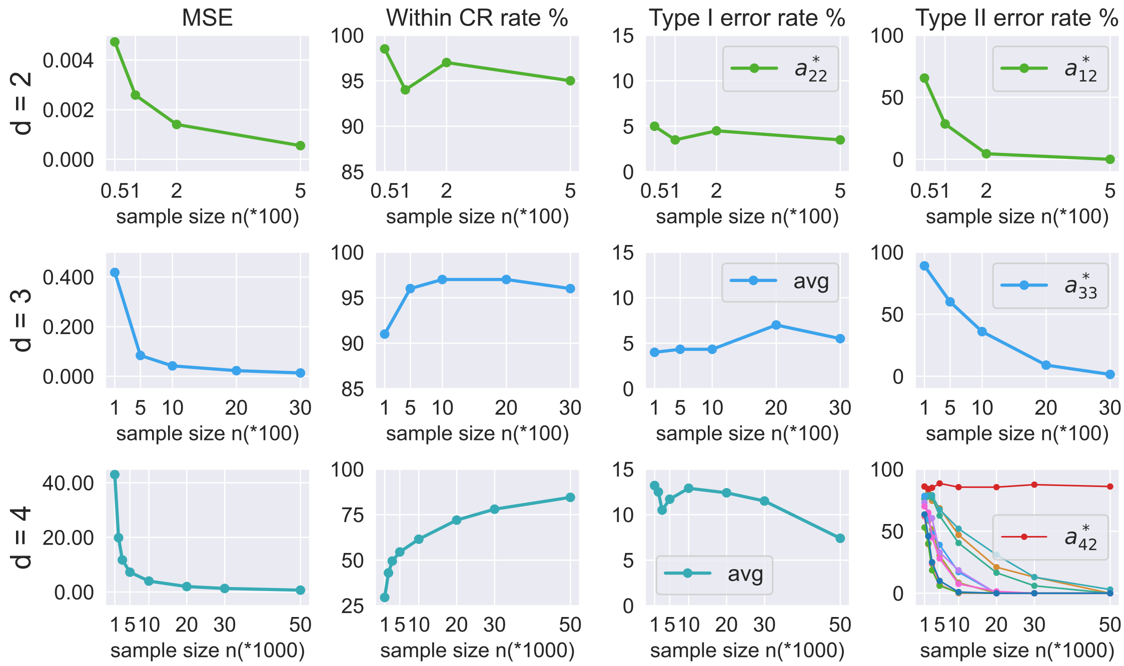

FIGURE 3.1: Simulation results for $d = 2, 3, 4$ dimensional ODE system, respectively

Within CR rate is defined as the rate of replications with the true parameters $\boldsymbol{\theta}^*$ included in the CR at 95% confidence level. Whether the 95% CR includes $\boldsymbol{\theta}^*$ at each replication can be calculated by Equation (3.7). Worth to be mentioned that here we use this metric aiming to test the correctness of our asymptotic normal theory, for example, the variance matrix Σ^*, not to infer the confidence region. Therefore, we use the true value of Σ^* here rather than the estimated one in Equation (3.7). Since we set the confidence level at 95%, if our theoretical results are correct, the within CR rate in our simulation results should approach 95%.

Type I/II error rate is calculated based on the hypothesis test introduced in subsection 3.3.2.2. We set the significance level $\alpha = 0.05$. Type I error rate is the rate of replications rejecting the null hypothesis, for $a^*_{jk} = 0$. And type II error rate is the rate of replications not rejecting the null hypothesis, for $a^*_{jk} \neq 0$. Whether we reject the null hypothesis or not is calculated by Equation (3.10). Given the significance level set at 0.05, the anticipated outcome is that if our theoretical results are correct, the Type I error rate in our simulation results should approximate 5%, while the Type II error rate should tend towards zero. It is noteworthy that the lower the type II error rate is, the more powerful our causal structure inference test is.

3.4.3 Results analysis

The simulation results are presented in Figure 3.1. We first explain the legend in the figure. Since A_2^* only has one zero entry a_{22}^*, the type I error rate is based on a_{22}^*. However, A_3^* and A_4^* have multiple zero entries, and the type I error rate for each zero entry is similar. Therefore, we show the average value of all zero entries in A_3^* and A_4^*, labelled as avg. For type II error rate in cases with $d = 2$ and $d = 3$, we only show the value of a_{12}^* and a_{33}^*, respectively. Because all other non-zero entries in A_2^* and A_3^* have zero or close to zero type II error rate since the sample size n is small. For $d = 4$, we present the results of all the 11 non-zero entries in A_4^*, but due to space limitations, we only label entry a_{42}^*, which has a different trend from others.

It can be seen from the first column in Figure 3.1 that for all three cases where $d = 2, 3$ and 4, MSE decreases and approaches zero with the increase of sample size n, which indicates the consistency of the estimators. As can be seen from the figure, for $d = 2$ and $d = 3$ cases, the within CR rate is around 95%, and the type I error rate is around 5% for all different sample sizes n. Moreover, the type II error rate reduces as the sample size increases and attains or approaches zero when the sample size is large enough. This result implies the correctness of our asymptotic normality theory and indicates that the test of causal structure inference for the ODE system is powerful.

For the 4-dimensional case, we can see that with the increase of sample size, the within CR rate increases, and the type I error rate decreases, which implies that as the parameter estimates approach their true parameter values, the within CR rate and type I error rate is closer to 95% and 5% respectively. They do not attain their ideal values in our simulation because the parameter estimates are not precise enough under the current sample sizes due to the high dimension of the system parameters. For the same reason, the type II error rate of entry a_{42}^* keeps high. This result is reasonable because $a_{42}^* = 0.12$ is close to zero. When the parameter estimate is not accurate enough, it is easy to get a wrong result that does not reject the null hypothesis. The type II error rates for $d = 2$ and $d = 3$ cases also support this conclusion. We can see that the absolute values of a_{12}^* in A_2^* and a_{33}^* in A_3^* are also small. Nevertheless, as sample size increases, with the help of sufficiently accurate parameter estimates, their type II error rates approach zero.

It is worth noting that in the 4-dimensional case, the type II error rates of other entries approach zero when the sample size is much smaller compared to the case of a_{42}^*. This result implies that when the causal effect between variables, that is, $|a^*_{jk}|$, is significant, the causal structure can be easily and correctly discovered using our method. However, for cases where the causal effect is small or negligible, we need a sufficiently large sample to discover the causal relationship.

3.4.4 Simulation results from degraded observations

In this subsection, we illustrate the corollaries built on aggregated/time-scaled observations in Section 3.3 by simulation.

We chose the $d = 3$ case with the true system parameters $(\boldsymbol{x}^*_{03}, A^*_3)$ the same as the one we used in subsection 3.4.1. The original noisy observations are generated using the same way we introduced in subsection 3.4.1. And then the aggregated/time-scaled observations are generated from the original ones based on Definition 3.5/Definition 3.6 with various k.

3.4.4.1 Simulation results from aggregated observations

In the following, we show the simulation results for aggregated observations with $k = 5, 10,$ and 20, respectively. Moreover, to compare the results from the aggregated and the original observations, we also present the simulation results from the original observations in Table 3.1. We use n and $\tilde{n}$ to denote the sample size of the original observations and the aggregated observations, respectively.

TABLE 3.1: Simulation results from original observations

Sample Size		MSE	CR Rate%	Type I Error Rate%			Type II Error Rate%					
n	$\tilde{n}$			a_{12}	a_{22}	a_{32}	a_{11}	a_{13}	a_{21}	a_{23}	a_{31}	a_{33}
100	-	0.480	94	3	2	5.5	0	0	0	0	0	82.5
200	-	0.243	97.5	5.5	4.5	4	0	0	0	0	0	75.5
500	-	0.093	97	3.5	4	2.5	0	0	0	0	0	38.5
1000	-	0.045	95	5	4.5	2	0	0	0	0	0	8.5
2000	-	0.023	98	7	4	3.5	0	0	0	0	0	0

The results show that for all three cases where $k = 5, 10$ and 20, MSE decreases and approaches zero with the increase of sample size $\tilde{n}$, which indicates the consistency of the estimators. As can be seen from the tables, the within CR rate is around 95% and the

Table 3.2: Simulation results from aggregated observations with $k = 5$

Sample Size		MSE	CR Rate%	Type I Error Rate%			Type II Error Rate%					
n	$\tilde{n}$			a_{12}	a_{22}	a_{32}	a_{11}	a_{13}	a_{21}	a_{23}	a_{31}	a_{33}
100	20	0.498	95.5	1.5	2	2.5	0	0	0	0	0	88.5
200	40	0.247	99	4.5	3.5	3	0	0	0	0	0	81
500	100	0.092	97.5	3.5	4	2.5	0	0	0	0	0	42
1000	200	0.045	95.5	5	4.5	2	0	0	0	0	0	9.5
2000	400	0.023	98.5	6.5	4	3.5	0	0	0	0	0	0

Table 3.3: Simulation results from aggregated observations with $k = 10$

Sample Size		MSE	CR Rate%	Type I Error Rate%			Type II Error Rate%					
n	$\tilde{n}$			a_{12}	a_{22}	a_{32}	a_{11}	a_{13}	a_{21}	a_{23}	a_{31}	a_{33}
100	10	0.536	84	1	0.5	2.5	0	0	0	0	0	94.5
200	20	0.262	98	3	2.5	2.5	0	0	0	0	0	85.5
500	50	0.093	98	3	3	1.5	0	0	0	0	0	47
1000	100	0.045	96	5	4.5	2	0	0	0	0	0	8.5
2000	200	0.023	98.5	6.5	3.5	3.5	0	0	0	0	0	0

Table 3.4: Simulation results from aggregated observations with $k = 20$

Sample Size		MSE	CR Rate%	Type I Error Rate%			Type II Error Rate%					
n	$\tilde{n}$			a_{12}	a_{22}	a_{32}	a_{11}	a_{13}	a_{21}	a_{23}	a_{31}	a_{33}
100	5	0.858	30.5	0.5	1	0.5	0	0	0	0	3.5	99
200	10	0.301	86	0.5	3	1.5	0	0	0	0	0	91
500	25	0.093	98	3.5	2.5	1	0	0	0	0	0	53.5
1000	50	0.045	96.5	4	3.5	2	0	0	0	0	0	11.5
2000	100	0.023	99	6	4	3.5	0	0	0	0	0	0

type I error rate for each of the zero entries in A is around 5% when the sample size $\tilde{n}$ is large enough in all $k = 5, 10$ and 20 cases. In addition, the type II error rate reduces to zero as the sample size $\tilde{n}$ increases. This result implies the correctness of our asymptotic normality theory and indicates the test of causal structure inference for the ODE system is powerful.

As can be seen from the tables, the type II error rate from the aggregated observations tend to be slightly greater than that of the original observations for each n. Specifically, with the greater the k being, the greater the type II error rate is. This is reasonable, because the sample size of the aggregated observations ($\tilde{n}$) is much smaller than that of the original ones (n), which causes the lack of accuracy of the parameter estimates when the sample size $\tilde{n}$ is not large enough. Thus leading to a higher MSE and a higher type

II error rate.

However, it can be seen from the last two rows in each of the four tables, the MSEs and type II error rates are almost same for each case, which implies that when the sample size of the aggregated observations $\tilde{n}$ is large enough, the parameter estimates of the aggregated observations can reach the same level of accuracy as that of the original observations. In addition, the power of inferring the causal structure of the ODE system from aggregated observations can be as good as that from the original observations. The reason why the aggregated observations with a much smaller sample size $\tilde{n} = n/k$ can still perform as good as the original observations with the corresponding size n is that, the variance of the aggregated noise term $\tilde{\boldsymbol{\epsilon}}_i$ becomes k times smaller than that of the original one $\boldsymbol{\epsilon}_i$, that is

$$\tilde{\Sigma} = \Sigma/k$$

based on the generation rules of the aggregated observations. Therefore, with a much smaller noise variance, the parameter estimates from aggregated observations can reach the same level of accuracy as that of the original observations with a much smaller sample size.

3.4.4.2 Simulation results from time-scaled observations

In the following, we show the simulation results for time-scaled observations with $k = 0.01, 0.1, 1, 10,$ and 100, respectively. Since for all the metrics except MSE, under all cases the simulation results are identical, we present their results in one table. The MSE for each k are the same up to 10^{-5}, therefore, the differences of MSE are negligible and one can safely conclude that the simulation results for time-scaled observations with various k is the same as that of the original observations (that is, $k = 1$). This implies the correctness of our theoretical results of the time-scaled observations established in Section 3.3.

3.5 Related work

In this section, we introduce the related work from three closely related aspects.

TABLE 3.5: Simulation results from time-scaled observations with $k = 0.01, 0.1, 1, 10, 100$

Sample Size		MSE	CR Rate%	Type I Error Rate%			Type II Error Rate%					
n	$\tilde{n}$			a_{12}	a_{22}	a_{32}	a_{11}	a_{13}	a_{21}	a_{23}	a_{31}	a_{33}
100	100	0.419	96	2	2.5	2	0	0	0	0	0	85.5
200	200	0.269	94	7	6	4.5	0	0	0	0	0	72.5
500	500	0.104	91.5	5.5	5.5	6	0	0	0	0	0	38.5
1000	1000	0.050	92.5	5	6	3	0	0	0	0	0	9
2000	2000	0.022	95	3.5	4.5	0.5	0	0	0	0	0	1

Identifiability Analysis of Linear ODE systems Most current studies for identifiability analysis of parameters in linear dynamical systems are in the control theory [23–27]. In the applied mathematics area, Stanhope et al. [28], Qiu et al. [29] provided systematic studies of the identifiability analysis of linear ODE systems from a **single** trajectory. Our identifiability analysis is built upon [28]. However, instead of building identifiability based on the entire continuous trajectory, we extend the identifiability work to more practical scenarios where we only have discrete observations sampled from the trajectory and do not know the initial conditions. The authors in [28] also discussed using equally-spaced error-free observations to calculate the system parameters explicitly. Our work's main distinction is that we build the identifiability condition entirely on the parameter of interest, that is, (A and $\boldsymbol{x}_0$), without further linearly independent assumption on the observations. With the help of our identifiability condition, the parameter estimator's asymptotic properties can be established with mild assumptions. Specifically, the covariance matrix of the asymptotic normality distribution can be explicitly expressed. Thus, we can perform statistical inference for the unknown parameters. Moreover, we treat the initial condition $\boldsymbol{x}_0$ as a parameter while $\boldsymbol{x}_0$ is a given fixed value in their setting. In the most recent work [29], the authors proposed several quantitative scores for identifiability analysis of linear ODEs in practice.

Parameter Estimation Methods for ODE systems The NLS method is applied to estimate parameters in ODE systems in [30–32]. However, to our knowledge, no existing work provides a systematic asymptotic analysis of the NLS estimator for the ODE system (3.1). The closest related work is that of Xue et al. [32], who studied asymptotic properties of the NLS estimator based on approximating ODEs' solutions by using the Runge-Kutta algorithm [88]. Nonetheless, their work requires strong assumptions, which are complex and cumbersome to verify. In contrast, our approach necessitates milder assumptions and yields an analytic covariance matrix for the asymptotic normal distribution. In addition to the NLS method, the two-stage smoothing-based estimation method is also well used for parameter estimation in ODE systems and its asymptotic properties have been extensively explored [5, 33, 34, 89, 90]. This method usually applies smoothing approaches such as penalized splines to estimate the state variables and their derivatives at the first stage. Thus a large number of observations are needed to ensure the estimates' accuracy. Principal differential analysis [35, 36, 68, 91] and Bayesian approaches [37] were also proposed to estimate unknown parameters in ODE systems. In recent years, several neural-network-based parameter estimation methods for ODE systems have been proposed [38, 39, 92]. In these works, the authors use multiple (usually a large number) trajectories instead of a single trajectory to train the neural network model and aim to perform trajectory prediction. No identifiability is guaranteed.

Connection between Causality and Differential Equations Aalen et al. [93] suggested, differential equations allow for a natural interpretation of causality in dynamic systems. The authors in [18–20] built an explicit bridge between the differential equations and the causal models by establishing the relationship between ODEs/Random Differential Equations (RDEs) and structural causal models. Hansen and Sokol [21] and Wang et al. [3] proposed causal interpretations and identifiability analysis of Stochastic Differential Equations (SDEs). Bellot et al. [48] proposed a method to consistently discover the causal structure of SDE systems based on penalized neural ODEs [62]. These works aim to build a theoretical connection between causality and differential equations in various ways. Our work contributes to this body of literature by proposing a method for inferring the causal structure of ODEs from a statistical perspective. To our knowledge, our approach represents the first application of hypothesis testing in ODE systems for the inference of causal structure.

3.6 Conclusion of chapter

In this chapter, we derived a sufficient condition for identifiability of homogeneous linear ODE systems from a sequence of equally-spaced error-free observations. Specifically, the observations lie on a single trajectory. Furthermore, we studied the consistency and asymptotic normality of the NLS estimator. The inference of unknown parameters based on the established theoretical results was also investigated. In particular, we proposed a new method to infer the causal structure of the ODE system. Finally, we extended the identifiability and asymptotic properties results to cases with aggregated and time-scaled observations.

A time series is most commonly collected at equally-spaced time points in practice, which motivates us to focus on the study over equally-spaced observations from a single trajectory in this chapter. However, as the authors pointed out in [94], using irregularly-spaced observations can be advantageous in obtaining more information from the dynamical system. Therefore, extending the study to cases with irregularly-spaced observations from a single trajectory is a possible direction for future work.

Chapter 4

Identifiability Analysis of Linear ODE Systems in the Presence of Hidden Confounders

The identifiability analysis of linear Ordinary Differential Equation (ODE) systems is a necessary prerequisite for making reliable causal inferences about these systems. While identifiability has been well studied in scenarios where the system is fully observable, the conditions for identifiability remain unexplored when latent variables interact with the system. This chapter aims to address this gap by presenting a systematic analysis of identifiability in linear ODE systems incorporating hidden confounders. Specifically, we investigate two cases of such systems. In the first case, latent confounders exhibit no causal relationships, yet their evolution adheres to specific functional forms, such as polynomial functions of time t. Subsequently, we extend this analysis to encompass scenarios where hidden confounders exhibit causal dependencies, with the causal structure of latent variables described by a Directed Acyclic Graph (DAG). The second case represents a more intricate variation of the first case, prompting a more comprehensive identifiability analysis. Accordingly, we conduct detailed identifiability analyses of the second system under various observation conditions, including both continuous and discrete observations from single or multiple trajectories. To validate our theoretical results, we perform a series of simulations, which support and substantiate our findings.

This chapter is derived from the following publication: Identifiability Analysis of Linear ODE Systems with Hidden Confounders [2]

4.1 Introduction

Understanding the dynamics of systems governed by Ordinary Differential Equations (ODEs) is fundamental in various scientific disciplines, from physics [95–98], biology [99–103] to economics [59, 60, 104, 105]. These ODE systems provide a natural framework for modeling causal relationships among system variables, enabling us to make reliable interpretations and interventions [18, 19, 22]. Central to unraveling the causal mechanisms of such systems is the concept of identifiability analysis, which aims to uncover conditions under which system parameters can be uniquely determined from error-free observations. Identifiability is crucial for ensuring reliable parameter estimates, thereby guaranteeing reliable causal inferences about the system [3]. The motivation for our research on the identifiability analysis of ODE systems arises from the necessity of making reliable causal inferences about these systems.

Our research focuses on the homogeneous linear ODE system, represented as:

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t)\,, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0\,, \tag{4.1}$$

where $t \in [0, \infty)$ denotes time, $\boldsymbol{x}(t) \in \mathbb{R}^d$ represents the system's state at time t, $\dot{\boldsymbol{x}}(t)$ denotes the first derivative of $\boldsymbol{x}(t)$ w.r.t. time, and $\boldsymbol{x}_0$ represents the initial condition of the system. The solution (trajectory) of the system, denoted as $\boldsymbol{x}(t; \boldsymbol{x}_0, A)$ for $t \in [0, \infty)$, is a single d-dimensional trajectory initialized with $\boldsymbol{x}_0$.

Existing literature has extensively examined the identifiability of linear ODE systems under the assumption of complete observability, where all state variables are directly observable [1, 23–26, 28, 29]. Specifically, researchers have investigated identifiability of the ODE system (4.1) from a single whole trajectory [28, 29], and extended analysis to discrete observations sampled from the trajectory [1]. However, practical scenarios often entail systems with latent variables, rendering them not entirely observable. In this chapter, we explore the identifiability analysis of this ODE system under latent confounders, particularly examining cases where no causal relationships exist from observable variables to latent variables, a commonly assumed condition in causality analysis with hidden variables [106–111].

In this chapter, we focus on two scenarios:

1. **Independent latent confounders:** Latent variables exhibit no causal relationships among themselves, leading to the following linear ODE system:

$$\begin{bmatrix}\dot{\boldsymbol{x}}(t)\\ \dot{\boldsymbol{z}}(t)\end{bmatrix} = \begin{bmatrix}A & B\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}(t)\\ \boldsymbol{z}(t)\end{bmatrix} + \begin{bmatrix}\mathbf{0}\\ \boldsymbol{f}(t)\end{bmatrix}, \quad \begin{bmatrix}\boldsymbol{x}(0)\\ \boldsymbol{z}(0)\end{bmatrix} = \begin{bmatrix}\boldsymbol{x}_0\\ \boldsymbol{z}_0\end{bmatrix}. \tag{4.2}$$

2. **Causally related latent confounders:** Latent variables exhibit causal relationships among themselves, specifically, they follow a DAG structure, represented as:

$$\begin{bmatrix}\dot{\boldsymbol{x}}(t)\\ \dot{\boldsymbol{z}}(t)\end{bmatrix} = \begin{bmatrix}A & B\\ \mathbf{0} & G\end{bmatrix}\begin{bmatrix}\boldsymbol{x}(t)\\ \boldsymbol{z}(t)\end{bmatrix}, \quad \begin{bmatrix}\boldsymbol{x}(0)\\ \boldsymbol{z}(0)\end{bmatrix} = \begin{bmatrix}\boldsymbol{x}_0\\ \boldsymbol{z}_0\end{bmatrix}. \tag{4.3}$$

In these two ODE systems, $\boldsymbol{x}(t) \in \mathbb{R}^d$ denotes the state of observable variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$, while $\boldsymbol{z}(t) \in \mathbb{R}^p$ denotes the state of latent variables $\boldsymbol{z} = (z_1, z_2, \ldots, z_p)$. Example causal structures of these two ODE systems are illustrated in Figure 4.1. It is noteworthy that the structure may include cycles and self-loops within the observable variables. Additionally, two real-world examples are provided in Appendix B.2.

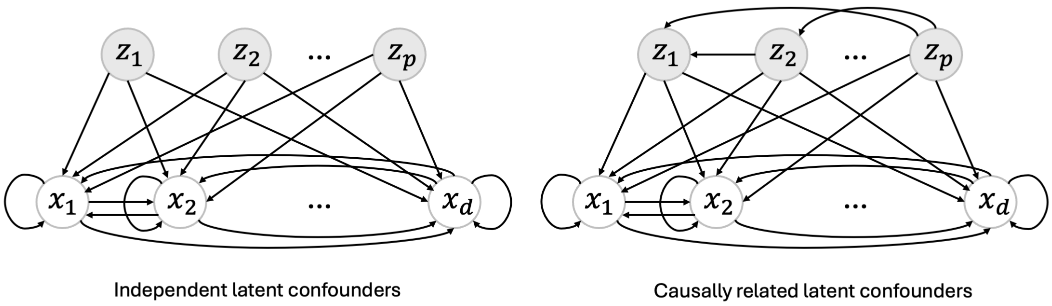

FIGURE 4.1: Example causal structures of the ODE system (4.2) and (4.3).

This chapter provides an identifiability analysis for the ODE system (4.2) under specific latent variable evolutions, such as polynomial functions of time t. Additionally, we conduct a systematic identifiability analysis of the ODE system (4.3) when the causal structure of the latent variables can be described by a DAG.

4.2 Background

4.2.1 Causal interpretation of ODE systems

When an ODE system describes the underlying causal mechanisms governing a dynamic system, it provides a natural framework for modeling causal relationships among system variables. The causal structure inherent in such systems can be directly read off [18, 22]. For instance, in the ODE system (4.1), where the ij-th entry of the parameter matrix A is denoted as A_{ij}, the presence of $A_{ij} \neq 0$ signifies that the derivative of $x_i(t)$ is influenced by $x_j(t)$, thus indicating a causal link from x_j to x_i. Here, x_i denotes the i-th variable of the ODE system (4.1), and $x_i(t)$ represents its state at time t. Since the system matrix A in the ODE system (4.1) is constant and does not explicitly depend on time t, the causal structure of this system is time-invariant.

An essential prerequisite for reliably inferring the causal structure and effects of an ODE system, for purposes of interpretation or intervention, is the identifiability analysis of such systems. To underscore this necessity, we provide an illustrative example. Consider the ODE system (4.3). Set

$$\boldsymbol{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{z}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A' = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad M = \begin{bmatrix} A & B \\ \mathbf{0} & G \end{bmatrix}, \quad M' = \begin{bmatrix} A' & B \\ \mathbf{0} & G \end{bmatrix}.$$

Calculations reveal that the solutions (trajectory) of the ODE system (4.3) with parameter matrices M or M' are identical, i.e.,

$$\begin{bmatrix} \boldsymbol{x}(t) \\ \boldsymbol{z}(t) \end{bmatrix} = e^{Mt} \begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{z}_0 \end{bmatrix} = e^{M't} \begin{bmatrix} \boldsymbol{x}_0 \\ \boldsymbol{z}_0 \end{bmatrix}.$$

This indicates that using observations sampled from this trajectory to estimate parameter matrix M may end up yielding M', which exhibits a fundamentally distinct causal relationship between x_1 and x_2, see Figure 4.2. This discrepancy in parameter estimation, wherein M' is obtained instead of the true underlying parameter matrix M, may lead to misleading interpretations and causal inferences, potentially influencing decision-making,

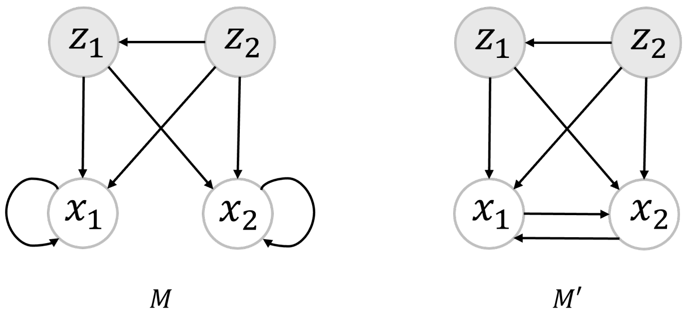

FIGURE 4.2: Causal structures of the ODE system (4.3) with parameter matrix M and M'.

particularly regarding interventions. For instance, intervention with $x_1(t) = 1$, under the true underlying parameter matrix M, yields the trajectory $x_2(t) = 4e^t - t - 3$ (post-intervention), whereas under matrix M', the trajectory becomes $x_2(t) = t^2/2 + 3t + 1$ (post-intervention). Detailed calculations are provided in Appendix B.3.

4.2.2 Identifiability analysis of the linear ODE system (4.1)

The identifiability analysis of the ODE system (4.1) has been well studied. Here, we present a fundamental definition and theorem essential for understanding identifiability in the ODE system (4.1). Denoting its solution as $\boldsymbol{x}(t; \boldsymbol{x}_0, A)$, it is noteworthy that the system is fully observable, without latent variables interacting with it. We present the identifiability definition and theorem as follows.

Definition 4.1. For $\boldsymbol{x}_0 \in \mathbb{R}^d, A \in \mathbb{R}^{d\times d}$, the ODE system (4.1) is said to be $(\boldsymbol{x}_0, A)$-identifiable, if for all $\boldsymbol{x}_0' \in \mathbb{R}^d$ and all $A' \in \mathbb{R}^{d\times d}$, with $(\boldsymbol{x}_0, A) \neq (\boldsymbol{x}_0', A')$, it holds that $\boldsymbol{x}(\cdot; \boldsymbol{x}_0, A) \neq \boldsymbol{x}(\cdot; \boldsymbol{x}_0', A')$.[1]

Lemma 4.2. *For $\boldsymbol{x}_0 \in \mathbb{R}^d, A \in \mathbb{R}^{d\times d}$, the ODE system (4.1) is $(\boldsymbol{x}_0, A)$-identifiable if and only if condition **A0** is satisfied.*

A0 *the set of vectors $\{\boldsymbol{x}_0, A\boldsymbol{x}_0, \dots, A^{d-1}\boldsymbol{x}_0\}$ is linearly independent.*

Definition 4.1 and Theorem 4.2 are adapted from [1, Definition 1] and [1, Lemma2]. We use $\boldsymbol{x}_0'$ and A' to distinguish other system parameters from the true system parameters $\boldsymbol{x}_0$ and A; $\boldsymbol{x}_0'$ and A' can represent any d-dimensional initial conditions and any $d \times d$

[1] $\boldsymbol{x}(\cdot; \boldsymbol{x}_0, A) = \{\boldsymbol{x}(t; \boldsymbol{x}_0, A) : 0 \leqslant t < \infty\}$, this inequation means that there exists at least one $t \geqslant 0$ such that $\boldsymbol{x}(t; \boldsymbol{x}_0, A) \neq \boldsymbol{x}(t; \boldsymbol{x}_0', A')$.

parameter matrices, respectively. Here instead of describing a collective property of a set of systems, we describe an intrinsic property of a single system with parameters $(\boldsymbol{x}_0, A)$. In practice, the aim is to ascertain whether the true underlying system parameter $(\boldsymbol{x}_0, A)$ is uniquely determined by observations. Hence, $(\boldsymbol{x}_0, A)$-identifiability offers a more intuitive description of the identifiability of the ODE system from a practical perspective.

From a geometric perspective, condition **A0** stated in Lemma 4.2 indicates that the initial condition $\boldsymbol{x}_0$ is not contained in an A-invariant **proper** subspace of $\mathbb{R}^d$. Intuitively, this means the trajectory of this system started from $\boldsymbol{x}_0$ spans the entire d-dimensional state space. That is, our observations cover information on all dimensions of the state space, thus rendering the identifiability of the system. Additionally, condition **A0** is generic, as noted in [3], meaning that the set of system parameters violating this condition has Lebesgue measure zero. Thus, condition **A0** is satisfied for almost all combinations of $\boldsymbol{x}_0$ and A.

4.3 Identifiability analysis of the linear ODE system (4.2)

In this section, we present the identifiability condition for the linear ODE system (4.2). We consider the function $\boldsymbol{f}(t)$ in (4.2) as a specific function of time t. Here we first define $\boldsymbol{f}(t)$ as a r-degree polynomial function of time t, expressed as follows:

$$\boldsymbol{f}(t) = \sum_{k=0}^{r} \boldsymbol{v}_k t^k\,, \quad \boldsymbol{v}_k \in \mathbb{R}^p\,. \tag{4.4}$$

Simple calculations show that

$$\boldsymbol{z}(t) = \sum_{k=0}^{r} \frac{\boldsymbol{v}_k}{k+1} t^{k+1} + \boldsymbol{z}_0\,.$$

Thus,

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) + B\boldsymbol{z}(t) = A\boldsymbol{x}(t) + \sum_{k=0}^{r} \frac{B\boldsymbol{v}_k}{k+1} t^{k+1} + B\boldsymbol{z}_0\,. \tag{4.5}$$

We denote the unknown parameters of the ODE system (4.2) as $\boldsymbol{\theta}$, specifically, $\boldsymbol{\theta} := (\boldsymbol{x}_0, \boldsymbol{z}_0, A, B, \{\boldsymbol{v}_k\}_0^r)$, where $\{\boldsymbol{v}_k\}_0^r$ denotes all the $\boldsymbol{v}_k$'s for $k = 0, \ldots, r$. Let $[\boldsymbol{x}^T(t;\boldsymbol{\theta}), \boldsymbol{z}^T(t;\boldsymbol{\theta})]^T$

denote the solution of the ODE system (4.2). It is important to note that under our hidden variables setting, only $\boldsymbol{x}(t;\boldsymbol{\theta})$ is observable. Based on Equation (4.5), we present the following identifiability definition.

Definition 4.3. For $\boldsymbol{x}_0 \in \mathbb{R}^d, \boldsymbol{z}_0 \in \mathbb{R}^p, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$ and $\{\boldsymbol{v}_k\}_0^r \in \mathbb{R}^p$, for all $\boldsymbol{x}_0' \in \mathbb{R}^d$, all $\boldsymbol{z}_0' \in \mathbb{R}^p$, all $A' \in \mathbb{R}^{d\times d}$, all $B' \in \mathbb{R}^{d\times p}$, and all $\{\boldsymbol{v}_k'\}_0^r \in \mathbb{R}^p$, we denote $\boldsymbol{\theta}' := (\boldsymbol{x}_0', \boldsymbol{z}_0', A', B', \{\boldsymbol{v}_k'\}_1^r)$, we say the ODE system (4.2) is $\boldsymbol{\theta}$-identifiable: if $(\boldsymbol{x}_0, A, B\boldsymbol{z}_0, \{B\boldsymbol{v}_k\}_0^r) \neq (\boldsymbol{x}_0', A', B'\boldsymbol{z}_0', \{B'\boldsymbol{v}_k'\}_0^r)$, it holds that $\boldsymbol{x}(\cdot;\boldsymbol{\theta}) \neq \boldsymbol{x}(\cdot;\boldsymbol{\theta}')$.

In the ODE system (4.2), where only variables $\boldsymbol{x}$ are observable, we will, with some terminological leniency, refer to $\boldsymbol{x}(\cdot;\boldsymbol{\theta})$ as the trajectory of the ODE system (4.2) with parameters $\boldsymbol{\theta}$. According to Definition 4.3, if the ODE system (4.2) with a polynomial $\boldsymbol{f}(t)$ is $\boldsymbol{\theta}$-identifiable, then the trajectory of the system can uniquely determine the values of $(\boldsymbol{x}_0, A, B\boldsymbol{z}_0, \{B\boldsymbol{v}_k\}_0^r)$. This determination is sufficient to identify the causal relationships between observable variables $\boldsymbol{x}$ as described by Equation (4.5). Consequently, one can safely intervene in the observable variables of the ODE system and make reliable causal inferences, despite the fact that matrix B cannot be identified under this definition.

Theorem 4.4. *For $\boldsymbol{x}_0 \in \mathbb{R}^d, \boldsymbol{z}_0 \in \mathbb{R}^p, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$, $\{\boldsymbol{v}_k\}_0^r \in \mathbb{R}^p$, the ODE system (4.2) is $\boldsymbol{\theta}$-identifiable if and only if assumption **A1** is satisfied.*

- ***A1*** *the set of vectors $\{\boldsymbol{\beta}, A\boldsymbol{\beta}, \ldots, A^{d-1}\boldsymbol{\beta}\}$ is linearly independent, where $\boldsymbol{\beta} = A^{r+1}(A\boldsymbol{x}_0 + B\boldsymbol{z}_0) + \sum_{j=0}^{r} j! A^{r-j} B\boldsymbol{v}_j$, and $j!$ denotes the factorial of j.*

The proof of Theorem 4.4 can be found in Appendix B.4.1. Condition **A1** is both sufficient and necessary, indicating, from a geometric perspective, that the vector $\boldsymbol{\beta}$ is not contained in an A-invariant proper subspace of $\mathbb{R}^d$ [28, Lemma 3.1].

The key point of the proof is the introduction of an augmented state $\boldsymbol{y}(t) = [\boldsymbol{x}^T(t), 1, t, t^2, \ldots, t^{r+1}]^T$ with a corresponding ODE system:

$$\begin{aligned}
\dot{\boldsymbol{y}}(t) &= \underbrace{\begin{bmatrix} A & B\boldsymbol{z}_0 & B\boldsymbol{v}_0 & \ldots & B\boldsymbol{v}_{r-1}/r & B\boldsymbol{v}_r/(r+1) \\ \mathbf{0}_d & 0 & 0 & \ldots & 0 & 0 \\ \mathbf{0}_d & 1 & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_d & 0 & 0 & \ldots & r+1 & 0 \end{bmatrix}}_{\text{denoted as } F} \boldsymbol{y}(t)\,, \\
\boldsymbol{y}(0) &= [\boldsymbol{x}_0^T, 1, 0, \ldots, 0]^T := \boldsymbol{y}_0\,,
\end{aligned} \tag{4.6}$$

where $\mathbf{0}_d$ is a d-dimensional zero row vector, and matrix $F \in \mathbb{R}^{(d+r+2)\times(d+r+2)}$. The ODE system (4.6) is a homogeneous linear ODE system analogous to (4.1) but with fully observable variables $\boldsymbol{y}$. In other words, we transform our system of interest, (4.2), which includes hidden confounders, into a fully observable ODE system (4.6). This allows us to leverage existing identifiability results for homogeneous linear ODE systems, specifically Lemma 4.2, to derive the identifiability condition for the ODE system (4.2).

Based on this approach, if the state of the hidden variables $\boldsymbol{z}(t)$, as determined by the function $\boldsymbol{f}(t)$ in the ODE system (4.2), can be described by some linear combinations of observable functions of time t, then the identifiability condition of the ODE system (4.2) can be derived. For an illustration, in the Appendix B.5, we provide identifiability conditions for the ODE system (4.2) when $\boldsymbol{f}(t) = \boldsymbol{v}e^t$ and $\boldsymbol{f}(t) = \boldsymbol{v}_1 sin(t) + \boldsymbol{v}_2 cos(t)$. While we do not enumerate all functions $\boldsymbol{f}(t)$ that meet this condition, our primary objective is to demonstrate a method for deriving the identifiability condition for the ODE (4.2) when the evolution of its hidden variables conforms to certain specific functions. Researchers can apply this approach to find appropriate functions $\boldsymbol{f}(t)$ according to their specific requirements.

4.4 Identifiability analysis of the linear ODE system (4.3)

In this section, we extend the identifiability analysis to linear ODE systems with causally related latent confounders. Specifically, we assume that the causal structure of latent variables satisfies the following latent DAG assumption.

Latent DAG: *the causal structure of latent variables can be described by a DAG.*

The DAG assumption is commonly employed in causality studies [106, 107, 109, 110, 112, 113]. Under the latent DAG assumption, the matrix G can be permuted to be a strictly upper triangular matrix, i.e., an upper triangular matrix with zeros along the main diagonal [107, 114]. Without loss of generality, we set G as a strictly upper triangular matrix.

Since G is a strictly upper triangular matrix, by the Cayley–Hamilton theorem [115], G is a nilpotent matrix with an index $\leqslant p$. Consequently, $G^k = 0$ for all $k \geqslant p$.

Based on [28, 116], the solution of $\boldsymbol{z}(t)$ can be expressed as:

$$\boldsymbol{z}(t) = e^{Gt}\boldsymbol{z}_0 = \sum_{k=0}^{\infty} \frac{G^k \boldsymbol{z}_0}{k!} t^k = \sum_{k=0}^{p-1} \frac{G^k \boldsymbol{z}_0}{k!} t^k \,.$$

Thus,

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) + B\boldsymbol{z}(t) = A\boldsymbol{x}(t) + \sum_{k=0}^{p-1} \frac{BG^k \boldsymbol{z}_0}{k!} t^k \,. \tag{4.7}$$

We observe that Equation (4.7) has the same function form as Equation (4.5), but with different coefficients (system parameters) for the polynomial of time t. Therefore, the ODE system (4.3) under the latent DAG assumption can be considered a more complex version of the ODE system (4.2) when $\boldsymbol{f}(t)$ follows a polynomial function of time t. Since the ODE system (4.3) incorporates causally related latent confounders, which is a more interesting and practical case, we will provide a more comprehensive identifiability analysis of the ODE system (4.3). The derived identifiability results can be easily generated to the case of the ODE system (4.2).

4.4.1 Identifiability condition from a single whole trajectory

We denote the unknown parameters of the ODE system (4.3) as $\boldsymbol{\eta}$, that is, $\boldsymbol{\eta} := (\boldsymbol{x}_0, \boldsymbol{z}_0, A, B, G)$. We further denote the solution of the ODE system (4.3) as $[\boldsymbol{x}^T(t;\boldsymbol{\eta}), \boldsymbol{z}^T(t;\boldsymbol{\eta})]^T$; note that under our latent variables setting, only $\boldsymbol{x}(t;\boldsymbol{\eta})$ is observable. Thus, based on Equation (4.7), we present the following identifiability definition.

Definition 4.5. For $\boldsymbol{x}_0 \in \mathbb{R}^d, \boldsymbol{z}_0 \in \mathbb{R}^p, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$ and $G \in \mathbb{R}^{p\times p}$, under the latent DAG assumption, for all $\boldsymbol{x}_0' \in \mathbb{R}^d$, all $\boldsymbol{z}_0' \in \mathbb{R}^p$, all $A' \in \mathbb{R}^{d\times d}$, all $B' \in \mathbb{R}^{d\times p}$, and all $G' \in \mathbb{R}^{p\times p}$, we denote $\boldsymbol{\eta}' := (\boldsymbol{x}_0', \boldsymbol{z}_0', A', B', G')$, we say the ODE system (4.3) is $\boldsymbol{\eta}$-identifiable: if $(\boldsymbol{x}_0, A, B\boldsymbol{z}_0, BG\boldsymbol{z}_0, \ldots, BG^{p-1}\boldsymbol{z}_0) \neq (\boldsymbol{x}_0', A', B'\boldsymbol{z}_0', B'G'\boldsymbol{z}_0', \ldots, B'G'^{p-1}\boldsymbol{z}_0')$, it holds that $\boldsymbol{x}(\cdot;\boldsymbol{\eta}) \neq \boldsymbol{x}(\cdot;\boldsymbol{\eta}')$.

Similar to the case of the ODE system (4.2), we refer to $\boldsymbol{x}(\cdot;\boldsymbol{\eta})$ as the trajectory of the ODE system (4.3) with parameters $\boldsymbol{\eta}$. Definition 4.5 defines the identifiability of the ODE system (4.3) from a single whole trajectory $\boldsymbol{x}(\cdot;\boldsymbol{\eta})$. Once the ODE system (4.3) is $\boldsymbol{\eta}$-identifiable, the causal relationships among the observable variables $\boldsymbol{x}$ can be determined through Equation (4.7). We then establish the condition for the identifiability of the ODE system (4.3) based on Definition 4.5.

Theorem 4.6. *For $\boldsymbol{x}_0 \in \mathbb{R}^d, \boldsymbol{z}_0 \in \mathbb{R}^p, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$ and $G \in \mathbb{R}^{p\times p}$, under the latent DAG assumption, the ODE system (4.3) is $\boldsymbol{\eta}$-identifiable if and only if assumption **B1** is satisfied.*

***B1**: the set of vectors $\{\boldsymbol{\gamma}, A\boldsymbol{\gamma}, \ldots, A^{d-1}\boldsymbol{\gamma}\}$ is linearly independent, where $\boldsymbol{\gamma} = A^p\boldsymbol{x}_0 + \sum_{j=0}^{p-1} A^{p-1-j}BG^j\boldsymbol{z}_0$.*

The proof of Theorem 4.6 can be found in Appendix B.4.2. Condition **B1** is both sufficient and necessary, and from a geometric perspective, it indicates that the vector $\boldsymbol{\gamma}$ is not contained in an A-invariant proper subspace of $\mathbb{R}^d$ [28, Lemma 3.1].

4.4.2 Identifiability condition from discrete observations sampled from a single trajectory

In practice, we often have access only to a sequence of discrete observations sampled from a trajectory rather than knowing the whole trajectory. Therefore, we also derive the identifiability conditions under the scenario where only discrete observations from a trajectory are available. Firstly, we extend the identifiability definition of the ODE system (4.3) as follows.

Definition 4.7. For $\boldsymbol{x}_0 \in \mathbb{R}^d, \boldsymbol{z}_0 \in \mathbb{R}^p, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$ and $G \in \mathbb{R}^{p\times p}$. For any $n \geqslant 1$, let $t_j, j = 1, \ldots, n$ be any n time points and $\boldsymbol{x}_j := \boldsymbol{x}(t_j;\boldsymbol{\eta})$ be the error-free observation of the trajectory $\boldsymbol{x}(\cdot;\boldsymbol{\eta})$ at time t_j. Under the latent DAG assumption, we say the ODE system (4.3) is $\boldsymbol{\eta}$-identifiable from $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, if for all $\boldsymbol{x}_0' \in \mathbb{R}^d$, all $\boldsymbol{z}_0' \in \mathbb{R}^p$, all

$A' \in \mathbb{R}^{d\times d}$, all $B' \in \mathbb{R}^{d\times p}$, and all $G' \in \mathbb{R}^{p\times p}$ with $(\boldsymbol{x}_0, A, B\boldsymbol{z}_0, BG\boldsymbol{z}_0, \ldots, BG^{p-1}\boldsymbol{z}_0) \neq (\boldsymbol{x}'_0, A', B'\boldsymbol{z}'_0, B'G'\boldsymbol{z}'_0, \ldots, B'G'^{p-1}\boldsymbol{z}'_0)$, it holds that $\exists j \in \{1, \ldots, n\}$ such that $\boldsymbol{x}(t_j;\boldsymbol{\eta}) \neq \boldsymbol{x}(t_j;\boldsymbol{\eta}')$.

Definition 4.7 defines the identifiability of the ODE system (4.3) from n observations sampled from the trajectory $\boldsymbol{x}(\cdot;\boldsymbol{\eta})$. Then we establish the condition for the identifiability of the ODE system (4.3) from discrete observations based on Definition 4.7.

Theorem 4.8. *For $\boldsymbol{x}_0 \in \mathbb{R}^d, \boldsymbol{z}_0 \in \mathbb{R}^p, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$ and $G \in \mathbb{R}^{p\times p}$. We define new observation $\boldsymbol{y}_j := [\boldsymbol{x}_j^T, 1, t_j, t_j^2, \ldots, t_j^{p-1}]^T \in \mathbb{R}^{d+p}$, for $j = 1, \ldots, n$. Under the latent DAG assumption, the ODE system (4.3) is $\boldsymbol{\eta}$-identifiable from discrete observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, if and only if assumption **C1** is satisfied.*

***C1**: there exists $(d+p)$ $\boldsymbol{y}_j$'s with indices denoting as $\{j_1, j_2, \ldots, j_{d+p}\} \subseteq \{1, 2, \ldots, n\}$, such that the set of vectors $\{\boldsymbol{y}_{j_1}, \boldsymbol{y}_{j_2}, \ldots, \boldsymbol{y}_{j_{d+p}}\}$ is linearly independent.*

The proof of Theorem 4.8 can be found in Appendix B.4.3. Condition **C1** is both sufficient and necessary. This theorem states that as long as there are $d+p$ observations $\boldsymbol{x}_j$'s such that the corresponding augmented new observations $\boldsymbol{y}_j$'s are linearly independent, the ODE system (4.3) is $\boldsymbol{\eta}$-identifiable from these discrete observations. Under the latent DAG assumption, we can transfer the ODE system (4.3), which includes hidden confounders, into a $(d+p)$-dimensional fully observable ODE system (4.1) through the augmented state $\boldsymbol{y}(t)$. Condition **C1** indicates that our observations span the entire $(d+p)$-dimensional state space, thus rendering the system identifiable.

Both Definition 4.5 and Definition 4.7 define the identifiability of the ODE system (4.3) to some extent of the unknown parameters. In other words, given the available observations, under Definition 4.5 and Definition 4.7, one can only identify the values of $(\boldsymbol{x}_0, A, B\boldsymbol{z}_0, BG\boldsymbol{z}_0, \ldots, BG^{p-1}\boldsymbol{z}_0)$, but not the values of $(\boldsymbol{z}_0, B, G)$. Based on Equation (4.7), this level of identifiability is sufficient to identify the causal relationships between observable variables $\boldsymbol{x}$, enabling safe intervention on the observable variables with reliable causal inferences. However, in scenarios where practitioners can intervene in the latent variables and require inferring the causal effects of the intervened system, identifying the matrices B and G becomes essential for reliable causal references. For instance, in chemical kinetics, where the evolution of chemical concentrations over time can often be modeled by an ODE system [117, 118], some chemicals may not be measurable during the reaction, rendering them latent variables. Nonetheless, practitioners can intervene in these latent variables by setting specific initial concentrations. Therefore, we provide an identifiability analysis of the linear ODE system (4.3) when practitioners can control the initial condition of the latent variables: $\boldsymbol{z}_0$.

4.4.3 Identifiability condition from p controllable whole trajectories

Assuming the initial condition of the latent variables $\boldsymbol{z}_0$ is controllable, which means that the values of $\boldsymbol{z}_0$ can be treated as given values, we denote it as $\boldsymbol{z}_0^*$. In the following, we provide the identifiability condition of the ODE system (4.3) when we are given p initial conditions $\boldsymbol{z}_0^*$, denoting as $\boldsymbol{z}_0^{*i}$. We first present the definition.

Definition 4.9. Given $\boldsymbol{z}_0^{*i} \in \mathbb{R}^p$ for $i = 1, \ldots, p$, for $\boldsymbol{x}_0 \in \mathbb{R}^d, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$ and $G \in \mathbb{R}^{p\times p}$, under the latent DAG assumption, for all $\boldsymbol{x}_0' \in \mathbb{R}^d$, all $A' \in \mathbb{R}^{d\times d}$, all $B' \in \mathbb{R}^{d\times p}$, and all $G' \in \mathbb{R}^{p\times p}$, we denote $\boldsymbol{\eta}_i := (\boldsymbol{x}_0, \boldsymbol{z}_0^{*i}, A, B, G)$ and $\boldsymbol{\eta}_i' := (\boldsymbol{x}_0', \boldsymbol{z}_0^{*i}, A', B', G')$, we say the ODE system (4.3) is $\{\boldsymbol{\eta}_i\}_1^p$-identifiable: if $(\boldsymbol{x}_0, A, B, G) \neq (\boldsymbol{x}', A', B', G')$, it holds that $\exists i$ such that $\boldsymbol{x}(\cdot; \boldsymbol{\eta}_i) \neq \boldsymbol{x}(\cdot; \boldsymbol{\eta}_i')$.

Definition 4.9 defines the identifiability of the ODE system (4.3) from p whole trajectories $\boldsymbol{x}(\cdot; \boldsymbol{\eta}_i)$ with $i = 1, \ldots, p$, and under this definition, matrix B and G are also identifiable. Based on this definition, we provide the identifiability condition.

Theorem 4.10. *Given $\boldsymbol{z}_0^{*i} \in \mathbb{R}^p$ for $i = 1, \ldots, p$, for $\boldsymbol{x}_0 \in \mathbb{R}^d, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$ and $G \in \mathbb{R}^{p\times p}$, under the latent DAG assumption, the ODE system (4.3) is $\{\boldsymbol{\eta}_i\}_1^p$-identifiable if assumptions $\boldsymbol{B}_2$, $\boldsymbol{B}_3$ and $\boldsymbol{B}_4$ are all satisfied.*

- ***B2**: each $\boldsymbol{z}_0^{*i}$ for $i = 1, \ldots, p$, satisfies assumption **B1**. That is, if we set $\boldsymbol{\gamma}_i = A^p \boldsymbol{x}_0 + \sum_{j=0}^{p-1} A^{p-1-j} B G^j \boldsymbol{z}_0^{*i}$, then the set of vectors $\{\boldsymbol{\gamma}_i, A\boldsymbol{\gamma}_i, \ldots, A^{d-1}\boldsymbol{\gamma}_i\}$ is linearly independent for all $i = 1, \ldots, p$.*
- ***B3**: the set of vectors $\{\boldsymbol{z}_0^{*1}, \boldsymbol{z}_0^{*2}, \ldots, \boldsymbol{z}_0^{*p}\}$ is linearly independent.*
- ***B4**: the matrix composed by vertically stack the matrices $\{B, BG, \ldots, BG^{p-1}\}$ has rank p.*

The proof of Theorem 4.10 can be found in Appendix B.4.4. Assumption **B2** ensures that the ODE system (4.3) is $\boldsymbol{\eta}_i$-identifiable for all $i = 1, \ldots, p$. Consequently,

$(\boldsymbol{x}_0, A, B\boldsymbol{z}_0^{*i}, BG\boldsymbol{z}_0^{*i}, \ldots, BG^{p-1}\boldsymbol{z}_0^{*i})$ for all $i = 1, \ldots, p$ is identifiable. Then, under assumption **B3**, the identifiability of matrix B is established. To identify matrix G, assumption **B4** is required. While the ability to control the initial condition of the latent variables may appear strict, it is a reasonable assumption in our context. This is because identifying matrices B and G is necessary only when practitioners can intervene in the latent variables, thereby allowing control over their initial conditions. An alternative approach to identifying B and G involves intervening in the initial condition of each latent variable z_i independently, rather than controlling the initial condition of all latent variables $\boldsymbol{z}$ simultaneously. This method draws inspiration from the "genetic single-node intervention" proposed by [119], where one can intervene at each latent node individually. Further details of this method can be found in Appendix B.6.

4.4.4 Identifiability condition from discrete observations sampled from p controllable trajectories

We also extend the identifiability analysis of the ODE system (4.3) to cases where only discrete observations from p controllable trajectories are available.

Definition 4.11. Given $\boldsymbol{z}_0^{*i} \in \mathbb{R}^p$ for $i = 1, \ldots, p$, for $\boldsymbol{x}_0 \in \mathbb{R}^d, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$ and $G \in \mathbb{R}^{p\times p}$. For any $n \geqslant 1$, let $t_j, j = 1, \ldots, n$ be any n time points and $\boldsymbol{x}_{ij} := \boldsymbol{x}(t_j; \boldsymbol{\eta}_i)$ be the error-free observation of the trajectory $\boldsymbol{x}(\cdot; \boldsymbol{\eta}_i)$ at time t_j. Under the latent DAG assumption, we say the ODE system (4.3) is $\{\boldsymbol{\eta}_i\}_1^p$-identifiable from $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in}$, $i = 1, \ldots, p$, if for all $\boldsymbol{x}_0' \in \mathbb{R}^d$, all $A' \in \mathbb{R}^{d\times d}$, all $B' \in \mathbb{R}^{d\times p}$, and all $G' \in \mathbb{R}^{p\times p}$ with $(\boldsymbol{x}_0, A, B, G) \neq (\boldsymbol{x}_0', A', B', G')$, it holds that $\exists i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, n\}$ such that $\boldsymbol{x}(t_j; \boldsymbol{\eta}_i) \neq \boldsymbol{x}(t_j; \boldsymbol{\eta}_i')$.

Based on Definition 4.11 we present the identifiability condition.

Theorem 4.12. *Given $\boldsymbol{z}_0^{*i} \in \mathbb{R}^p$ for $i = 1, \ldots, p$, for $\boldsymbol{x}_0 \in \mathbb{R}^d, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$ and $G \in \mathbb{R}^{p\times p}$. We define new observation $\boldsymbol{y}_{ij} := [\boldsymbol{x}_{ij}^T, 1, t_j, t_j^2, \ldots, t_j^{p-1}]^T \in \mathbb{R}^{d+p}$, for $i = 1, \ldots, p$ and $j = 1, \ldots, n$. Under the latent DAG assumption, the ODE system (4.3) is $\{\boldsymbol{\eta}_i\}_1^p$-identifiable from discrete observations $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in}$, $i = 1, \ldots, p$, if assumptions **C2**, **B3** and **B4** are all satisfied.*

C2: *for each* $i \in \{1, \ldots, p\}$ *there exists* $(d{+}p)$ $\boldsymbol{y}_{ij}$*'s with indexes denoting as* $\{j_{i1}, j_{i2}, \ldots, j_{i,d+p}\} \subseteq \{1, 2, \ldots, n\}$*, such that the set of vectors* $\{\boldsymbol{y}_{ij_{i1}}, \boldsymbol{y}_{ij_{i2}}, \ldots, \boldsymbol{y}_{ij_{i,d+p}}\}$ *is linearly independent.*

The proof of Theorem 4.12 can be found in Appendix B.4.5. Assumption **C2** ensures that the ODE system (4.3) is $\boldsymbol{\eta}_i$-identifiable from discrete observations $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in}$ for all $i = 1, \ldots, p$. As in Subsection 4.4.3, under assumptions **B3** and **B4**, the matrices B and G are also identifiable.

4.5 Simulations

To evaluate the validity of the identifiability conditions established in Section 3 and 4, we present the results of simulations. As previously indicated, the ODE system (4.3) can be treated as a more intricate version of the ODE system (4.2); hence, our simulation experiments are centered on the former.

Simulation design. We conduct four sets of simulations, which include one identifiable case and one unidentifiable case for both the $\boldsymbol{\eta}$-identifiable check and the $\{\boldsymbol{\eta}_i\}_1^p$-identifiable check. The dimensions of both observable variables, d, and latent variables, p, are set to 3. The true underlying parameters of the systems are provided below. Observations are simulated from the true ODE systems for each case, with n equally-spaced observations generated from the time interval $[0, 1]$ for each trajectory, and we only keep the values of the observable variables $\boldsymbol{x}$.

$$A = \begin{bmatrix} 2 & -2 & 1 \\ 1 & 1 & -1 \\ 1 & 0 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} -2 & -2 & 2 \\ 0 & -1 & -2 \\ -1 & -1 & -2 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & 2 & 1 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \end{bmatrix}, \quad A' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$
$$\boldsymbol{x}_0 = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{z}_0 = \begin{bmatrix} 1 \\ -2 \\ -1 \end{bmatrix}, \quad \boldsymbol{z}_0^{*1} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{z}_0^{*2} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{z}_0^{*3} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

$\boldsymbol{\eta}$-identifiable: $\boldsymbol{\eta} = (\boldsymbol{x}_0, \boldsymbol{z}_0, A, B, G)$, unidentifiable: $\boldsymbol{\eta} = (\boldsymbol{x}_0, \boldsymbol{z}_0, A', B, G)$.

$\{\boldsymbol{\eta}_i\}_1^p$-identifiable: $\boldsymbol{\eta}_i = (\boldsymbol{x}_0, \boldsymbol{z}_0^{*i}, A, B, G)$, unidentifiable: $\boldsymbol{\eta}_i = (\boldsymbol{x}_0, \boldsymbol{z}_0^{*i}, A', B, G), i = 1, 2, 3$.

Parameter estimation. The Nonlinear Least Squares (NLS) method is employed for parameter estimation, a widely used technique for estimating parameters in nonlinear regression models, including ODEs [30, 32, 81]. The *"least_squares"* function from the *"scipy.optimize"* Python module, with default hyperparameter settings, is utilized for implementation. Given that the NLS loss function for our simulation is non-convex, parameter initialization is performed near the true values to promote convergence to the global minimum. Specifically, for the $\boldsymbol{\eta}$-(un)identifiable cases, initial parameter values are set to the true parameters plus a random value drawn from a uniform distribution $U(-0.1, 0.1)$ for each replication. For $\{\boldsymbol{\eta}_i\}_1^p$-(un)identifiable cases, initial parameter values are set to the true values plus a random value from $U(-0.3, 0.3)$.

Evaluation metric. Mean Squared Error (MSE) is adopted as the metric to assess the accuracy of the parameter estimator. To ensure the reliability of the estimation results, 100 independent random replications are run for each configuration, and we report the mean and variance of the squared error.

Results analysis. Table 4.1 and Table 4.2 present the simulation results for the $\boldsymbol{\eta}$-(un)identifiable cases and the $\{\boldsymbol{\eta}_i\}_1^p$-(un)identifiable cases, respectively. According to Definition 4.5 and Definition 4.9, for the $\boldsymbol{\eta}$-(un)identifiable cases, the identifiability of $(\boldsymbol{x}_0, A, B\boldsymbol{z}_0, BG\boldsymbol{z}_0, BG^2\boldsymbol{z}_0)$ needs to be checked, while for the $\{\boldsymbol{\eta}_i\}_1^p$-(un)identifiable cases, we need to check the identifiability of $(\boldsymbol{x}_0, A, B, G)$. Since $\boldsymbol{x}_0$ is consistently identifiable (with MSE less than 1.00E-10) across all (un)identifiable cases, its results are not presented.

In both Tables, for identifiable cases, as the number of samples n increases, the MSEs for all parameters of interest decrease and approach zero. However, in the unidenfiable cases, where the identifiability condition **B1/B2** stated in Theorem 4.6/4.10 is unmet, the MSEs for certain parameters remain high irrespective of sample size. These results offer strong empirical support for the validity of the identifiability conditions outlined in Theorem 4.6 and Theorem 4.10. It is noteworthy that in the $\{\boldsymbol{\eta}_i\}_1^p$ case, where observations are sampled from $p = 3$ controllable trajectories, remarkably accurate parameter estimates can be obtained even with a limited number of samples.

For the $\boldsymbol{\eta}$-(un)identifiable cases, assumption **C1** stated in Theorem 4.8 holds true for all values of n in the identifiable cases, while it is violated across all n in the unidentifiable cases. In the $\{\boldsymbol{\eta}_i\}_1^p$-(un)identifiable cases, condition **C2** stated in Theorem 4.12 is satisfied

Table 4.1: MSEs of the $\boldsymbol{\eta}$-(un)identifiable cases of the ODE (4.3)

$\boldsymbol{n}$	**Identifiable**				**Unidentifiable**			
	A	$B\boldsymbol{z}_0$	$BG\boldsymbol{z}_0$	$BG^2\boldsymbol{z}_0$	A	$B\boldsymbol{z}_0$	$BG\boldsymbol{z}_0$	$BG^2\boldsymbol{z}_0$
10	6.00E-05 (±5.40E-08)	0.0004 (±3.45E-06)	0.0044 (±0.0004)	0.0007 (±3.91E-06)	0.0994 (±0.0157)	0.0494 (±0.1243)	0.9185 (±8.3148)	0.6482 (±1.4306)
100	4.15E-05 (±1.62E-08)	0.0003 (±8.52E-07)	0.0029 (±9.42E-05)	0.0005 (±2.90E-06)	0.0372 (±0.0032)	0.0174 (±0.0087)	0.3517 (±0.3460)	0.5767 (±1.4055)
500	2.65E-05 (±8.71E-09)	0.0002 (±4.38E-07)	0.0019 (±4.84E-05)	0.0002 (±8.38E-07)	0.0461 (±0.0099)	0.1071 (±0.1768)	0.5783 (±2.5747)	0.3648 (±0.4507)

Table 4.2: MSEs of the $\{\boldsymbol{\eta}_i\}_1^p$-(un)identifiable cases of the ODE (4.3)

$\boldsymbol{n}$	**Identifiable**			**Unidentifiable**		
	A	B	G	A	B	G
10	5.83E-22 (±7.41E-42)	2.85E-21 (±2.75E-40)	2.27E-21 (±5.69E-41)	0.6349 (±0.7464)	0.1913 (±0.0686)	0.0044 (±0.0011)
30	1.50E-22 (±3.23E-43)	7.80E-22 (±1.14E-41)	5.76E-22 (±5.28E-42)	0.6169 (±0.7194)	0.1850 (±0.0657)	0.0045 (±0.0007)
50	5.16E-23 (±6.20E-44)	3.01E-22 (±3.27E-42)	2.39E-22 (±8.46E-43)	0.5876 (±0.6895)	0.1761 (±0.0627)	0.0045 (±0.0008)

for all values of n in the identifiable cases, but is found to be violated for all values of n in the unidentifiable cases. These findings provide strong empirical evidence supporting the validity of the identifiability conditions proposed in Theorem 4.8 and Theorem 4.12.

In Appendix B.7, we present additional simulation results for higher-dimensional cases, along with simulations that incorporate a variety of ground-truth parameter configurations. These results consistently affirm the validity of our proposed identifiability conditions. For further details, please refer to Appendix B.7.

4.6 Related work

Identifiability analysis of linear ODE systems. Within control theory, extensive research has been conducted on the identifiability analysis of linear dynamical systems governed by ODEs [23–26]. In the applied mathematics area, Stanhope et al. [28] and Qiu et al. [29] have systematically investigated the identifiability of linear ODE systems based on a single trajectory. Furthermore, Wang et al. [1] have extended these findings to scenarios where only discrete observations sampled from a single trajectory are available. However, existing studies primarily concentrate on linear ODE systems with fully observable variables. To the best of our knowledge, our work represents the inaugural endeavor to systematically analyze the identifiability of linear ODE systems in the presence of hidden confounders.

Connection between causality and differential equations. Differential equations provide a natural framework for understanding causality within dynamic systems, particularly in the context of continuous-time processes [22, 93]. Consequently, significant efforts have been directed towards establishing a theoretical link between causality and differential equations. In the deterministic case, Mooij et al. [18] and Rubenstein et al. [19] have established a mathematical connection between ODEs and Structural Causal Models (SCMs). Wang et al. [1] have proposed a method for inferring the causal structure of linear ODEs. In the domain of neural ODEs, Aliee et al. [46, 47] have applied various regularization techniques to enhance the recovery of the causal relationships. Turning to the stochastic case, Hansen et al. [21] and Wang et al. [3] have proposed causal interpretations and identifiability analysis of Stochastic Differential Equations (SDEs). Additionally, Bellot et al. [48] have introduced a method for consistently discovering the causal structure of SDE systems using penalized neural ODEs. These works aim to establish a theoretical connection between causality and differential equations in various ways. Our contribution to this scholarly landscape lies in the systematic analysis of the identifiability of linear ODEs, particularly in the presence of hidden confounders.

4.7 Conclusion of chapter

This chapter presents a systematic identifiability analysis of linear ODE systems incorporating hidden confounders. Specifically, we establish a sufficient and necessary identifiability condition for the linear ODE system with independent latent confounders. Additionally, we provide four identifiability conditions for the linear ODE system with causally related latent confounders, wherein the causal structure of the latent confounders adheres to a DAG.

A notable limitation of our work lies in the practical verification of these identifiability conditions, given that the true underlying system parameters are often unavailable in real-world scenarios. However, our study significantly contributes to the understanding of the intrinsic structure of linear ODE systems with hidden confounders. By providing insights into the identifiability aspects, our findings empower practitioners to utilize models that adhere to the proposed conditions (e.g., through constrained parameter estimation) for learning from real-world data while ensuring identifiability.

Chapter 5

Generator Identification for Linear SDE Systems with Additive and Multiplicative Noise

In this chapter, we present conditions for identifying the ***generator*** *of a linear stochastic differential equation (SDE) from the distribution of its solution process with a given fixed initial state. These identifiability conditions are crucial in causal inference using linear SDEs as they enable the identification of the post-intervention distributions from its observational distribution. Specifically, we derive a sufficient and necessary condition for identifying the generator of linear SDEs with additive noise, as well as a sufficient condition for identifying the generator of linear SDEs with multiplicative noise. We show that the conditions derived for both types of SDEs are generic. Moreover, we offer geometric interpretations of the derived identifiability conditions to enhance their understanding. To validate our theoretical results, we perform a series of simulations, which support and substantiate the established findings.*

5.1 Introduction

Stochastic differential equations (SDEs) are a powerful mathematical tool for modelling dynamic systems subject to random fluctuations. These equations are widely used in

This chapter is derived from the following publication: Generator Identification for Linear SDEs with Additive and Multiplicative Noise [3] various scientific disciplines, including finance [120–122], physics [123–125], biology [126–128] and engineering [124, 129, 130]. In recent years, SDEs have garnered growing interest in the machine learning research community. Specifically, they have been used for tasks such as modelling time series data [131–133] and estimating causal effects [48, 134, 135].

To enhance understanding we first introduce the SDEs of our interest, which are multidimensional linear SDEs with additive and multiplicative noise, respectively. Consider an m-dimensional standard Brownian motion defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\})$, denoted by $W := \{W_t = [W_{1,t}, \ldots, W_{m,t}]^\top : 0 \leqslant t < \infty\}$. Let $X_t \in \mathbb{R}^d$ be the state at time t and let $\boldsymbol{x}_0 \in \mathbb{R}^d$ be a constant vector denoting the initial state of the system, we present the forms of the aforementioned two linear SDEs.

1. Linear SDEs with additive noise.

$$dX_t = AX_t dt + GdW_t\,, \quad X_0 = \boldsymbol{x}_0\,, \tag{5.1}$$

   where $0 \leqslant t < \infty$, $A \in \mathbb{R}^{d\times d}$ and $G \in \mathbb{R}^{d\times m}$ are some constant matrices.

2. Linear SDEs with multiplicative noise.

$$dX_t = AX_t dt + \textstyle\sum_{k=1}^{m} G_k X_t dW_{k,t}\,, \quad X_0 = \boldsymbol{x}_0\,, \tag{5.2}$$

   where $0 \leqslant t < \infty$, $A, G_k \in \mathbb{R}^{d\times d}$ for $k = 1, \ldots, m$ are some constant matrices.

Linear SDEs are wildly used in financial modeling for tasks like asset pricing, risk assessment, and portfolio optimization [136–138], where they are used to model the evolution of financial variables, such as stock prices and interest rates. Furthermore, linear SDEs are also used in genomic research, for instance, they are used for modeling the gene expression in the yeast microorganism Saccharomyces Cerevisiae [21]. The identifiability analysis of linear SDEs is essential for reliable causal inference of dynamic systems governed by these equations. For example, in the case of Saccharomyces Cerevisiae, one aims to identify the system such that making reliable causal inference when interventions are introduced. Such interventions may involve deliberate knockout of specific genes to achieve optimal growth of an organism. In this regard, identifiability analysis plays a pivotal role in ensuring reliable predictions concerning the impact of interventions on the system.

Previous studies on identifiability analysis of linear SDEs have primarily focused on Gaussian diffusions, as described by the SDE (5.1) [41–45, 139]. These studies are typically based on observations located on one trajectory of the system and thus require restrictive identifiability conditions, such as the ergodicity of the diffusion or other restrictive requirements on the eigenvalues of matrix A. However, in practical applications, multiple trajectories of the dynamic system can often be accessed [38, 39, 140, 141]. In particular, these multiple trajectories may start from the same initial state, e.g., in experimental studies where repeated trials or experiments are conducted under the same conditions [142–145] or when the same experiment is performed on multiple identical units [146]. To this end, this work presents an identifiability analysis for linear SDEs based on the distribution of the observational process with a given fixed initial state. Furthermore, our study is not restricted to Gaussian diffusions (5.1), but also encompasses linear SDEs with multiplicative noise (5.2). Importantly, the conditions derived for both types of SDEs are generic, meaning that the set of system parameters that violate the proposed conditions has Lebesgue measure zero.

Traditional identifiability analysis of dynamic systems focuses on deriving conditions under which a unique set of parameters can be obtained from error-free observational data. However, our analysis of dynamic systems that are described by SDEs aims to uncover conditions that would enable a unique generator to be obtained from its observational distribution. Our motivation for identifying generators of SDEs is twofold. Firstly, obtaining a unique set of parameters from the distribution of a stochastic process described by an SDE is generally unfeasible. For example, in the SDE (5.1), parameter G cannot be uniquely identified since one can only identify $GG^\top$ based on the distribution of its solution process [21, 41]. Secondly, the identifiability of an SDE's generator suffices for reliable causal inferences for this system. Note that, in the context of SDEs, the main task of causal analysis is to identify the post-intervention distributions from the observational distribution. As proposed in [21], for an SDE satisfying specific criteria, the post-intervention distributions are identifiable from the generator of this SDE. Consequently, the intricate task of unraveling causality can be decomposed into two constituent components through the generator. This chapter aims to uncover conditions under which the generator of a linear SDE attains identifiability from the observational distribution. By establishing these identifiability conditions, we can effectively address the causality task for linear SDEs.

In this chapter, we present a sufficient and necessary identifiability condition for the generator of linear SDEs with additive noise (5.1), along with a sufficient identifiability condition for the generator of linear SDEs with multiplicative noise (5.2).

5.2 Background knowledge

In this section, we introduce some background knowledge of linear SDEs. In addition, we provide a concise overview of the causal interpretation of SDEs, which is a critical aspect of understanding the nature and dynamics of these equations. This interpretation also forms a strong basis for the motivation of this research.

5.2.1 Background knowledge of linear SDE systems

The solution to the SDE (5.1) can be explicitly expressed as (cf. [49]):

$$X_t := X(t; \boldsymbol{x}_0, A, G) = e^{At}\boldsymbol{x}_0 + \int_0^t e^{A(t-s)} G dW_s \,. \tag{5.3}$$

Note that in the context of our study, the solution stands for the strong solution, refer to [50] for its detailed definition.

In general, obtaining an explicit expression for the solution to the SDE (5.2) is not feasible. In fact, an explicit solution can be obtained when the matrices $A, G_1, \ldots, G_k$ commute, that is when

$$AG_k = G_k A \quad \text{and} \quad G_k G_l = G_l G_k \tag{5.4}$$

holds for all $k, l = 1, \ldots, m$ (cf. [51]). However, the conditions described in (5.4) are too restrictive and impractical. Therefore, this study will focus on the general case of the SDE (5.2).

We know that both the SDE (5.1) and the SDE (5.2) admit unique solutions that manifest as continuous stochastic processes [50]. A d-dimensional stochastic process is a collection of $\mathbb{R}^d$-valued random variables, denoted as $X = \{X_t; 0 \leqslant t < \infty\}$ defined on some probability space. When comparing two stochastic processes, X and $\tilde{X}$, that are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$, various notions of equality may be considered. In this study, we adopt the notion of equality with respect to their distributions, which is a weaker requirement than strict equivalence, see [50] for relevant notions. We now present the definition of the distribution of a stochastic process.

Definition 5.1. Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a measurable space $(S, \mathcal{B}(S))$, i.e., the function $X : \Omega \to S$ is $\mathcal{F}/\mathcal{B}(S)$-measurable. Then, the distribution of the random variable X is the probability measure P^X on $(S, \mathcal{B}(S))$ given by

$$P^X(B) = \mathbb{P}(X \in B) = \mathbb{P}\{\omega \in \Omega : X(\omega) \in B\}\,, \quad B \in \mathcal{B}(S)\,.$$

When $X := \{X_t; 0 \leqslant t < \infty\}$ is a continuous stochastic process on $(\Omega, \mathcal{F}, \mathbb{P})$, and $S = C[0, \infty)$, such an X can be regarded as a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $(C[0, \infty), \mathcal{B}(C[0, \infty)))$, and P^X is called the distribution of X. Here $C[0, \infty)$ stands for the space of all continuous, real-valued functions on $[0, \infty]$.

It is noteworthy that the distribution of a continuous process can be uniquely determined by its finite-dimensional distributions. Hence, if two stochastic processes, labelled as X and $\tilde{X}$, share identical finite-dimensional distributions, they are regarded as equivalent in distribution, denoted by $X \overset{\mathrm{d}}{=} \tilde{X}$. Relevant concepts and theories regarding this property can be found in [50].

The generator of a stochastic process is typically represented by a differential operator that acts on functions. It provides information about how a function evolves over time in the context of the underlying stochastic process. Mathematically, the generator of a stochastic process X_t can be defined as

$$(\mathcal{L}f)(\boldsymbol{x}) = \lim_{s \to 0} \frac{\mathbb{E}[f(X_{t+s}) - f(X_t) | X_t = \boldsymbol{x}]}{s}\,,$$

where f is a suitably regular function.

In the following, we present the generator of the SDEs under consideration. Obviously, both the SDE (5.1) and the SDE (5.2) conform to the general form:

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t\,, \quad X_0 = \boldsymbol{x}_0\,. \tag{5.5}$$

where b and σ are locally Lipschitz continuous in the space variable $\boldsymbol{x}$. The generator $\mathcal{L}$ of the SDE (5.5) can be explicitly computed by utilizing Itô's formula (cf. [49]).

Proposition 5.2.1. Let X be a stochastic process defined by the SDE (5.5). The generator $\mathcal{L}$ of X on $C_b^2(\mathbb{R}^d)$ is given by

$$(\mathcal{L}f)(\boldsymbol{x}) := \sum_{i=1}^{d} b_i(\boldsymbol{x}) \frac{\partial f(\boldsymbol{x})}{\partial x_i} + \frac{1}{2} \sum_{i,j=1}^{d} c_{ij}(\boldsymbol{x}) \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j} \tag{5.6}$$

for $f \in C_b^2(\mathbb{R}^d)$ and $\boldsymbol{x} \in \mathbb{R}^d$, where $c(\boldsymbol{x}) = \sigma(\boldsymbol{x}) \cdot \sigma(\boldsymbol{x})^\top$ is a $d \times d$ matrix, and $C_b^2(\mathbb{R}^d)$ denotes the space of continuous functions on $\mathbb{R}^d$ that have bounded derivatives up to order two.

5.2.2 Causal interpretation of SDE systems

An important motivation for the identification of the generator of an SDE lies in the desire to infer reliable causality within dynamic models described by SDEs. In this subsection, we aim to provide some necessary background knowledge on the causal interpretation of SDEs. Consider the general SDE framework described as:

$$dX_t = a(X_t)dZ_t\,, \quad X_0 = \boldsymbol{x}_0\,, \tag{5.7}$$

where Z is a p-dimensional semimartingale and $a : \mathbb{R}^d \to \mathbb{R}^{d\times p}$ is a continuous mapping. By writing the SDE (5.7) in integral form

$$X_t^i = \boldsymbol{x}_0^i + \textstyle\sum_{j=1}^p \int_0^t a_{ij}(X_s)dZ_s^j\,, \quad i \leqslant d\,. \tag{5.8}$$

The authors of [21] proposed a mathematical definition of the SDE resulting from an intervention to the SDE (5.8). In the following, $X^{(-l)}$ denotes the $(d-1)$-dimensional vector that results from the removal of the l-th coordinate of $X \in \mathbb{R}^d$.

Definition 5.2. [21, Definition 2.4.] Consider some $l \leqslant d$ and $\zeta : \mathbb{R}^{d-1} \to \mathbb{R}$. The SDE arising from (5.8) under the intervention $X_t^l := \zeta(X_t^{(-l)})$ is the $(d-1)$-dimensional equation

$$(Y^{(-l)})_t^i = \boldsymbol{x}_0^i + \textstyle\sum_{j=1}^p \int_0^t b_{ij}(Y_s^{(-l)})dZ_s^j\,, \quad i \neq l\,, \tag{5.9}$$

where $b : \mathbb{R}^{d-1} \to \mathbb{R}^{(d-1)\times p}$ is defined by $b_{ij}(\boldsymbol{y}) = a_{ij}(y_1, \ldots, \zeta(\boldsymbol{y}), \ldots, y_d)$ for $i \neq l$ and $j \leqslant p$ and the $\zeta(\boldsymbol{y})$ is on the l-th coordinate.

Definition 5.2 presents a natural approach to defining how interventions should affect dynamic systems governed by SDEs. We adopt the same notations as used in [21]. Assuming (5.8) and (5.9) have unique solutions for all interventions, we refer to (5.8) as the observational SDE, to its solution as the observational process, to the distribution of its solution as observational distribution, to (5.9) as the post-intervention SDE, to the solution of (5.9) as the post-intervention process, and to the distribution of the solution of (5.9) as the post-intervention distribution. The authors in [21] related Definition 5.2 to mainstream causal concepts by establishing a mathematical connection between SDEs and structural equation models (SEMs). Specifically, the authors showed that under regularity assumptions, the solution to the post-intervention SDE is equal to the limit of a sequence of interventions in SEMs based on the Euler scheme of the observational SDE. Despite the fact that the parameters of the SDEs are generally not identifiable from the observational distribution, the post-intervention distributions can be identified, thus enabling causal inference of the system. To this end, Sokol and Hansen [21] derived a condition under which the generator associated with the observational SDE allows for the identification of the post-intervention distributions. We present the corresponding theory as follows.

Lemma 5.3. *[21, Theorem 5.3.] Consider the SDEs*

$$dX_t = a(X_t)dZ_t\,, \quad X_0 = \boldsymbol{x}_0\,, \tag{5.10}$$

$$d\tilde{X}_t = \tilde{a}(\tilde{X}_t)d\tilde{Z}_t\,, \quad \tilde{X}_0 = \tilde{\boldsymbol{x}}_0\,, \tag{5.11}$$

where Z is a p-dimensional Lévy process and $\tilde{Z}$ is a $\tilde{p}$-dimensional Lévy process. Assume that (5.10) and (5.11) have the same generator, that $a : \mathbb{R}^d \to \mathbb{R}^{d\times p}$ and $\zeta : \mathbb{R}^{d-1} \to \mathbb{R}$ are Lipschitz and that the initial values have the same distribution. Then the post-intervention distributions of doing $X^l := \zeta(X^{(-l)})$ in (5.10) and doing $\tilde{X}^l := \zeta(\tilde{X}^{(-l)})$ in (5.11) are equal for any choice of ζ and l.

A main task in the causality research community is to uncover the conditions under which the post-intervention distributions are identifiable from the observational distribution. In the context of dynamic systems modelled in SDEs, similar conditions need to be derived. Lemma 5.3 establishes that, for SDEs with a Lévy process as the driving noise, the post-intervention distributions can be identifiable from the generator. Nevertheless, a gap remains between the observational distribution and the SDE generator's identifiability. This chapter aims to address this gap by providing conditions under which the generator is identifiable from the observational distribution.

5.3 Main results

In this section, we present some prerequisites first, and then we present the main theoretical results of our study, which include the condition for the identifiability of generator that is associated with the SDE (5.1) / SDE (5.2) from the distribution of the corresponding solution process.

5.3.1 Prerequisites

We first show that both the SDE (5.1) and the SDE (5.2) satisfy the conditions stated in Lemma 5.3.

Lemma 5.4. *Both the SDE* (5.1) *and the SDE* (5.2) *can be expressed as the form of* (5.10)*, with Z being a p-dimensional Lévy process, and $a : \mathbb{R}^d \to \mathbb{R}^{d\times p}$ being Lipschitz.*

The proof of Lemma 5.4 can be found in Appendix C.1.1. This lemma suggests that Lemma 5.3 applies to both the SDE (5.1) and the SDE (5.2), given that they meet the specified conditions. Therefore, for either SDE, deriving the conditions that allow for the generator to be identifiable from the observational distribution is sufficient. By applying Lemma 5.3, when the intervention function ζ is Lipschitz, the post-intervention distributions can be identified from the observational distribution under these conditions.

We then address the identifiability condition of the generator $\mathcal{L}$ defined by (5.6).

Proposition 5.3.1. Let $\mathcal{L}$ and $\tilde{\mathcal{L}}$ be generators of stochastic processes defined by the form of the SDE (5.5) on $C_b^2(\mathbb{R}^d)$, where $\mathcal{L}$ is given by (5.6) and $\tilde{\mathcal{L}}$ is given by the same expression, with $\tilde{b}(\boldsymbol{x})$ and $\tilde{c}(\boldsymbol{x})$ substituted for $b(\boldsymbol{x})$ and $c(\boldsymbol{x})$. It then holds that the two generators $\mathcal{L} = \tilde{\mathcal{L}}$ if and only if $b(\boldsymbol{x}) = \tilde{b}(\boldsymbol{x})$ and $c(\boldsymbol{x}) = \tilde{c}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

The proof of Proposition 5.3.1 can be found in Appendix C.1.2. This proposition states that for stochastic processes defined by the SDE (5.5), the generator is identifiable from functions associated with its coefficients: $b(\boldsymbol{x})$ and $c(\boldsymbol{x}) = \sigma(\boldsymbol{x}) \cdot \sigma(\boldsymbol{x})^\top$.

5.3.2 Conditions for identifying generators of linear SDE systems with additive noise

Expressing the SDE (5.1) in the form given by (5.5) yields $b(\boldsymbol{x}) = A\boldsymbol{x}$ and $c(\boldsymbol{x}) = GG^\top$. By defining $\tilde{b}(\boldsymbol{x}) = \tilde{A}\boldsymbol{x}$ and $\tilde{c}(\boldsymbol{x}) = \tilde{G}\tilde{G}^\top$, it can be easily checked that $b(\boldsymbol{x}) = \tilde{b}(\boldsymbol{x})$ and $c(\boldsymbol{x}) = \tilde{c}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$ if and only if $(A, GG^\top) = (\tilde{A}, \tilde{G}\tilde{G}^\top)$. Therefore, based on Proposition 5.3.1, we define the identifiability of the generator of the SDE (5.1) as follows.

Definition 5.5 ($(\boldsymbol{x}_0, A, G)$-identifiability)**.** For $\boldsymbol{x}_0 \in \mathbb{R}^d$, $A \in \mathbb{R}^{d\times d}$ and $G \in \mathbb{R}^{d\times m}$, the generator of the SDE (5.1) is said to be identifiable from $\boldsymbol{x}_0$, if for all $\tilde{A} \in \mathbb{R}^{d\times d}$ and all $\tilde{G} \in \mathbb{R}^{d\times m}$, with $(A, GG^\top) \neq (\tilde{A}, \tilde{G}\tilde{G}^\top)$, it holds that $X(\cdot; \boldsymbol{x}_0, A, G)$[1] $\overset{\mathrm{d}}{\neq} X(\cdot; \boldsymbol{x}_0, \tilde{A}, \tilde{G})$.

In the following, we begin by introducing two lemmas that serve as the foundation for deriving our main identifiability theorem.

Lemma 5.6. *For $\boldsymbol{x}_0 \in \mathbb{R}^d$, $A, \tilde{A} \in \mathbb{R}^{d\times d}$ and $G, \tilde{G} \in \mathbb{R}^{d\times m}$, let $X_t := X(t; \boldsymbol{x}_0, A, G)$, $\tilde{X}_t := X(t; \boldsymbol{x}_0, \tilde{A}, \tilde{G})$, then $X(\cdot; \boldsymbol{x}_0, A, G) \overset{\mathrm{d}}{=} X(\cdot; \boldsymbol{x}_0, \tilde{A}, \tilde{G})$ if and only if the mean $\mathbb{E}[X_t] = \mathbb{E}[\tilde{X}_t]$ and the covariance $\mathbb{E}\{(X_{t+h} - \mathbb{E}[X_{t+h}])(X_t - \mathbb{E}[X_t])^\top\} = \mathbb{E}\{(\tilde{X}_{t+h} - \mathbb{E}[\tilde{X}_{t+h}])(\tilde{X}_t - \mathbb{E}[\tilde{X}_t])^\top\}$ for all $0 \leqslant t < \infty$ and $0 \leqslant h < \infty$.*

The proof of Lemma 5.6 can be found in Appendix C.1.3. This lemma states that for stochastic processes modelled by the SDE (5.1), the equality of the distribution of two processes can be deconstructed as the equality of the mean and covariance of the state variables at all time points. Calculation shows

$$
\begin{aligned}
\mathbb{E}[X_t] &= e^{At}\boldsymbol{x}_0\,, \\
V(t, t+h) &:= \mathbb{E}\{(X_{t+h} - \mathbb{E}[X_{t+h}])(X_t - \mathbb{E}[X_t])^\top\} \\
&= e^{Ah}V(t)\,,
\end{aligned}
\tag{5.12}
$$

where $V(t) := V(t, t)$. Please refer to the proof C.1.5 of Theorem 5.8 for the detailed calculations. It can be easily checked that $\mathbb{E}[X_t]$ follows the linear ordinary differential equation (ODE)

$$
\dot{\boldsymbol{m}}(t) = A\boldsymbol{m}(t), \quad \boldsymbol{m}(0) = \boldsymbol{x}_0\,, \tag{5.13}
$$

where $\dot{\boldsymbol{m}}(t)$ denotes the first derivative of function $\boldsymbol{m}(t)$ with respect to time t. Similarly, each column of the covariance $V(t, t+h)$ also follows the linear ODE (5.13) but with a

[1] $X(\cdot; \boldsymbol{x}_0, A, G) = \{X(t; \boldsymbol{x}_0, A, G) : 0 \leqslant t < \infty\}$

different initial state: the corresponding column of $V(t)$. This observation allows us to leverage not only the characteristics of the SDE (5.1), but also the established theories [28, 29] on identifiability analysis for the ODE (5.13), to derive the identifiability conditions for the generator of the SDE (5.1).

We adopt the same setting as in [29], discussing the case where A has distinct eigenvalues. Because random matrix theory suggests that almost every $A \in \mathbb{R}^{d\times d}$ has d distinct eigenvalues with respect to the Lebesgue measure on $\mathbb{R}^{d\times d}$ [29]. And the Jordan decomposition of such a matrix A follows a straightforward form which is helpful for deriving the geometric interpretation of the proposed identifiability condition. The Jordan decomposition can be expressed as $A = Q\Lambda Q^{-1}$, where

$$\Lambda = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_K \end{bmatrix}, \text{with } J_k = \begin{cases} \lambda_k, & \text{if } k = 1, \dots, K_1\,, \\ \begin{bmatrix} a_k & -b_k \\ b_k & a_k \end{bmatrix}, & \text{if } k = K_1 + 1, \dots, K\,. \end{cases}$$

$$Q = [Q_1| \dots |Q_K] = [v_1| \dots |v_d]\,,$$

$$Q_k = \begin{cases} v_k, & \text{if } k = 1, \dots, K_1\,, \\ \left[v_{2k-K_1-1}|v_{2k-K_1}\right], & \text{if } k = K_1 + 1, \dots, K\,, \end{cases}$$

where λ_k is a real eigenvalue of A and v_k is the corresponding eigenvector of λ_k, for $k = 1, \dots, K_1$. For $k = K_1+1, \dots, K$, $[v_{2k-K_1-1}|v_{2k-K_1}]$ are the corresponding "eigenvectors" of complex eigenvalues $a_k \pm b_k i$. Inspired by [29, Definition 2.3., Lemma 2.3.], we establish the following Lemma.

Lemma 5.7. *Assuming $A \in \mathbb{R}^{d\times d}$ has d distinct eigenvalues, with Jordan decomposition $A = Q\Lambda Q^{-1}$. Let $\boldsymbol{\gamma}_j \in \mathbb{R}^d$ and $\tilde{\boldsymbol{\gamma}}_j := Q^{-1}\boldsymbol{\gamma}_j \in \mathbb{R}^d$ for all $j = 1, \dots, n$ with $n \geqslant 2$. We define*

$$w_{j,k} := \begin{cases} \tilde{\gamma}_{j,k} \in \mathbb{R}^1, & \textit{for } k = 1, \dots, K_1\,, \\ (\tilde{\gamma}_{j,2k-K_1-1}, \tilde{\gamma}_{j,2k-K_1})^\top \in \mathbb{R}^2, & \textit{for } k = K_1 + 1, \dots, K\,, \end{cases}$$

where $\tilde{\gamma}_{j,k}$ denotes the k-th entry of $\tilde{\boldsymbol{\gamma}}_j$. $\operatorname{rank}([\boldsymbol{\gamma}_1|A\boldsymbol{\gamma}_1|\dots|A^{d-1}\boldsymbol{\gamma}_1|\dots|\boldsymbol{\gamma}_n|A\boldsymbol{\gamma}_n|\dots|A^{d-1}\boldsymbol{\gamma}_n]) < d$ if and only if there exists $k \in \{1, \dots, K\}$, such that $|w_{j,k}| = 0$ for all $j = 1, \dots, n$, where $|w_{j,k}|$ is the absolute value of $w_{j,k}$ for $k = 1, \dots, K_1$, and the Euclidean norm of $w_{j,k}$ for $k = K_1 + 1, \dots, K$.

The proof of Lemma 5.7 can be found in Appendix C.1.4. From a geometric perspective, $\boldsymbol{\gamma}_j$ can be decomposed into a linear combination of Q_k's

$$\boldsymbol{\gamma}_j = Q\tilde{\boldsymbol{\gamma}}_j = \textstyle\sum_{k=1}^{K} Q_k w_{j,k} \,.$$

Let $L_k := \text{span}(Q_k)$. According to [29, Theorem 2.2], each L_k is an A-invariant subspace of $\mathbb{R}^d$. Recall that a space L is called A-invariant, if for all $\boldsymbol{\gamma} \in L$, $A\boldsymbol{\gamma} \in L$. We say L is a proper subspace of $\mathbb{R}^d$ if $L \subset \mathbb{R}^d$ and $L \neq \mathbb{R}^d$. If $|w_{j,k}| = 0$ (i.e., $w_{j,k} = 0$ in $\mathbb{R}^1$ or $\mathbb{R}^2$), then $\boldsymbol{\gamma}_j$ does not contain any information from L_k. In this case, $\boldsymbol{\gamma}_j$ is contained in an A-invariant proper subspace of $\mathbb{R}^d$ that excludes L_k, denoted as L_{-k}. It is worth emphasizing that $L_{-k} \subset \mathbb{R}^d$ is indeed a **proper** subspace of $\mathbb{R}^d$. This further implies that the trajectory of the ODE (5.13) generated from initial state $\boldsymbol{\gamma}_j$ is confined to L_{-k} [28, Lemma 3.2]. Lemma 5.7 indicates that if $\text{rank}([\boldsymbol{\gamma}_1|A\boldsymbol{\gamma}_1|\ldots|A^{d-1}\boldsymbol{\gamma}_1|\ldots|\boldsymbol{\gamma}_n|A\boldsymbol{\gamma}_n|\ldots|A^{d-1}\boldsymbol{\gamma}_n]) < d$ then all $\boldsymbol{\gamma}_j$ for $j = 1, \ldots, n$ are confined to an A-invariant proper subspace of $\mathbb{R}^d$, denoted as L. Therefore, all trajectories of the ODE (5.13) generated from initial states $\boldsymbol{\gamma}_j$ are also confined to L. Furthermore, based on the identifiability conditions proposed in [28], the ODE (5.13) is not identifiable from observational data collected in these trajectories. This lemma provides an approach to interpreting our identifiability conditions from a geometric perspective.

Now we are ready to present our main theorem.

Theorem 5.8. *Let $\boldsymbol{x}_0 \in \mathbb{R}^d$ be fixed. Assuming that the matrix A in the SDE (5.1) has d distinct eigenvalues. The generator of the SDE (5.1) is identifiable from $\boldsymbol{x}_0$ if and only if*

$$\text{rank}([\boldsymbol{x}_0|A\boldsymbol{x}_0|\ldots|A^{d-1}\boldsymbol{x}_0|H_{\cdot 1}|AH_{\cdot 1}|\ldots|A^{d-1}H_{\cdot 1}|\ldots|H_{\cdot d}|AH_{\cdot d}|\ldots|A^{d-1}H_{\cdot d}]) = d\,, \tag{5.14}$$

where $H := GG^T$, and $H_{\cdot j}$ stands for the j-th column vector of matrix H, for all $j = 1, \cdots, d$.

The proof of Theorem 5.8 can be found in Appendix C.1.5. The condition in Theorem 5.8 is both sufficient and necessary when the matrix A has distinct eigenvalues. It is worth noting that almost every $A \in \mathbb{R}^{d\times d}$ has d distinct eigenvalues concerning the Lebesgue measure on $\mathbb{R}^{d\times d}$. Hence, this condition is both sufficient and necessary for almost every

A in $\mathbb{R}^{d\times d}$. However, in cases where A has repetitive eigenvalues, this condition is solely sufficient and not necessary.

Remark. The identifiability condition stated in Theorem 5.8 is generic, that is, let

$$S := \{(\boldsymbol{x}_0, A, G) \in \mathbb{R}^{d+d^2+dm} : \text{condition (5.14) is violated}\}\,,$$

S has Lebesgue measure zero in $\mathbb{R}^{d+d^2+dm}$. Refer to Appendix C.2.2 for the detailed proof.

From the geometric perspective, suppose matrix A has distinct eigenvalues, the generator of the SDE (5.1) is identifiable from $\boldsymbol{x}_0$ when not all of the vectors: $\boldsymbol{x}_0, H_{\cdot 1}, \ldots, H_{\cdot d}$ are confined to an A-invariant **proper** subspace of $\mathbb{R}^d$. A key finding is that when all the vectors $H_{\cdot j}$, $j = 1, \ldots, d$ are confined to an A-invariant proper subspace L of $\mathbb{R}^d$, each column of the covariance matrix $V(t)$ in Equation (5.12) is also confined to L, for all $0 \leqslant t < \infty$. Thus, the identifiability of the generator of the SDE (5.1) can be fully determined by $\boldsymbol{x}_0$ and the system parameters $(A, GG^\top)$. Further details can be found in the proof C.1.5 of Theorem 5.8.

By rearranging the matrix in (5.14), the identifiability condition can also be expressed as

$$\operatorname{rank}([\boldsymbol{x}_0|A\boldsymbol{x}_0|\ldots|A^{d-1}\boldsymbol{x}_0|GG^\top|AGG^\top|\ldots|A^{d-1}GG^\top]) = d\,. \tag{5.15}$$

Based on the identifiability condition (5.15), we derive the following corollary.

Corollary 5.9. *Let $\boldsymbol{x}_0 \in \mathbb{R}^d$ be fixed. If $\operatorname{rank}([G|AG|\ldots|A^{d-1}G]) = d$, then the generator of the SDE (5.1) is identifiable from $\boldsymbol{x}_0$.*

The proof of Corollary 5.9 can be found in Appendix C.1.6. This corollary indicates that the generator of the SDE (5.1) is identifiable from **any** initial state $\boldsymbol{x}_0 \in \mathbb{R}^d$ when the pair $[A, G]$ is controllable ($\operatorname{rank}([G|AG|\ldots|A^{d-1}G]) = d$). Notably, this identifiability condition is stricter than that proposed in Theorem 5.8, as it does not use the information of $\boldsymbol{x}_0$.

5.3.3 Conditions for identifying generators of linear SDE systems with multiplicative noise

Expressing the SDE (5.2) in the form given by (5.5) yields $b(\boldsymbol{x}) = A\boldsymbol{x}$ and $\sigma(\boldsymbol{x}) = [G_1\boldsymbol{x}|\dots|G_m\boldsymbol{x}] \in \mathbb{R}^{d\times m}$, thus, $c(\boldsymbol{x}) = \sigma(\boldsymbol{x})\sigma(\boldsymbol{x})^\top = \sum_{k=1}^m G_k\boldsymbol{x}\boldsymbol{x}^\top G_k^\top$. By defining $\tilde{b}(\boldsymbol{x}) = \tilde{A}\boldsymbol{x}$ and $\tilde{c}(\boldsymbol{x}) = \sum_{k=1}^m \tilde{G}_k\boldsymbol{x}\boldsymbol{x}^\top\tilde{G}_k^\top$, it can be easily checked that $b(\boldsymbol{x}) = \tilde{b}(\boldsymbol{x})$ and $c(\boldsymbol{x}) = \tilde{c}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$ if and only if $(A, \sum_{k=1}^m G_k\boldsymbol{x}\boldsymbol{x}^\top G_k^\top) = (\tilde{A}, \sum_{k=1}^m \tilde{G}_k\boldsymbol{x}\boldsymbol{x}^\top\tilde{G}_k^\top)$ for all $\boldsymbol{x} \in \mathbb{R}^d$. Let $X(t;\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m)$ denote the solution to the SDE (5.2), then based on Proposition 5.3.1, we define the identifiability of the generator of the SDE (5.2) as follows.

Definition 5.10 ($(\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m)$-identifiability)**.** For $\boldsymbol{x}_0 \in \mathbb{R}^d$, $A, G_k \in \mathbb{R}^{d\times d}$ for all $k = 1,\dots,m$, the generator of the SDE (5.2) is said to be identifiable from $\boldsymbol{x}_0$, if for all $\tilde{A}, \tilde{G}_k \in \mathbb{R}^{d\times d}$, there exists an $\boldsymbol{x} \in \mathbb{R}^d$, such that $(A, \sum_{k=1}^m G_k\boldsymbol{x}\boldsymbol{x}^\top G_k^\top) \neq (\tilde{A}, \sum_{k=1}^m \tilde{G}_k\boldsymbol{x}\boldsymbol{x}^\top\tilde{G}_k^\top)$, it holds that $X(\cdot;\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m) \overset{\mathrm{d}}{\neq} X(\cdot;\boldsymbol{x}_0, \tilde{A}, \{\tilde{G}_k\}_{k=1}^m)$.

Based on Definition 5.10, we present the identifiability condition for the generator of the SDE (5.2).

Theorem 5.11. *Let $\boldsymbol{x}_0 \in \mathbb{R}^d$ be fixed. The generator of the SDE (5.2) is identifiable from $\boldsymbol{x}_0$ if the following conditions are satisfied:*

A1 $\operatorname{rank}([\boldsymbol{x}_0|A\boldsymbol{x}_0|\dots|A^{d-1}\boldsymbol{x}_0]) = d\,,$

A2 $\operatorname{rank}([\boldsymbol{v}|\mathcal{A}\boldsymbol{v}|\dots|\mathcal{A}^{(d^2+d-2)/2}\boldsymbol{v}]) = (d^2+d)/2\,,$

where $\mathcal{A} = A \oplus A + \sum_{k=1}^m G_k \otimes G_k \in \mathbb{R}^{d^2\times d^2}$, $\oplus$ denotes Kronecker sum and $\otimes$ denotes Kronecker product, $\boldsymbol{v}$ is a d^2-dimensional vector defined by $\boldsymbol{v} := \operatorname{vec}(\boldsymbol{x}_0\boldsymbol{x}_0^\top)$, where $\operatorname{vec}(M)$ denotes the vectorization of matrix M.

The proof of Theorem 5.11 can be found in Appendix C.1.7. This condition is only sufficient but not necessary. Specifically, condition A1 guarantees that matrix A is identifiable, and once A is identifiable, condition A2 ensures that the identifiability of $\sum_{k=1}^m G_k\boldsymbol{x}\boldsymbol{x}^\top G_k^\top$ holds for all $\boldsymbol{x} \in \mathbb{R}^d$.

Remark. The identifiability condition stated in Theorem 5.11 is generic, that is, let

$$S := \{(\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m) \in \mathbb{R}^{d+(m+1)d^2} : \text{ either condition A1 or A2 in Theorem 5.11 is violated}\}\,,$$

S has Lebesgue measure zero in $\mathbb{R}^{d+(m+1)d^2}$. This signifies that the conditions are satisfied for most of the combinations of $\boldsymbol{x}_0$, A and G_k's, except for those that lie in a set of Lebesgue measure zero. The corresponding proposition and detailed proof can be found in Appendix C.2.1.

Since obtaining an explicit solution for the SDE (5.2) is generally infeasible, we resort to utilizing the first- and second-order moments of this SDE to derive the identifiability conditions. Let $m(t) := \mathbb{E}[X_t]$ and $P(t) := \mathbb{E}[X_t X_t^\top]$, it is known that these moments satisfy ODE systems. Specifically, $m(t)$ satisfies the ODE (5.13), while $P(t)$ satisfies the following ODE (cf. [147]):

$$\dot{P}(t) = AP(t) + P(t)A^\top + \textstyle\sum_{k=1}^m G_k P(t) G_k^\top \,, \quad P(0) = \boldsymbol{x}_0 \boldsymbol{x}_0^\top \,. \tag{5.16}$$

An important trick to deal with the ODE (5.16) is to vectorize $P(t)$, then it can be expressed as:

$$\mathrm{vec}(\dot{P}(t)) = \mathcal{A}\mathrm{vec}(P(t)) \,, \quad \mathrm{vec}(P(0)) = \boldsymbol{v} \,, \tag{5.17}$$

where $\mathcal{A}$ and $\boldsymbol{v}$ are defined in Theorem 5.11. In fact, the ODE (5.17) follows the same mathematical structure as that of the ODE (5.13), which is known as homogeneous linear ODEs. Thus, in addition to the inherent properties of the SDE (5.2), we also employ some existing identifiability theories for homogeneous linear ODEs to establish the identifiability condition for the generator of the SDE (5.2).

From the geometric perspective, condition A1 indicates that the initial state $\boldsymbol{x}_0$ is not confined to an A-invariant **proper** subspace of $\mathbb{R}^d$ [28, Lemma 3.1.]. And condition A2 implies that the vectorization of $\boldsymbol{x}_0 \boldsymbol{x}_0^\top$ is not confined to an $\mathcal{A}$-invariant **proper** subspace of W, with $W \subset \mathbb{R}^{d^2}$, and $\dim(W) = (d^2 + d)/2$, where $\dim(W)$ denotes the dimension of the subspace W, that is the number of vectors in any basis for W. In particular, one can construct a basis for W as follows:

$$\{\mathrm{vec}(E_{11}), \mathrm{vec}(E_{21}), \mathrm{vec}(E_{22}), \ldots, \mathrm{vec}(E_{dd})\} \,,$$

where E_{ij} denotes a $d \times d$ matrix whose ij-th and ji-th elements are 1, and all other elements are 0, for all $i, j = 1, \ldots, d$ and $i \geqslant j$. Refer to the proof C.1.7 of Theorem 5.11 for more details.

5.4 Simulations and examples

In order to assess the validity of the identifiability conditions established in Section 5.3, we present the results of simulations. Specifically, we consider SDEs with system parameters that either satisfy or violate the proposed identifiability conditions. We then apply the maximum likelihood estimation (MLE) method to estimate the system parameters from discrete observations sampled from the corresponding SDE. The accuracy of the resulting parameter estimates serves as an indicator of the validity of the proposed identifiability conditions.

Simulations. We conduct five sets of simulations, which include one identifiable case and one unidentifiable case for the SDE (5.1), and one identifiable case and two unidentifiable cases with either condition A1 or A2 in Theorem 5.11 unsatisfied for the SDE (5.2). We set both the system dimension, d, and the Brownian motion dimension, m, to 2. Details on the true underlying system parameters for the SDEs can be found in Appendix C.3. We simulate observations from the true SDEs for each of the five cases under investigation. Specifically, the simulations are carried out for different numbers of trajectories (N), with 50 equally-spaced observations sampled on each trajectory from the time interval $[0, 1]$. We employ the Euler-Maruyama (EM) method [148], a widely used numerical scheme for simulating SDEs, to generate the observations.

Estimation. We use MLE [49, 149] to estimate the system parameters. The MLE method requires knowledge of the transition probability density function (pdf) that governs the evolution of the system. For the specific case of the SDE (5.1), the transition density follows a Gaussian distribution, which can be computed analytically based on the system's drift and diffusion coefficients (cf. [49]). To compute the covariance, we employ the commonly used matrix fraction decomposition method [49, 150, 151]. However, in general, the transition pdf of the SDE (5.2) cannot be obtained analytically due to the lack of a closed-form solution. To address this issue, we implement the Euler-Maruyama approach [148, 152], which has been shown to be effective in approximating the transition pdf of SDEs.

Metric. We adopt the commonly used metric, mean squared error (MSE), to assess the accuracy of parameter estimates. For each configuration, we conduct 100 independent random replications and present the MSEs along with their variances to evaluate estimation reliability.

TABLE 5.1: Simulation results of the SDE (5.1)

N	Identifiable		Unidentifiable	
	MSE-A	MSE-$GG^\top$	MSE-A	MSE-$GG^\top$
5	0.0117 ± 0.0115	5.28E-05 $\pm$ 4.39E-05	3.66 ± 0.10	0.05 ± 0.03
10	0.0063 ± 0.0061	2.39E-05 $\pm$ 1.82E-05	3.88 ± 0.06	0.64 ± 0.59
20	0.0029 ± 0.0027	1.87E-05 $\pm$ 1.51E-05	3.70 ± 0.06	0.09 ± 0.07
50	0.0013 ± 0.0010	8.00E-06 $\pm$ 5.68E-06	3.76 ± 0.07	0.11 ± 0.08
100	0.0007 ± 0.0004	4.34E-06 $\pm$ 2.70E-06	3.66 ± 0.02	2.09 ± 1.98

Results analysis. Table 5.1 and Table 5.2 present the simulation results for the SDE (5.1) and the SDE (5.2), respectively. In Table 5.1, the simulation results demonstrate that in the identifiable case, as the number of trajectories N increases, the MSE for both A and $GG^\top$ decreases and approaches zero. However, in the unidentifiable case, where the identifiable condition (5.14) stated in Theorem 5.8 is not satisfied, the MSE for both A and $GG^\top$ remains high regardless of the number of trajectories. These findings provide strong empirical evidence supporting the validity of the identifiability condition proposed in Theorem 5.8. The simulation results presented in Table 5.2 show that in the identifiable case, the MSE for both A and $Gs\boldsymbol{x}$ decreases and approaches zero with the increase of the number of trajectories N . Here, $Gs\boldsymbol{x} := \sum_{k=1}^{m} G_k \boldsymbol{x}\boldsymbol{x}^\top G_k^\top$, where $\boldsymbol{x}$ is a randomly generated vector from $\mathbb{R}^2$ (in these simulations, $\boldsymbol{x} = [1.33, 0.72]^\top$). Interestingly, even in unidentifiable case 1, the MSE for both A and $Gs\boldsymbol{x}$ decreases with an increasing number of trajectories N, indicating that the generator of the SDE utilized in this particular case is still identifiable, although a larger number of trajectories is required compared to the identifiable case to achieve the same level of accuracy. This result is reasonable, because it aligns with our understanding that condition A1 is only sufficient but not necessary for identifying A, as the lack of an explicit solution for the SDE (5.2) results in condition A1 not incorporating any information from G_k's. The identifiability condition derived for the SDE (5.1) in Theorem 5.8 leverages the information of G, similarly, if information regarding G_k's is available, a weaker condition for identifying A could be obtained. For illustration, in Appendix C.5, we present such a condition assuming the SDE (5.2) has a closed-form solution. In the case of unidentifiable case 2, the MSE for A decreases with an increasing number of trajectories N; however, the MSE for $Gs\boldsymbol{x}$ remains high, indicating that A is identifiable, while $Gs\boldsymbol{x}$ is not, albeit requiring more trajectories compared to the identifiable case to achieve the same level of accuracy of A (since the $Gs\boldsymbol{x}$ is far away from its true underlying value). This finding is consistent with the derived identifiability condition, as condition A1 is sufficient to identify A, whereas condition A2 governs the identifiability of $Gs\boldsymbol{x}$. Worth noting that in cases where neither condition A1 nor condition A2 is satisfied, the estimated parameters barely deviate from their initial values, implying poor estimation of both A and $Gs\boldsymbol{x}$. These results indicate the validity of the identifiability condition stated in Theorem 5.11.

TABLE 5.2: Simulation results of the SDE (5.2)

N	**Identifiable**		**Unidentifiable**			
			case1: A1-False, A2-True		**case2: A1-True, A2-False**	
	MSE-A	MSE-$Gs\boldsymbol{x}$	MSE-A	MSE-$Gs\boldsymbol{x}$	MSE-A	MSE-$Gs\boldsymbol{x}$
10	0.069 ± 0.061	0.3647 ± 0.3579	0.509 ± 0.499	0.194 ± 0.140	2.562 ± 2.522	9763 ± 8077
20	0.047 ± 0.045	0.1769 ± 0.1694	0.195 ± 0.180	0.088 ± 0.058	0.967 ± 0.904	8353 ± 6839
50	0.018 ± 0.018	0.1703 ± 0.1621	0.132 ± 0.131	0.081 ± 0.045	0.423 ± 0.410	4779 ± 4032
100	0.006 ± 0.006	0.0015 ± 0.0012	0.065 ± 0.065	0.068 ± 0.036	0.207 ± 0.198	3569 ± 3150
500	0.001 ± 0.001	0.0004 ± 0.0001	0.008 ± 0.008	0.059 ± 0.004	0.046 ± 0.046	4490 ± 3991

Illustrative instances of causal inference for linear SDEs (with interventions). To illustrate how our proposed identifiability conditions can guarantee reliable causal inference for linear SDEs, we present examples corresponding to both the SDE (5.1) and the SDE (5.2). In these examples, we show that under our proposed identifiability conditions, the post-intervention distributions are identifiable from their corresponding observational distributions. Please refer to Appendix C.4.1 and C.4.2 for the details of the examples.

5.5 Related work

Most current studies on the identifiability analysis of SDEs are based on the Gaussian diffusion processes that conform to the form described in the SDE (5.1). In particular, the authors of [40–42] have conducted research on the identifiability or asymptotic properties of parameter estimators of Gaussian diffusions in view of continuous observations of one trajectory, and have highlighted the need for the diffusion to be ergodic. A considerable amount of effort has also been directed towards the identifiability analysis of Gaussian diffusions, relying on the exact discrete models of the SDEs [43–45, 139, 153]. Typically, these studies involve transferring the continuous-time system described in the SDE (5.1) to a discrete-time model such as a vector autoregressive model, based on equally-spaced observations sampled from one trajectory, and then attempting to determine conditions under which $(A, GG^\top)$ is identifiable from the parameters of the corresponding exact discrete models. These conditions often have requirements on eigenvalues of A among other conditions, such as requiring the eigenvalues to have only negative real parts, or the eigenvalues to be strictly real. Due to the limitation of the available observations (continuous or discrete observations located on one trajectory of the SDE system), the identifiability conditions proposed in these works are restrictive.

Causal modelling theories have been well-developed based on directed acyclic graphs (DAGs), which do not explicitly incorporate a time component [112]. In recent years, similar concepts of causality have been developed for dynamic systems operating in both discrete and continuous time. Discrete-time models, such as autoregressive processes, can be readily accommodated within the DAG-based framework [154, 155]. On the other hand, differential equations offer a natural framework for understanding causality in dynamic systems within the context of continuous-time processes [22, 93]. Consequently, considerable effort has been devoted to establishing a theoretical connection between causality and differential equations. In the deterministic case, Mooij et al. [18] and Rubenstein et al. [19] have established a mathematical link between ODEs and structural causal models (SCMs). Wang et al. [1] have proposed a method to infer the causal structure of linear ODEs. Turning to the stochastic case, Boogers and Mooij have built a bridge from random differential equations (RDEs) to SCMs [20], while Hansen and Sokol have proposed a causal interpretation of SDEs by establishing a connection between SDEs and SEMs [21].

5.6 Conclusion of chapter

In this chapter, we present an investigation into the identifiability of the generators of linear SDEs under additive and multiplicative noise. Specifically, we derive the conditions that are fully built on system parameters and the initial state $\boldsymbol{x}_0$, which enables the identification of a linear SDE's generator from the distribution of its solution process with a given fixed initial state. We establish that, under the proposed conditions, the post-intervention distribution is identifiable from the corresponding observational distribution for any Lipschitz intervention ζ.

The main limitation of our work is that the practical verification of these identifiability conditions poses a challenge, as the true underlying system parameters are typically unavailable in real-world applications. Nevertheless, our study contributes to the understanding of the intrinsic structure of linear SDEs. By offering valuable insights into the identifiability aspects, our findings empower researchers and practitioners to employ models that satisfy the proposed conditions (e.g., through constrained parameter estimation) to learn real-world data while ensuring identifiability. We believe the paramount significance of this work lies in providing a systematic and rigorous causal interpretation of linear SDEs, which facilitates reliable causal inference for dynamic systems governed by such equations. It is worth noting that in our simulations, we employed the MLE method to estimate the system parameters. This necessitates the calculation of the transition pdf from one state to the successive state at each discrete temporal increment. Consequently, as the state dimension and Brownian motion dimension increase, the computational time is inevitably significantly increased, rendering the process quite time-consuming. To expedite parameter estimation for scenarios involving high dimensions, alternative estimation approaches are required. The development of a more efficient parameter estimation approach remains an important task in the realm of SDEs, representing a promising direction for our future research.

Chapter 6

Conclusion and Discussion

6.1 Summary of contributions

This thesis investigates fundamental identifiability problems in continuous-time causal models, focusing on linear ODE and SDE systems under various observational settings. The contributions span three complementary dimensions:

1. **Identifiability of linear ODE systems from discrete observations.** In Chapter 3, we establish that a homogeneous linear ODE system is identifiable from a single sequence of equally spaced observations, even when the initial condition is unknown. We prove that under mild regularity conditions, the Nonlinear Least Squares (NLS) estimator is consistent and asymptotically normal. These results provide a rigorous statistical foundation for constructing confidence intervals and performing hypothesis testing on causal structures.

2. **Identifiability of linear ODE systems with hidden confounders.** Chapter 4 extends the analysis to settings where part of the dynamical system is unobserved. We consider two classes of latent confounding: (i) independent latent variables evolving according to a known functional form, and (ii) dynamically evolving latent states structured by a DAG. For both cases, we establish identifiability conditions—sufficient, and in some instances also necessary—that guarantee partial/full recovery of the causal mechanism from observed data. These results enable valid causal inference even in the presence of latent confounders.

3. **Generator identification for linear SDE systems.** Chapter 5 addresses identifiability in stochastic dynamics. For linear SDEs with additive noise, we prove that the generator is generically identifiable under a structural condition on the drift and diffusion matrices, which is both sufficient and necessary. For multiplicative noise, we establish a generic sufficient identifiability condition. Both conditions ensure that the infinitesimal generator—which determines all Lipschitz-continuous post-intervention distributions—can be uniquely recovered from observational data. In both cases, we provide geometric insights to aid interpretation and understanding.

6.2 Future directions

The results in this thesis open several promising directions for further research:

1. **Beyond linearity.** Extending identifiability analysis to nonlinear ODE and SDE systems remains an important challenge. Approaches such as local linearization, Lie algebraic methods, or kernel-based representations may offer tractable paths forward.

2. **Finite-sample and robust inference.** The theoretical results in this thesis assume idealized noise-free observations or asymptotic regimes. Bridging these with robust finite-sample estimators that retain identifiability guarantees under bounded noise or model misspecification is essential for real-world applications.

3. **Empirical validation and benchmarks.** Developing benchmark datasets and empirical protocols for validating identifiability conditions in scientific domains (e.g., systems biology, finance, climate science) will provide a testbed for assessing practical applicability.

4. **Causal discovery under SDE generators.** In stochastic settings, identifiability of the generator opens avenues for causal discovery frameworks that directly operate on observational distributions. Leveraging recent developments in score-based or generative modeling may prove fruitful.

5. **Interventional design and control.** Building on identifiability, one can explore how to optimally design interventions or control inputs to improve inference efficiency or disambiguate latent causal structure—especially in adaptive or online settings.

6. **Irregularly spaced discrete observations.** In our analysis of identifiability for linear ODE systems from discrete observations, we focused on the case of equally spaced sampling from a single trajectory. While this setting is common in practice, extending the results to irregularly spaced discrete observations would be valuable. As noted in [94], irregular sampling can sometimes provide richer information about the underlying dynamics.

7. **Applications to real-world datasets.** While this thesis has focused on theoretical results supported by simulation studies, an important next step is to apply the proposed methods to suitable real-world datasets. Such applications would help assess the empirical utility of the identifiability conditions and demonstrate their relevance in scientific domains.

6.3 Concluding remarks

Reliable causal inference in continuous-time systems hinges critically on whether the underlying dynamical model is uniquely recoverable from data. This thesis provides theoretical and statistical tools to address this challenge for linear ODE and SDE systems, even in the presence of latent confounding and stochasticity.

By formalizing parameter-identifiability or generator-identifiability under various observational regimes, we offer a principled framework for causal reasoning and prediction in time-dependent domains. We hope these results contribute to a deeper understanding of causal mechanisms in dynamical systems, and serve as a foundation for future methodological and applied research.

Appendix A

Appendix of Chapter 3

A.1 Detailed proofs

In this appendix, we present the detailed proofs of all our lemmas, theorems and corollaries.

A.1.1 Proof of Lemma 3.2

Proof. For $A, A', \boldsymbol{x}_0, \boldsymbol{x}_0'$ defined in Definition 3.1, if $\boldsymbol{x}_0 \neq \boldsymbol{x}_0'$, one sees that $\boldsymbol{x}(\cdot; \boldsymbol{x}_0, A) \neq \boldsymbol{x}(\cdot; \boldsymbol{x}_0', A')$, because $\boldsymbol{x}(0; \boldsymbol{x}_0, A) \neq \boldsymbol{x}(0; \boldsymbol{x}_0', A')$. Therefore, one only needs to consider the case where $\boldsymbol{x}_0 = \boldsymbol{x}_0'$ and $A \neq A'$. Under this circumstance, the Definition 3.1 and Lemma 3.2 are similar to [28, Definition 2.3, Theorem 2.5], with the only difference is that the identifiability of ODE system (3.1) not only applies to a fixed initial condition $\boldsymbol{x}_0$ but an open set $M^0 \subset \mathbb{R}^d$. According to [28, Proof of Theorem 2.5], we can directly get the proof for Lemma 3.2. $\square$

A.1.2 Proof of Theorem 3.4

Proof. Let $\Phi(t)$ denote the principal matrix solution of model (3.1), that is, $\Phi(t) := e^{At}$. Let $\boldsymbol{X}_q$ denote the matrix $(\boldsymbol{x}_q, \boldsymbol{x}_{q+1}, \ldots, \boldsymbol{x}_{q+d-1}) \in \mathbb{R}^{d\times d}$, for $q = 1, 2$. As defined in the statement of Theorem 3.4, we have $d + 1$ equally-spaced observations $\boldsymbol{x}_j = \boldsymbol{x}(t_j; \boldsymbol{x}_0, A)$, with the t_j's denoting equally-spaced time points, for $j = 1, \ldots, d+1$, we denote the equal time space as $\Delta t := t_{j+1} - t_j$, for all $j = 1, \dots, d$. According to the solution function (3.3), one obtains

$$\boldsymbol{x}_{j+1} = \boldsymbol{x}(t_{j+1}; \boldsymbol{x}_0, A) = \boldsymbol{x}(t_j + \Delta t; \boldsymbol{x}_0, A) = e^{A\Delta t} e^{At_j} \boldsymbol{x}_0 = e^{A\Delta t} \boldsymbol{x}(t_j; \boldsymbol{x}_0, A) = \Phi(\Delta t) \boldsymbol{x}_j \,,$$

for all $j = 1, \dots, d$. Therefore,

$$\boldsymbol{X}_2 = \Phi(\Delta t) \boldsymbol{X}_1 \,.$$

Then the proof can be broken down into three steps.

Step i: We show that $\boldsymbol{X}_1$ is invertible, thus $\Phi(\Delta t) = \boldsymbol{X}_2 \boldsymbol{X}_1^{-1}$.

Suppose that $\boldsymbol{X}_1 = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_d) \in \mathbb{R}^{d \times d}$ is singular, that is, $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_d$ are linearly dependent, then there exists a non-zero vector $\boldsymbol{c} = [c_1, c_2, \dots, c_d]^\top \neq \mathbf{0}_d$ satisfying $\boldsymbol{X}_1 \boldsymbol{c} = \mathbf{0}_d$, that is,

$$c_1 \boldsymbol{x}_1 + c_2 \boldsymbol{x}_2 + \cdots + c_d \boldsymbol{x}_d = \mathbf{0}_d \,,$$

where $\mathbf{0}_d$ denotes the d-dimensional zero vector. By plugging the solution of $\boldsymbol{x}_j$ in the above equation, one obtains that

$$c_1 e^{At_1} \boldsymbol{x}_0 + c_2 e^{A\Delta t} e^{At_1} \boldsymbol{x}_0 + \cdots + c_d e^{A(d-1)\Delta t} e^{At_1} \boldsymbol{x}_0 = \mathbf{0}_d \,,$$

which is

$$(c_1 I + c_2 e^{A\Delta t} + \cdots + c_d e^{A(d-1)\Delta t}) e^{At_1} \boldsymbol{x}_0 = \mathbf{0}_d \,. \tag{A.1}$$

Under condition A2 stated in Theorem3.4, one gets the Jordan decomposition of parameter matrix A, denoting as $A = Q \Lambda Q^{-1}$, with $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_d)$, a d-dimensional diagonal matrix, where $\lambda_1, \lambda_2, \dots, \lambda_d$ are d distinct real eigenvalues of the parameter matrix A. Then Equation (A.1) is equivalent to

$$Q(c_1 I + c_2 e^{\Lambda\Delta t} + \cdots + c_d e^{\Lambda(d-1)\Delta t}) e^{\Lambda t_1} Q^{-1} \boldsymbol{x}_0 = \mathbf{0}_d \,.$$

Since matrix Q is invertible, by multiplying Q^{-1} in both hand sides of the equation, one sees that

$$(c_1 I + c_2 e^{\Lambda\Delta t} + \cdots + c_d e^{\Lambda(d-1)\Delta t}) e^{\Lambda t_1} Q^{-1} \boldsymbol{x}_0 = \mathbf{0}_d \,,$$

which is

$$\begin{bmatrix} c_1 + c_2 e^{\lambda_1 \Delta t} + \cdots + c_d e^{\lambda_1 (d-1)\Delta t} & & \\ & \ddots & \\ & & c_1 + c_2 e^{\lambda_d \Delta t} + \cdots + c_d e^{\lambda_d (d-1)\Delta t} \end{bmatrix} e^{\Lambda t_1} Q^{-1} \boldsymbol{x}_0 = \mathbf{0}_d \,,$$

Let

$$u_i := e^{\lambda_i t_1}(c_1 + c_2 e^{\lambda_i \Delta t} + \cdots + c_d e^{\lambda_i (d-1)\Delta t}) \,,$$

for all $i = 1, \ldots, d$ and matrix $U := \mathrm{diag}(u_1, \ldots, u_d)$, Equation (A.1) is equivalent to

$$UQ^{-1}\boldsymbol{x}_0 = \mathbf{0}_d \,. \tag{A.2}$$

We next show that if Equation (A.2) holds, then U is a zero matrix. Let

$$\tilde{\Lambda} = \begin{bmatrix} 1 & \lambda_1 & \cdots & \lambda_1^{d-1} \\ 1 & \lambda_2 & \cdots & \lambda_2^{d-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_d & \cdots & \lambda_d^{d-1} \end{bmatrix} .$$

Since, by condition A2, $\tilde{\Lambda}$ is a square Vandermonde matrix with all the λ_i's distinct, $\tilde{\Lambda}$ is invertible. Thus, for any $\boldsymbol{c}$, there always exists a unique vector $\boldsymbol{l} = [l_1, l_2, \ldots, l_d]^\top \in \mathbb{R}^d$ such that

$$\boldsymbol{l} = \tilde{\Lambda}^{-1} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{bmatrix} .$$

That is $l_1 + l_2\lambda_i + \ldots + l_d\lambda_i^{d-1} = u_i$ for all $i = 1, \ldots, d$. Using the Jordan decomposition form of A, we then have

$$l_1\boldsymbol{x}_0 + l_2 A\boldsymbol{x}_0 + \cdots + l_d A^{d-1}\boldsymbol{x}_0 = Q(l_1 I + l_2\Lambda + \cdots + l_d\Lambda^{d-1})Q^{-1}\boldsymbol{x}_0 \;= QUQ^{-1}\boldsymbol{x}_0 \,.$$

Thus, if Equation (A.2) holds, then we have

$$l_1\boldsymbol{x}_0 + l_2 A\boldsymbol{x}_0 + \cdots + l_d A^{d-1}\boldsymbol{x}_0 = \mathbf{0}_d \,. \tag{A.3}$$

Under condition A1 stated in Theorem 3.4, if Equation (A.3) holds, one obtains $\boldsymbol{l} = \mathbf{0}_d$ since $\{\boldsymbol{x}_0, A\boldsymbol{x}_0, \ldots, A^{d-1}\boldsymbol{x}_0\}$ are linearly independent. Then $u_i = 0$ for all $i = 1, \ldots, d$ and U is a zero matrix, which means

$$\begin{cases} c_1 + c_2 e^{\lambda_1 \Delta t} + \cdots + c_d e^{\lambda_1 (d-1)\Delta t} = 0\,, \\ \cdots \\ c_1 + c_2 e^{\lambda_d \Delta t} + \cdots + c_d e^{\lambda_d (d-1)\Delta t} = 0\,. \end{cases} \tag{A.4}$$

Equation (A.4) can be written in matrix form as:

$$\begin{bmatrix} 1 & e^{\lambda_1 \Delta t} & e^{2\lambda_1 \Delta t} & \ldots & e^{(d-1)\lambda_1 \Delta t} \\ 1 & e^{\lambda_2 \Delta t} & e^{2\lambda_2 \Delta t} & \ldots & e^{(d-1)\lambda_2 \Delta t} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & e^{\lambda_d \Delta t} & e^{2\lambda_d \Delta t} & \ldots & e^{(d-1)\lambda_d \Delta t} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_d \end{bmatrix} = \mathbf{0}_d\,,$$

where, by condition A2, the first matrix on the left-hand side is again an invertible square Vandermonde matrix. This implies that $\boldsymbol{c} = [c_1, c_2, \ldots, c_d]^\top = \mathbf{0}_d$, which contradicts the assumption ($\boldsymbol{c} \neq \mathbf{0}_d$) made at the beginning of the proof. Therefore, one concludes that $\boldsymbol{X}_1$ is invertible, and thus

$$\Phi(\Delta t) = \boldsymbol{X}_2 \boldsymbol{X}_1^{-1}\,.$$

Step ii: We prove that under conditions A1 and A2, there always exists a unique logarithm of matrix $\Phi(\Delta t)(= e^{A\Delta t})$, thus, A is identifiable from $\Phi(\Delta t)$.

We first present a lemma we will use for our proof.

Lemma A.1. [28, Theorem 6.3] *Let C be a real square matrix. Then there exists a unique real solution Y to the equation $C = e^Y$ if and only if all the eigenvalues of C are positive real and no Jordan block of C belonging to any eigenvalue appears more than once.*

The Jordan decomposition of $e^{A\Delta t}$ is

$$e^{A\Delta t} = Q e^{\Lambda \Delta t} Q^{-1} = Q \begin{bmatrix} e^{\lambda_1 \Delta t} & & \\ & \ddots & \\ & & e^{\lambda_d \Delta t} \end{bmatrix} Q^{-1}\,.$$

Under condition A2, $\lambda_1, \lambda_2, \ldots, \lambda_d$ are d distinct real values, therefore, $e^{A\Delta t}$ has d distinct positive real eigenvalues $e^{\lambda_j \Delta t}$, for all $j = 1, 2, \ldots, d$. Then by Lemma A.1, one obtains $A\Delta t$ by taking logarithm of $e^{A\Delta_t}$, thus one obtains A.

Step iii: We show the initial condition $\boldsymbol{x}_0$ is identifiable.

One sees that

$$\begin{aligned}\det(e^{At}) &= \det(Qe^{\Lambda t}Q^{-1}) = \det(Q^{-1})\det(e^{\Lambda t})\det(Q) \\ &= \det(Q^{-1}Q)\det(e^{\Lambda t}) = \det(e^{\Lambda t}) = e^{\sum_{i=1}^{d}\lambda_i t} \neq 0\,,\end{aligned}$$

for any $t > 0$. Therefore, e^{At} is nonsingular for any $t > 0$. Since $\boldsymbol{x}_1 = e^{At_1}\boldsymbol{x}_0$, one obtains that $\boldsymbol{x}_0 = e^{-At_1}\boldsymbol{x}_1$.

Therefore, we have proved that both A and $\boldsymbol{x}_0$ can be explicitly calculated by using any $d+1$ equally-spaced error-free observations, which means the ODE system (3.1) is identifiable from these observations. □

A.1.3 Proof of Corollary 3.7

Proof. By definition of the aggregated observations in Definition 3.5, one sees that

$$\begin{aligned}\tilde{\boldsymbol{x}}_j &= (\boldsymbol{x}_{(j-1)k+1} + \boldsymbol{x}_{(j-1)k+2} + \cdots + \boldsymbol{x}_{jk})/k \\ &= (e^{At_{(j-1)k+1}}\boldsymbol{x}_0 + e^{At_{(j-1)k+2}}\boldsymbol{x}_0 + \cdots + e^{At_{jk}}\boldsymbol{x}_0)/k \\ &= e^{At_{(j-1)k+1}}(I + e^{A\Delta_t} + \cdots + e^{A(k-1)\Delta_t})\boldsymbol{x}_0/k\,,\end{aligned} \tag{A.5}$$

where $\Delta t = t_{j+1} - t_j$, for all $j = 1, \ldots, n-1$, denotes the equal time space of the original observations.

If we set the time of the aggregated observation $\tilde{\boldsymbol{x}}_j$ as the time of the first original observation (that is, $\boldsymbol{x}_{(j-1)k+1}$) of these k consecutive, non-overlapping observations. Then according to Equation (A.5), one sees that the solution function of the aggregated observations follows the structure of

$$\tilde{\boldsymbol{x}}(t; \tilde{\boldsymbol{x}}_0, \tilde{A}) = e^{\tilde{A}t}\tilde{\boldsymbol{x}}_0\,, \tag{A.6}$$

with

$$\tilde{\boldsymbol{x}}_0 = (I + e^{A\Delta_t} + \cdots + e^{A(k-1)\Delta_t})\boldsymbol{x}_0/k \text{ and } \tilde{A} = A\,.$$

This implies that the aggregated observations follow a new ODE system as (3.1) but with a different system parameter $(\tilde{\boldsymbol{x}}_0, \tilde{A})$. To see whether $(\tilde{\boldsymbol{x}}_0, \tilde{A})$ are the true parameters corresponding to the new ODE system, one has to prove the identifiability of the new ODE system from the aggregated observations.

According to Theorem 3.4, one needs to prove $\{\tilde{\boldsymbol{x}}_0, \tilde{A}\tilde{\boldsymbol{x}}_0, \ldots, \tilde{A}^{d-1}\tilde{\boldsymbol{x}}_0\}$ are linearly independent first, that is, proving $\{\tilde{\boldsymbol{x}}_0, A\tilde{\boldsymbol{x}}_0, \ldots, A^{d-1}\tilde{\boldsymbol{x}}_0\}$ are linearly independent. Let $\boldsymbol{l} = [l_1, l_2, \ldots, l_d]^\top$ be such that

$$l_1\tilde{\boldsymbol{x}}_0 + l_2A\tilde{\boldsymbol{x}}_0 + \cdots + l_dA^{d-1}\tilde{\boldsymbol{x}}_0 = \mathbf{0}_d\,, \tag{A.7}$$

we want to show that $\boldsymbol{l} = \mathbf{0}_d$. If one sets

$$B := (I + e^{A\Delta_t} + \cdots + e^{A(k-1)\Delta_t})/k\,,$$

then one sees $\tilde{\boldsymbol{x}}_0 = B\boldsymbol{x}_0$. Taking Jordan decomposition of A in B, one obtains that

$$B = Q(I + e^{\Lambda\Delta_t} + \cdots + e^{\Lambda(k-1)\Delta_t})Q^{-1}/k\,.$$

If one sets $s_j = 1 + e^{\lambda_j\Delta t} + \cdots + e^{\lambda_j(k-1)\Delta t}$, for all $j = 1, \ldots, d$, and denotes S as the diagonal matrix with the jth element being s_j, one obtains that

$$B = QSQ^{-1}/k\,. \tag{A.8}$$

Substituting $\tilde{\boldsymbol{x}}_0$ in Equation (A.7) with $B\boldsymbol{x}_0$, one obtains that

$$(l_1I + l_2A + \cdots + l_dA^{d-1})B\boldsymbol{x}_0 = \mathbf{0}_d\,.$$

By further taking Jordan decomposition of A, one obtains that

$$Q(l_1I + l_2\Lambda + \cdots + l_d\Lambda^{d-1})Q^{-1}B\boldsymbol{x}_0 = \mathbf{0}_d\,.$$

By plugging B expressed in Equation (A.8) into the previous equation, and multiply kQ^{-1} in both-hand sides of the equation, one obtains that

$$(l_1I + l_2\Lambda + \cdots + l_d\Lambda^{d-1})SQ^{-1}\boldsymbol{x}_0 = \mathbf{0}_d\,,$$

that is,

$$\begin{bmatrix} s_1(l_1 + l_2\lambda_1 + \cdots + l_d\lambda_1^{(d-1)}) & & \\ & \ddots & \\ & & s_d(l_1 + l_2\lambda_d + \cdots + l_d\lambda_d^{(d-1)}) \end{bmatrix} Q^{-1}\boldsymbol{x}_0 = \mathbf{0}_d. \quad \text{(A.9)}$$

According to the proof process of Theorem 3.4 in Appendix A.1.2, one obtains that

$$\begin{cases} s_1(l_1 + l_2\lambda_1 + \cdots + l_d\lambda_1^{d-1}) = 0\,, \\ \cdots \\ s_d(l_1 + l_2\lambda_d + \cdots + l_d\lambda_d^{d-1}) = 0\,. \end{cases} \tag{A.10}$$

Since $s_j > 0$ for all $j = 1, \ldots, d$, the system of equations (A.10) is equivalent to

$$\begin{cases} l_1 + l_2\lambda_1 + \cdots + l_d\lambda_1^{d-1} = 0\,, \\ \cdots \\ l_1 + l_2\lambda_d + \cdots + l_d\lambda_d^{d-1} = 0\,. \end{cases} \tag{A.11}$$

Now Equation (A.11) can be written in matrix form as:

$$\begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{d-1} \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{d-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_d & \lambda_d^2 & \cdots & \lambda_d^{d-1} \end{bmatrix} \begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_d \end{bmatrix} = \mathbf{0}_d\,,$$

where, by condition A2, the first matrix on the left-hand side is an invertible square Vandermonde matrix. This implies that $\boldsymbol{l} = \mathbf{0}_d$. Therefore, $\{\tilde{\boldsymbol{x}}_0, A\tilde{\boldsymbol{x}}_0, \ldots, A^{d-1}\tilde{\boldsymbol{x}}_0\}$ are linearly independent.

Obviously, the aggregated observations are equally-spaced, since

$$\tilde{t}_{j+1} - \tilde{t}_j = t_{1+jk} - t_{1+(j-1)k} = k\Delta t\,, \tag{A.12}$$

for all $j = 1, \ldots, \tilde{n} - 1$. Under condition A2, $\tilde{A}$(that is, A) has d distinct real eigenvalues, and under condition A3, $\tilde{n} > d$. Therefore, by Theorem 3.4, one concludes that the new ODE system is identifiable from aggregated observations with $\tilde{\boldsymbol{x}}_0 = (I + e^{A\Delta_t} + \cdots +$

$e^{A(k-1)\Delta_t})\boldsymbol{x}_0/k$ and $\tilde{A} = A$. Since

$$I + e^{A\Delta_t} + \cdots + e^{A(k-1)\Delta_t} = Q(I + e^{\Lambda\Delta_t} + \cdots + e^{\Lambda(k-1)\Delta_t})Q^{-1}$$

is invertible, one obtains that

$$\boldsymbol{x}_0 = k(I + e^{A\Delta_t} + \cdots + e^{A(k-1)\Delta_t})^{-1}\tilde{\boldsymbol{x}}_0\,,$$

and obviously,

$$A = \tilde{A}\,.$$

Therefore, the original ODE system with initial condition $\boldsymbol{x}_0$ and parameter matrix A is fully identifiable from the aggregated observations. □

A.1.4 Proof of Corollary 3.8

Proof. By Definition of the time-scaled observations in Definition 3.6, one sees that

$$\tilde{\boldsymbol{x}}_j = \boldsymbol{x}_j = \boldsymbol{x}(t_j; \boldsymbol{x}_0, A) = e^{At_j}\boldsymbol{x}_0 = e^{(A/k)\cdot kt_j}\boldsymbol{x}_0\,. \tag{A.13}$$

By definition the time of the time-scaled observation $\tilde{\boldsymbol{x}}_j$ is kt_j. Therefore, according to Equation A.13, one sees that the solution function of the time-scaled observations follows the structure of

$$\tilde{\boldsymbol{x}}(t; \tilde{\boldsymbol{x}}_0, \tilde{A}) = e^{\tilde{A}t}\tilde{\boldsymbol{x}}_0\,,$$

with $\tilde{\boldsymbol{x}}_0 = \boldsymbol{x}_0$, and $\tilde{A} = A/k$. This implies that the time-scaled observations follow a new ODE system as (3.1) but with a different system parameter $(\tilde{\boldsymbol{x}}_0, \tilde{A})$. To see whether $(\tilde{\boldsymbol{x}}_0, \tilde{A})$ are the true parameters corresponding to the new ODE system, one has to prove the identifiability of the new ODE system from the time-scaled observations.

Obviously, $\{\tilde{\boldsymbol{x}}_0, \tilde{A}\tilde{\boldsymbol{x}}_0, \ldots, \tilde{A}^{d-1}\tilde{\boldsymbol{x}}_0\} = \{\boldsymbol{x}_0, A\boldsymbol{x}_0/k, \ldots, A^{d-1}\boldsymbol{x}_0/k^{d-1}\}$ are linearly independent, since $\{\boldsymbol{x}_0, A\boldsymbol{x}_0, \ldots, A^{d-1}\boldsymbol{x}_0\}$ are linearly independent under condition A1. Moreover, $\tilde{A}$ has d distinct real eigenvalues since A has d distinct real eigenvalues under condition A2.

Based on the generation rules of the time-scaled observations, one sees that the new time-scaled observations are also equally-spaced. And under condition A3, $\tilde{n} = n > d$.

Then by Theorem 3.4, one concludes that the new ODE system is identifiable from the time-scaled observations, with $\tilde{\boldsymbol{x}}_0 = \boldsymbol{x}_0$ and $\tilde{A} = A/k$.

Since $k \neq 0$, simple calculation shows that

$$\boldsymbol{x}_0 = \tilde{\boldsymbol{x}}_0,\ A = k\tilde{A}\,.$$

Therefore, the original ODE system with initial condition $\boldsymbol{x}_0$ and parameter matrix A is fully identifiable from the time-scaled observations. □

A.1.5 Proof of Theorem 3.9

We first present two Lemmas we will use for our proof.

Lemma A.2. [82, Lemma 2.9] *Suppose M_1, M_n: $\mathbb{R}^{d_p} \to \mathbb{R}^{d_p'}$ for some finite integer $d_p, d_p' \geq 1$. If Θ is compact, $M_1(\boldsymbol{\theta})$ is continuous,*

$$M_n(\boldsymbol{\theta}) \xrightarrow{p} M_1(\boldsymbol{\theta})\,,\ \text{as } n \to \infty \tag{A.14}$$

for all $\boldsymbol{\theta} \in \Theta$, and there exist $\alpha > 0$ and $B_n = O_p(1)$ such that for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$,

$$\| M_n(\boldsymbol{\theta}_1) - M_n(\boldsymbol{\theta}_2) \|_2 \leq B_n \| \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \|_2^{\alpha} \tag{A.15}$$

almost surely, then

$$\sup_{\boldsymbol{\theta}\in\Theta} \| M_n(\boldsymbol{\theta}) - M_1(\boldsymbol{\theta}) \|_2 \xrightarrow{p} 0\,,\ \text{as } n \to \infty\,.$$

Lemma A.3. [156, Theorem 5.7] *Let M_n be random functions and let M_1 be a fixed function of $\boldsymbol{\theta}$ such that for every $\epsilon > 0$*

$$\sup_{\boldsymbol{\theta}\in\Theta} \| M_n(\boldsymbol{\theta}) - M_1(\boldsymbol{\theta}) \|_2 \xrightarrow{p} 0\,, \tag{A.16}$$

$$\sup_{\boldsymbol{\theta}:d(\boldsymbol{\theta},\boldsymbol{\theta}^*)\geq\varepsilon} M_1(\boldsymbol{\theta}) > M_1(\boldsymbol{\theta}^*)\,. \tag{A.17}$$

Then any sequence of estimators $\hat{\boldsymbol{\theta}}_n$ with

$$M_n(\hat{\boldsymbol{\theta}}_n) \leq M_n(\boldsymbol{\theta}^*) + o_p(1) \tag{A.18}$$

converges in probability to $\boldsymbol{\theta}^*$. $d(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ *denotes the distance between* $\boldsymbol{\theta}$ *and* $\boldsymbol{\theta}^*$.

Proof. Recall that we have defined $M(\boldsymbol{\theta})$ and $M_n(\boldsymbol{\theta})$ in Equations (3.5) and (3.6), respectively. Here, we set

$$M_1(\boldsymbol{\theta}) := M(\boldsymbol{\theta}) + E[\| \boldsymbol{\epsilon} \|_2^2] = \frac{1}{T}\int_0^T \| e^{A^* t}\boldsymbol{x}_0^* - e^{At}\boldsymbol{x}_0 \|_2^2 \, dt + E[\| \boldsymbol{\epsilon} \|_2^2]\,. \tag{A.19}$$

In order to prove the NLS estimator $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}^*$, as $n \to \infty$, by Lemma A.3, we need to prove that all three conditions (A.16), (A.17) and (A.18) are satisfied w.r.t $M_n(\boldsymbol{\theta})$ and $M_1(\boldsymbol{\theta})$. Therefore, the proof can be broken down into three steps based on the proofs of each of these three conditions. In the following, we will show that all these three conditions are satisfied.

Step i: We prove that condition (A.16) in Lemma A.3 is satisfied based on Lemma A.2.

According to Lemma A.2, to prove the uniform convergence of $M_n(\boldsymbol{\theta})$ to $M_1(\boldsymbol{\theta})$ in parameter space Θ, that is, condition (A.16), we first need to prove the point-wise convergence of $M_n(\boldsymbol{\theta})$ to $M_1(\boldsymbol{\theta})$, that is, condition (A.14).

According to Equation (3.6), one sees that

$$\begin{aligned} M_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^n \| \boldsymbol{y}_i - e^{At_i}\boldsymbol{x}_0 \|_2^2 \\ &= \frac{1}{n}\sum_{i=1}^n \| e^{A^* t_i}\boldsymbol{x}_0^* - e^{At_i}\boldsymbol{x}_0 \|_2^2 + \frac{1}{n}\sum_{i=1}^n \| \boldsymbol{\epsilon}_i \|_2^2 + \frac{2}{n}\sum_{i=1}^n \boldsymbol{\epsilon}_i^\top (e^{A^* t_i}\boldsymbol{x}_0^* - e^{At_i}\boldsymbol{x}_0)\,. \end{aligned} \tag{A.20}$$

Let n tend to infinity, one obtains that

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \| e^{A^* t_i}\boldsymbol{x}_0^* - e^{At_i}\boldsymbol{x}_0 \|_2^2 \\ =& \lim_{n\to\infty} \frac{n-1}{nT}\sum_{i=1}^n \frac{T}{n-1} \| e^{A^* t_i}\boldsymbol{x}_0^* - e^{At_i}\boldsymbol{x}_0 \|_2^2 = \frac{1}{T}\int_0^T \| e^{A^* t}\boldsymbol{x}_0^* - e^{At}\boldsymbol{x}_0 \|_2^2 \, dt\,. \end{aligned} \tag{A.21}$$

By weak law of large numbers, one sees that

$$\frac{1}{n}\sum_{i=1}^{n} \| \boldsymbol{\epsilon}_i \|_2^2 \xrightarrow{p} E[\| \boldsymbol{\epsilon} \|_2^2]\,, \text{ as } n \to \infty\,, \tag{A.22}$$

where we set $E[\| \boldsymbol{\epsilon} \|_2^2] := E[\| \boldsymbol{\epsilon}_i \|_2^2] = \sum_{j=1}^{d} \sigma_j^2$, for all $i = 1, \ldots, n$. Since

$$E\left[\frac{2}{n}\sum_{i=1}^{n} \boldsymbol{\epsilon}_i^\top (e^{A^* t_i}\boldsymbol{x}_0^* - e^{At_i}\boldsymbol{x}_0)\right] = 0\,,$$

then by Chebyshev's inequality, for any $\varepsilon > 0$, one has

$$P\left(\left|\frac{2}{n}\sum_{i=1}^{n} \boldsymbol{\epsilon}_i^\top (e^{A^* t_i}\boldsymbol{x}_0^* - e^{At_i}\boldsymbol{x}_0) - 0\right| \geq \varepsilon\right) \leq var\left(\frac{2}{n}\sum_{i=1}^{n} \boldsymbol{\epsilon}_i^\top (e^{A^* t_i}\boldsymbol{x}_0^* - e^{At_i}\boldsymbol{x}_0)\right)/\varepsilon^2\,. \tag{A.23}$$

$$\begin{aligned} var\left(\frac{2}{n}\sum_{i=1}^{n} \boldsymbol{\epsilon}_i^\top (e^{A^* t_i}\boldsymbol{x}_0^* - e^{At_i}\boldsymbol{x}_0)\right) &= \frac{4}{n^2}\sum_{i=1}^{n} var\big(\boldsymbol{\epsilon}_i^\top (e^{A^* t_i}\boldsymbol{x}_0^* - e^{At_i}\boldsymbol{x}_0)\big) \\ = \frac{4}{n^2}\sum_{i=1}^{n} E[\{\boldsymbol{\epsilon}_i^\top (e^{A^* t_i}\boldsymbol{x}_0^* - e^{At_i}\boldsymbol{x}_0)\}^2] &\leq \frac{4}{n^2}\sum_{i=1}^{n} E[\| \boldsymbol{\epsilon}_i \|_2^2 \| e^{A^* t_i}\boldsymbol{x}_0^* - e^{At_i}\boldsymbol{x}_0 \|_2^2]\,, \end{aligned} \tag{A.24}$$

by Cauchy-Schwarz inequality. Specifically,

$$\begin{aligned} \| e^{A^* t_i}\boldsymbol{x}_0^* - e^{At_i}\boldsymbol{x}_0 \|_2 &= \| e^{A^* t_i}\boldsymbol{x}_0^* - e^{A^* t_i}\boldsymbol{x}_0 + e^{A^* t_i}\boldsymbol{x}_0 - e^{At_i}\boldsymbol{x}_0 \|_2 \\ &= \| e^{A^* t_i}(\boldsymbol{x}_0^* - \boldsymbol{x}_0) + (e^{A^* t_i} - e^{At_i})\boldsymbol{x}_0 \|_2 \\ &\leq \| e^{A^* t_i}(\boldsymbol{x}_0^* - \boldsymbol{x}_0) \|_2 + \| (e^{A^* t_i} - e^{At_i})\boldsymbol{x}_0 \|_2 \\ &\leq \| e^{A^* t_i} \|_2 \| \boldsymbol{x}_0^* - \boldsymbol{x}_0 \|_2 + \| e^{A^* t_i} - e^{At_i} \|_2 \| \boldsymbol{x}_0 \|_2 \\ &\leq \underbrace{\| e^{A^* t_i} \|_F}_{\text{item 1}} \underbrace{\| \boldsymbol{x}_0^* - \boldsymbol{x}_0 \|_2}_{\text{item 2}} + \underbrace{\| e^{A^* t_i} - e^{At_i} \|_F}_{\text{item 3}} \underbrace{\| \boldsymbol{x}_0 \|_2}_{\text{item 4}}\,, \end{aligned} \tag{A.25}$$

where $\| M \|_2$ denotes the subordinate matrix norm induced by the norm $\| \cdot \|_2$ of a matrix $M \in \mathbb{R}^{m\times n}$, with

$$\| M \|_2 = \sup_{\boldsymbol{v}\in\mathbb{R}^n, \boldsymbol{v}\neq\boldsymbol{0}} \frac{\| M\boldsymbol{v} \|_2}{\| \boldsymbol{v} \|_2}\,,$$

and $\| M \|_F$ is the Frobenius norm of matrix M, with

$$\| M \|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |M_{ij}|^2}\,,$$

where M_{ij} is the ij-th entry of matrix M.

Under assumption A4, parameter space Θ is a compact subset of $\mathbb{R}^{d+d^2}$, therefore, Θ can be enclosed with a $(d+d^2)$-dimensional box. We denote it as $\Theta \subset [-l, l]^{d+d^2}$ with $0 < l < \infty$. Then we will analyse each of the four items in Equation (A.25).

$$\begin{aligned}\text{item 1} &= \parallel e^{A^* t_i} \parallel_F = \parallel e^{A^* t_i} - e^{I_d t_i} + e^{I_d t_i} \parallel_F \leq \parallel e^{A^* t_i} - e^{I_d t_i} \parallel_F + \parallel e^{I_d t_i} \parallel_F \\ &\leq \parallel A^* t_i - I_d t_i \parallel_F e^{\|I_d t_i\|_F} e^{\|A^* t_i - I_d t_i\|_F} + \parallel e^{I_d t_i} \parallel_F ,\end{aligned}$$

where I_d denotes the d-dimensional identity matrix. Simple calculation shows that

$$\parallel A^* - I_d \parallel_F \ \leq \sqrt{(l+1)^2 \times d + l^2 \times (d^2 - d)} = \sqrt{l^2 d^2 + 2ld + d}\,,$$

$$\parallel I_d \parallel_F = \sqrt{d}\,,$$

and

$$\parallel e^{I_d t_i} \parallel_F = \sqrt{e^{2t_i} \times d} = e^{t_i}\sqrt{d}\,.$$

Since $t_i \leq T$ for all $i = 1, \ldots, n$ and $0 < T < \infty$ by condition A6, therefore, one obtains that

$$\text{item 1} = \parallel e^{A^* t_i} \parallel_F \leq C'_d\,, \tag{A.26}$$

where $0 < C'_d < \infty$ is a constant only depending on d. Similarly,

$$\text{item 2} = \parallel \boldsymbol{x}_0^* - \boldsymbol{x}_0 \parallel_2 \leq \sqrt{(2l)^2 \times d} = 2l\sqrt{d}\,, \tag{A.27}$$

and

$$\begin{aligned}\text{item 3} &= \parallel e^{A^* t_i} - e^{A t_i} \parallel_F \leq \parallel A^* t_i - A t_i \parallel_F e^{\|A t_i\|_F} e^{\|A^* t_i - A t_i\|_F} \\ &= \parallel A^* - A \parallel_F t_i e^{\|A\|_F t_i} e^{\|A^* - A\|_F t_i}\,,\end{aligned}$$

where

$$\parallel A \parallel_F \leq \sqrt{l^2 \times d^2} = ld\,,$$

$$\parallel A^* - A \parallel_F \leq \sqrt{(2l)^2 \times d^2} = 2ld\,,$$

therefore, by simple calculation one obtains that

$$\text{item 3} = \parallel e^{A^* t_i} - e^{A t_i} \parallel_F \leq C''_d\,, \tag{A.28}$$

with $0 < C_d'' < \infty$, is a constant only depends on d. One sees that

$$\text{item 4} = \| \boldsymbol{x}_0 \|_2 \leq \sqrt{l^2 \times d} = l\sqrt{d}\,. \tag{A.29}$$

Combining (A.26), (A.27), (A.28), (A.29) and (A.25), one sees that

$$\| e^{A^* t_i} \boldsymbol{x}_0^* - e^{A t_i} \boldsymbol{x}_0 \|_2 \leq C_d\,,$$

where $0 < C_d < \infty$ is a constant only depends on d. Then (A.24) can be expressed as

$$var\left(\frac{2}{n}\sum_{i=1}^{n} \boldsymbol{\epsilon}_i^\top (e^{A^* t_i} \boldsymbol{x}_0^* - e^{A t_i} \boldsymbol{x}_0)\right) \leq \frac{4}{n^2}\sum_{i=1}^{n} E[\| \boldsymbol{\epsilon}_i \|_2^2 \| e^{A^* t_i} \boldsymbol{x}_0^* - e^{A t_i} \boldsymbol{x}_0 \|_2^2]$$
$$\leq 4\, C_d^2 E[\| \boldsymbol{\epsilon} \|_2^2]/n = 4\, C_d^2 \sum_{j=1}^{d} \sigma_j^2 / n\,.$$

Therefore, (A.23) can be expressed as

$$P\left(\left|\frac{2}{n}\sum_{i=1}^{n} \boldsymbol{\epsilon}_i^\top (e^{A^* t_i} \boldsymbol{x}_0^* - e^{A t_i} \boldsymbol{x}_0) - 0\right| \geq \varepsilon\right) \leq \frac{4\, C_d^2}{n\varepsilon^2}\sum_{j=1}^{d} \sigma_j^2 \to 0\,, \text{ as } n \to \infty.$$

That is

$$\frac{2}{n}\sum_{i=1}^{n} \boldsymbol{\epsilon}_i^\top (e^{A^* t_i} \boldsymbol{x}_0^* - e^{A t_i} \boldsymbol{x}_0) \xrightarrow{p} 0\,, \text{ as } n \to \infty\,. \tag{A.30}$$

Combining (A.21), (A.22), (A.30) and (A.20), one obtains that

$$M_n(\boldsymbol{\theta}) \xrightarrow{p} \frac{1}{T}\int_0^T \| e^{A^* t} \boldsymbol{x}_0^* - e^{A t} \boldsymbol{x}_0 \|_2^2 \, dt + E[\| \boldsymbol{\epsilon} \|_2^2]\,, \text{ as } n \to \infty\,. \tag{A.31}$$

By the definition of $M_1(\boldsymbol{\theta})$ in Equation (A.19), the right-hand side of (A.31) is $M_1(\theta)$, that is,

$$M_n(\boldsymbol{\theta}) \xrightarrow{p} M_1(\boldsymbol{\theta})\,, \text{ as } n \to \infty\,.$$

Now that we have proved that condition (A.14) in Lemma A.2 is satisfied. Next, we will find a $\alpha > 0$ and a $B_n = O_p(1)$ such that for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, condition (A.15) is met.

For any $\boldsymbol{\theta}_1 = (\boldsymbol{x}_{01}, A_1), \boldsymbol{\theta}_2 = (\boldsymbol{x}_{02}, A_2) \in \Theta$,

$$
\begin{aligned}
&\| M_n(\boldsymbol{\theta}_1) - M_n(\boldsymbol{\theta}_2) \|_2 \\
=&\left| \frac{1}{n}\sum_{i=1}^{n} \| \boldsymbol{y}_i - e^{A_1 t_i}\boldsymbol{x}_{01} \|_2^2 - \frac{1}{n}\sum_{i=1}^{n} \| \boldsymbol{y}_i - e^{A_2 t_i}\boldsymbol{x}_{02} \|_2^2 \right| \\
=&\left| \frac{1}{n}\sum_{i=1}^{n} \{ \| e^{A_1 t_i}\boldsymbol{x}_{01} \|_2^2 - \| e^{A_2 t_i}\boldsymbol{x}_{02} \|_2^2 - 2(e^{A^* t_i}\boldsymbol{x}_0^* + \boldsymbol{\epsilon}_i)^\top (e^{A_1 t_i}\boldsymbol{x}_{01} - e^{A_2 t_i}\boldsymbol{x}_{02}) \} \right| \\
\leq& \frac{1}{n}\sum_{i=1}^{n} \Big| \| e^{A_1 t_i}\boldsymbol{x}_{01} \|_2^2 - \| e^{A_2 t_i}\boldsymbol{x}_{02} \|_2^2 \\
&- 2(e^{A^* t_i}\boldsymbol{x}_0^*)^\top (e^{A_1 t_i}\boldsymbol{x}_{01} - e^{A_2 t_i}\boldsymbol{x}_{02}) - 2\boldsymbol{\epsilon}_i^\top (e^{A_1 t_i}\boldsymbol{x}_{01} - e^{A_2 t_i}\boldsymbol{x}_{02}) \Big| \\
\leq& \frac{1}{n}\sum_{i=1}^{n} \Big\{ \Big| \| e^{A_1 t_i}\boldsymbol{x}_{01} \|_2^2 - \| e^{A_2 t_i}\boldsymbol{x}_{02} \|_2^2 \Big| \\
&+ \left| 2(e^{A^* t_i}\boldsymbol{x}_0^*)^\top (e^{A_1 t_i}\boldsymbol{x}_{01} - e^{A_2 t_i}\boldsymbol{x}_{02}) \right| + \left| 2\boldsymbol{\epsilon}_i^\top (e^{A_1 t_i}\boldsymbol{x}_{01} - e^{A_2 t_i}\boldsymbol{x}_{02}) \right| \Big\} \\
\leq& \frac{1}{n}\sum_{i=1}^{n} \{ \| e^{A_1 t_i}\boldsymbol{x}_{01} - e^{A_2 t_i}\boldsymbol{x}_{02} \|_2 (\| e^{A_1 t_i}\boldsymbol{x}_{01} \|_2 + \| e^{A_2 t_i}\boldsymbol{x}_{02} \|_2) \\
&+ 2 \| e^{A^* t_i}\boldsymbol{x}_0^* \|_2 \cdot \| e^{A_1 t_i}\boldsymbol{x}_{01} - e^{A_2 t_i}\boldsymbol{x}_{02} \|_2 + 2 \| \boldsymbol{\epsilon}_i \|_2 \cdot \| e^{A_1 t_i}\boldsymbol{x}_{01} - e^{A_2 t_i}\boldsymbol{x}_{02} \|_2 \} \\
=& \frac{1}{n}\sum_{i=1}^{n} \| e^{A_1 t_i}\boldsymbol{x}_{01} - e^{A_2 t_i}\boldsymbol{x}_{02} \|_2 (\| e^{A_1 t_i}\boldsymbol{x}_{01} \|_2 + \| e^{A_2 t_i}\boldsymbol{x}_{02} \|_2 + 2 \| e^{A^* t_i}\boldsymbol{x}_0^* \|_2 + 2 \| \boldsymbol{\epsilon}_i \|_2).
\end{aligned}
\tag{A.32}
$$

Similar to the process of analysing $\| e^{A^* t_i}\boldsymbol{x}_0^* - e^{A t_i}\boldsymbol{x}_0 \|_2$ in Equation (A.25), by some calculation one obtains that

$$
\begin{aligned}
&\| e^{A_1 t_i}\boldsymbol{x}_{01} - e^{A_2 t_i}\boldsymbol{x}_{02} \|_2 \leq \| e^{A_1 t_i} \|_F \| \boldsymbol{x}_{01} - \boldsymbol{x}_{02} \|_2 + \| e^{A_1 t_i} - e^{A_2 t_i} \|_F \| \boldsymbol{x}_{02} \|_2 \\
\leq& C_d' \| \boldsymbol{x}_{01} - \boldsymbol{x}_{02} \|_2 + C_d'''' \| A_1 - A_2 \|_F \leq \max(C_d', C_d'''')(\| \boldsymbol{x}_{01} - \boldsymbol{x}_{02} \|_2 + \| A_1 - A_2 \|_F),
\end{aligned}
$$

where $0 < C_d', C_d'''' < \infty$ are constants only depending on d. Since

$$
(\| \boldsymbol{x}_{01} - \boldsymbol{x}_{02} \|_2 + \| A_1 - A_2 \|_F)^2 \leq 2(\| \boldsymbol{x}_{01} - \boldsymbol{x}_{02} \|_2^2 + \| A_1 - A_2 \|_F^2) = 2 \| \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \|_2^2,
$$

one obtains that

$$
\| \boldsymbol{x}_{01} - \boldsymbol{x}_{02} \|_2 + \| A_1 - A_2 \|_F \leq \sqrt{2} \| \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \|_2 .
$$

Therefore,

$$\| e^{A_1 t_i}\boldsymbol{x}_{01} - e^{A_2 t_i}\boldsymbol{x}_{02} \|_2 \leq \sqrt{2}\max(C'_d, C'''_d) \| \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \|_2 \ . \tag{A.33}$$

Then we analyse the second item in (A.32), some simple calculation shows that

$$\| e^{A_1 t_i}\boldsymbol{x}_{01} \|_2 \leq \| e^{A_1 t_i} \|_2 \| \boldsymbol{x}_{01} \|_2 \leq \| e^{A_1 t_i} \|_F \| \boldsymbol{x}_{01} \|_2 \leq C'_d l\sqrt{d}\,. \tag{A.34}$$

Similarly, one sees that

$$\| e^{A_2 t_i}\boldsymbol{x}_{02} \|_2 \,, \| e^{A^* t_i}\boldsymbol{x}^*_0 \|_2 \leq C'_d l\sqrt{d}\,. \tag{A.35}$$

Combining (A.33), (A.34), (A.35) and (A.32), one obtains that

$$\| M_n(\boldsymbol{\theta}_1) - M_n(\boldsymbol{\theta}_2) \|_2 \leq \frac{1}{n}\sum_{i=1}^{n} \sqrt{2}\max(C'_d, C'''_d)(4C'_d ld + 2 \| \boldsymbol{\epsilon}_i \|_2) \| \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \|_2$$

$$\leq \left(\tilde{C}_d^2 + \frac{2\tilde{C}_d}{n}\sum_{i=1}^{n} \| \boldsymbol{\epsilon}_i \|_2 \right) \| \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \|_2 \,,$$

where $\tilde{C}_d = \max(\sqrt{2}\max(C'_d, C'''_d), 4C'_d ld)$ and $0 < \tilde{C}_d < \infty$ is a constant that only depends on d. If one sets

$$B_n := \tilde{C}_d^2 + \frac{2\tilde{C}_d}{n}\sum_{i=1}^{n} \| \boldsymbol{\epsilon}_i \|_2 \,,$$

one sees

$$\| M_n(\boldsymbol{\theta}_1) - M_n(\boldsymbol{\theta}_2) \|_2 \leq B_n \| \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2 \|_2 \,,$$

for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$. Let

$$F_n := \frac{1}{n}\sum_{i=1}^{n} \| \boldsymbol{\epsilon}_i \|_2 \,,$$

by Chebyshev's inequality, one sees that

$$F_n = O_p(E[F_n] + \sqrt{var(F_n)})\,,$$

where

$$E[F_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} \| \boldsymbol{\epsilon}_i \|_2 \right] = \frac{1}{n}\sum_{i=1}^{n} E[\| \boldsymbol{\epsilon}_i \|_2] = E[\| \boldsymbol{\epsilon} \|_2] = O(1)\,,$$

because $E[\| \boldsymbol{\epsilon} \|_2^2] = \sum_{i=1}^d \sigma_i^2 < \infty$ implies that $E[\| \boldsymbol{\epsilon} \|_2] < \infty$. And

$$\begin{aligned} var(F_n) =& var\left(\frac{1}{n}\sum_{i=1}^n \| \boldsymbol{\epsilon}_i \|_2\right) = \frac{1}{n^2}\sum_{i=1}^n var(\| \boldsymbol{\epsilon}_i \|_2) \\ =& \frac{1}{n^2}\sum_{i=1}^n \{E[\| \boldsymbol{\epsilon}_i \|_2^2] - (E[\| \boldsymbol{\epsilon}_i \|_2])^2\} = O(1)\,, \end{aligned}$$

therefore, one obtains $F_n = O_p(1)$, which implies $B_n = O_p(1)$. Thus, the condition (A.15) is satisfied with $\alpha = 1$. Since $M_1(\boldsymbol{\theta})$ is continuous w.r.t $\boldsymbol{\theta}$, and under assumption A4, parameter space Θ is compact, therefore, by Lemma A.2, one sees that

$$\sup_{\boldsymbol{\theta}\in\Theta} \| M_n(\boldsymbol{\theta}) - M_1(\boldsymbol{\theta}) \|_2 \xrightarrow{p} 0\,,\ \text{as } n \to \infty\,,$$

that is, the condition (A.16) in Lemma A.3 is satisfied.

Step ii: We show that condition (A.17) in Lemma A.3 is satisfied.

Recall that

$$M_1(\boldsymbol{\theta}) = \frac{1}{T}\int_0^T \| e^{A^*t}\boldsymbol{x}_0^* - e^{At}\boldsymbol{x}_0 \|_2^2\, dt + E[\| \boldsymbol{\epsilon} \|_2^2]\,,$$

when assumptions A1 and A2 are satisfied with respect to $\boldsymbol{\theta}^*$, according to Theorem 3.4, the ODE system (3.1) is identifiable at $\boldsymbol{\theta}^* = (\boldsymbol{x}_0^*, A^*)$ from any $d+1$ equally-spaced error-free observations, which implies the ODE system is also identifiable at $\boldsymbol{\theta}^*$ from the corresponding trajectory at $[0, T]$, which further implies that $M_1(\boldsymbol{\theta})$ attains its unique global minimum at $\boldsymbol{\theta}^*$. Therefore, one sees that the condition (A.17) in Lemma A.3 is satisfied.

Step iii: We show that condition (A.18) in Lemma A.3 is satisfied.

The definition of $\hat{\boldsymbol{\theta}}_n$, that is

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} M_n(\boldsymbol{\theta})\,,$$

implies that the condition (A.18) is satisfied.

Now that we have proved that all the three conditions in Lemma A.3 are satisfied, therefore, one concludes that

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}^*\,,\ \text{as } n \to \infty\,.$$

□

A.1.6 Proof of Corollary 3.10

Proof. The main task in this proof is to prove that the NLS parameter estimator $\hat{\tilde{\boldsymbol{\theta}}} := (\hat{\tilde{\boldsymbol{x}}}_0, \hat{\tilde{A}})$ from aggregated observations is consistent to the true system parameters corresponding to the new ODE system, that is, $\tilde{\boldsymbol{\theta}}^* := (\tilde{\boldsymbol{x}}_0^*, \tilde{A}^*)$. Once one has proved this result, one can reach the conclusion in Corollary 3.10 by taking the function $g(\cdot)$ with respect to $\hat{\tilde{\boldsymbol{\theta}}}$ and $\tilde{\boldsymbol{\theta}}^*$, respectively.

In order to prove

$$\hat{\tilde{\boldsymbol{\theta}}} \xrightarrow{p} \tilde{\boldsymbol{\theta}}^*, \text{ as } \tilde{n} \to \infty,$$

base on Theorem 3.9, one needs to prove that the assumptions A1, A2 and A4-A6 are satisfied with respect to the new ODE system, the new parameter $\tilde{\boldsymbol{\theta}}^*$ and the new error terms $\tilde{\boldsymbol{\epsilon}}_i$ corresponding to the aggregated observations $\tilde{\boldsymbol{Y}}$.

Under assumptions A1-A3, according to the proof of Corollary 3.7 in Appendix A.1.3, one sees that assumptions A1-A2 are satisfied with respect to $\tilde{\boldsymbol{\theta}}^*$, where

$$\tilde{\boldsymbol{\theta}}^* = (\tilde{\boldsymbol{x}}_0^*, \tilde{A}^*) = \big((I + e^{A^*\Delta_t} + \cdots + e^{A^*(k-1)\Delta_t})\boldsymbol{x}_0^*/k, A^*\big). \tag{A.36}$$

Since under assumption A4, Θ is compact, and according to the relationship between $\tilde{\boldsymbol{\theta}}^*$ and $\boldsymbol{\theta}^* = (\boldsymbol{x}_0^*, A^*)$ in Equation (A.36), one obtains that the new parameter space $\tilde{\Theta}$ is compact.

Based on the generation rules of the aggregated observations, assumption A5 is satisfied with the new error terms $\{\tilde{\boldsymbol{\epsilon}}_i\}$ being independent and identically distributed random vectors with mean zero and covariance matrix

$$\tilde{\Sigma} = \Sigma/k = diag(\frac{\sigma_1^2}{k}, \ldots, \frac{\sigma_d^2}{k}).$$

By aggregated observations generation rules, assumption A6 is satisfied with $\tilde{t}_i = t_{(i-1)k+1}$, $\Delta\tilde{t} = k\Delta t$ and

$$\tilde{T} = (\lfloor n/k \rfloor - 1)kT/(n-1),$$

where $\lfloor \cdot \rfloor$ stands for the floor function.

Then by Theorem 3.9, one concludes that $\hat{\tilde{\boldsymbol{\theta}}} \xrightarrow{p} \tilde{\boldsymbol{\theta}}^*$, as $\tilde{n} \to \infty$.

By definition of $\hat{\boldsymbol{\theta}}_{\tilde{n}}$ in Corollary 3.10, that is

$$\hat{\boldsymbol{\theta}}_{\tilde{n}} := g(\hat{\tilde{\boldsymbol{\theta}}}) := \left(k(I + e^{\hat{\tilde{A}}\Delta_t} + \cdots + e^{\hat{\tilde{A}}(k-1)\Delta_t})^{-1}\hat{\tilde{\boldsymbol{x}}}_0, \hat{\tilde{A}}\right),$$

one obtains that

$$g(\tilde{\boldsymbol{\theta}}^*) = \left(k(I + e^{\tilde{A}^*\Delta_t} + \cdots + e^{\tilde{A}^*(k-1)\Delta_t})^{-1}\tilde{\boldsymbol{x}}_0^*, \tilde{A}^*\right) = (\boldsymbol{x}_0^*, A^*) = \boldsymbol{\theta}^*.$$

By multivariate continuous mapping theorem, one concludes that

$$g(\hat{\tilde{\boldsymbol{\theta}}}) \xrightarrow{p} g(\tilde{\boldsymbol{\theta}}^*), \text{ as } \tilde{n} \to \infty,$$

that is

$$\hat{\boldsymbol{\theta}}_{\tilde{n}} \xrightarrow{p} \boldsymbol{\theta}^*, \text{ as } \tilde{n} \to \infty.$$

□

A.1.7 Proof of Corollary 3.11

Proof. Similar to the proof of Corollary 3.10 in Appendix A.1.6, one first needs to prove that the NLS parameter estimator $\hat{\tilde{\boldsymbol{\theta}}} := (\hat{\tilde{\boldsymbol{x}}}_0, \hat{\tilde{A}})$ from time-scaled observations is consistent to the true system parameters corresponding to the new ODE system, that is, $\tilde{\boldsymbol{\theta}}^* := (\tilde{\boldsymbol{x}}_0^*, \tilde{A}^*)$. Once one has proved this result, then the conclusion in Corollary 3.11 can be reached by taking the function $g(\cdot)$ with respect to $\hat{\tilde{\boldsymbol{\theta}}}$ and $\tilde{\boldsymbol{\theta}}^*$, respectively.

In order to prove

$$\hat{\tilde{\boldsymbol{\theta}}} \xrightarrow{p} \tilde{\boldsymbol{\theta}}^*, \text{ as } \tilde{n} \to \infty,$$

base on Theorem 3.9, one needs to prove that the assumptions A1, A2 and A4-A6 are satisfied with respect to the new ODE system, the new parameter $\tilde{\boldsymbol{\theta}}^*$ and the new error terms $\tilde{\boldsymbol{\epsilon}}_i$ corresponding to the time-scaled observations $\tilde{\boldsymbol{Y}}$.

Under assumptions A1-A3, according to the proof of Corollary 3.8 in Appendix A.1.4, one sees that assumptions A1-A2 are satisfied with respect to $\tilde{\boldsymbol{\theta}}^*$, where

$$\tilde{\boldsymbol{\theta}}^* = (\tilde{\boldsymbol{x}}_0^*, \tilde{A}^*) = (\boldsymbol{x}_0^*, A^*/k). \tag{A.37}$$

Since under assumption A4, Θ is compact, and according to the relationship between $\tilde{\boldsymbol{\theta}}^*$ and $\boldsymbol{\theta}^* = (\boldsymbol{x}_0^*, A^*)$ in Equation (A.37), one obtains that the new parameter space $\tilde{\Theta}$ is compact.

Based on the generation rules of the time-scaled observations, assumption A5 is satisfied with the new error terms $\{\tilde{\boldsymbol{\epsilon}}_i\}$ being the original error terms $\{\boldsymbol{\epsilon}_i\}$.

Assumption A6 is satisfied with

$$\tilde{t}_i = kt_i \,, \Delta\tilde{t} = k\Delta t \,, \tilde{T} = kT \,.$$

Then by Theorem 3.9, one concludes that $\hat{\tilde{\boldsymbol{\theta}}} \xrightarrow{p} \tilde{\boldsymbol{\theta}}^*$, as $\tilde{n} \to \infty$.

By definition of $\hat{\boldsymbol{\theta}}_{\tilde{n}}$ in Corollary 3.11, that is

$$\hat{\boldsymbol{\theta}}_{\tilde{n}} := g(\hat{\tilde{\boldsymbol{\theta}}}) := (\hat{\tilde{\boldsymbol{x}}}_0, k\hat{\tilde{A}}) \,,$$

one obtains that

$$g(\tilde{\boldsymbol{\theta}}^*) = (\tilde{\boldsymbol{x}}_0^*, k\tilde{A}^*) = (\boldsymbol{x}_0^*, A^*) = \boldsymbol{\theta}^* \,.$$

By multivariate continuous mapping theorem, one concludes that

$$g(\hat{\tilde{\boldsymbol{\theta}}}) \xrightarrow{p} g(\tilde{\boldsymbol{\theta}}^*) \,, \text{ as } \tilde{n} \to \infty \,,$$

that is

$$\hat{\boldsymbol{\theta}}_{\tilde{n}} \xrightarrow{p} \boldsymbol{\theta}^* \,, \text{ as } \tilde{n} \to \infty \,.$$

□

A.1.8 Proof of Theorem 3.12

Proof. Recall that we have defined $M_n(\boldsymbol{\theta})$ in Equation (3.6), one sees that $M_n(\boldsymbol{\theta})$ is twice differentiable at $\boldsymbol{\theta} \in \Theta$. Then by the mean value theorem, one obtains

$$\nabla_{\boldsymbol{\theta}} M_n(\hat{\boldsymbol{\theta}}_n) = \nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}^*) + \nabla^2_{\boldsymbol{\theta}} M_n(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \,, \tag{A.38}$$

where $\nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}^*)$ denotes the gradient of $M_n(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta}^*$, and $\nabla^2_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}^*)$ denotes the Hessian matrix of $M_n(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta}^*$, and $\tilde{\boldsymbol{\theta}}$ is in the line joining $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^*$.

Since assumptions A1 and A2 are satisfied with respect to $\boldsymbol{\theta}^*$ and assumptions A4-A6 hold, then by Theorem 3.9, one obtains that

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}^* \, , \text{ as } n \to \infty \, .$$

Assumption A7 implies that $\hat{\boldsymbol{\theta}}_n$ is an interior point of Θ as $n \to \infty$, and by definition,

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}) \, ,$$

therefore, one obtains that

$$\nabla_{\boldsymbol{\theta}} M_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0} \, .$$

Suppose that $\nabla^2_{\boldsymbol{\theta}} M_n(\tilde{\boldsymbol{\theta}})$ is nonsingular, by rearranging Equation (A.38), one obtains that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) = -\{\nabla^2_{\boldsymbol{\theta}} M_n(\tilde{\boldsymbol{\theta}})\}^{-1} \sqrt{n} \nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}^*) \, .$$

By definition of $\tilde{\boldsymbol{\theta}}$ and the convergence of $\hat{\boldsymbol{\theta}}_n$ to $\boldsymbol{\theta}^*$, one obtains that

$$\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^* \, , \text{ as } n \to \infty \, ,$$

and according to the proof of Theorem 3.9 in Appendix A.1.5,

$$\sup_{\boldsymbol{\theta} \in \Theta} \| \, M_n(\boldsymbol{\theta}) - M_1(\boldsymbol{\theta}) \, \|_2 \xrightarrow{p} 0 \, , \text{ as } n \to \infty \, ,$$

one obtains that

$$\nabla^2_{\boldsymbol{\theta}} M_n(\tilde{\boldsymbol{\theta}}) \xrightarrow{p} \nabla^2_{\boldsymbol{\theta}} M_1(\boldsymbol{\theta}^*) \, , \text{ as } n \to \infty \, ,$$

By the relationship between $M_1(\boldsymbol{\theta})$ and $M(\boldsymbol{\theta})$ defined in Equation (A.19), one sees that

$$\nabla^2_{\boldsymbol{\theta}} M_1(\boldsymbol{\theta}^*) = \nabla^2_{\boldsymbol{\theta}} M(\boldsymbol{\theta}^*) \, ,$$

therefore,

$$\nabla^2_{\boldsymbol{\theta}} M_n(\tilde{\boldsymbol{\theta}}) \xrightarrow{p} \nabla^2_{\boldsymbol{\theta}} M(\boldsymbol{\theta}^*) \, , \text{ as } n \to \infty \, .$$

For the simplicity of notation, we set

$$H := \nabla^2_{\boldsymbol{\theta}} M(\boldsymbol{\theta}^*)\,.$$

We will show that H is positive definite, thus invertible. In addition, if one shows that

$$-\sqrt{n}\nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}^*) \xrightarrow{d} N(\mathbf{0}, V)\,, \text{ as } n \to \infty\,, \tag{A.39}$$

then, based on the Slutsky's theorem one obtains that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{d} N(\mathbf{0}, H^{-1}VH^{-1})\,, \text{ as } n \to \infty\,.$$

In the following, we will first prove $-\sqrt{n}\nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}^*)$ converges in distribution to a normal distribution, that is, (A.39), and calculate matrix V. Then we will calculate matrix H.

Step i: We prove that $-\sqrt{n}\nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}^*)$ converges in distribution to a normal distribution, and we calculate matrix V.

By definition of the system parameter $\boldsymbol{\theta}$, one sees that

$$\boldsymbol{\theta} = (\boldsymbol{x}_0, A) = [\boldsymbol{x}_0^\top, a_{11}, \ldots, a_{1d}, \ldots, a_{dd}]^\top \in \mathbb{R}^{d+d^2}\,,$$

where a_{jk} is the jk-th entry of parameter matrix A, for all $j, k = 1, \ldots, d$.

Therefore, $\nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}) \in \mathbb{R}^{d+d^2}$, with

$$\nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}) = \frac{\partial M_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left[\left\{\frac{\partial M_n(\boldsymbol{\theta})}{\partial \boldsymbol{x}_0}\right\}^\top, \frac{\partial M_n(\boldsymbol{\theta})}{\partial a_{11}}, \ldots, \frac{\partial M_n(\boldsymbol{\theta})}{\partial a_{1d}}, \ldots, \frac{\partial M_n(\boldsymbol{\theta})}{\partial a_{dd}}\right]^\top.$$

Recall that

$$M_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \|\, \boldsymbol{y}_i - e^{At_i}\boldsymbol{x}_0 \,\|_2^2\,, \tag{A.40}$$

obviously, if one wants to calculate the partial derivative of $M_n(\boldsymbol{\theta})$ with respect to a_{jk}, that is, $\partial M_n(\boldsymbol{\theta})/\partial a_{jk}$, one needs to calculate $\partial e^{At}/\partial a_{jk}$ first. Suppose that matrix A has d distinct eigenvalues $\lambda_1, \ldots, \lambda_d$, and A has the Jordan decomposition $A = Q\Lambda Q^{-1}$, where

$\Lambda = diag(\lambda_1, \ldots, \lambda_d)$. Then according to [157, 158], one obtains that

$$Z_{jk}(t) := \frac{\partial e^{At}}{\partial a_{jk}} = Q\{(Q^{-1}_{\cdot j} Q_{k\cdot}) \circ U(t)\}Q^{-1} \,,$$

here, for notational simplicity, we denote $\partial e^{At}/\partial a_{jk}$ as $Z_{jk}(t)$. The column vector $Q^{-1}_{\cdot j}$ stands for the jth column of matrix Q^{-1} and the row vector $Q_{k\cdot}$ denotes the kth row of matrix Q. Let $B \circ C$ denote the Hadamard product, with each element

$$(B \circ C)_{ij} = (B)_{ij}(C)_{ij} \,,$$

where matrices B and C are of the same dimension. $U(t)$ has the form:

$$U(t) = \begin{bmatrix} te^{\lambda_1 t} & \dfrac{e^{\lambda_1 t} - e^{\lambda_2 t}}{\lambda_1 - \lambda_2} & \cdots & \dfrac{e^{\lambda_1 t} - e^{\lambda_d t}}{\lambda_1 - \lambda_d} \\ \dfrac{e^{\lambda_2 t} - e^{\lambda_1 t}}{\lambda_2 - \lambda_1} & te^{\lambda_2 t} & \cdots & \dfrac{e^{\lambda_2 t} - e^{\lambda_d t}}{\lambda_2 - \lambda_d} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{e^{\lambda_d t} - e^{\lambda_1 t}}{\lambda_d - \lambda_1} & \dfrac{e^{\lambda_d t} - e^{\lambda_2 t}}{\lambda_d - \lambda_2} & \cdots & te^{\lambda_d t} \end{bmatrix} .$$

In the following, we will calculate $\nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta})$. Simple calculation shows that

$$\begin{aligned} M_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} \| \boldsymbol{y}_i - e^{At_i}\boldsymbol{x}_0 \|_2^2 = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{y}_i - e^{At_i}\boldsymbol{x}_0)^\top (\boldsymbol{y}_i - e^{At_i}\boldsymbol{x}_0) \\ &= \frac{1}{n}\sum_{i=1}^{n} \{\boldsymbol{y}_i^\top \boldsymbol{y}_i - \boldsymbol{x}_0^\top (e^{At_i})^\top \boldsymbol{y}_i - \boldsymbol{y}_i^\top e^{At_i}\boldsymbol{x}_0 + \boldsymbol{x}_0^\top (e^{At_i})^\top e^{At_i}\boldsymbol{x}_0\} \,. \end{aligned}$$

Then one obtains that

$$\frac{\partial M_n(\boldsymbol{\theta})}{\partial \boldsymbol{x}_0} = \frac{1}{n}\sum_{i=1}^{n} \{-2(e^{At_i})^\top \boldsymbol{y}_i + 2(e^{At_i})^\top e^{At_i}\boldsymbol{x}_0\} \,,$$

$$\begin{aligned}
\frac{\partial M_n(\boldsymbol{\theta})}{\partial a_{jk}} &= \frac{1}{n}\sum_{i=1}^{n} Tr\Big\{\Big(\frac{\partial \parallel \boldsymbol{y}_i - e^{At_i}\boldsymbol{x}_0 \parallel_2^2}{\partial e^{At_i}}\Big)^\top \frac{\partial e^{At_i}}{\partial a_{jk}}\Big\} \\
&= \frac{1}{n}\sum_{i=1}^{n} Tr\{(-\boldsymbol{y}_i\boldsymbol{x}_0^\top - \boldsymbol{y}_i\boldsymbol{x}_0^\top + 2e^{At_i}\boldsymbol{x}_0\boldsymbol{x}_0^\top)^\top Z_{jk}(t_i)\} \\
&= \frac{1}{n}\sum_{i=1}^{n} Tr[\{-2\boldsymbol{x}_0\boldsymbol{y}_i^\top + 2\boldsymbol{x}_0\boldsymbol{x}_0^\top(e^{At_i})^\top\}Z_{jk}(t_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n}\{-2\boldsymbol{y}_i^\top Z_{jk}(t_i)\boldsymbol{x}_0 + 2\boldsymbol{x}_0^\top(e^{At_i})^\top Z_{jk}(t_i)\boldsymbol{x}_0\}\,.
\end{aligned}$$

Therefore, one obtains that

$$\begin{aligned}
\frac{\partial M_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{x}_0} :&= \frac{\partial M_n(\boldsymbol{\theta})}{\partial \boldsymbol{x}_0}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \frac{1}{n}\sum_{i=1}^{n}\{-2(e^{A^*t_i})^\top(e^{A^*t_i}\boldsymbol{x}_0^* + \boldsymbol{\epsilon}_i) + 2(e^{A^*t_i})^\top e^{A^*t_i}\boldsymbol{x}_0^*\} \\
&= -\frac{2}{n}\sum_{i=1}^{n}(e^{A^*t_i})^\top\boldsymbol{\epsilon}_i\,,
\end{aligned}$$

$$\begin{aligned}
\frac{\partial M_n(\boldsymbol{\theta}^*)}{\partial a_{jk}} :&= \frac{\partial M_n(\boldsymbol{\theta})}{\partial a_{jk}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \frac{1}{n}\sum_{i=1}^{n}\{-2(e^{A^*t_i}\boldsymbol{x}_0^* + \boldsymbol{\epsilon}_i)^\top Z_{jk}^*(t_i)\boldsymbol{x}_0^* + 2(\boldsymbol{x}_0^*)^\top(e^{A^*t_i})^\top Z_{jk}^*(t_i)\boldsymbol{x}_0^*\} \\
&= -\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{\epsilon}_i^\top Z_{jk}^*(t_i)\boldsymbol{x}_0^*\,,
\end{aligned}$$

where

$$Z_{jk}^*(t) := \frac{\partial e^{At}}{\partial a_{jk}}\bigg|_{A=A^*} = Q^*[\{(Q^*)^{-1}_{.j}Q^*_{k.}\}\circ U^*(t)](Q^*)^{-1}\,,$$

with $Q^*, U^*(t)$ corresponding to the true parameter matrix A^*. That is, the Jordan decomposition of A^* is $A^* = Q^*\Lambda^*(Q^*)^{-1}$, where $\Lambda^* = diag(\lambda_1^*, \ldots, \lambda_d^*)$, with $\lambda_1^*, \lambda_2^*, \ldots, \lambda_d^*$ being the eigenvalues of A^*, and under assumption A2, these eigenvalues are distinct real values. Set $V := \lim_{n\to\infty} var(\sqrt{n}\nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}^*))$, then we will calculate V in the following. By some calculation, one obtains that

$$\begin{aligned}
var\Big\{\frac{\sqrt{n}\partial M_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{x}_0}\Big\} &= n\cdot var\Big\{-\frac{2}{n}\sum_{i=1}^{n}(e^{A^*t_i})^\top\boldsymbol{\epsilon}_i\Big\} \\
&= \frac{4}{n}\sum_{i=1}^{n} var\{(e^{A^*t_i})^\top\boldsymbol{\epsilon}_i\} = \frac{4}{n}\sum_{i=1}^{n}(e^{A^*t_i})^\top\Sigma e^{A^*t_i} \\
&\to \frac{4}{T}\int_0^T (e^{A^*t})^\top\Sigma e^{A^*t}\,dt\,,\ \text{as } n\to\infty\,.
\end{aligned}$$

Recall that Σ is the covariance matrix of error terms $\boldsymbol{\epsilon}_i$ for all $i = 1, \ldots, n$ under Assumption A5. Similarly, one obtains that

$$
\begin{aligned}
cov\Big\{\frac{\sqrt{n}\partial M_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{x}_0}, \frac{\sqrt{n}\partial M_n(\boldsymbol{\theta}^*)}{\partial a_{jk}}\Big\} &= nE\left[\frac{\partial M_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{x}_0} \cdot \frac{\partial M_n(\boldsymbol{\theta}^*)}{\partial a_{jk}}\right] - nE\left[\frac{\partial M_n(\boldsymbol{\theta}^*)}{\partial \boldsymbol{x}_0}\right] E\left[\frac{\partial M_n(\boldsymbol{\theta}^*)}{\partial a_{jk}}\right] \\
&= nE\left[\frac{2}{n}\sum_{i=1}^{n}(e^{A^*t_i})^\top \boldsymbol{\epsilon}_i \cdot \frac{2}{n}\sum_{i=1}^{n} \boldsymbol{\epsilon}_i^\top Z^*_{jk}(t_i)\boldsymbol{x}^*_0\right] \\
&= \frac{4}{n}\sum_{i=1}^{n} E[(e^{A^*t_i})^\top \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top Z^*_{jk}(t_i)\boldsymbol{x}^*_0] \\
&= \frac{4}{n}\sum_{i=1}^{n}(e^{A^*t_i})^\top E[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top] Z^*_{jk}(t_i)\boldsymbol{x}^*_0 \\
&= \frac{4}{n}\sum_{i=1}^{n}(e^{A^*t_i})^\top \Sigma Z^*_{jk}(t_i)\boldsymbol{x}^*_0 \\
&\to \frac{4}{T}\int_0^T (e^{A^*t})^\top \Sigma Z^*_{jk}(t)\boldsymbol{x}^*_0 \, dt\,, \text{ as } n \to \infty\,,
\end{aligned}
$$

and

$$
\begin{aligned}
cov\Big\{\frac{\sqrt{n}\partial M_n(\boldsymbol{\theta}^*)}{\partial a_{jk}}, \frac{\sqrt{n}\partial M_n(\boldsymbol{\theta}^*)}{\partial a_{pq}}\Big\} &= \frac{4}{n}\sum_{i=1}^{n}\{Z^*_{jk}(t_i)\boldsymbol{x}^*_0\}^\top \Sigma Z^*_{pq}(t_i)\boldsymbol{x}^*_0 \\
&\to \frac{4}{T}\int_0^T \{Z^*_{jk}(t)\boldsymbol{x}^*_0\}^\top \Sigma Z^*_{pq}(t)\boldsymbol{x}^*_0 \, dt\,, \text{ as } n \to \infty\,.
\end{aligned}
$$

If one denotes

$$
R(\boldsymbol{\theta}^*, t) := \Sigma^{1/2}\big(e^{A^*t}, Z^*_{11}(t)\boldsymbol{x}^*_0, \ldots, Z^*_{1d}(t)\boldsymbol{x}^*_0, \ldots, Z^*_{dd}(t)\boldsymbol{x}^*_0\big)\,, \tag{A.41}
$$

one sees that

$$
V = 4\int_0^T R(\boldsymbol{\theta}^*, t)^\top R(\boldsymbol{\theta}^*, t)/T \, dt \in \mathbb{R}^{(d+d^2)\times(d+d^2)}\,. \tag{A.42}
$$

Now that we have calculated V, then we will prove that $-\sqrt{n}\nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}^*)$ converges in distribution to a normal distribution. We first present a Lemma we will use for our proof.

Lemma A.4. [Lindeberg-Feller Central Limit Theorem] *Suppose* $\{\boldsymbol{w}_{ni}\}$ *is a triangular array of* $p \times 1$ *random vectors such that* $\boldsymbol{s}_n = \sum_{i=1}^n \boldsymbol{w}_{ni}/n$ *and*

$$V_n = \frac{1}{n}\sum_{i=1}^n var(\boldsymbol{w}_{ni}) \to V\,,$$

where V *is positive definite. If for every* $\varepsilon > 0$,

$$\frac{1}{n}\sum_{i=1}^n E[\| \boldsymbol{w}_{ni} \|_2^2 \, \mathbb{1}(\| \boldsymbol{w}_{ni} \|_2 \geq \varepsilon\sqrt{n})] \to 0\,, \tag{A.43}$$

then

$$\sqrt{n}\boldsymbol{s}_n \xrightarrow{d} N(0, V)\,.$$

If one sets

$$\boldsymbol{s}_n := -\nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}^*)\,,$$

$$\boldsymbol{w}_{ni} := 2[\{(e^{A^* t_i})^\top \boldsymbol{\epsilon}_i\}^\top, \boldsymbol{\epsilon}_i^\top Z_{11}^*(t_i)\boldsymbol{x}_0^*, \ldots, \boldsymbol{\epsilon}_i^\top Z_{1d}^*(t_i)\boldsymbol{x}_0^*, , \cdots, \boldsymbol{\epsilon}_i^\top Z_{dd}^*(t_i)\boldsymbol{x}_0^*]^\top\,,$$

where $\boldsymbol{w}_{ni} \in \mathbb{R}^{d+d^2}$, then V in Equation (A.42) corresponds the V in Lemma A.4. Then the proof of the asymptotic normality of $-\sqrt{n}\nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}^*)$ can be broken down into two tasks. Proving condition (A.43) is satisfied and proving variance matrix V is positive definite.

To this end, we first prove the condition (A.43) is satisfied. Simple calculation show that

$$\| \boldsymbol{w}_{ni} \|_2^2 = 4 \| (e^{A^* t_i})^\top \boldsymbol{\epsilon}_i \|_2^2 + 4\sum_{j=1}^d \sum_{k=1}^d \{\boldsymbol{\epsilon}_i^\top Z_{jk}^*(t_i)\boldsymbol{x}_0^*\}^2 \in (-\infty, \infty)\,,$$

therefore, one obtains that

$$\lim_{n\to\infty} \| \boldsymbol{w}_{ni} \|_2^2 \, \mathbb{1}(\| \boldsymbol{w}_{ni} \|_2 \geq \varepsilon\sqrt{n}) = 0\,,$$

almost surely. By calculation, one obtains that

$$\begin{aligned} E[\| (e^{A^* t_i})^\top \boldsymbol{\epsilon}_i \|_2^2] &\leq E[\| e^{A^* t_i} \|_F^2 \| \boldsymbol{\epsilon}_i \|_2^2] = \| e^{A^* t_i} \|_F^2 \, E[\| \boldsymbol{\epsilon}_i \|_2^2] \\ &= \| e^{A^* t_i} \|_F^2 \sum_{j=1}^d \sigma_j^2 < \infty\,, \end{aligned}$$

and

$$
\begin{aligned}
E[\{\boldsymbol{\epsilon}_i^\top Z_{jk}^*(t_i)\boldsymbol{x}_0^*\}^2] &= E[\{Z_{jk}^*(t_i)\boldsymbol{x}_0^*\}^\top \boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top Z_{jk}^*(t_i)\boldsymbol{x}_0^*] \\
&= \{Z_{jk}^*(t_i)\boldsymbol{x}_0^*\}^\top E[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top] Z_{jk}^*(t_i)\boldsymbol{x}_0^* \\
&= \{Z_{jk}^*(t_i)\boldsymbol{x}_0^*\}^\top \Sigma Z_{jk}^*(t_i)\boldsymbol{x}_0^* < \infty\,.
\end{aligned}
$$

Therefore, $E[\| \boldsymbol{w}_{ni} \|_2^2] < \infty$. Since

$$
\| \boldsymbol{w}_{ni} \|_2^2 \, \mathbb{1}(\| \boldsymbol{w}_{ni} \|_2 \geq \varepsilon\sqrt{n}) \leq \| \boldsymbol{w}_{ni} \|_2^2\,,
$$

then by Lebesgue's dominated convergence theorem, one obtains that

$$
\lim_{n\to\infty} E[\| \boldsymbol{w}_{ni} \|_2^2 \, \mathbb{1}(\| \boldsymbol{w}_{ni} \|_2 \geq \varepsilon\sqrt{n})] = E[\lim_{n\to\infty} \| \boldsymbol{w}_{ni} \|_2^2 \, \mathbb{1}(\| \boldsymbol{w}_{ni} \|_2 \geq \varepsilon\sqrt{n})] = 0\,.
$$

Now that we have proved that the condition (A.43) is satisfied, we will then prove V is positive definite. If one denotes

$$
W(t) := \left(e^{A^* t}, Z_{11}^*(t)\boldsymbol{x}_0^*, \ldots, Z_{1d}^*(t)\boldsymbol{x}_0^*, \ldots, Z_{dd}^*(t)\boldsymbol{x}_0^*\right),
$$

then according to Equation (A.41) and Equation (A.42), one sees that

$$
V := \frac{4}{T}\int_0^T V(t)\,dt = \frac{4}{T}\int_0^T W(t)^\top \Sigma W(t)\,dt\,. \tag{A.44}
$$

In the following, we will show that $\int_0^T V(t)\,dt \in \mathbb{R}^{(d+d^2)\times(d+d^2)}$ is positive definite. Let $\boldsymbol{\xi} \in \mathbb{R}^{d+d^2}$ be such that

$$
\boldsymbol{\xi}^T \cdot \Big(\int_0^T V(t)dt\Big) \cdot \boldsymbol{\xi} = 0.
$$

We want to show that $\boldsymbol{\xi} = \mathbf{0}$. Since $V(t)$ is non-negative definitely for every t, this implies that

$$
\boldsymbol{\xi}^T V(t)\boldsymbol{\xi} = \boldsymbol{\xi}^T W(t)^\top \Sigma W(t)\boldsymbol{\xi} = 0\,\forall t
$$

which further implies that $W(t)\boldsymbol{\xi} = \mathbf{0}$ for all t. By differentiation, one sees that

$$
W^{(m)}(0)\boldsymbol{\xi} = \mathbf{0}\ \forall m = 0, 1, 2, \ldots,
$$

where $W^{(m)}(t)$ denotes the mth deivative of $W(t)$. As we will see, the first $d+1$ equations are sufficient to yield that $\boldsymbol{\xi} = \mathbf{0}$. Recall that

$$Z_{jk}^*(t) = \frac{\partial e^{At}}{\partial a_{jk}}\bigg|_{A=A^*},$$

by denoting the first d components of $\boldsymbol{\xi}$ as $\boldsymbol{\xi}_d$ and the (j,k)-component of the last d^2 components of $\boldsymbol{\xi}$ as $\boldsymbol{\xi}_{d+jk}$, one obtains that

$$\begin{aligned}
W^{(m)}(0)\boldsymbol{\xi} &= \left(\frac{\partial^{(m)} e^{A^*t}}{\partial t^m}\bigg|_{t=0}, \frac{\partial^{(m+1)} e^{At}}{\partial t^m \partial a_{11}}\bigg|_{t=0,A=A^*} \cdot \boldsymbol{x}_0^*, \ldots, \frac{\partial^{(m+1)} e^{At}}{\partial t^m \partial a_{dd}}\bigg|_{t=0,A=A^*} \cdot \boldsymbol{x}_0^*\right)\boldsymbol{\xi} \\
&= \left((A^*)^m, \frac{\partial A^m}{\partial a_{11}}\bigg|_{A=A^*} \cdot \boldsymbol{x}_0^*, \ldots, \frac{\partial A^m}{\partial a_{dd}}\bigg|_{A=A^*} \cdot \boldsymbol{x}_0^*\right)\boldsymbol{\xi} \\
&= \left((A^*)^m, \sum_{l=1}^{m}(A^*)^{m-l}E_{11}(A^*)^{l-1}\boldsymbol{x}_0^*, \ldots, \sum_{l=1}^{m}(A^*)^{m-l}E_{dd}(A^*)^{l-1}\boldsymbol{x}_0^*\right)\boldsymbol{\xi} \\
&= (A^*)^m\boldsymbol{\xi}_d + \sum_{l=1}^{m}\sum_{j=1}^{d}\sum_{k=1}^{d}\{(A^*)^{m-l}E_{jk}(A^*)^{l-1}\boldsymbol{x}_0^*\}\boldsymbol{\xi}_{d+jk}\,,
\end{aligned} \tag{A.45}$$

where $\sum_{i=1}^{0} a_i = 0$ for any sequence $\{a_i, i \in \mathbb{Z}\}$ denotes the empty sum, E_{jk} is a $d \times d$ matrix with the jk-th entry being 1 and all the other entries being 0.

If one identifies the last d^2 elements of $\boldsymbol{\xi}$ with a $d \times d$ matrix Ξ (the jk-th entry of Ξ being $\boldsymbol{\xi}_{d+jk}$), then (A.45) becomes

$$(A^*)^m\boldsymbol{\xi}_d + \sum_{l=1}^{m}(A^*)^{m-l} \cdot \Xi \cdot (A^*)^{l-1}\boldsymbol{x}_0^* = \mathbf{0} \quad \forall m \geqslant 0\,.$$

Taking $m = 0, 1, \ldots, d$ respectively, one obtains the following system of equations:

$$\begin{cases}
\boldsymbol{\xi}_d = \mathbf{0}\,, \\
A^*\boldsymbol{\xi}_d + \Xi \cdot \boldsymbol{x}_0^* = \mathbf{0}\,, \\
(A^*)^2\boldsymbol{\xi}_d + \Xi \cdot A^*\boldsymbol{x}_0^* + (A^*\Xi) \cdot \boldsymbol{x}_0^* = \mathbf{0}\,, \\
(A^*)^3\boldsymbol{\xi}_d + \Xi \cdot (A^*)^2\boldsymbol{x}_0^* + (A^*\Xi) \cdot A^*\boldsymbol{x}_0^* + ((A^*)^2\Xi) \cdot \boldsymbol{x}_0^* = \mathbf{0}\,, \\
\cdots \\
(A^*)^d\boldsymbol{\xi}_d + \Xi \cdot (A^*)^{d-1}\boldsymbol{x}_0^* + (A^*\Xi) \cdot (A^*)^{d-2}\boldsymbol{x}_0^* + \cdots + ((A^*)^{d-1}\Xi) \cdot \boldsymbol{x}_0^* = \mathbf{0}\,.
\end{cases}$$

From the first equation, we have $\boldsymbol{\xi}_d = \mathbf{0}$. If one multiplies the second equation by A^* on the left and subtract it from the third equation, one obtains that

$$\Xi \cdot A^* \boldsymbol{x}_0^* = \mathbf{0} .$$

Similarly,

$$\text{“}(l+2)\text{-th eqn ”} - \ A^* \times \text{“}(l+1)\text{-th eqn”} \implies \Xi \cdot (A^*)^l \boldsymbol{x}_0^* = \mathbf{0} .$$

As a result,

$$\Xi \cdot \left(\boldsymbol{x}_0^*, A^* \boldsymbol{x}_0^*, \ldots, (A^*)^{d-1} \boldsymbol{x}_0^*\right) = \mathbf{0} .$$

Since the matrix $\left(\boldsymbol{x}_0^*, A^* \boldsymbol{x}_0^*, \ldots, (A^*)^{d-1} \boldsymbol{x}_0^*\right)$ is invertible by assumption A2, one concludes that $\Xi = \mathbf{0}$, that is, the last d^2 components of $\boldsymbol{\xi}$ are all zeros.

Therefore, we have proved that $\boldsymbol{\xi} = \mathbf{0}$, which means, $\int_0^T V(t)\, dt$ is positive definite. Thus, V is positive definite. Since V is symmetric, V is also nonsingular. By Lemma A.4, one concludes that

$$-\sqrt{n} \nabla_{\boldsymbol{\theta}} M_n(\boldsymbol{\theta}^*) \xrightarrow{d} N(\mathbf{0}, V)\,, \ \text{as } n \to \infty ,$$

where V is defined in Equation (A.42).

Step ii: We calculate matrix H.

Recall that $H = \nabla^2_{\boldsymbol{\theta}} M(\boldsymbol{\theta}^*)$, that is, the Hessian matrix of $M(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$. And

$$\begin{aligned} M(\boldsymbol{\theta}) &= \frac{1}{T} \int_0^T \| \, e^{A^* t} \boldsymbol{x}_0^* - e^{At} \boldsymbol{x}_0 \, \|_2^2 \ dt \\ &= \frac{1}{T} \int_0^T \| \, e^{A^* t} \boldsymbol{x}_0^* \, \|_2^2 + \| \, e^{At} \boldsymbol{x}_0 \, \|_2^2 \, - 2 (\boldsymbol{x}_0^*)^\top (e^{A^* t})^\top e^{At} \boldsymbol{x}_0 \, dt . \end{aligned}$$

If one sets

$$h(\boldsymbol{x}_0, A) := \| \, e^{At} \boldsymbol{x}_0 \, \|_2^2 \, - 2 (\boldsymbol{x}_0^*)^\top (e^{A^* t})^\top e^{At} \boldsymbol{x}_0 ,$$

by taking derivative of $h(\boldsymbol{x}_0, A)$ with respect to $\boldsymbol{x}_0$ one obtains

$$\frac{\partial h(\boldsymbol{x}_0, A)}{\partial \boldsymbol{x}_0} = 2 (e^{At})^\top e^{At} \boldsymbol{x}_0 - 2 (e^{At})^\top e^{A^* t} \boldsymbol{x}_0^* ,$$

by further taking derivative with respect to $\boldsymbol{x}_0^\top$ one obtains that

$$\frac{\partial^2 h(\boldsymbol{x}_0, A)}{\partial \boldsymbol{x}_0 \partial \boldsymbol{x}_0^\top} = 2(e^{At})^\top e^{At}\,.$$

Therefore,

$$\frac{\partial^2 M(\boldsymbol{\theta}^*)}{\partial \boldsymbol{x}_0 \partial \boldsymbol{x}_0^\top} = \frac{\partial^2 M(\boldsymbol{\theta})}{\partial \boldsymbol{x}_0 \partial \boldsymbol{x}_0^\top}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \frac{2}{T}\int_0^T (e^{A^*t})^\top e^{A^*t}\,dt\,.$$

Taking derivative of $h(\boldsymbol{x}_0, A)$ with respect to a_{jk} one obtains that

$$\begin{aligned}
\frac{\partial h(\boldsymbol{x}_0, A)}{\partial a_{jk}} &= Tr\left\{\left(\frac{\partial h(\boldsymbol{x}_0, A)}{\partial e^{At}}\right)^\top \frac{\partial e^{At}}{\partial a_{jk}}\right\} \\
&= Tr\left\{(2e^{At}\boldsymbol{x}_0\boldsymbol{x}_0^\top - 2e^{A^*t}\boldsymbol{x}_0^*\boldsymbol{x}_0^\top)^\top \frac{\partial e^{At}}{\partial a_{jk}}\right\} \\
&= Tr\left[\{2\boldsymbol{x}_0\boldsymbol{x}_0^\top(e^{At})^\top - 2\boldsymbol{x}_0(\boldsymbol{x}_0^*)^\top(e^{A^*t})^\top\}\frac{\partial e^{At}}{\partial a_{jk}}\right] \\
&= Tr\left\{2\boldsymbol{x}_0^\top(e^{At})^\top\frac{\partial e^{At}}{\partial a_{jk}}\boldsymbol{x}_0\right\} - Tr\left\{2(\boldsymbol{x}_0^*)^\top(e^{A^*t})^\top\frac{\partial e^{At}}{\partial a_{jk}}\boldsymbol{x}_0\right\} \\
&= 2\boldsymbol{x}_0^\top(e^{At})^\top\frac{\partial e^{At}}{\partial a_{jk}}\boldsymbol{x}_0 - 2(\boldsymbol{x}_0^*)^\top(e^{A^*t})^\top\frac{\partial e^{At}}{\partial a_{jk}}\boldsymbol{x}_0\,,
\end{aligned}$$

by further taking derivative with respect to $\boldsymbol{x}_0^\top$ one obtains that

$$\frac{\partial^2 h(\boldsymbol{x}_0, A)}{\partial a_{jk}\partial \boldsymbol{x}_0^\top} = 2\boldsymbol{x}_0^\top\left\{(e^{At})^\top\frac{\partial e^{At}}{\partial a_{jk}} + \left(\frac{\partial e^{At}}{\partial a_{jk}}\right)^\top e^{At}\right\} - 2(\boldsymbol{x}_0^*)^\top(e^{A^*t})^\top\frac{\partial e^{At}}{\partial a_{jk}}\,.$$

Therefore,

$$\frac{\partial^2 M(\boldsymbol{\theta}^*)}{\partial a_{jk}\partial \boldsymbol{x}_0^\top} = \frac{\partial^2 M(\boldsymbol{\theta})}{\partial a_{jk}\partial \boldsymbol{x}_0^\top}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \frac{2}{T}\int_0^T \{Z_{jk}^*(t)\boldsymbol{x}_0^*\}^\top e^{A^*t}\,dt\,.$$

Similarly,

$$\frac{\partial^2 h(\boldsymbol{x}_0, A)}{\partial a_{jk}\partial a_{pq}} = 2\boldsymbol{x}_0^\top\left(\frac{\partial e^{At}}{\partial a_{pq}}\right)^\top\frac{\partial e^{At}}{\partial a_{jk}}\boldsymbol{x}_0 + 2\boldsymbol{x}_0^\top(e^{At})^\top\frac{\partial^2 e^{At}}{\partial a_{jk}\partial a_{pq}}\boldsymbol{x}_0 - 2(\boldsymbol{x}_0^*)^\top(e^{A^*t})^\top\frac{\partial^2 e^{At}}{\partial a_{jk}\partial a_{pq}}\boldsymbol{x}_0\,,$$

then by some calculation, one obtains that

$$\frac{\partial^2 M(\boldsymbol{\theta}^*)}{\partial a_{jk}\partial a_{pq}} = \frac{\partial^2 M(\boldsymbol{\theta})}{\partial a_{jk}\partial a_{pq}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \frac{2}{T}\int_0^T (Z_{pq}^*(t)\boldsymbol{x}_0^*)^\top Z_{jk}^*(t)\boldsymbol{x}_0^*\,dt\,.$$

If one denotes

$$S(\boldsymbol{\theta}^*, t) := \left(e^{A^*t}, Z_{11}^*(t)\boldsymbol{x}_0^*, \ldots, Z_{1d}^*(t)\boldsymbol{x}_0^*, \ldots, Z_{dd}^*(t)\boldsymbol{x}_0^*\right), \tag{A.46}$$

one sees that

$$H = 2\int_0^T S(\boldsymbol{\theta}^*, t)^\top S(\boldsymbol{\theta}^*, t)/T\,dt\,, \tag{A.47}$$

and $H \in \mathbb{R}^{(d+d^2)\times(d+d^2)}$.

Rearranging V in Equation (A.42), one obtains that

$$V = 4\int_0^T S(\boldsymbol{\theta}^*, t)^\top \Sigma S(\boldsymbol{\theta}^*, t)/T\,dt\,, \tag{A.48}$$

obviously, V and H have similar forms. Therefore, by using the same way of proving V is positive definite, one can prove that H is positive definite. Since H is symmetric, H is nonsingular.

Therefore, we have completed the proof and provided the explicit forms of both V and H. Moreover, we have shown that both V and H are positive definite and nonsingular. □

A.1.9 Proof of Corollary 3.13

Proof. According to the proof of Corollary 3.10 in Appendix A.1.6, one sees that assumptions A1, A2 and A4-A6 hold with respect to the new ODE system, the new parameter $\tilde{\boldsymbol{\theta}}^*$ and the new error terms $\tilde{\boldsymbol{\epsilon}}_i$ corresponding to the aggregated observations $\tilde{\boldsymbol{Y}}$.

Therefore, by Theorem 3.12, one obtains that

$$\sqrt{\tilde{n}}(\hat{\tilde{\boldsymbol{\theta}}} - \tilde{\boldsymbol{\theta}}^*) \xrightarrow{d} N(\mathbf{0}, \tilde{H}^{-1}\tilde{V}\tilde{H}^{-1})\,,\ \text{as } \tilde{n} \to \infty\,,$$

where

$$\tilde{\boldsymbol{\theta}}^* = (\tilde{\boldsymbol{x}}_0^*, \tilde{A}^*) = \left((I + e^{A^*\Delta_t} + \cdots + e^{A^*(k-1)\Delta_t})\boldsymbol{x}_0^*/k, A^*\right),$$

recall that we we have built the relationship between $\tilde{\boldsymbol{\theta}}^*$ and $\boldsymbol{\theta}^*$ in Corollary 3.7.

Using the same way we calculate V and H in the proof of Theorem 3.12 in Appendix A.1.8, one obtains that

$$\tilde{V} = 4\int_0^{\tilde{T}} R(\tilde{\boldsymbol{\theta}}^*, t)^\top R(\tilde{\boldsymbol{\theta}}^*, t)/k\tilde{T}\, dt\,,$$

where $R(\tilde{\boldsymbol{\theta}}^*, t)$ is defined in Equation (A.41), and $\tilde{T} = (\lfloor n/k \rfloor - 1)kT/(n-1)$. Note that, one devides k in the formula of $\tilde{V}$, since the covariance matrix of the new error terms $\{\tilde{\boldsymbol{\epsilon}}_i\}$ is Σ/k based on the generation rules of the aggregated observations.

Similarly, one obtains that

$$\tilde{H} = 2\int_0^{\tilde{T}} S(\tilde{\boldsymbol{\theta}}^*, t)^\top S(\tilde{\boldsymbol{\theta}}^*, t)/\tilde{T}\, dt\,,$$

where $S(\tilde{\boldsymbol{\theta}}^*, t)$ is defined in Equation (A.46), and $\tilde{T} = (\lfloor n/k \rfloor - 1)kT/(n-1)$.

Moreover, by using the same way we prove that both V and H are positive definite and nonsingular in the proof of Theorem 3.12 in Appendix A.1.8, one obtains that both $\tilde{V}$ and $\tilde{H}$ are positive definite and nonsingular.

The definition of $\hat{\boldsymbol{\theta}}_{\tilde{n}}$ in Corollary 3.10, that is,

$$\hat{\boldsymbol{\theta}}_{\tilde{n}} := g(\hat{\tilde{\boldsymbol{\theta}}}) := \left(k(I + e^{\hat{\tilde{A}}\Delta_t} + \cdots + e^{\hat{\tilde{A}}(k-1)\Delta_t})^{-1}\hat{\tilde{\boldsymbol{x}}}_0, \hat{\tilde{A}}\right),$$

shows that the function $g(\boldsymbol{\theta})$ has the following form

$$g(\boldsymbol{\theta}) = g(\boldsymbol{x}_0, A) = \left(k(I + e^{A\Delta_t} + \cdots + e^{A(k-1)\Delta_t})^{-1}\boldsymbol{x}_0, A\right).$$

By plugging $\tilde{\boldsymbol{\theta}}^*$ in, one obtains that

$$g(\tilde{\boldsymbol{\theta}}^*) = \left(k(I + e^{\tilde{A}^*\Delta_t} + \cdots + e^{\tilde{A}^*(k-1)\Delta_t})^{-1}\tilde{\boldsymbol{x}}_0^*, \tilde{A}^*\right) = (\boldsymbol{x}_0^*, A^*) = \boldsymbol{\theta}^*.$$

Taking derivative of $g(\boldsymbol{\theta})$ with respect to $\boldsymbol{x}_0$ one obtains that

$$\begin{aligned}\frac{\partial g(\tilde{\boldsymbol{\theta}}^*)}{\partial \boldsymbol{x}_0^\top} &= \left.\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{x}_0^\top}\right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}^*} \\ &= \Big(k\{(I + e^{A^*\Delta_t} + \cdots + e^{A^*(k-1)\Delta_t})^{-1}\}^\top, \underbrace{\mathbf{0}_d, \ldots, \mathbf{0}_d}_{d^2 \text{ entries}}\Big)^\top \in \mathbb{R}^{(d+d^2)\times d}.\end{aligned}$$

Since

$$\begin{aligned}
&\frac{\partial k(I + e^{A\Delta_t} + \cdots + e^{A(k-1)\Delta_t})^{-1}\boldsymbol{x}_0}{\partial a_{pq}} \\
=& - k(I + e^{A\Delta_t} + \cdots + e^{A(k-1)\Delta_t})^{-2}\frac{\partial (I + e^{A\Delta_t} + \cdots + e^{A(k-1)\Delta_t})}{\partial a_{pq}}\boldsymbol{x}_0 \\
=& - k(I + e^{A\Delta_t} + \cdots + e^{A(k-1)\Delta_t})^{-2}[Z_{pq}(\Delta_t) + \cdots + Z_{pq}\{(k-1)\Delta_t\}]\boldsymbol{x}_0\,,
\end{aligned}$$

one obtains that

$$\begin{aligned}
&\frac{\partial g(\tilde{\boldsymbol{\theta}}^*)}{\partial a_{pq}} = \frac{\partial g(\boldsymbol{\theta})}{\partial a_{pq}}\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}^*} \\
=& \Big[- k\big\{(I + e^{A^*\Delta_t} + \cdots + e^{A^*(k-1)\Delta_t})^{-2}[Z^*_{pq}(\Delta_t) + \cdots + Z^*_{pq}\{(k-1)\Delta_t\}]\tilde{\boldsymbol{x}}^*_0\big\}^\top, \\
&\underbrace{0, \ldots, 0, 1, 0, \ldots, 0}_{d^2 \text{ entries}}\Big]^\top \\
\in& \mathbb{R}^{(d+d^2)\times 1},
\end{aligned}$$

where 1 in the last d^2 entries corresponds to the pq-th entry, where $p, q = 1, \ldots, d$.

Therefore, the gradient of $g(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ at $\tilde{\boldsymbol{\theta}}^*$ is

$$G := \nabla g(\tilde{\boldsymbol{\theta}}^*) = \left(\frac{\partial g(\tilde{\boldsymbol{\theta}}^*)}{\partial \boldsymbol{x}_0^\top}, \frac{\partial g(\tilde{\boldsymbol{\theta}}^*)}{\partial a_{11}}, \cdots, \frac{\partial g(\tilde{\boldsymbol{\theta}}^*)}{\partial a_{1d}}, \cdots, \frac{\partial g(\tilde{\boldsymbol{\theta}}^*)}{\partial a_{dd}}\right) \in \mathbb{R}^{(d+d^2)\times(d+d^2)}\,.$$

By the multivariate Delta method, one obtains that

$$\sqrt{\tilde{n}}\{g(\hat{\tilde{\boldsymbol{\theta}}}) - g(\tilde{\boldsymbol{\theta}}^*)\} \xrightarrow{d} N(\mathbf{0}, G\tilde{H}^{-1}\tilde{V}\tilde{H}^{-1}G^\top)\,, \text{ as } \tilde{n} \to \infty\,,$$

which is

$$\sqrt{\tilde{n}}(\hat{\boldsymbol{\theta}}_{\tilde{n}} - \boldsymbol{\theta}^*) \xrightarrow{d} N(\mathbf{0}, G\tilde{H}^{-1}\tilde{V}\tilde{H}^{-1}G^\top)\,, \text{ as } \tilde{n} \to \infty\,.$$

□

A.1.10 Proof of Corollary 3.14

Proof. According to the proof of Corollary 3.11 in Appendix A.1.7, one sees that assumptions A1, A2 and A4-A6 hold with respect to the new ODE system, the new parameter $\tilde{\boldsymbol{\theta}}^*$ and the new error terms $\tilde{\boldsymbol{\epsilon}}_i$ corresponding to the time-scaled observations $\tilde{\boldsymbol{Y}}$.

Therefore, by Theorem 3.12, one obtains that

$$\sqrt{\tilde{n}}(\hat{\tilde{\boldsymbol{\theta}}} - \tilde{\boldsymbol{\theta}}^*) \xrightarrow{d} N(\mathbf{0}, \tilde{H}^{-1}\tilde{V}\tilde{H}^{-1}), \text{ as } \tilde{n} \to \infty,$$

where

$$\tilde{\boldsymbol{\theta}}^* = (\tilde{\boldsymbol{x}}_0^*, \tilde{A}^*) = (\boldsymbol{x}_0^*, A^*/k),$$

recall that we we have built the relationship between $\tilde{\boldsymbol{\theta}}^*$ and $\boldsymbol{\theta}^*$ in Corollary 3.8.

Using the same way we calculate V and H in the proof of Theorem 3.12 in Appendix A.1.8, one obtains that

$$\tilde{V} = 4\int_0^{kT} R(\tilde{\boldsymbol{\theta}}^*, t)^\top R(\tilde{\boldsymbol{\theta}}^*, t)/(kT)\, dt,$$

where $R(\tilde{\boldsymbol{\theta}}^*, t)$ is defined in Equation (A.41).

Similarly, one obtains that

$$\tilde{H} = 2\int_0^{kT} S(\tilde{\boldsymbol{\theta}}^*, t)^\top S(\tilde{\boldsymbol{\theta}}^*, t)/(kT)\, dt,$$

where $S(\tilde{\boldsymbol{\theta}}^*, t)$ is defined in Equation (A.46).

Moreover, by using the same way we prove that both V and H are positive definite and nonsingular in the proof of Theorem 3.12 in Appendix A.1.8, one obtains that both $\tilde{V}$ and $\tilde{H}$ are positive definite and nonsingular.

The definition of $\hat{\boldsymbol{\theta}}_{\tilde{n}}$ in Corollary 3.11, that is,

$$\hat{\boldsymbol{\theta}}_{\tilde{n}} := g(\hat{\tilde{\boldsymbol{\theta}}}) := (\hat{\tilde{\boldsymbol{x}}}_0, k\hat{\tilde{A}}),$$

shows that the function $g(\boldsymbol{\theta})$ has the following form

$$g(\boldsymbol{\theta}) = g(\boldsymbol{x}_0, A) = (\boldsymbol{x}_0, kA).$$

By plugging $\tilde{\boldsymbol{\theta}}^*$ in, one obtains that

$$g(\tilde{\boldsymbol{\theta}}^*) = (\tilde{\boldsymbol{x}}_0^*, k\tilde{A}^*) = (\boldsymbol{x}_0^*, A^*) = \boldsymbol{\theta}^*.$$

Following the same way we calculate the gradient of $g(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ at $\tilde{\boldsymbol{\theta}}^*$ in the proof of Corollary 3.13 in Appendix A.1.9, one obtains that

$$G := \nabla g(\tilde{\boldsymbol{\theta}}^*) = \frac{\partial g(\tilde{\boldsymbol{\theta}}^*)}{\partial \boldsymbol{\theta}} = \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}^*} = \begin{bmatrix} I_d & \mathbf{0}_d & \mathbf{0}_d & \cdots & \mathbf{0}_d \\ \mathbf{0}_d^\top & k & 0 & \cdots & 0 \\ \mathbf{0}_d^\top & 0 & k & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \mathbf{0}_d^\top & 0 & 0 & \cdots & k \end{bmatrix} \in \mathbb{R}^{(d+d^2)\times(d+d^2)} .$$

By the multivariate Delta method, one obtains that

$$\sqrt{\tilde{n}}\{g(\hat{\tilde{\boldsymbol{\theta}}}) - g(\tilde{\boldsymbol{\theta}}^*)\} \xrightarrow{d} N(\mathbf{0}, G\tilde{H}^{-1}\tilde{V}\tilde{H}^{-1}G^\top)\,, \text{ as } \tilde{n} \to \infty\,,$$

which is

$$\sqrt{\tilde{n}}(\hat{\boldsymbol{\theta}}_{\tilde{n}} - \boldsymbol{\theta}^*) \xrightarrow{d} N(\mathbf{0}, G\tilde{H}^{-1}\tilde{V}\tilde{H}^{-1}G^\top)\,, \text{ as } \tilde{n} \to \infty\,.$$

□

Appendix B

Appendix of Chapter 4

B.1 Notations and proposed identifiability conditions

TABLE B.1: Summary of proposed identifiability conditions

ODEs	Conds.	# Traj.	Obs.	Def./Thm.	Necessity
Eq.(4.2)+(4.4)	A1	1	continuous	3.1	Yes
Eq.(4.3)	latent DAG, B1	1	continuous	4.1	Yes
Eq.(4.3)	latent DAG, C1	1	discrete	4.2	Yes
Eq.(4.3)	latent DAG, B2, B3, B4	p	continuous	4.3	No
Eq.(4.3)	latent DAG, C2, B3, B4	p	discrete	4.4	No

B.2 Real world examples

In this section, we present two real-world examples that correspond to the ODE models (4.2) and (4.3). These examples initially assume fully observable systems, with latent variables introduced by us based on prior experience or established physical laws.

B.2.1 Damped harmonic oscillators model

Consider a one-dimensional system comprising D point masses m_i for $i = 1, \ldots, D$ with positions $Q_i(t) \in \mathbb{R}$ and momenta $P_i(t) \in \mathbb{R}$. These masses are interconnected by springs characterized by spring constants k_i and equilibrium lengths l_i, and each mass is subject

TABLE B.2: Summary of notations

Notation	Description
$\boldsymbol{x}/\boldsymbol{z}$	observable/latent variables
x_i/z_i	the i-th observable/latent variable
t	time
t_j	the j-th time point
$\boldsymbol{x}(t)/\boldsymbol{z}(t)$	state of observable/latent variable at time t
$\boldsymbol{x}_j$	$\boldsymbol{x}(t_j)$, observable state at time t_j
$\boldsymbol{x}_0/\boldsymbol{z}_0$	initial condition of observable/latent variable
$\dot{\boldsymbol{x}}(t)$	first derivative of $\boldsymbol{x}(t)$ w.r.t. time t
d	dimension of observable variables
p	dimension of latent variables
A, B, G	constant parameter matrices defined in Eq.(4.2) and (4.3)
$\boldsymbol{f}(t)$	Function of time t defined in Eq.(4.2)
$\boldsymbol{v}_k$	constant parameter vector defined in Eq.(4.4)
$\{\boldsymbol{v}_k\}_0^r$	all the $\boldsymbol{v}_k$'s for $k = 0, \dots, r$
$\boldsymbol{\theta}$	$:= (\boldsymbol{x}_0, \boldsymbol{z}_0, A, B, \{\boldsymbol{v}_k\}_0^r)$, the system parameter of ODE system (4.2)
$\boldsymbol{\beta}$	a vector defined in Thm.4.4 **A1**
$\boldsymbol{y}(t)$	augmented state
$\boldsymbol{y}_0$	initial condition of augmented variable
$\boldsymbol{\eta}$	$:= (\boldsymbol{x}_0, \boldsymbol{z}_0, A, B, G)$, the system parameter of ODE system (4.3)
$\boldsymbol{\gamma}$	a vector defined in Thm.4.6 **B1**
$\boldsymbol{z}_0^*$	given initial condition of latent variable
$\boldsymbol{z}_0^{*i}$	the i-th given initial condition of latent variable
$\boldsymbol{\eta}_i$	$:= (\boldsymbol{x}_0, \boldsymbol{z}_0^{*i}, A, B, G)$, the system parameter of ODE system (4.3)
$\boldsymbol{\gamma}_i$	a vector defined in Thm 4.10 **B2**
$\boldsymbol{x}_{ij}$	$:= \boldsymbol{x}(t_j; \boldsymbol{\eta}_i)$, observable state of ODE system (4.3) with parameter $\boldsymbol{\eta}_i$ at time t_j
$\boldsymbol{y}_{ij}$	augmented state of $\boldsymbol{x}_{ij}$ at time t_j
$A', \boldsymbol{x}_0', \dots$	the alternative counterpart corresponding to $A, \boldsymbol{x}_0, \dots$

to friction with coefficient b_i. The system's boundary conditions are fixed at $Q_0(t) = 0$ and $Q_{D+1}(t) = L$.

The dynamics of this system are described by the following linear ODE system [18]:

$$\begin{aligned} \dot{P}_i(t) &= k_i(Q_{i+1}(t) - Q_i(t) - l_i) - k_{i-1}(Q_i(t) - Q_{i-1}(t) - l_{i-1}) - b_i P_i(t)/m_i \\ \dot{Q}_i(t) &= P_i(t)/m_i \end{aligned} \tag{B.1}$$

where $Q_0(t) = 0$ and $Q_{D+1}(t) = L$ represent the fixed boundary conditions. External forces $F_j(t)$ (e.g., wind force or a varying magnetic field) may influence the entire system of coupled oscillators. These external forces can be modeled here as latent variables with constant derivatives. Consequently, the system can be reformulated as follows:

$$\dot{P}_i(t) = k_i(Q_{i+1}(t) - Q_i(t) - l_i) - k_{i-1}(Q_i(t) - Q_{i-1}(t) - l_{i-1}) - b_i P_i(t)/m_i + \sum_j \alpha_{ij} F_j(t)$$
$$\dot{Q}_i(t) = P_i(t)/m_i$$
$$\dot{F}_j(t) = c_j \tag{B.2}$$

where α_{ij} is a constant determining the effect of the external force $F_j(t)$ on the i-th mass, and c_j is the constant rate of change of the external force $F_j(t)$. This model aligns with our ODE system (2), and an illustrative causal structure for this model is provided in Figure B.1.

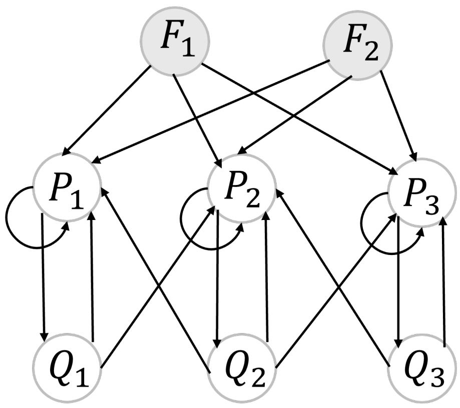

FIGURE B.1: Example causal structure of the damped harmonic oscillators model with 3 oscillators and 2 latent variables.

In regions with predictable wind patterns, such as during monsoon seasons or in controlled experimental settings, wind force can be approximated with a constant rate, making this an ideal context for modeling external forces with constant derivatives. Furthermore, constant forces or those represented as polynomial functions of time align well with our ODE system structure. For instance, a uniform magnetic field acting on the system would produce a constant force. These examples demonstrate that various latent factors can effectively fit within our ODE structure.

B.2.2 Population model

The growth of a population $P(t)$ can be described by a linear ODE [76]:

$$\dot{P}(t) = aP(t),$$

where a is a constant representing the population growth rate. This system may also be influenced by latent variables L_i, such as environmental factors or food supply. By incorporating these latent influences, the system can be expressed as:

$$\begin{aligned}
\dot{P}(t) &= aP(t) + bL_1(t) + cL_2(t) \\
\dot{L_1}(t) &= lL_2(t) \\
\dot{L_2}(t) &= m
\end{aligned}$$

where a, b, c, l and m are constants. Here, $L_1(t)$ represents the food supply, which is influenced by the environmental factor $L_2(t)$. $L_2(t)$ corresponds to an environmental factor, such as temperature or pollution level, that changes steadily over time. This model aligns well with our ODE system (3), and an illustrative causal structure for this model is provided in Figure B.2.

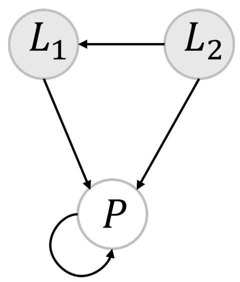

FIGURE B.2: Causal structure of the population model.

An example of an environmental factor changing at a constant rate is pollution from an industrial plant that continuously releases a fixed amount of pollutants, or from a wastewater treatment plant that discharges a specified amount of treated wastewater into a river on an hourly basis.

B.3 An example of an unidentifiable case of the linear ODE system (4.3)

Recall that the parameters of the ODE system (4.3) are:

$$\boldsymbol{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{z}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix},$$

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A' = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad M = \begin{bmatrix} A & B \\ \mathbf{0} & G \end{bmatrix}, \quad M' = \begin{bmatrix} A' & B \\ \mathbf{0} & G \end{bmatrix}.$$

We first calculate the solution of $\boldsymbol{z}(t)$,

$$\begin{aligned} \boldsymbol{z}(t) &= e^{Gt}\boldsymbol{z}_0 \\ &= \sum_{k=0}^{\infty} \frac{G^k \boldsymbol{z}_0}{k!} t^k = \sum_{k=0}^{1} \frac{G^k \boldsymbol{z}_0}{k!} t^k = \begin{bmatrix} 1+t \\ 1 \end{bmatrix} \end{aligned}$$

We intervene $x_1(t) = 1$, then under matrix M:

$$\begin{aligned} \dot{x}_2(t) &= x_2(t) + z_1(t) + z_2(t) \\ &= x_2(t) + t + 2\,. \end{aligned}$$

To solve this differential equation, we rewrite it in the standard linear form and multiply both sides by the integrating factor e^{-t},

$$e^{-t}\dot{x}_2(t) - e^{-t}x_2(t) = (t+2)e^{-t}\,.$$

The left-hand side of this equation is the derivative of $e^{-t}x_2(t)$:

$$\frac{d}{dt}(e^{-t}x_2(t)) = (t+2)e^{-t}\,.$$

Next, integrate both sides w.r.t. t:

$$\int \frac{d}{dt}(e^{-t}x_2(t))dt = \int (t+2)e^{-t}dt\,.$$

The left-hand side integrates to:

$$e^{-t}x_2(t)\,.$$

Next, we use integration by parts to find the integral on the right-hand side:

$$\begin{aligned}\int (t+2)e^{-t}dt &= -(t+2)e^{-t} - \int -e^{-t}dt \\ &= -(t+2)e^{-t} - e^{-t} \\ &= -(t+3)e^{-t}\,.\end{aligned}$$

Thus:

$$e^{-t}x_2(t) = -(t+3)e^{-t} + C\,,$$

where C is the constant of the integration.

Multiplying both sides by e^t to solve for $x_2(t)$:

$$x_2(t) = -t - 3 + Ce^t\,.$$

Now, use the initial condition $x_2(0) = 1$, we get

$$C = 4\,.$$

Therefore,

$$x_2(t) = 4e^t - t - 3\,.$$

Whereas under matrix M':

$$\begin{aligned}\dot{x}_2(t) &= x_1(t) + z_1(t) + z_2(t) \\ &= t + 3\,.\end{aligned}$$

Simple calculations show that

$$\boldsymbol{x}_2(t) = t^2/2 + 3t + 1\,.$$

B.4 Detailed proofs

B.4.1 Proof of Theorem 4.4

Proof. Recall that the first derivative of $\boldsymbol{x}(t)$ can be expressed as:

$$\begin{aligned}\dot{\boldsymbol{x}}(t) &= A\boldsymbol{x}(t) + B\boldsymbol{z}(t) \\ &= A\boldsymbol{x}(t) + \sum_{k=0}^{r} \frac{B\boldsymbol{v}_k}{k+1} t^{k+1} + B\boldsymbol{z}_0\,.\end{aligned}$$

Set

$$\boldsymbol{y}(t) = \begin{bmatrix} \boldsymbol{x}(t) \\ 1 \\ t \\ t^2 \\ \vdots \\ t^{r+1} \end{bmatrix},$$

we see that $\boldsymbol{y}(t) \in \mathbb{R}^{d+r+2}$, and the first derivative of $\boldsymbol{y}(t)$ w.r.t. time t can be expressed as

$$\begin{aligned}\dot{\boldsymbol{y}}(t) &= \begin{bmatrix} \dot{\boldsymbol{x}}(t) \\ 0 \\ 1 \\ 2t \\ \vdots \\ (r+1)t^r \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} A & B\boldsymbol{z}_0 & B\boldsymbol{v}_0 & \frac{B\boldsymbol{v}_1}{2} & \dots & \frac{B\boldsymbol{v}_{r-1}}{r} & \frac{B\boldsymbol{v}_r}{r+1} \\ \mathbf{0}_d & 0 & 0 & 0 & \dots & 0 & 0 \\ \mathbf{0}_d & 1 & 0 & 0 & \dots & 0 & 0 \\ \mathbf{0}_d & 0 & 2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_d & 0 & 0 & 0 & \dots & r+1 & 0 \end{bmatrix}}_{\text{denoted as } F} \underbrace{\begin{bmatrix} \boldsymbol{x}(t) \\ 1 \\ t \\ t^2 \\ \vdots \\ t^{r+1} \end{bmatrix}}_{\boldsymbol{y}(t)},\end{aligned}$$

where $\mathbf{0}_d$ denotes a d dimensional zero row vector. Obviously,

$$\boldsymbol{y}(0) = [\boldsymbol{x}_0^T, 1, 0, 0, \ldots, 0]^\top ,$$

we denote it as $\boldsymbol{y}_0$. Therefore, $\boldsymbol{y}(t)$ follows a homogeneous linear ODE system that can be expressed as:

$$\begin{aligned} \dot{\boldsymbol{y}}(t) &= F\boldsymbol{y}(t) \,, \\ \boldsymbol{y}(0) &= \boldsymbol{y}_0 \,, \end{aligned} \tag{B.3}$$

where $F \in \mathbb{R}^{(d+r+2)\times(d+r+2)}$. Worth noting that all state variables in the ODE system (B.3) are observable. Then according to Lemma 4.2, the identifiability of the dynamical system described by the ODE system (B.3) is contingent upon the linear independence of the vectors $\{\boldsymbol{y}_0, F\boldsymbol{y}_0, F^2\boldsymbol{y}_0, \ldots, F^{d+r+1}\boldsymbol{y}_0\}$. Specifically, the system is $(\boldsymbol{y}_0, F)$-identifiable if and only if this set of vectors is linearly independent, indicating that the matrix formed by these vectors, denoted by H, has a rank of $d+r+2$. In the following, we will elucidate that if and only assumption **A1** is satisfied, the rank of this matrix H equals $d+r+2$.

Some calculations show that,

$$F^k\boldsymbol{y}_0 = \begin{bmatrix} A^{k-1}(A\boldsymbol{x}_0 + B\boldsymbol{z}_0) + \sum_{j=0}^{k-2} j! A^{k-2-j} B\boldsymbol{v}_j \\ 0 \\ \vdots \\ 0 \\ k! \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{for } k = 1, 2, \ldots, r+1 \,, \tag{B.4}$$

where $k!$ is the $(d+k+1)$-th element.

And

$$F^k \boldsymbol{y}_0 = \begin{bmatrix} A^{k-(r+2)}(A^{r+2}\boldsymbol{x}_0 + A^{r+1}B\boldsymbol{z}_0 + \sum_{j=0}^{r} j!A^{r-j}B\boldsymbol{v}_j) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{for } k = r+2, \ldots, r+d+1\,. \tag{B.5}$$

According to assumption **A1** in Theorem 4.4,

$$\boldsymbol{\beta} = A^{r+2}\boldsymbol{x}_0 + A^{r+1}B\boldsymbol{z}_0 + \sum_{j=0}^{r} j!A^{r-j}B\boldsymbol{v}_j\,,$$

therefore, $F^k \boldsymbol{y}_0$ can also be expressed as

$$F^k \boldsymbol{y}_0 = \begin{bmatrix} A^{k-(r+2)}\boldsymbol{\beta} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{for } k = r+2, \ldots, r+d+1\,. \tag{B.6}$$

We denote the matrix

$$\begin{aligned} H :&= \begin{bmatrix} \boldsymbol{y}_0 & F\boldsymbol{y}_0 & F^2\boldsymbol{y}_0 & \ldots & F^{r+1}\boldsymbol{y}_0 & F^{r+2}\boldsymbol{y}_0 & \ldots & F^{d+r+1}\boldsymbol{y}_0 \end{bmatrix} \\ :&= \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \end{aligned}$$

as a block matrix. Then, based on Equations (B.4) and (B.6), one obtains that

$$
\begin{aligned}
H_{11} &= \begin{bmatrix} \boldsymbol{x}_0 & A\boldsymbol{x}_0 + B\boldsymbol{z}_0 & A^2\boldsymbol{x}_0 + AB\boldsymbol{z}_0 + B\boldsymbol{v}_0 & \dots & A^{r+1}\boldsymbol{x}_0 + A^r B\boldsymbol{z}_0 + \sum_{j=0}^{r-1} j! A^{r-1-j} B\boldsymbol{v}_j \end{bmatrix} \\
&\in \mathbb{R}^{d\times(r+2)}\,, \\
H_{12} &= \begin{bmatrix} \boldsymbol{\beta} & A\boldsymbol{\beta} & \dots & A^{d-1}\boldsymbol{\beta} \end{bmatrix} \\
&\in \mathbb{R}^{d\times d}\,, \\
H_{21} &= \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 2! & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & (r+1)! \end{bmatrix} \in \mathbb{R}^{(r+2)\times(r+2)}\,, \\
H_{22} &= \mathbf{0}_{(r+2)\times d} \in \mathbb{R}^{(r+2)\times d}\,.
\end{aligned}
$$

Some calculations show that

$$\operatorname{rank}(H) = \operatorname{rank}(H_{12}) + \operatorname{rank}(H_{21})\,.$$

It is apparent that

$$\operatorname{rank}(H_{21}) = r + 2\,.$$

To achieve $\operatorname{rank}(H) = d + r + 2$, the rank of H_{12} must be d. The rank of H_{12} equals d if and only if the set of vectors $\{\boldsymbol{\beta}, A\boldsymbol{\beta}, \dots, A^{d-1}\boldsymbol{\beta}\}$ is linearly independent, that is, assumption **A1** is satisfied.

Now that we have proved that the ODE system (B.3) is $(\boldsymbol{y}_0, F)$-identifiable if and only if assumption **A1** is satisfied. That is, under assumption **A1**, the trajectory $\boldsymbol{y}(\cdot; \boldsymbol{y}_0, F)$ uniquely determines both $\boldsymbol{y}_0$ and matrix F. Consequently, it also uniquely determines $(\boldsymbol{x}_0, A, B\boldsymbol{z}_0, B\boldsymbol{v}_0, \dots, B\boldsymbol{v}_r)$, thus establishing that the ODE system (4.2) is $\boldsymbol{\theta}$-identifiable if and only if assumption **A1** is satisfied. $\square$

B.4.2 Proof of Theorem 4.6

Proof. Recall that the first derivative of $\boldsymbol{x}(t)$ can be expressed as:

$$\begin{aligned}\dot{\boldsymbol{x}}(t) &= A\boldsymbol{x}(t) + B\boldsymbol{z}(t) \\ &= A\boldsymbol{x}(t) + \sum_{k=0}^{p-1} \frac{BG^k \boldsymbol{z}_0}{k!} t^k .\end{aligned}$$

Set

$$\boldsymbol{y}(t) = \begin{bmatrix} \boldsymbol{x}(t) \\ 1 \\ t \\ t^2 \\ \vdots \\ t^{p-1} \end{bmatrix},$$

we see that $\boldsymbol{y}(t) \in \mathbb{R}^{d+p}$, and the first derivative of $\boldsymbol{y}(t)$ w.r.t. time t can be expressed as

$$\begin{aligned}\dot{\boldsymbol{y}}(t) &= \begin{bmatrix} \dot{\boldsymbol{x}}(t) \\ 0 \\ 1 \\ 2t \\ \vdots \\ (p-1)t^{p-2} \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} A & B\boldsymbol{z}_0 & BG\boldsymbol{z}_0 & \frac{BG^2\boldsymbol{z}_0}{2!} & \dots & \frac{BG^{p-2}\boldsymbol{z}_0}{(p-2)!} & \frac{BG^{p-1}\boldsymbol{z}_0}{(p-1)!} \\ \mathbf{0}_d & 0 & 0 & 0 & \dots & 0 & 0 \\ \mathbf{0}_d & 1 & 0 & 0 & \dots & 0 & 0 \\ \mathbf{0}_d & 0 & 2 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_d & 0 & 0 & 0 & \dots & p-1 & 0 \end{bmatrix}}_{\text{denoted as } F} \underbrace{\begin{bmatrix} \boldsymbol{x}(t) \\ 1 \\ t \\ t^2 \\ \vdots \\ t^{p-1} \end{bmatrix}}_{\boldsymbol{y}(t)},\end{aligned}$$

where $\mathbf{0}_d$ denotes a d dimensional zero row vector. Obviously,

$$\boldsymbol{y}(0) = [\boldsymbol{x}_0^T, 1, 0, 0, \dots, 0]^\top ,$$

we denote it as $\boldsymbol{y}_0$. Therefore, $\boldsymbol{y}(t)$ follows a homogeneous linear ODE system that can be expressed as:

$$\begin{aligned} \dot{\boldsymbol{y}}(t) &= F\boldsymbol{y}(t)\,, \\ \boldsymbol{y}(0) &= \boldsymbol{y}_0\,, \end{aligned} \tag{B.7}$$

where $F \in \mathbb{R}^{(d+p)\times(d+p)}$. Worth noting that all state variables in the ODE system (B.7) are observable. Then according to Lemma 4.2, the identifiability of the dynamical system described by the ODE system (B.7) is contingent upon the linear independence of the vectors $\{\boldsymbol{y}_0, F\boldsymbol{y}_0, F^2\boldsymbol{y}_0, \dots, F^{d+p-1}\boldsymbol{y}_0\}$. Specifically, the system is $(\boldsymbol{y}_0, F)$-identifiable if and only if this set of vectors is linearly independent, indicating that the matrix formed by these vectors, denoted by H, has a rank of $d+p$. In the following, we will elucidate that if and only assumption **B1** is satisfied, the rank of this matrix H equals $d+p$.

Some calculations show that,

$$F^k\boldsymbol{y}_0 = \begin{bmatrix} A^k\boldsymbol{x}_0 + \sum_{j=0}^{k-1} A^{k-1-j}BG^j\boldsymbol{z}_0 \\ 0 \\ \vdots \\ 0 \\ k! \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{for } k = 1, 2, \dots, p-1\,, \tag{B.8}$$

where $k!$ is the $(d+k+1)$-th element.

And

$$F^k\boldsymbol{y}_0 = \begin{bmatrix} A^{k-p}(A^p\boldsymbol{x}_0 + \sum_{j=0}^{p-1} A^{p-1-j}BG^j\boldsymbol{z}_0) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{for } k = p, p+1, \dots, p+d-1\,. \tag{B.9}$$

According to assumption **B1** in Theorem 4.4,

$$\boldsymbol{\gamma} = A^p \boldsymbol{x}_0 + \sum_{j=0}^{p-1} A^{p-1-j} B G^j \boldsymbol{z}_0 \,,$$

therefore, $F^k \boldsymbol{y}_0$ can also be expressed as

$$F^k \boldsymbol{y}_0 = \begin{bmatrix} A^{k-p}\boldsymbol{\gamma} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{for } k = p, p+1, \dots, p+d-1\,. \tag{B.10}$$

We denote the matrix

$$\begin{aligned} H : &= \begin{bmatrix} \boldsymbol{y}_0 & F\boldsymbol{y}_0 & F^2\boldsymbol{y}_0 & \dots & F^{p-1}\boldsymbol{y}_0 & F^p\boldsymbol{y}_0 & \dots & F^{p+d-1}\boldsymbol{y}_0 \end{bmatrix} \\ : &= \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \end{aligned}$$

as a block matrix. Then, based on Equations (B.8) and (B.10), one obtains that

$$\begin{aligned} H_{11} &= \begin{bmatrix} \boldsymbol{x}_0 & A\boldsymbol{x}_0 + B\boldsymbol{z}_0 & A^2\boldsymbol{x}_0 + AB\boldsymbol{z}_0 + BG\boldsymbol{z}_0 & \dots & A^{p-1}\boldsymbol{x}_0 + \sum_{j=0}^{p-2} A^{p-2-j}BG^j\boldsymbol{z}_0 \end{bmatrix} \\ &\in \mathbb{R}^{d\times p}\,, \\ H_{12} &= \begin{bmatrix} \boldsymbol{\gamma} & A\boldsymbol{\gamma} & \dots & A^{d-1}\boldsymbol{\gamma} \end{bmatrix} \\ &\in \mathbb{R}^{d\times d}\,, \\ H_{21} &= \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 2! & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & (p-1)! \end{bmatrix} \in \mathbb{R}^{p\times p}\,, \\ H_{22} &= \mathbf{0}_{p\times d} \in \mathbb{R}^{p\times d}\,. \end{aligned}$$

Some calculations show that

$$\operatorname{rank}(H) = \operatorname{rank}(H_{12}) + \operatorname{rank}(H_{21})\,.$$

It is apparent that

$$\mathrm{rank}(H_{21}) = p\,.$$

To achieve $\mathrm{rank}(H) = d + p$, the rank of H_{12} must be d. The rank of H_{12} equals d if and only if the set of vectors $\{\boldsymbol{\gamma}, A\boldsymbol{\gamma}, \ldots, A^{d-1}\boldsymbol{\gamma}\}$ is linearly independent, that is, assumption **B1** is satisfied.

Now that we have proved that the ODE system (B.7) is $(\boldsymbol{y}_0, F)$-identifiable if and only if assumption **B1** is satisfied. That is, under assumption **B1**, the trajectory $\boldsymbol{y}(\cdot; \boldsymbol{y}_0, F)$ uniquely determines both $\boldsymbol{y}_0$ and the matrix F. Consequently, it also uniquely determines $(\boldsymbol{x}_0, A, B\boldsymbol{z}_0, BG\boldsymbol{z}_0, \ldots, BG^{p-1}\boldsymbol{z}_0)$, thus establishing that the ODE system (4.3) is $\boldsymbol{\eta}$-identifiable if and only if assumption **B1** is satisfied. □

B.4.3 Proof of Theorem 4.8

Before providing the main proof, we first present two lemmas we will use for our proof.

Lemma B.1. *[28, Theorem 3.4] The ODE system (4.1) is $(\boldsymbol{x}_0, A)$-identifiable if and only if the trajectory $\boldsymbol{x}(\cdot; \boldsymbol{x}_0, A)$ is not confined to a proper subspace of $\mathbb{R}^d$.*

Lemma B.2. *[28, Lemma 6.1] Trajectory $\boldsymbol{x}(\cdot; \boldsymbol{x}_0, A)$ is not confined to a proper subspace of $\mathbb{R}^d$ if and only if there exists $t_1, t_2, \ldots, t_d$ such that $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_d$ are linearly independent.*

Proof. In the proof of Theorem 4.6, we demonstrated that the ODE system (4.3), under latent DAG assumption, can be transformed into a fully observable homogeneous linear ODE system (B.7). According to Lemma B.1, the ODE system (B.7) is $(\boldsymbol{y}_0, F)$-identifiable if and only if trajectory $\boldsymbol{y}(\cdot; \boldsymbol{y}_0, F)$ is not confined to a proper subspace of $\mathbb{R}^{d+p}$. Furthermore, based on Lemma B.2, this condition holds if and only if there exists time points $t_1, t_2, \ldots, t_{d+p}$ such that the vectors $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{d+p}$ are linearly independent (i.e., assumption **C1**). Therefore, if and only if assumption **C1** is satisfied, the trajectory $\boldsymbol{y}(\cdot; \boldsymbol{y}_0, F)$ is not confined to a proper subspace of $\mathbb{R}^{d+p}$, ensuring that the ODE system (B.7) is $(\boldsymbol{y}_0, F)$-identifiable. Consequently, the ODE system (4.3) is $\boldsymbol{\eta}$-identifiable. □

B.4.4 Proof of Theorem 4.10

Proof. Under assumption **B2**, since each $\boldsymbol{z}_0^{*i}$ satisfies assumption **B1**, Theorem 4.6 implies that the ODE system (4.3) is $\boldsymbol{\eta}_i$-identifiable for all $i = 1, \ldots, p$. That is, one can identify

$$(\boldsymbol{x}_0, A, B\boldsymbol{z}_0^{*i}, BG\boldsymbol{z}_0^{*i}, \ldots, BG^{p-1}\boldsymbol{z}_0^{*i})$$

for all $i = 1, \ldots, p$.

Next, we will prove that matrix B is identifiable under assumption **B3**.

Define the matrix

$$S := \begin{bmatrix} B\boldsymbol{z}_0^{*1} & B\boldsymbol{z}_0^{*2} & \ldots & B\boldsymbol{z}_0^{*p} \end{bmatrix} ,$$

we know that $S \in \mathbb{R}^{d \times p}$, and S is identifiable. The matrix S can also be expressed as:

$$\begin{aligned} S &= B \begin{bmatrix} \boldsymbol{z}_0^{*1} & \boldsymbol{z}_0^{*2} & \ldots & \boldsymbol{z}_0^{*p} \end{bmatrix} \\ &:= BZ , \end{aligned}$$

where under assumption **B3**, the matrix Z is invertible. Therefore,

$$B = SZ^{-1} .$$

Since Z is a known matrix, B is identifiable.

Similarly, we can prove that BG^j for $j = 1, \ldots, p-1$ is also identifiable.

We now show that, under assumption **B4**, the matrix G is identifiable.

Define the matrix

$$W := \begin{bmatrix} B \\ BG \\ \vdots \\ BG^{p-1} \end{bmatrix} ,$$

we know that $W \in \mathbb{R}^{dp \times p}$, and W is identifiable.

Since G is a $p \times p$ nilpotent matrix, $G^p = \mathbf{0}$, thus $BG^p = \mathbf{0}$. If we define the matrix

$$V := \begin{bmatrix} BG \\ BG^2 \\ \vdots \\ BG^p \end{bmatrix},$$

then $V \in \mathbb{R}^{dp \times p}$, and V is identifiable. The matrix V can also be expressed as:

$$V = \begin{bmatrix} B \\ BG \\ \vdots \\ BG^{p-1} \end{bmatrix} G = WG\,. \tag{B.11}$$

Under assumption **B4**, one can find p linearly independent rows in matrix W. Denote the matrix composed of these p linearly independent rows as W_p, which is invertible. Denote the matrix composed of the corresponding p rows of V as V_p, we have

$$V_p = W_p G\,.$$

Since W_p is invertible, then

$$G = W_p^{-1} V_p\,.$$

Because both V_p and W_p are identifiable, G is also identifiable. □

B.4.5 Proof of Theorem 4.12

Proof. Under assumption **C2**, for each $i \in \{1, \ldots, p\}$, the corresponding observations satisfy assumption **C1**. Based on Theorem 4.8, the ODE system (4.3) is $\boldsymbol{\eta}_i$-identifiable for all $i = 1, \ldots, p$. This implies that one can identify

$$(\boldsymbol{x}_0, A, B\boldsymbol{z}_0^{*i}, BG\boldsymbol{z}_0^{*i}, \ldots, BG^{p-1}\boldsymbol{z}_0^{*i})$$

for all $i = 1, \ldots, p$.

According to the proof of Theorem 4.10, under assumptions **B3** and **B4**, matrices B and G are also identifiable. □

B.5 Identifiability conditions of the linear ODE system (4.2) with other *f(t)*

In this section, we provide identifiability conditions for the linear ODE system (4.2) with $\boldsymbol{f}(t) = \boldsymbol{v}e^t$ and $\boldsymbol{f}(t) = \boldsymbol{v}_1 sin(t) + \boldsymbol{v}_2 cos(t)$. For notational simplicity, we slightly abuse notation by using the same symbols as in Section 4.3.

B.5.1 When *f(t)* follows an exponenial function of time *t*

We define $\boldsymbol{f}(t)$ in the ODE system (4.2) as:

$$\boldsymbol{f}(t) = \boldsymbol{v}e^t\,, \quad \boldsymbol{v} \in \mathbb{R}^p\,.$$

Simple calculations show that

$$\boldsymbol{z}(t) = \boldsymbol{v}e^t + \boldsymbol{z}_0 - \boldsymbol{v}\,.$$

Thus,

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= A\boldsymbol{x}(t) + B\boldsymbol{z}(t) \\ &= A\boldsymbol{x}(t) + B\boldsymbol{v}e^t + B\boldsymbol{z}_0 - B\boldsymbol{v}\,. \end{aligned} \tag{B.12}$$

We denote the unknown parameters of the ODE system (4.2) with this $\boldsymbol{f}(t)$ as $\boldsymbol{\theta}$, specifically, $\boldsymbol{\theta} := (\boldsymbol{x}_0, \boldsymbol{z}_0, A, B, \boldsymbol{v})$. Let $[\boldsymbol{x}^T(t;\boldsymbol{\theta}), \boldsymbol{z}^T(t;\boldsymbol{\theta})]^T$ denote the solution of the ODE system (4.2). It is important to note that under our hidden variables setting, only $\boldsymbol{x}(t;\boldsymbol{\theta})$ is observable. Based on Equation (B.12), we present the following identifiability definition.

Definition B.3. For $\boldsymbol{x}_0 \in \mathbb{R}^d, \boldsymbol{z}_0 \in \mathbb{R}^p, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$ and $\boldsymbol{v} \in \mathbb{R}^p$, for all $\boldsymbol{x}_0' \in \mathbb{R}^d$, all $\boldsymbol{z}_0' \in \mathbb{R}^p$, all $A' \in \mathbb{R}^{d\times d}$, all $B' \in \mathbb{R}^{d\times p}$, and all $\boldsymbol{v}' \in \mathbb{R}^p$, we denote $\boldsymbol{\theta}' := (\boldsymbol{x}_0', \boldsymbol{z}_0', A', B', \boldsymbol{v}')$, we say the ODE system (4.2) is $\boldsymbol{\theta}$-identifiable: if $(\boldsymbol{x}_0, A, B\boldsymbol{z}_0, B\boldsymbol{v}) \neq (\boldsymbol{x}_0', A', B'\boldsymbol{z}_0', B'\boldsymbol{v}')$, it holds that $\boldsymbol{x}(\cdot;\boldsymbol{\theta}) \neq \boldsymbol{x}(\cdot;\boldsymbol{\theta}')$.

According to Definition B.3, if the ODE system (4.2) with an exponential $\boldsymbol{f}(t)$ is $\boldsymbol{\theta}$-identifiable, then the trajectory of the system can uniquely determine the values of

$(\boldsymbol{x}_0, A, B\boldsymbol{z}_0, B\boldsymbol{v})$. This determination is sufficient to identify the causal relationships between observable variables $\boldsymbol{x}$ as described by Equation (B.12). Consequently, one can safely intervene in the observable variables of the ODE system and make reliable causal inferences, despite the fact that matrix B cannot be identified under this definition.

Theorem B.4. *For $\boldsymbol{x}_0 \in \mathbb{R}^d, \boldsymbol{z}_0 \in \mathbb{R}^p, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$, and $\boldsymbol{v} \in \mathbb{R}^p$, the ODE system (4.2) is $\boldsymbol{\theta}$-identifiable if and only if assumption **D1** is satisfied.*

D1 *the set of vectors $\{\boldsymbol{y}_0, F\boldsymbol{y}_0, \ldots, F^{d+1}\boldsymbol{y}_0\}$ is linearly independent, where $\boldsymbol{y}_0 = [\boldsymbol{x}_0^T, 1, 1]^T$, and*

$$F = \begin{bmatrix} A & B\boldsymbol{v} & B\boldsymbol{z}_0 - B\boldsymbol{v} \\ \mathbf{0}_d & 1 & 0 \\ \mathbf{0}_d & 0 & 0 \end{bmatrix},$$

$\mathbf{0}_d$ denotes a d dimensional zero row vector.

The proof of Theorem B.4 is presented below. Condition **D1** is both sufficient and necessary, indicating, from a geometric perspective, that the vector $\boldsymbol{y}_0$ is not contained in an F-invariant proper subspace of $\mathbb{R}^{d+2}$.

Proof. Set

$$\boldsymbol{y}(t) = \begin{bmatrix} \boldsymbol{x}(t) \\ e^t \\ 1 \end{bmatrix},$$

we see that $\boldsymbol{y}(t) \in \mathbb{R}^{d+2}$, and the first derivative of $\boldsymbol{y}(t)$ w.r.t. time t can be expressed as

$$\dot{\boldsymbol{y}}(t) = \begin{bmatrix} \dot{\boldsymbol{x}}(t) \\ e^t \\ 0 \end{bmatrix} = \underbrace{\begin{bmatrix} A & B\boldsymbol{v} & B\boldsymbol{z}_0 - B\boldsymbol{v} \\ \mathbf{0}_d & 1 & 0 \\ \mathbf{0}_d & 0 & 0 \end{bmatrix}}_{F} \underbrace{\begin{bmatrix} \boldsymbol{x}(t) \\ e^t \\ 1 \end{bmatrix}}_{\boldsymbol{y}(t)},$$

where $\mathbf{0}_d$ denotes a d dimensional zero row vector. Obviously,

$$\boldsymbol{y}(0) = [\boldsymbol{x}_0^T, 1, 1]^T = \boldsymbol{y}_0 .$$

Therefore, $\boldsymbol{y}(t)$ follows a homogeneous linear ODE system that can be expressed as:

$$\begin{aligned} \dot{\boldsymbol{y}}(t) &= F\boldsymbol{y}(t)\,, \\ \boldsymbol{y}(0) &= \boldsymbol{y}_0\,, \end{aligned} \tag{B.13}$$

where $F \in \mathbb{R}^{(d+2)\times(d+2)}$. Worth noting that all state variables in the ODE system (B.13) are observable. Then according to Lemma 4.2, the system (B.13) is $(\boldsymbol{y}_0, F)$-identifiable if and only if condition **D1** stated in Theorem B.4 is satisfied. That is, under assumption **D1**, the trajectory $\boldsymbol{y}(\cdot;\boldsymbol{y}_0, F)$ uniquely determines both $\boldsymbol{y}_0$ and matrix F. Consequently, it also uniquely determines $(\boldsymbol{x}_0, A, B\boldsymbol{z}_0, B\boldsymbol{v})$, thus establishing that the ODE system (4.2) is $\boldsymbol{\theta}$-identifiable if and only if assumption **D1** is satisfied. □

B.5.2 When *f(t)* follows an trigonometric function of time *t*

We define $\boldsymbol{f}(t)$ in the ODE system (4.2) as:

$$\boldsymbol{f}(t) = \boldsymbol{v}_1 sin(t) + \boldsymbol{v}_2 cos(t)\,, \quad \boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^p\,.$$

Simple calculations show that

$$\boldsymbol{z}(t) = \boldsymbol{v}_2 sin(t) - \boldsymbol{v}_1 cos(t) + \boldsymbol{z}_0 + \boldsymbol{v}_1\,.$$

Thus,

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= A\boldsymbol{x}(t) + B\boldsymbol{z}(t) \\ &= A\boldsymbol{x}(t) + B\boldsymbol{v}_2 sin(t) - B\boldsymbol{v}_1 cos(t) + B\boldsymbol{z}_0 + B\boldsymbol{v}_1\,. \end{aligned} \tag{B.14}$$

We denote the unknown parameters of the ODE system (4.2) with this $\boldsymbol{f}(t)$ as $\boldsymbol{\theta}$, specifically, $\boldsymbol{\theta} := (\boldsymbol{x}_0, \boldsymbol{z}_0, A, B, \boldsymbol{v}_1, \boldsymbol{v}_2)$. Let $[\boldsymbol{x}^T(t;\boldsymbol{\theta}), \boldsymbol{z}^T(t;\boldsymbol{\theta})]^T$ denote the solution of the ODE system (4.2). It is important to note that under our hidden variables setting, only $\boldsymbol{x}(t;\boldsymbol{\theta})$ is observable. Based on Equation (B.14), we present the following identifiability definition.

Definition B.5. For $\boldsymbol{x}_0 \in \mathbb{R}^d, \boldsymbol{z}_0 \in \mathbb{R}^p, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$ and $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^p$, for all $\boldsymbol{x}_0' \in \mathbb{R}^d$, all $\boldsymbol{z}_0' \in \mathbb{R}^p$, all $A' \in \mathbb{R}^{d\times d}$, all $B' \in \mathbb{R}^{d\times p}$, and all $\boldsymbol{v}_1', \boldsymbol{v}_2' \in \mathbb{R}^p$, we denote $\boldsymbol{\theta}' := (\boldsymbol{x}_0', \boldsymbol{z}_0', A', B', \boldsymbol{v}_1', \boldsymbol{v}_2')$, we say the ODE system (4.2) is $\boldsymbol{\theta}$-identifiable: if $(\boldsymbol{x}_0, A, B\boldsymbol{z}_0, B\boldsymbol{v}_1, B\boldsymbol{v}_2) \neq (\boldsymbol{x}_0', A', B'\boldsymbol{z}_0', B'\boldsymbol{v}_1', B'\boldsymbol{v}_2')$, it holds that $\boldsymbol{x}(\cdot; \boldsymbol{\theta}) \neq \boldsymbol{x}(\cdot; \boldsymbol{\theta}')$.

According to Definition B.5, if the ODE system (4.2) with a trigonometric $\boldsymbol{f}(t)$ is $\boldsymbol{\theta}$-identifiable, then the trajectory of the system can uniquely determine the values of $(\boldsymbol{x}_0, A, B\boldsymbol{z}_0, B\boldsymbol{v}_1, B\boldsymbol{v}_2)$. This determination is sufficient to identify the causal relationships between observable variables $\boldsymbol{x}$ as described by Equation (B.14). Consequently, one can safely intervene in the observable variables of the ODE system and make reliable causal inferences, despite the fact that matrix B cannot be identified under this definition.

Theorem B.6. *For $\boldsymbol{x}_0 \in \mathbb{R}^d, \boldsymbol{z}_0 \in \mathbb{R}^p, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$, and $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^p$, the ODE system (4.2) is $\boldsymbol{\theta}$-identifiable if and only if assumption **E1** is satisfied.*

E1 *the set of vectors $\{\boldsymbol{y}_0, F\boldsymbol{y}_0, \ldots, F^{d+2}\boldsymbol{y}_0\}$ is linearly independent, where $\boldsymbol{y}_0 = [\boldsymbol{x}_0^T, 0, 1, 1]^T$, and*

$$F = \begin{bmatrix} A & B\boldsymbol{v}_2 & -B\boldsymbol{v}_1 & B\boldsymbol{z}_0 + B\boldsymbol{v}_1 \\ \mathbf{0}_d & 0 & 1 & 0 \\ \mathbf{0}_d & -1 & 0 & 0 \\ \mathbf{0}_d & 0 & 0 & 0 \end{bmatrix},$$

$\mathbf{0}_d$ denotes a d dimensional zero row vector.

The proof of Theorem B.6 is presented below. Condition **E1** is both sufficient and necessary, indicating, from a geometric perspective, that the vector $\boldsymbol{y}_0$ is not contained in an F-invariant proper subspace of $\mathbb{R}^{d+3}$.

Proof. Set

$$\boldsymbol{y}(t) = \begin{bmatrix} \boldsymbol{x}(t) \\ sin(t) \\ cos(t) \\ 1 \end{bmatrix},$$

we see that $\boldsymbol{y}(t) \in \mathbb{R}^{d+3}$, and the first derivative of $\boldsymbol{y}(t)$ w.r.t. time t can be expressed as

$$\dot{\boldsymbol{y}}(t) = \begin{bmatrix} \dot{\boldsymbol{x}}(t) \\ cos(t) \\ -sin(t) \\ 0 \end{bmatrix} = \underbrace{\begin{bmatrix} A & B\boldsymbol{v}_2 & -B\boldsymbol{v}_1 & B\boldsymbol{z}_0 + B\boldsymbol{v}_1 \\ \mathbf{0}_d & 0 & 1 & 0 \\ \mathbf{0}_d & -1 & 0 & 0 \\ \mathbf{0}_d & 0 & 0 & 0 \end{bmatrix}}_{F} \underbrace{\begin{bmatrix} \boldsymbol{x}(t) \\ sin(t) \\ cos(t) \\ 1 \end{bmatrix}}_{\boldsymbol{y}(t)},$$

where $\mathbf{0}_d$ denotes a d dimensional zero row vector. Obviously,

$$\boldsymbol{y}(0) = [\boldsymbol{x}_0^T, 0, 1, 1]^T = \boldsymbol{y}_0 \,.$$

Therefore, $\boldsymbol{y}(t)$ follows a homogeneous linear ODE system that can be expressed as:

$$\begin{aligned} \dot{\boldsymbol{y}}(t) &= F\boldsymbol{y}(t)\,, \\ \boldsymbol{y}(0) &= \boldsymbol{y}_0\,, \end{aligned} \tag{B.15}$$

where $F \in \mathbb{R}^{(d+3)\times(d+3)}$. Worth noting that all state variables in the ODE system (B.15) are observable. Then according to Lemma 4.2, the system (B.15) is $(\boldsymbol{y}_0, F)$-identifiable if and only if condition **E1** stated in Theorem B.6 is satisfied. That is, under assumption **E1**, the trajectory $\boldsymbol{y}(\cdot; \boldsymbol{y}_0, F)$ uniquely determines both $\boldsymbol{y}_0$ and matrix F. Consequently, it also uniquely determines $(\boldsymbol{x}_0, A, B\boldsymbol{z}_0, B\boldsymbol{v}_1, B\boldsymbol{v}_2)$, thus establishing that the ODE system (4.2) is $\boldsymbol{\theta}$-identifiable if and only if assumption **E1** is satisfied. □

B.6 An alternative approach to identifying matrices B and G in the ODE system (4.3)

B.6.1 Identifiability condition from *2p* controllable whole trajectories

Recall that $\boldsymbol{z}_0$ denotes the initial condition of the latent variables in the ODE system (4.3). We further specify the initial condition of the latent variable z_j as z_{0j} for $j = 1, \ldots, p$. Assume that it is possible to control the initial condition of each latent variable, z_{0j}, independently. Specifically, for each experiment, researchers can intervene in the initial condition of a latent variable, denoted as z_{0j}^*. The value of z_{0j}^* is treated as a given value. Under this intervention, the initial conditions of the latent variables are adjusted to $[z_{01}, \ldots, z_{0j}^{*}, \ldots, z_{0p}]^T$, which we denote as $\tilde{\boldsymbol{z}}_{0j}$.

To identify matrices B and G, it is necessary to have at least two intervened initial conditions for each latent variable, denoted as z_{0j}^{*1} and z_{0j}^{*2} for the latent variable z_j. Consequently, the corresponding intervened initial conditions for all latent variables can be represented as $\tilde{\boldsymbol{z}}_{0j}^1$ and $\tilde{\boldsymbol{z}}_{0j}^2$. Under these conditions, we present the definition of the identifiability of the ODE system (4.3).

Definition B.7. Given $z_{0j}^{*1}, z_{0j}^{*2} \in \mathbb{R}$ for $j = 1, \ldots, p$, for $\boldsymbol{x}_0 \in \mathbb{R}^d, \boldsymbol{z}_0 \in \mathbb{R}^p, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$ and $G \in \mathbb{R}^{p\times p}$, under the latent DAG assumption, for all $\boldsymbol{x}_0' \in \mathbb{R}^d$, all $\boldsymbol{z}_0' \in \mathbb{R}^p$, all $A' \in \mathbb{R}^{d\times d}$, all $B' \in \mathbb{R}^{d\times p}$, and all $G' \in \mathbb{R}^{p\times p}$, we denote $\tilde{\boldsymbol{z}}_{0j}^i = [z_{01}, \ldots, z_{0j}^{*i}, \ldots, z_{0p}]^T$ and $(\tilde{\boldsymbol{z}}_{0j}')^i = [z_{01}', \ldots, z_{0j}^{*i}, \ldots, z_{0p}']^T$, we further denote $\boldsymbol{\eta}_j^i := (\boldsymbol{x}_0, \tilde{\boldsymbol{z}}_{0j}^i, A, B, G)$ and $(\boldsymbol{\eta}_j')^i := (\boldsymbol{x}_0', (\tilde{\boldsymbol{z}}_{0j}')^i, A', B', G')$ for $i = 1, 2$, we say the ODE system (4.3) is $\{\boldsymbol{\eta}_j^{1,2}\}_1^p$-identifiable: if $(\boldsymbol{x}_0, A, B, G) \neq (\boldsymbol{x}', A', B', G')$, it holds that $\exists i \in \{1, 2\}$ and $j \in \{1, \ldots, p\}$ such that $\boldsymbol{x}(\cdot; \boldsymbol{\eta}_j^i) \neq \boldsymbol{x}(\cdot; (\boldsymbol{\eta}_j')^i)$.

Definition B.7 establishes the identifiability of the ODE system (4.3) from $2p$ whole trajectories $\boldsymbol{x}(\cdot; \boldsymbol{\eta}_j^i)$ with $i = 1, 2$ and $j = 1, \ldots, p$. According to this definition, both matrices B and G are identifiable. Based on this definition, we present the identifiability condition.

Theorem B.8. *Given $z_{0j}^{*1}, z_{0j}^{*2} \in \mathbb{R}$ with $z_{0j}^{*1} \neq z_{0j}^{*2}$ for $j = 1, \ldots, p$, for $\boldsymbol{x}_0 \in \mathbb{R}^d, \boldsymbol{z}_0 \in \mathbb{R}^p, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$ and $G \in \mathbb{R}^{p\times p}$, under the latent DAG assumption, the ODE system (4.3) is $\{\boldsymbol{\eta}_j^{1,2}\}_1^p$-identifiable if assumptions $\boldsymbol{B}_5$ and $\boldsymbol{B}_4$ are both satisfied.*

***B5**: each $\tilde{\boldsymbol{z}}_{0j}^i$ for $i = 1, 2$ and $j = 1, \ldots, p$, satisfies assumption **B1**. That is, if we set $\boldsymbol{\gamma}_j^i = A^p \boldsymbol{x}_0 + \sum_{k=0}^{p-1} A^{p-1-k} B G^k \tilde{\boldsymbol{z}}_{0j}^i$, then the set of vectors $\{\boldsymbol{\gamma}_j^i, A\boldsymbol{\gamma}_j^i, \ldots, A^{d-1}\boldsymbol{\gamma}_j^i\}$ is linearly independent for all $i = 1, 2$ and $j = 1, \ldots, p$.*

The proof of Theorem B.8 is presented below. Assumption **B5** ensures that the ODE system (4.3) is $\boldsymbol{\eta}_j^i$-identifiable for all $i = 1, 2$ and $j = 1, \ldots, p$. Consequently, $(\boldsymbol{x}_0, A, B\tilde{\boldsymbol{z}}_{0j}^i, BG\tilde{\boldsymbol{z}}_{0j}^i, \ldots, BG^{p-1}\tilde{\boldsymbol{z}}_{0j}^i)$ for all $i = 1, 2$ and $j = 1, \ldots, p$ is identifiable. Through straightforward calculations, the identifiability of matrix B is established. To identify matrix G, assumption **B4** is required.

The assumption that the initial condition of each latent variable z_i can be controlled independently is inspired by the "genetic single-node intervention" proposed in [119], where interventions can be made at each latent node individually. This assumption is relatively more relaxed compared to controlling the initial condition of all latent variables $\boldsymbol{z}$ simultaneously, as discussed in Subsection 4.4.3. However, this method requires p more trajectories, totalling $2p$ trajectories, to identify matrices B and G.

Proof. Under assumption **B5**, since each $\tilde{\boldsymbol{z}}_{0j}^i$ satisfies assumption **B1**. By Theorem 4.6, the ODE system (4.3) is $\boldsymbol{\eta}_j^i$-identifiable for all $i = 1, 2$ and $j = 1, \dots, p$. Consequently,

$$(\boldsymbol{x}_0, A, B\tilde{\boldsymbol{z}}_{0j}^i, BG\tilde{\boldsymbol{z}}_{0j}^i, \dots, BG^{p-1}\tilde{\boldsymbol{z}}_{0j}^i)$$

for all $i = 1, 2$ and $j = 1, \dots, p$ is identifiable.

We express $B\tilde{\boldsymbol{z}}_{0j}^i$ as

$$B\tilde{\boldsymbol{z}}_{0j}^i = \begin{bmatrix} B_{11} & \dots & B_{1j} & \dots & B_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ B_{d1} & \dots & B_{dj} & \dots & B_{dp} \end{bmatrix} \begin{bmatrix} z_{01} \\ \vdots \\ z_{0j}^{*i} \\ \vdots \\ z_{0p} \end{bmatrix}.$$

We know that $B\tilde{\boldsymbol{z}}_{0j}^i \in \mathbb{R}^d$ is identifiable for $i = 1, 2$. Thus, the first entry of $B\tilde{\boldsymbol{z}}_{0j}^i$, denoted as $(B\tilde{\boldsymbol{z}}_{0j}^i)_1$, is identifiable and can be expressed as

$$\begin{aligned} (B\tilde{\boldsymbol{z}}_{0j}^1)_1 &= B_{11}z_{01} + \dots + B_{1j}z_{0j}^{*1} + \dots + B_{1p}z_{0p} \\ (B\tilde{\boldsymbol{z}}_{0j}^2)_1 &= B_{11}z_{01} + \dots + B_{1j}z_{0j}^{*2} + \dots + B_{1p}z_{0p} \end{aligned}.$$

Since z_{0j}^{*1} and z_{0j}^{*2} are given values, we can easily calculate the value of B_{1j}. Similarly, one can calculate the values of B_{mj} for all $m = 1, \dots, d$ and $j = 1, \dots, p$, thereby establishing the identifiability of matrix B.

In a similar manner, matrices $BG, BG^2, \dots, BG^{p-1}$ are also identifiable. Then, according to the proof B.4.4 of Theorem 4.10, the matrix G is identifiable under assumption **B4**. □

B.6.2 Identifiability condition from discrete observations sampled from *2p* controllable trajectories

We further extend the identifiability analysis of the ODE system (4.3) to cases where only discrete observations from $2p$ controllable trajectories are available.

Definition B.9. Given $z_{0j}^{*1}, z_{0j}^{*2} \in \mathbb{R}$ for $j = 1, \ldots, p$, for $\boldsymbol{x}_0 \in \mathbb{R}^d, \boldsymbol{z}_0 \in \mathbb{R}^p, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$ and $G \in \mathbb{R}^{p\times p}$. For any $n \geqslant 1$, let $t_k, k = 1, \ldots, n$ be any n time points and $\boldsymbol{x}_{jk}^i := \boldsymbol{x}(t_k; \boldsymbol{\eta}_j^i)$ be the error-free observation of the trajectory $\boldsymbol{x}(\cdot; \boldsymbol{\eta}_j^i)$ at time t_k. Under the latent DAG assumption, we say the ODE system (4.3) is $\{\boldsymbol{\eta}_j^{1,2}\}_1^p$-identifiable from $\boldsymbol{x}_{j1}^i, \ldots, \boldsymbol{x}_{jn}^i$, $i = 1, 2$ and $j = 1, \ldots, p$, if for all $\boldsymbol{x}_0' \in \mathbb{R}^d$, all $\boldsymbol{z}_0' \in \mathbb{R}^p$, all $A' \in \mathbb{R}^{d\times d}$, all $B' \in \mathbb{R}^{d\times p}$, and all $G' \in \mathbb{R}^{p\times p}$ with $(\boldsymbol{x}_0, A, B, G) \neq (\boldsymbol{x}_0', A', B', G')$, it holds that $\exists i \in \{1, 2\}, j \in \{1, \ldots, p\}$ and $k \in \{1, \ldots, n\}$ such that $\boldsymbol{x}(t_k; \boldsymbol{\eta}_j^i) \neq \boldsymbol{x}(t_k; (\boldsymbol{\eta}_j')^i)$.

Based on Definition B.9 we present the identifiability condition.

Theorem B.10. *Given $z_{0j}^{*1}, z_{0j}^{*2} \in \mathbb{R}$ with $z_{0j}^{*1} \neq z_{0j}^{*2}$ for $j = 1, \ldots, p$, for $\boldsymbol{x}_0 \in \mathbb{R}^d, \boldsymbol{z}_0 \in \mathbb{R}^p, A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{d\times p}$ and $G \in \mathbb{R}^{p\times p}$. We define new observation $\boldsymbol{y}_{jk}^i := [(\boldsymbol{x}_{jk}^i)^T, 1, t_k, t_k^2, \ldots, t_k^{p-1}]^T \in \mathbb{R}^{d+p}$, for $i = 1, 2, j = 1, \ldots, p$ and $k = 1, \ldots, n$. Under the latent DAG assumption, the ODE system (4.3) is $\{\boldsymbol{\eta}_j^{1,2}\}_1^p$-identifiable from discrete observations $\boldsymbol{x}_{j1}^i, \ldots, \boldsymbol{x}_{jn}^i$, $i = 1, 2$ and $j = 1, \ldots, p$, if assumptions **C3** and **B4** are both satisfied.*

***C3**: for each $i \in \{1, 2\}, j \in \{1, \ldots, p\}$ there exists $(d + p)$ $\boldsymbol{y}_{jk}^i$'s with indexes denoting as $\{k_{j1}^i, k_{j2}^i, \ldots, k_{j,d+p}^i\} \subseteq \{1, 2, \ldots, n\}$, such that the set of vectors $\{\boldsymbol{y}_{jk_{j1}^i}^i, \boldsymbol{y}_{jk_{j2}^i}^i, \ldots, \boldsymbol{y}_{jk_{j,d+p}^i}^i\}$ is linearly independent.*

The proof of Theorem B.10 is presented below. Assumption **C3** ensures that the ODE system (4.3) is $\boldsymbol{\eta}_j^i$-identifiable from discrete observations $\boldsymbol{x}_{j1}^i, \ldots, \boldsymbol{x}_{jn}^i$ for all $i = 1, 2$ and $j = 1, \ldots, p$. As in Subsection B.6.1, matrix B is identifiable. Then, under assumption **B4**, matrix G is also identifiable.

Proof. Under assumption **C3**, for each $i \in \{1, 2\}$ and $j \in \{1, \ldots, p\}$, the corresponding observations satisfy assumption **C1**. Based on Theorem 4.8, the ODE system (4.3) is $\boldsymbol{\eta}_j^i$-identifiable for all $i = 1, 2$ and $j = 1, \ldots, p$. Consequently,

$$(\boldsymbol{x}_0, A, B\tilde{\boldsymbol{z}}_{0j}^i, BG\tilde{\boldsymbol{z}}_{0j}^i, \ldots, BG^{p-1}\tilde{\boldsymbol{z}}_{0j}^i)$$

for all $i = 1, 2$ and $j = 1, \dots, p$ is identifiable.

Following the proof of Theorem B.8, matrix B is identifiable. Under assumption **B4**, matrix G is also identifiable. $\square$

B.7 More simulation results

In this section, we present additional simulation results for higher-dimensional cases, along with simulations that incorporate a variety of ground-truth parameter configurations.

B.7.1 Higher dimensional cases

In this subsection, for the $\boldsymbol{\eta}$-(un)identifiable cases of the ODE system (4.3), we provide a case with $d = 5$ and $p = 5$. The true underlying parameters of the systems are provided below. Initial parameter values are set to the true parameters plus a random value drawn from a uniform distribution $U(-0.14, 0.14)$ for each replication. To ensure reliability in the estimation results, we perform 50 independent random replications for each configuration, reporting the mean and variance of the squared error in Table B.3.

$$A = \begin{bmatrix} 2 & -2 & 1 & 1 & 1 \\ -1 & 1 & 0 & 2 & -2 \\ -2 & 2 & 0 & -1 & -2 \\ -1 & -1 & -2 & -1 & 2 \\ 1 & -2 & 1 & -2 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & -2 & -1 & 1 & 1 \\ 1 & -2 & -1 & -1 & -1 \\ -2 & 0 & 2 & 1 & 1 \\ 0 & 2 & 0 & -2 & -2 \\ 2 & -2 & 2 & -1 & 2 \end{bmatrix},$$

$$G = \begin{bmatrix} 0 & 0 & 0 & -2 & -1 \\ 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad A' = \boldsymbol{I}_5 \,, \quad \boldsymbol{x}_0 = \begin{bmatrix} 2 \\ -2 \\ 2 \\ 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{z}_0 = \begin{bmatrix} -2 \\ -1 \\ -1 \\ 1 \\ -2 \end{bmatrix},$$

$$\boldsymbol{\eta}\text{-identifiable: } \boldsymbol{\eta} = (\boldsymbol{x}_0, \boldsymbol{z}_0, A, B, G), \text{ unidentifiable: } \boldsymbol{\eta} = (\boldsymbol{x}_0, \boldsymbol{z}_0, A', B, G)\,.$$

$\boldsymbol{I}_j$ denotes a $j \times j$ identity matrix.

For $\{\boldsymbol{\eta}_i\}_1^p$-(un)identifiable cases of the ODE system (4.3), we consider a case with $d = 10$ and $p = 5$. To accelerate estimation, sparsity is introduced in the parameter

TABLE B.3: MSEs of the $\boldsymbol{\eta}$-(un)identifiable cases of the ODE (3) with $d = 5, p = 5$

	$\boldsymbol{n}$	A	$B\boldsymbol{z}_0$	$BG\boldsymbol{z}_0$	$BG^2\boldsymbol{z}_0$	$BG^3\boldsymbol{z}_0$	$BG^4\boldsymbol{z}_0$
Identifiable	10	0.0148 (±0.0006)	0.3911 (±0.5989)	0.9624 (±3.9249)	0.7316 (±1.8971)	0.1037 (±0.0374)	0.0096 (±0.0003)
	100	0.0059 (±4.01E-05)	0.1529 (±0.0277)	0.1726 (±0.0541)	0.2447 (±0.0748)	0.0212 (±0.0007)	0.0012 (±1.10E-05)
	1000	0.0053 (±2.92E-05)	0.1394 (±0.0200)	0.1241 (±0.0251)	0.2119 (±0.0479)	0.0164 (±0.0004)	0.0004 (±6.00E-07)
Unidentifiable	10	0.0853 (±0.0075)	1.0067 (±1.3518)	3.7422 (±55.8402)	2.7696 (±24.5043)	0.9229 (±2.7959)	0.0508 (±0.0111)
	100	0.0357 (±0.0019)	0.4091 (±0.3812)	1.0428 (±2.1792)	0.9782 (±5.3654)	0.3871 (±0.6747)	0.0256 (±0.0032)
	1000	0.0332 (±0.0017)	0.3286 (±0.1824)	0.7123 (±1.8836)	0.9782 (±2.3163)	0.5487 (±0.9240)	0.0393 (±0.0047)

matrices by randomly setting 70, 35, and 20 entries in matrices A, B and G, respectively, as zero. The true underlying parameters of the systems are provided below. Initial parameter values are set to the true parameters plus a random value drawn from a uniform distribution $U(-0.1, 0.1)$ for each replication. To ensure reliability in the estimation results, we perform 50 independent random replications for each configuration, reporting the mean and variance of the squared error in Table B.4.

$$A = \begin{bmatrix} 0 & 0 & -2 & -1 & 1 & 2 & 0 & -2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & -2 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & -2 & 0 \\ 2 & 0 & 0 & -1 & 0 & -2 & 0 & 0 & -1 & 1 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & -2 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & 0 & 0 & 0 & 2 \\ 0 & -1 & 0 & 2 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 & -1 \end{bmatrix},$$

$$G = \begin{bmatrix} 0 & 1 & -1 & 0 & 2 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad A' = \boldsymbol{I}_{10}\,,$$

$$\boldsymbol{x}_0 = \begin{bmatrix} -2 & 0 & 0 & -2 & 2 & -1 & 1 & 0 & 1 & 1 \end{bmatrix}^\top, \quad \boldsymbol{z}_0^{*i} = \boldsymbol{e}_i\,, \text{for } i = 1, \ldots, 5\,.$$

$$\{\boldsymbol{\eta}_i\}_1^p\text{-identifiable: } \boldsymbol{\eta}_i = (\boldsymbol{x}_0, \boldsymbol{z}_0^{*i}, A, B, G), \text{ unidentifiable: } \boldsymbol{\eta}_i = (\boldsymbol{x}_0, \boldsymbol{z}_0^{*i}, A', B, G)\,.$$

$\boldsymbol{e}_i$ stands for a p-dimensional vector, with the i-th entry being 1 and the other entries being 0.

Table B.4: MSEs of the $\{\boldsymbol{\eta}_i\}_1^p$-(un)identifiable cases of the ODE (3) with $d = 10, p = 5$

$\boldsymbol{n}$	**Identifiable**			**Unidentifiable**		
	A	B	G	A	B	G
10	1.53E-11 (±2.36E-21)	2.49E-10 (±6.30E-19)	3.01E-10 (±9.20E-19)	0.8345 (±0.6268)	0.2118 (±0.0260)	0.0037 (±0.0002)
30	9.15E-13 (±4.45E-24)	1.49E-11 (±1.18E-21)	1.80E-11 (±1.73E-21)	0.7216 (±0.4099)	0.1952 (±0.0156)	1.25E-21 (±5.18E-41)
50	9.64E-14 (±1.29E-25)	1.57E-12 (±3.43E-23)	1.90E-12 (±5.02E-23)	0.6510 (±0.2251)	0.2211 (±0.0278)	0.0042 (±0.0003)

Tables B.3 and B.4 present results similar to those in Tables 4.1 and 4.2, providing strong empirical support for the validity of our proposed identifiability conditions.

B.7.2 Various true parameters

To further support our proposed identifiability conditions, we conduct additional simulations incorporating a variety of ground-truth parameter configurations, rather than a fixed underlying parameter set. Specifically, for each simulation run, a unique ground-truth parameter configuration was generated using different random seeds, and we subsequently reported the mean and variance of the squared error across all results. For the low-dimensional $\boldsymbol{\eta}$ and $\{\boldsymbol{\eta}_i\}_1^p$ (un)identifiable cases, we perform 100 replications, while for the higher-dimensional cases, we perform 50 replications. Additionally, in the $\{\boldsymbol{\eta}_i\}_1^p$-(un)identifiable cases, we initialize the parameter values as the true parameters plus a random value drawn from $U(-0.1, 0.1)$ for the $d = 3, p = 3$ case and from $U(-0.05, 0.05)$ for the $d = 10, p = 5$ cases. For the $\boldsymbol{\eta}$-(un)identifiable cases, the initialization settings are the same as those used in the fixed-parameter configurations.

The simulation results are presented in Tables B.5, B.6, B.7, and B.8. Across all these tables, parameter estimates in the identifiable cases are notably more accurate than in the unidentifiable cases, providing strong empirical support for the validity of our proposed identifiability conditions.

It is noteworthy, however, that even in theoretically identifiable cases, certain scenarios emerge where parameter identification is challenging in practice; we refer to these as hard estimate cases. In these instances, estimates may deviate significantly from satisfactory values, similar to challenges encountered in fully observable ODE models (4.1) as discussed in [29]. Consequently, for identifiable cases with varying true parameter configurations, the results are less precise than those for corresponding fixed-parameter cases, due to the inclusion of some hard estimate instances. Investigating the practical identifiability of the ODE system (4.3) remains an intriguing direction for future research.

TABLE B.5: MSEs of the $\boldsymbol{\eta}$-(un)identifiable cases of the ODE (4.3) - with various true parameters

$\boldsymbol{n}$	**Identifiable**				**Unidentifiable**			
	A	$B\boldsymbol{z}_0$	$BG\boldsymbol{z}_0$	$BG^2\boldsymbol{z}_0$	A	$B\boldsymbol{z}_0$	$BG\boldsymbol{z}_0$	$BG^2\boldsymbol{z}_0$
10	0.0060 (±0.0008)	0.0157 (±0.0036)	0.1698 (±0.5665)	0.2297 (±1.1053)	0.0691 (±0.0203)	0.2720 (±0.5914)	1.3133 (±7.4471)	0.6622 (±8.5348)
100	0.0026 (±9.27E-05)	0.0108 (±0.0022)	0.0820 (±0.1159)	0.1287 (±0.7042)	0.0283 (±0.0031)	0.1003 (±0.0441)	0.4880 (±2.6547)	0.2649 (±1.6631)
500	0.0020 (±6.48E-05)	0.0092 (±0.0023)	0.0870 (±0.1941)	0.0705 (±0.1179)	0.0227 (±0.0018)	0.1061 (±0.0672)	0.5015 (±3.0811)	0.2574 (±2.0779)

TABLE B.6: MSEs of the $\{\boldsymbol{\eta}_i\}_1^p$-(un)identifiable cases of the ODE (4.3) - with various true parameters

$\boldsymbol{n}$	**Identifiable**			**Unidentifiable**		
	A	B	G	A	B	G
10	0.0006 (±2.21E-5)	1.89E-5 (±3.55E-8)	0.0009 (±6.71E-5)	0.0861 (±0.1773)	0.0088 (±0.0020)	0.0101 (±0.0045)
30	0.0006 (±2.20E-5)	1.87E-5 (±3.47E-8)	0.0010 (±6.64E-5)	0.0789 (±0.1280)	0.0092 (±0.0028)	0.0104 (±0.0046)
50	0.0006 (±2.21E-5)	1.88E-5 (±3.51E-8)	0.0009 (±6.67E-5)	0.0503 (±0.0430)	0.0063 (±0.0006)	0.0114 (±0.0047)

TABLE B.7: MSEs of the $\boldsymbol{\eta}$-(un)identifiable cases of the ODE (3) with $d = 5, p = 5$ - with various true parameters

	$\boldsymbol{n}$	A	$B\boldsymbol{z}_0$	$BG\boldsymbol{z}_0$	$BG^2\boldsymbol{z}_0$	$BG^3\boldsymbol{z}_0$	$BG^4\boldsymbol{z}_0$
Identifiable	10	0.0144 (±0.0004)	0.1215 (±0.0757)	1.4643 (±8.3976)	2.1890 (±54.9706)	1.8254 (±48.7033)	0.4826 (±5.7127)
	100	0.0041 (±4.55E-05)	0.0395 (±0.0092)	0.2850 (±0.1739)	0.3891 (±0.4936)	0.2078 (±0.2950)	0.0239 (±0.0024)
	1000	0.0032 (±3.26E-05)	0.0337 (±0.0049)	0.1934 (±0.0686)	0.2242 (±0.2180)	0.1197 (±0.0712)	0.0181 (±0.0014)
Unidentifiable	10	0.0740 (±0.0047)	0.4599 (±0.4841)	2.8628 (±9.5476)	1.8743 (±8.6653)	0.4834 (±1.2606)	0.0334 (±0.0147)
	100	0.0263 (±0.0031)	0.2142 (±0.1869)	1.1678 (±8.2277)	1.2354 (±9.4970)	0.2878 (±0.8655)	0.0193 (±0.0052)
	1000	0.0142 (±0.0003)	0.1389 (±0.0463)	0.6979 (±1.2080)	0.6701 (±1.5228)	0.0732 (±0.0336)	0.0062 (±0.0003)

Table B.8: MSEs of the $\{\boldsymbol{\eta}_i\}_1^p$-(un)identifiable cases of the ODE (3) with $d = 10, p = 5$ - with various true parameters

$\boldsymbol{n}$	**Identifiable**			**Unidentifiable**		
	A	B	G	A	B	G
10	0.0044 (±0.0001)	0.0350 (±0.0098)	0.0287 (±0.0053)	0.6266 (±0.1524)	0.1310 (±0.0269)	0.0054 (±0.0004)
30	0.0067 (±0.0005)	0.1258 (±0.5097)	0.0315 (±0.0104)	0.5833 (±0.2085)	0.1058 (±0.0114)	0.0021 (±8.79E-05)
50	0.0033 (±5.66E-05)	0.0323 (±0.0103)	0.0354 (±0.0084)	0.5193 (±0.0982)	0.1108 (±0.0146)	0.0021 (±9.02E-05)

Appendix C

Appendix of Chapter 5

C.1 Detailed proofs

C.1.1 Proof of Lemma 5.4

Proof. We start by presenting the mathematical definition of a Lévy process. (cf. [159])

Definition C.1. A stochastic process $X := \{X_t : 0 \leqslant t < \infty\}$ is said to be a Lévy process if it satisfies the following properties:

1. $X_0 = 0$ almost surely;
2. Independence of increments: For any $0 \leqslant t_1 < t_2 < \ldots < t_n < \infty$, $X_{t_2} - X_{t_1}$, $X_{t_3} - X_{t_2}$, $\ldots$, $X_{t_n} - X_{t_{n-1}}$ are independent;
3. Stationary increments: For any $s < t$, $X_t - X_s$ is equal in distribution to X_{t-s};
4. Continuity in probability: For any $\varepsilon > 0$ and $0 \leqslant t < \infty$ it holds that $\lim_{h \to 0} \mathbb{P}(|X_{t+h} - X_t| > \varepsilon) = 0$.

In the following, we first show that the SDE (5.1) can be expressed as the form of (5.10), with Z being a p-dimensional Lévy process and $a : \mathbb{R}^d \to \mathbb{R}^{d \times p}$ being Lipschitz.

The first equation in the SDE (5.1) can be rearranged as

$$\begin{aligned} dX_t &= AX_t dt + G dW_t \\ &= \begin{bmatrix} AX_t & G \end{bmatrix} \begin{bmatrix} dt \\ dW_t \end{bmatrix} \\ &= a(X_t) dZ_t \,, \end{aligned} \tag{C.1}$$

with

$$a(X_t) = \begin{bmatrix} AX_t & G \end{bmatrix} \in \mathbb{R}^{d\times(m+1)} \,,$$

and

$$dZ_t = \begin{bmatrix} dt \\ dW_t \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 0_{m\times 1} \end{bmatrix}}_{r} dt + \underbrace{\begin{bmatrix} 0_{1\times 1} & 0_{1\times m} \\ 0_{m\times 1} & I_{m\times m} \end{bmatrix}}_{E} \begin{bmatrix} dW_{0,t} \\ dW_t \end{bmatrix} \,, \tag{C.2}$$

where $0_{i\times j}$ denotes an $i \times j$ zero matrix, let $\tilde{W} := \{\tilde{W}_t : 0 \leqslant t < \infty\}$ with $\tilde{W}_t = [W_{0,t}, W_{1,t}, \ldots, W_{m,t}]^\top$ denote a $(m+1)$-dimensional standard Brownian motion, then one can find a process $Z := \{Z_t : 0 \leqslant t < \infty\}$ with

$$\begin{aligned} &Z_t = rt + E\tilde{W}_t \,, \\ &Z_0 = 0 \,, \end{aligned} \tag{C.3}$$

satisfying dZ_t described in Equation (C.2). Then we will show that the process Z described in (C.3) is a Lévy process, that is, it satisfies the four properties stated in Definition C.1.

Property 1: The first property is readily checked since $Z_0 = 0$.

Property 2: For any $0 \leqslant t_1 < t_2 < t_3 < \infty$,

$$\begin{aligned} Z_{t_2} - Z_{t_1} &= r(t_2 - t_1) + E(\tilde{W}_{t_2} - \tilde{W}_{t_1}) \\ &= \begin{bmatrix} t_2 - t_1 \\ 0_{m\times 1} \end{bmatrix} + \begin{bmatrix} 0 \\ W_{t_2} - W_{t_1} \end{bmatrix} \,. \end{aligned}$$

Similarly,

$$Z_{t_3} - Z_{t_2} = \begin{bmatrix} t_3 - t_2 \\ 0_{m\times 1} \end{bmatrix} + \begin{bmatrix} 0 \\ W_{t_3} - W_{t_2} \end{bmatrix} \,.$$

Since $W_{t_2} - W_{t_1}$ and $W_{t_3} - W_{t_2}$ are independent, $Z_{t_3} - Z_{t_2}$ and $Z_{t_2} - Z_{t_1}$ are independent.

Property 3: when $s < t$,

$$
\begin{aligned}
Z_t - Z_s &= \begin{bmatrix} t-s \\ 0_{m\times 1} \end{bmatrix} + \begin{bmatrix} 0 \\ W_t - W_s \end{bmatrix} \\
&\sim \mathcal{N}\left(\begin{bmatrix} t-s \\ 0_{m\times 1} \end{bmatrix}, \begin{bmatrix} 0_{1\times 1} & 0_{1\times m} \\ 0_{m\times 1} & (t-s)I_{m\times m} \end{bmatrix} \right).
\end{aligned}
$$

And

$$
\begin{aligned}
Z_{t-s} &= r(t-s) + E\tilde{W}_{t-s} \\
&= \begin{bmatrix} t-s \\ 0_{m\times 1} \end{bmatrix} + \begin{bmatrix} 0 \\ W_{t-s} \end{bmatrix} \\
&\sim \mathcal{N}\left(\begin{bmatrix} t-s \\ 0_{m\times 1} \end{bmatrix}, \begin{bmatrix} 0_{1\times 1} & 0_{1\times m} \\ 0_{m\times 1} & (t-s)I_{m\times m} \end{bmatrix} \right).
\end{aligned}
$$

Therefore, property 3 is checked.

Property 4: Obviously, process Z described in (C.3) is continuous with probability one at t for all $0 \leqslant t < \infty$, therefore, Z has continuity in probability.

Now that we have shown that process Z is a p-dimensional Lévy process with $p = m+1$. Then we will show that $a(X_t) = \begin{bmatrix} AX_t & G \end{bmatrix}$ is Lipschitz.

$$
\begin{aligned}
\| a(X_t) - a(X_s) \|_F &= \| \begin{bmatrix} A(X_t - X_s) & 0 \end{bmatrix} \|_F \\
&= \| A(X_t - X_s) \|_2 \\
&\leqslant \| A \|_F \| X_t - X_s \|_2
\end{aligned}
$$

where $\| M \|_F$ denotes the Frobenius norm of matrix M and $\| v \|_2$ denotes the Euclidean norm of vector v. Now it is readily checked that function $a : \mathbb{R}^d \to \mathbb{R}^{d\times p}$ is Lipschitz.

Similarly, we will show that the SDE (5.2) can also be expressed as the form of (5.10), with Z being a p-dimensional Lévy process and $a : \mathbb{R}^d \to \mathbb{R}^{d\times p}$ being Lipschitz. Let us rearrange the first equation in the SDE (5.2):

$$\begin{aligned} dX_t &= AX_t dt + \sum_{k=1}^{m} G_k X_t dW_{k,t} \\ &= \begin{bmatrix} AX_t & G_1 X_t & \dots & G_m X_t \end{bmatrix} \begin{bmatrix} dt \\ dW_{1,t} \\ \vdots \\ dW_{m,t} \end{bmatrix} \\ &= a(X_t) dZ_t \,. \end{aligned}$$

Since the dZ_t here has the same form as that of the SDE (5.1), we use the same process Z described in Equation (C.3), which has been shown to be a Lévy process.

As for the function $a(X_t)$,

$$\begin{aligned} \| a(X_t) - a(X_s) \|_F &= \| \begin{bmatrix} A(X_t - X_s) & G_1(X_t - X_s) & \dots & G_m(X_t - X_s) \end{bmatrix} \|_F \\ &\leqslant \| A(X_t - X_s) \|_2 + \sum_{k=1}^{m} \| G_k(X_t - X_s) \|_2 \\ &\leqslant \| A \|_F \| X_t - X_s \|_2 + \sum_{k=1}^{m} \| G_k \|_F \| X_t - X_s \|_2 \\ &= \left(\| A \|_F + \sum_{k=1}^{m} \| G_k \|_F \right) \| X_t - X_s \|_2 \,, \end{aligned}$$

it is readily checked that function $a : \mathbb{R}^d \to \mathbb{R}^{d \times p}$ is Lipschitz. □

C.1.2 Proof of Proposition 5.3.1

Proof. For the backward direction, when $b(\boldsymbol{x}) = \tilde{b}(\boldsymbol{x})$ and $c(\boldsymbol{x}) = \tilde{c}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$, it is obviously that $(\mathcal{L}f)(\boldsymbol{x}) = (\tilde{\mathcal{L}}f)(\boldsymbol{x})$ for all $f \in C_b^2(\mathbb{R}^d)$ and $\boldsymbol{x} \in \mathbb{R}^d$, that is $\mathcal{L} = \tilde{\mathcal{L}}$.

For the forward direction, since $(\mathcal{L}f)(\boldsymbol{x}') = (\tilde{\mathcal{L}}f)(\boldsymbol{x}')$ for all $f \in C_b^2(\mathbb{R}^d)$ and $\boldsymbol{x}' \in \mathbb{R}^d$.

We first set

$$f(\boldsymbol{x}) = x_p \,,$$

where x_p denotes the p-th component of variable $\boldsymbol{x}$. It is readily checked that

$$b_p(\boldsymbol{x}') = \tilde{b}_p(\boldsymbol{x}') \,,$$

for all $\boldsymbol{x}' \in \mathbb{R}^d$ and $p = 1, \ldots, d$. As a result,

$$b(\boldsymbol{x}) = \tilde{b}(\boldsymbol{x}), \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^d .$$

Then we set

$$f(\boldsymbol{x}) = (x_p - x'_p)(x_q - x'_q),$$

where x'_p denotes the p-th component of $\boldsymbol{x}'$. It is readily checked that

$$c_{pq}(\boldsymbol{x}') = \tilde{c}_{pq}(\boldsymbol{x}'),$$

for all $\boldsymbol{x}' \in \mathbb{R}^d$ and $p, q = 1, \ldots, d$. Consequently,

$$c(\boldsymbol{x}) = \tilde{c}(\boldsymbol{x}), \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^d .$$

□

C.1.3 Proof of Lemma 5.6

Proof. For the forward direction, since

$$X(\cdot; \boldsymbol{x}_0, A, G) \stackrel{\mathrm{d}}{=} X(\cdot; \boldsymbol{x}_0, \tilde{A}, \tilde{G}),$$

one has

$$\mathbb{E}[X_t] = \mathbb{E}[\tilde{X}_t], \quad \forall 0 \leqslant t < \infty .$$

Thus,

$$(X_t - \mathbb{E}[X_t])_{0 \leqslant t < \infty} \stackrel{\mathrm{d}}{=} (\tilde{X}_t - \mathbb{E}[\tilde{X}_t])_{0 \leqslant t < \infty},$$

in particular, one has

$$\mathbb{E}\{(X_{t+h} - \mathbb{E}[X_{t+h}])(X_t - \mathbb{E}[X_t])^\top\} = \mathbb{E}\{(\tilde{X}_{t+h} - \mathbb{E}[\tilde{X}_{t+h}])(\tilde{X}_t - \mathbb{E}[\tilde{X}_t])^\top\} \quad \text{for all } 0 \leqslant t, h < \infty .$$

For the backward direction, we know that the solution of the SDE (5.1) is a Gaussian process. The distribution of a Gaussian process can be fully determined by its mean and covariance functions. Therefore, the two processes have the same distribution when the mean and covariance are the same for both processes for all $0 \leqslant t, h < \infty$. $\square$

C.1.4 Proof of Lemma 5.7

Proof. For the forward direction, since

$$\operatorname{rank}([\boldsymbol{\gamma}_1|A\boldsymbol{\gamma}_1|\ldots|A^{d-1}\boldsymbol{\gamma}_1|\ldots|\boldsymbol{\gamma}_n|A\boldsymbol{\gamma}_n|\ldots|A^{d-1}\boldsymbol{\gamma}_n]) < d\,,$$

then for all $\boldsymbol{l} = [l_1, \ldots, l_n]^\top \in \mathbb{R}^n$,

$$\operatorname{rank}([\boldsymbol{\beta}|A\boldsymbol{\beta}|\ldots|A^{d-1}\boldsymbol{\beta}]) < d\,,$$

where $\boldsymbol{\beta} := l_1\boldsymbol{\gamma}_1 + \ldots + l_n\boldsymbol{\gamma}_n$. Consequently, the corresponding ODE system

$$\begin{aligned} \dot{\boldsymbol{x}}(t) &= A\boldsymbol{x}(t)\,, \\ \boldsymbol{x}(0) &= \boldsymbol{\beta}\,, \end{aligned} \tag{C.4}$$

is not identifiable from $\boldsymbol{\beta}$ by [28, Theorem 2.5.], where $\dot{\boldsymbol{x}}(t)$ denotes the first derivative of $\boldsymbol{x}(t)$ with respect to time t.

Let

$$\tilde{\boldsymbol{\beta}} := Q^{-1}\boldsymbol{\beta} \in \mathbb{R}^d\,,$$

and

$$w_k := \begin{cases} \tilde{\beta}_k \in \mathbb{R}^1, & \text{for } k = 1, \ldots, K_1\,, \\ (\tilde{\beta}_{2k-K_1-1}, \tilde{\beta}_{2k-K_1})^\top \in \mathbb{R}^2, & \text{for } k = K_1+1, \ldots, K\,. \end{cases}$$

Simple calculation shows that

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= Q^{-1}\boldsymbol{\beta} \\ &= Q^{-1}(l_1\boldsymbol{\gamma}_1 + \ldots + l_n\boldsymbol{\gamma}_n) \\ &= l_1\tilde{\boldsymbol{\gamma}}_1 + \ldots + l_n\tilde{\boldsymbol{\gamma}}_n\,, \end{aligned}$$

therefore, one has

$$w_k = l_1 w_{1,k} + \ldots + l_n w_{n,k}\,, \qquad \text{for all } k \in \{1, \ldots, K\}\,. \tag{C.5}$$

By [29, Theorem 2.4], we know that for any $\boldsymbol{l} \in \mathbb{R}^n$, there always exists $k \in \{1, \dots, K\}$ such that $w_k = 0$ ($\in \mathbb{R}^1$ or $\mathbb{R}^2$) since the ODE (C.4) is not identifiable from initial state $\boldsymbol{\beta}$. Next, we will show that this result is satisfied only when there exists a k such that $w_{j,k} = 0$ ($\in \mathbb{R}^1$ or $\mathbb{R}^2$) for all $j = 1, \dots, n$. Let us rearrange the Equation (C.5) as

$$\begin{bmatrix} w_{1,1} & \dots & w_{n,1} \\ \vdots & \ddots & \vdots \\ w_{1,K} & \dots & w_{n,K} \end{bmatrix} \begin{bmatrix} l_1 \\ \vdots \\ l_n \end{bmatrix} = \begin{bmatrix} w_1 \\ \vdots \\ w_K \end{bmatrix},$$

assume for any $k \in \{1, \dots, K\}$, $[w_{1,k}, \dots, w_{n,k}]^\top \neq \mathbf{0}$, then there always exists a $\boldsymbol{l} \in \mathbb{R}^n$ such that $w_k \neq 0$ for all $k = \{1, \dots, K\}$. The reason is that under this circumstance, for any $k \in \{1, \dots, K\}$, the set of $\boldsymbol{l}$'s such that $w_k = 0$ has Lebesgue measure zero in $\mathbb{R}^n$. Therefore, the set of $\boldsymbol{l}$'s such that there exists a k such that $w_k = 0$ has Lebesgue measure zero in $\mathbb{R}^n$. This result creates a contradiction. Thus, there must exist a k, such that $[w_{1,k}, \dots, w_{n,k}]^\top = \mathbf{0}$, that is $|w_{j,k}| = 0$ for all $j = 1, \dots, n$.

For the backward direction, there exists k such that $|w_{j,k}| = 0$ for all $j = 1, \dots, n$, that is $w_{j,k} = 0$ ($\in \mathbb{R}^1$ or $\mathbb{R}^2$) for all $j = 1, \dots, n$. Simple calculation shows that

$$\boldsymbol{\gamma}_j = Q\tilde{\boldsymbol{\gamma}}_j = \sum_{p=1}^{k-1} Q_p w_{j,p} + \sum_{p=k+1}^{K} Q_p w_{j,p}\,,$$

and

$$\begin{aligned} A^q \boldsymbol{\gamma}_j &= Q\Lambda^q Q^{-1} \boldsymbol{\gamma}_j \\ &= Q\Lambda^q \tilde{\boldsymbol{\gamma}}_j \\ &= \sum_{p=1}^{k-1} Q_p J_p^q w_{j,p} + \sum_{p=k+1}^{K} Q_p J_p^q w_{j,p}\,, \end{aligned}$$

recall that

$$J_k = \begin{cases} \lambda_k, & \text{if } k = 1, \dots, K_1\,, \\ \begin{bmatrix} a_k & -b_k \\ b_k & a_k \end{bmatrix}, & \text{if } k = K_1 + 1, \dots, K\,. \end{cases}$$

Then matrix

$$\begin{aligned} M :&= [\boldsymbol{\gamma}_1 | A\boldsymbol{\gamma}_1 | \dots | A^{d-1}\boldsymbol{\gamma}_1 | \dots | \boldsymbol{\gamma}_n | A\boldsymbol{\gamma}_n | \dots | A^{d-1}\boldsymbol{\gamma}_n] \\ &= Q_{-k} C\,, \end{aligned}$$

where

$$Q_{-k} = [Q_1|\dots|Q_{k-1}|Q_{k+1}|\dots|Q_K]\,,$$

and matrix C denotes:

$$\begin{bmatrix} w_{1,1} & J_1 w_{1,1} & \dots & J_1^{d-1} w_{1,1} & \dots & w_{n,1} & J_1 w_{n,1} & \dots & J_1^{d-1} w_{n,1} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ w_{1,k-1} & J_{k-1} w_{1,k-1} & \dots & J_{k-1}^{d-1} w_{1,k-1} & \dots & w_{n,k-1} & J_{k-1} w_{n,k-1} & \dots & J_{k-1}^{d-1} w_{n,k-1} \\ w_{1,k+1} & J_{k+1} w_{1,k+1} & \dots & J_{k+1}^{d-1} w_{1,k+1} & \dots & w_{n,k+1} & J_{k+1} w_{n,k+1} & \dots & J_{k+1}^{d-1} w_{n,k+1} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ w_{1,K} & J_K w_{1,K} & \dots & J_K^{d-1} w_{1,K} & \dots & w_{n,K} & J_K w_{n,K} & \dots & J_K^{d-1} w_{n,K} \end{bmatrix}$$

We know that

$$\mathrm{rank}(M) = \mathrm{rank}(Q_{-k}C) \leqslant \min(\mathrm{rank}(Q_{-k}), \mathrm{rank}(C))\,.$$

When $k \in \{1,\dots,K_1\}$, $Q_{-k} \in \mathbb{R}^{d\times(d-1)}$, and $\mathrm{rank}(Q_{-k}) = d-1$, while when $k \in \{K_1 + 1,\dots,K\}$, $Q_{-k} \in \mathbb{R}^{d\times(d-2)}$, and $\mathrm{rank}(Q_{-k}) = d-2$. In both cases, $\mathrm{rank}(Q_{-k}) < d$, thus $\mathrm{rank}(M) < d$. $\square$

C.1.5 Proof of Theorem 5.8

Proof. Let $\tilde{A} \in \mathbb{R}^{d\times d}$ and $\tilde{G} \in \mathbb{R}^{d\times m}$, such that $X(\cdot;\boldsymbol{x}_0, A, G) \overset{\mathrm{d}}{=} X(\cdot;\boldsymbol{x}_0, \tilde{A}, \tilde{G})$, we denote as $X \overset{\mathrm{d}}{=} \tilde{X}$. For simplicity of notation, in the following, we denote $A_1 := A$, $A_2 := \tilde{A}$, $G_1 := G$ and $G_2 := \tilde{G}$, and denote $X \overset{\mathrm{d}}{=} \tilde{X}$ as $X^1 \overset{\mathrm{d}}{=} X^2$.

Sufficiency. We will show that under the identifiability condition (5.14), one has $(A_1, G_1G_1^\top) = (A_2, G_2G_2^\top)$.

We first show that $H_1 = H_2$ ($H_i := G_iG_i^T$). Indeed, since X^1, X^2 have the same distribution, one has

$$\mathbb{E}[f(X_t^1)] = \mathbb{E}[f(X_t^2)] \tag{C.6}$$

for all $0 \leqslant t < \infty$ and $f \in C^\infty(\mathbb{R}^d)$. By differentiating (C.6) at $t = 0$, one finds that

$$(\mathcal{L}_1 f)(\boldsymbol{x}_0) = (\mathcal{L}_2 f)(\boldsymbol{x}_0)\,, \tag{C.7}$$

where $\mathcal{L}_i$ is the generator of X^i $(i = 1, 2)$. Based on the Proposition 5.2.1,

$$(\mathcal{L}_i f)(\boldsymbol{x}_0) = \sum_{k=1}^{d}\sum_{l=1}^{d}(A_i)_{kl}x_{0l}\frac{\partial f}{\partial x_k}(\boldsymbol{x}_0) + \frac{1}{2}\sum_{k,l=1}^{d}(H_i)_{kl}\frac{\partial^2 f}{\partial x_k \partial x_l}(\boldsymbol{x}_0)\,,$$

where $(M)_{kl}$ denotes the kl-entry of matrix M, and x_{0l} is the l-th component of $\boldsymbol{x}_0$. Since (C.7) is true for all f, by taking

$$f(\boldsymbol{x}) = (x_p - x_{0p})(x_q - x_{0q})\,,$$

it is readily checked that

$$(H_1)_{pq} = (H_2)_{pq}\,,$$

for all $p, q = 1, \ldots, d$. As a result, $H_1 = H_2$. Let us call this matrix H.

Next, we show that $A_1 = A_2$. We first show the relationship between A_i and $\boldsymbol{x}_0$, and then show the relationship between A_i and H. To this end, one first recalls that

$$X_t^i = e^{A_i t}\boldsymbol{x}_0 + \int_0^t e^{A_i(t-s)}G_i dW_s\,.$$

Set $\boldsymbol{m}_i(t) := \mathbb{E}[X_t^i]$, we know that $\boldsymbol{m}_i(t)$ satisfies the ODE

$$\begin{aligned} \dot{\boldsymbol{m}}_i(t) &= A_i \boldsymbol{m}_i(t)\,, \quad \forall 0 \leqslant t < \infty\,, \\ \boldsymbol{m}_i(0) &= \boldsymbol{x}_0\,, \end{aligned} \tag{C.8}$$

where $\dot{f}(t)$ denotes the first derivative of function $f(t)$ with respect to time t.

Simple calculation shows that

$$\boldsymbol{m}_i(t) = e^{A_i t}\boldsymbol{x}_0\,.$$

Since $X^1 \stackrel{\mathrm{d}}{=} X^2$, one has

$$\mathbb{E}[X_t^1] = \mathbb{E}[X_t^2]$$

for all $0 \leqslant t < \infty$. That is

$$e^{A_1 t}\boldsymbol{x}_0 = e^{A_2 t}\boldsymbol{x}_0\,, \quad \forall 0 \leqslant t < \infty\,.$$

Taking k-th derivative of $e^{A_i t}\boldsymbol{x}_0$ with respect to t, one finds that

$$\frac{d^k}{dt^k}\Big|_{t=0} e^{A_i t}\boldsymbol{x}_0 = A_i^k \boldsymbol{x}_0\,,$$

for all $k = 1, 2, \ldots$. Consequently,

$$A_1^k \boldsymbol{x}_0 = A_2^k \boldsymbol{x}_0\,.$$

Let us denote this vector $A^k \boldsymbol{x}_0$. Obviously, one gets

$$A_1 A^{k-1}\boldsymbol{x}_0 = A_2 A^{k-1}\boldsymbol{x}_0 \quad \text{for all } k = 1, 2, \ldots. \tag{C.9}$$

In the following, we show the relationship between A_i and H. Let us denote

$$Y_t^i := \int_0^t e^{A_i(t-s)} G_i dW_s = X_t^i - \mathbb{E}[X_t^i]$$

and

$$V_i(t, t+h) := \mathbb{E}[Y_{t+h}^i \cdot (Y_t^i)^T]\,.$$

Simple calculation shows that

$$\begin{aligned} V_i(t, t+h) &= e^{A_i h} \int_0^t e^{A_i(t-s)} H e^{A_i^\top (t-s)} ds \\ &= e^{A_i h} V_i(t)\,, \end{aligned} \tag{C.10}$$

where $V_i(t) := V_i(t, t)$.

Since $X^1 \overset{\mathrm{d}}{=} X^2$, by Lemma 5.6, one has

$$V_1(t, t+h) = V_2(t, t+h)\,, \quad \forall 0 \leqslant t, h < \infty\,.$$

To obtain information about A_i, let us fix t for now and take k-th derivative of (C.10) with respect to h. One finds that

$$\frac{d^k}{dh^k}\Big|_{h=0} V_i(t, t+h) = A_i^k V_i(t)\,, \tag{C.11}$$

for all $k = 1, 2, \ldots$.

On the other hand, the function $V_i(t)$ satisfies the ODE [147]

$$
\begin{aligned}
&\dot{V}_i(t) = A_i V_i(t) + V_i(t) A_i^\top + H, \quad 0 \leqslant t < \infty\,, \\
&V_i(0) = 0\,.
\end{aligned}
$$

In particular,

$$\dot{V}_i(0) = A_i V_i(0) + V_i(0) A_i + H = H.$$

By differentiating (C.11) at $t = 0$, it follows that

$$\frac{d}{dt}\Big|_{t=0} \frac{d^k}{dh^k}\Big|_{h=0} V_i(t, t+h) = A_i^k H\,,$$

for all $k = 1, 2, \ldots$. Consequently,

$$A_1^k H = A_2^k H\,.$$

Let us denote this matrix $A^k H$. Obviously, by rearranging this matrix, one gets

$$A_1 A^{k-1} H = A_2 A^{k-1} H \quad \text{for all } k = 1, 2, \ldots. \tag{C.12}$$

Recall our identifiability condition is that $\text{rank}(M) = d$ with

$$M := [\boldsymbol{x}_0 | A\boldsymbol{x}_0 | \ldots | A^{d-1}\boldsymbol{x}_0 | H_{\cdot 1} | AH_{\cdot 1} | \ldots | A^{d-1} H_{\cdot 1} | \ldots | H_{\cdot d} | AH_{\cdot d} | \ldots | A^{d-1} H_{\cdot d}]\,.$$

If we denote the j-th column in M as $M_{\cdot j}$, one gets

$$A_1 M_{\cdot j} = A_2 M_{\cdot j}\,,$$

for all $j = 1, \ldots, d + d^2$ by Equations (C.9) and (C.12).

This means one can find a full-rank matrix $B \in \mathbb{R}^{d \times d}$ by horizontally stacking d linearly independent columns of matrix M, such that $A_1 B = A_2 B$. Since B is invertible, one thus concludes that $A_1 = A_2$. Hence, the sufficiency of the condition is proved.

Necessity. In the following, we will show that when A has distinct eigenvalues. The condition (5.14) stated in Theorem 5.8 is also necessary. Specifically, we will show that when the identifiability condition (5.14) is not satisfied, one can always find a $\tilde{A}$ with

$(A, GG^\top) \neq (\tilde{A}, GG^\top)$ such that $X \overset{\mathrm{d}}{=} \tilde{X}$. Recall that for simplicity of notation, we denote $A_1 := A$, $A_2 := \tilde{A}$, and denote $X \overset{\mathrm{d}}{=} \tilde{X}$ as $X^1 \overset{\mathrm{d}}{=} X^2$, where process $X^i = \{X_t^i : 0 \leqslant t < \infty\}$, and $X_t^i = X(t; \boldsymbol{x}_0, A_i, G)$ following the form described in the solution process (5.3). In the following, we may use both A and A_1 interchangeably according to the context.

By Lemma 5.6, to guarantee $X^1 \overset{\mathrm{d}}{=} X^2$ one only needs to show that

$$\begin{aligned}
\mathbb{E}[X_t^1] = \mathbb{E}[X_t^2]\,, \quad &\forall 0 \leqslant t < \infty\,,\\
V_1(t, t+h) = V_2(t, t+h)\,, \quad &\forall 0 \leqslant t, h < \infty\,.
\end{aligned}$$

That is,

$$\begin{aligned}
e^{A_1 t}\boldsymbol{x}_0 = e^{A_2 t}\boldsymbol{x}_0\,, \quad &\forall 0 \leqslant t < \infty\,,\\
e^{A_1 h}V(t) = e^{A_2 h}V(t)\,, \quad &\forall 0 \leqslant t, h < \infty\,,\\
V_1(t) = V_2(t)\,, \quad &\forall 0 \leqslant t < \infty\,,
\end{aligned} \tag{C.13}$$

where $V(t) := V_1(t) = V_2(t)$.

Recall that $H = GG^\top$. For simplicity of notation, abusing notation a bit, we denote $H_{\cdot 0} := \boldsymbol{x}_0$. Let

$$\tilde{H}_{\cdot j} := Q^{-1}H_{\cdot j}\,, \quad \text{for all } j = 0, \ldots, d\,,$$

and

$$w_{j,k} := \begin{cases} \tilde{H}_{\cdot j,k} \in \mathbb{R}^1, & \text{for } k = 1, \ldots, K_1\,,\\ (\tilde{H}_{\cdot j,2k-K_1-1}, \tilde{H}_{\cdot j,2k-K_1})^\top \in \mathbb{R}^2, & \text{for } k = K_1+1, \ldots, K\,. \end{cases}$$

When the identifiability condition (5.14) is not satisfied, that is

$$\text{rank}([H_{\cdot 0}|AH_{\cdot 0}|\ldots|A^{d-1}H_{\cdot 0}|H_{\cdot 1}|AH_{\cdot 1}|\ldots|A^{d-1}H_{\cdot 1}|\ldots|H_{\cdot d}|AH_{\cdot d}|\ldots|A^{d-1}H_{\cdot d}]) < d\,,$$

by Lemma 5.7, there exists k such that $|w_{j,k}| = 0$, i.e., $w_{j,k} = 0$ ($\in \mathbb{R}^1$ or $\mathbb{R}^2$), for all $j = 0, \ldots, d$. Recall that

$$\begin{aligned} V(t) = V_1(t) &= \int_0^t e^{A(t-s)} H e^{A^\top (t-s)} ds \\ &= \int_0^t Q e^{\Lambda(t-s)} Q^{-1} Q [\tilde{H}_{\cdot 1}| \ldots |\tilde{H}_{\cdot d}] e^{A^\top (t-s)} ds \\ &= Q \int_0^t e^{\Lambda(t-s)} [\tilde{H}_{\cdot 1}| \ldots |\tilde{H}_{\cdot d}] e^{A^\top (t-s)} ds \\ &:= Q \int_0^t W e^{A^\top (t-s)} ds , \end{aligned} \tag{C.14}$$

where

$$W = e^{\Lambda(t-s)} [\tilde{H}_{\cdot 1}| \ldots |\tilde{H}_{\cdot d}] ,$$

and some calculation shows that

$$W = \begin{bmatrix} e^{J_1(t-s)} w_{1,1} & \ldots & e^{J_1(t-s)} w_{d,1} \\ \vdots & \ddots & \vdots \\ e^{J_K(t-s)} w_{1,K} & \ldots & e^{J_K(t-s)} w_{d,K} \end{bmatrix} ,$$

recall that

$$J_k = \begin{cases} \lambda_k, & \text{if } k = 1, \ldots, K_1 , \\ \begin{bmatrix} a_k & -b_k \\ b_k & a_k \end{bmatrix}, & \text{if } k = K_1 + 1, \ldots, K . \end{cases}$$

Since $w_{j,k} = 0$ ($\in \mathbb{R}^1$ or $\mathbb{R}^2$), for all $j = 0, 1, \ldots, d$, then if $k \in \{1, \ldots, K_1\}$, the k-th row of W

$$W_{k\cdot} = 0 ;$$

and if $k \in \{K_1 + 1, \ldots, K\}$, then the $(2k - K_1 - 1)$-th and the $(2k - K_1)$-th rows

$$W_{(2k-K_1-1)\cdot} = W_{(2k-K_1)\cdot} = 0 ,$$

where $W_{k\cdot}$ denotes the k-th row vector of matrix W.

If we denote

$$\tilde{V}(t)_{\cdot j} := Q^{-1} V(t)_{\cdot j} , \quad \text{for all } j = 1, \ldots, d ,$$

and

$$w(t)_{j,k} := \begin{cases} \tilde{V}(t)_{\cdot j,k} \in \mathbb{R}^1, & \text{for } k = 1, \dots, K_1\,, \\ (\tilde{V}(t)_{\cdot j,2k-K_1-1}, \tilde{V}(t)_{\cdot j,2k-K_1})^\top \in \mathbb{R}^2, & \text{for } k = K_1+1, \dots, K\,. \end{cases}$$

Then by multiplying Q^{-1} in both sides of Equation (C.14), one obtains that

$$w(t)_{j,k} = 0 \ (\in \mathbb{R}^1 \text{ or } \mathbb{R}^2)$$

for all $j = 1, \dots, d$ and all $0 \leqslant t < \infty$. This indicates that when all the vectors $H_{\cdot j}$ for $j = 1, \dots, d$ are confined to an A-invariant proper subspace of $\mathbb{R}^d$, denotes as L, then each column of the covariance matrix $V(t)$ in Equation (C.13) is also confined to L, for all $0 \leqslant t < \infty$. Therefore, under condition (5.14), $\boldsymbol{x}_0, H_{\cdot j}$ (for all $j = 1, \dots, d$) and each column of the covariance matrix $V(t)$ (for all $0 \leqslant t < \infty$) are confined to an A-invariant proper subspace of $\mathbb{R}^d$. Thus, a matrix A_2 exists, with $A_2 \neq A_1$ such that the first two equations in Equation (C.13) are satisfied.

In particular, by [29, Theorem 2.5], when $k \in \{1, \dots, K_1\}$, there exists matrix $D \in \mathbb{R}^{d\times d}$, with the kk-th element $D_{kk} = c \neq 0$ and all the other elements of D are zeros. Let

$$A_2 = A_1 + QDQ^{-1} \neq A_1\,,$$

then A_1 and A_2 satisfy the first two equations in Equation (C.13). Then we will show that such a A_2 also satisfy the third equation in Equation (C.13).

Some calculation shows that

$$\begin{aligned} V_1(t) &= \int_0^t e^{A_1(t-s)} H e^{A_1^\top (t-s)} ds \\ &= \int_0^t Q e^{\Lambda(t-s)} Q^{-1} H (Q^T)^{-1} e^{\Lambda(t-s)} Q^T ds \\ :&= \int_0^t Q e^{\Lambda(t-s)} P_1 e^{\Lambda(t-s)} Q^T ds\,, \end{aligned} \tag{C.15}$$

where $P_1 := Q^{-1} H (Q^T)^{-1}$. And

$$
\begin{aligned}
V_2(t) &= \int_0^t e^{A_2(t-s)} H e^{A_2^\top (t-s)} ds \\
&= \int_0^t e^{(A_1+QDQ^{-1})(t-s)} H e^{(A_1+QDQ^{-1})^\top (t-s)} ds \\
&= \int_0^t Q e^{\Lambda(t-s)} e^{D(t-s)} Q^{-1} H (Q^T)^{-1} e^{D(t-s)} e^{\Lambda(t-s)} Q^T ds \\
:&= \int_0^t Q e^{\Lambda(t-s)} P_2 e^{\Lambda(t-s)} Q^T ds \,,
\end{aligned}
\tag{C.16}
$$

where $P_2 := e^{D(t-s)} Q^{-1} H (Q^T)^{-1} e^{D(t-s)}$. If one can show that $P_1 = P_2$, then it is readily checked that $V_1(t) = V_2(t)$ for all $0 \leqslant t < \infty$. Recall that

$$
Q^{-1} H = \tilde{H},
$$

where $\tilde{H} = [\tilde{H}_{\cdot 1}| \dots |\tilde{H}_{\cdot d}]$. And when condition (5.14) is not satisfied, the k-th row of $\tilde{H}$:

$$
\tilde{H}_{k\cdot} = 0 \,.
$$

Since

$$
P_1 = Q^{-1} H (Q^T)^{-1} = \tilde{H} (Q^T)^{-1} \,,
$$

therefore, the k-th row of P_1:

$$
(P_1)_{k\cdot} = 0 \,.
$$

Simple calculation shows that matrix P_1 is symmetric, thus, the k-th column of P_1:

$$
(P_1)_{\cdot k} = 0 \,.
$$

It is easy to obtain that $e^{D(t-s)}$ is a diagonal matrix expressed as

$$
e^{D(t-s)} = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & e^{c(t-s)} & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}
$$

where $e^{c(t-s)}$ is the kk-th entry. Then, simple calculation shows that

$$
P_2 = e^{D(t-s)} Q^{-1} H (Q^T)^{-1} e^{D(t-s)} = e^{D(t-s)} P_1 e^{D(t-s)} = P_1 \,.
$$

Therefore, one obtains that

$$V_1(t) = V_2(t)\,, \quad \forall 0 \leqslant t < \infty\,.$$

Hence, when $k \in \{1, \ldots, K_1\}$, we find a A_2, with $A_2 \neq A_1$ such that Equation (C.13) is satisfied.

When $k \in \{K_1 + 1, \ldots, K\}$, there exists matrix $D' \in \mathbb{R}^{d \times d}$, with

$$\begin{bmatrix} D'_{2k-K_1-1,2k-K_1-1} & D'_{2k-K_1-1,2k-K_1} \\ D'_{2k-K_1,2k-K_1-1} & D'_{2k-K_1,2k-K_1} \end{bmatrix} := \begin{bmatrix} c_1 & c_2 \\ c_3 & c_4 \end{bmatrix},$$

where $M_{i,j}$ denotes the ij-th entry of matrix M, $c = [c_1, c_2, c_3, c_4]^\top \neq 0$, and all the other elements of D' are zeros. Let

$$A_2 = A_1 + QD'Q^{-1} \neq A_1\,,$$

then A_1 and A_2 satisfy the first two equations in Equation (C.13). Similar to the case where $k \in \{1, \ldots, K_1\}$, one can also show that such a A_2 also satisfies the third equation in Equation (C.13).

Therefore, assuming A has distinct eigenvalues, then when the identifiability condition (5.14) is not satisfied, one can always find a A_2 with $(A_1, GG^\top) \neq (A_2, GG^\top)$ such that Equation (C.13) is satisfied, i.e., $X^1 \overset{\mathrm{d}}{=} X^2$. Hence, the necessity of the condition is proved. □

C.1.6 Proof of Corollary 5.9

Proof. There are two ways to prove this corollary, we will present both of them in the following.

Way1. By [160, Lemma 2.2],

$$\text{span}([G|AG|\ldots|A^{d-1}G]) = \text{span}([GG^\top|AGG^\top|\ldots|A^{d-1}GG^\top])\,,$$

where span(M) denotes the linear span of the columns of the matrix M. therefore, when

$$\text{rank}([G|AG|\dots|A^{d-1}G]) = d\,,$$

then

$$\text{span}([G|AG|\dots|A^{d-1}G]) = \mathbb{R}^d\,,$$

thus,

$$\text{span}([GG^\top|AGG^\top|\dots|A^{d-1}GG^\top]) = \mathbb{R}^d\,.$$

Therefore,

$$\text{rank}([GG^\top|AGG^\top|\dots|A^{d-1}GG^\top]) = d\,,$$

since the rank of a matrix is the dimension of its span. Then by Theorem 5.8, the generator of the SDE (5.1) is identifiable from $\boldsymbol{x}_0$.

Way2. Let $\tilde{A} \in \mathbb{R}^{d\times d}$ and $\tilde{G} \in \mathbb{R}^{d\times m}$, such that $X(\cdot;\boldsymbol{x}_0, A, G) \overset{\text{d}}{=} X(\cdot;\boldsymbol{x}_0, \tilde{A}, \tilde{G})$, we denote as $X \overset{\text{d}}{=} \tilde{X}$, we will show that under our identifiability condition $(A, GG^\top) = (\tilde{A}, \tilde{G}\tilde{G}^\top)$. By applying the same notations used in the proof of Theorem 5.8, in the following, we denote $A_1 := A$, $A_2 := \tilde{A}$, $G_1 := G$ and $G_2 := \tilde{G}$.

In the proof of Theorem 5.8, we have shown that $G_1G_1^\top = G_2G_2^\top$, thus, we only need to show that under the condition stated in this corollary, $A_1 = A_2$. According to the proof of Theorem 5.8, for all $0 \leqslant t < \infty$, we have

$$\begin{aligned} V_1(t) &= V_2(t)\,,\\ A_1V_1(t) &= A_2V_2(t)\,. \end{aligned}$$

Let $V(t) := V_i(t)(i = 1, 2)$, one gets

$$A_1V(t) = A_2V(t)\,,\quad \forall 0 \leqslant t < \infty\,.$$

Therefore, if there exists a $0 \leqslant t < \infty$, such that $V(t)$ is nonsingular, then one can conclude that $A_1 = A_2$.

By [161, Theorem 3.2], the covariance $V(t)$ is nonsingular for all $t > 0$, if and only if

$$\text{rank}([G|AG|\dots|A^{d-1}G]) = d\,,$$

that is the pair $[A, G]$ is controllable. Therefore, under the condition stated in this corollary, $A_1 = A_2$, thus the generator of the SDE (5.1) is identifiable from $\boldsymbol{x}_0$. $\square$

C.1.7 Proof of Theorem 5.11

Proof. Let $\tilde{A}, \tilde{G}_k \in \mathbb{R}^{d\times d}$ for all $k = 1, \ldots, m$, such that $X(\cdot; \boldsymbol{x}_0, A, \{G_k\}_{k=1}^m) \overset{\mathrm{d}}{=} X(\cdot; \boldsymbol{x}_0, \tilde{A}, \{\tilde{G}_k\}_{k=1}^m)$, we denote as $X \overset{\mathrm{d}}{=} \tilde{X}$, we will show that under our identifiability condition, for all $\boldsymbol{x} \in \mathbb{R}^d$, $(A, \sum_{k=1}^m G_k \boldsymbol{x}\boldsymbol{x}^\top G_k^\top) = (\tilde{A}, \sum_{k=1}^m \tilde{G}_k \boldsymbol{x}\boldsymbol{x}^\top \tilde{G}_k^\top)$. For simplicity of notation, in the following, we denote $A_1 := A$, $A_2 := \tilde{A}$, $G_{1,k} := G_k$ and $G_{2,k} := \tilde{G}_k$, and denote $X \overset{\mathrm{d}}{=} \tilde{X}$ as $X^1 \overset{\mathrm{d}}{=} X^2$.

We first show that $A_1 = A_2$. Set $\boldsymbol{m}_i(t) := \mathbb{E}[X_t^i]$, we know that $\boldsymbol{m}_i(t)$ satisfies the ODE

$$\begin{aligned} \dot{\boldsymbol{m}}_i(t) &= A_i \boldsymbol{m}_i(t)\,, \quad \forall 0 \leqslant t < \infty\,, \\ \boldsymbol{m}_i(0) &= \boldsymbol{x}_0\,, \end{aligned} \tag{C.17}$$

where $\dot{f}(t)$ denotes the first derivative of function $f(t)$ with respect to time t.

Simple calculation shows that

$$\boldsymbol{m}_i(t) = e^{A_i t}\boldsymbol{x}_0\,.$$

Since $X^1 \overset{\mathrm{d}}{=} X^2$, one has

$$\mathbb{E}[X_t^1] = \mathbb{E}[X_t^2]$$

for all $0 \leqslant t < \infty$. That is

$$e^{A_1 t}\boldsymbol{x}_0 = e^{A_2 t}\boldsymbol{x}_0\,, \quad \forall 0 \leqslant t < \infty\,.$$

Taking j-th derivative of $e^{A_i t}\boldsymbol{x}_0$ with respect to t, one finds that

$$\frac{d^j}{dt^j}\Big|_{t=0} e^{A_i t}\boldsymbol{x}_0 = A_i^j \boldsymbol{x}_0\,,$$

for all $j = 1, 2, \ldots$. Consequently,

$$A_1^j \boldsymbol{x}_0 = A_2^j \boldsymbol{x}_0 \,.$$

Let us denote this vector $A^j \boldsymbol{x}_0$. Obviously, one gets

$$A_1 A^{j-1} \boldsymbol{x}_0 = A_2 A^{j-1} \boldsymbol{x}_0 \quad \text{for all } j = 1, 2, \ldots \,. \tag{C.18}$$

By condition A1, it is readily checked that $A_1 = A_2$ from Equation (C.18).

In the following, we show that under condition A2, for all $\boldsymbol{x} \in \mathbb{R}^d$,

$$\sum_{k=1}^{m} G_{1,k} \boldsymbol{x}\boldsymbol{x}^\top G_{1,k}^\top = \sum_{k=1}^{m} G_{2,k} \boldsymbol{x}\boldsymbol{x}^\top G_{2,k}^\top \,.$$

We know the function $P_i(t) := \mathbb{E}[X_t^i (X_t^i)^\top]$ satisfies the ODE

$$\begin{aligned} \dot{P}_i(t) &= A_i P_i(t) + P_i(t) A_i^\top + \sum_{k=1}^{m} G_{i,k} P_i(t) G_{i,k}^\top \,, \quad \forall 0 \leqslant t < \infty \,, \\ P_i(0) &= \boldsymbol{x}_0 \boldsymbol{x}_0^\top \,. \end{aligned} \tag{C.19}$$

Since $X^1 \overset{\mathrm{d}}{=} X^2$,

$$P_1(t) = P_2(t) \,, \quad \forall 0 \leqslant t < \infty \,,$$

let us call it $P(t)$. By differentiating $P_i(t)$ one also gets that

$$\dot{P}_1(t) = \dot{P}_2(t) \,, \quad \forall 0 \leqslant t < \infty \,.$$

Since we have shown that $A_1 = A_2$ under condition A1, from Equation (C.19) one observes that

$$\sum_{k=1}^{m} G_{1,k} P(t) G_{1,k}^\top = \sum_{k=1}^{m} G_{2,k} P(t) G_{2,k}^\top \,, \quad \forall 0 \leqslant t < \infty \,. \tag{C.20}$$

By vectorizing $P(t)$, some calculation shows that the ODE (C.19) can be expressed as

$$\begin{aligned} \operatorname{vec}(\dot{P}(t)) &= \mathcal{A} \operatorname{vec}(P(t)) \,, \\ \operatorname{vec}(P(0)) &= \operatorname{vec}(\boldsymbol{x}_0 \boldsymbol{x}_0^\top) \,, \end{aligned} \tag{C.21}$$

with an explicit solution

$$\mathrm{vec}(P(t)) = e^{\mathcal{A}t}\mathrm{vec}(\boldsymbol{x}_0\boldsymbol{x}_0^\top)\,,$$

where $\mathcal{A} = A \oplus A + \sum_{k=1}^m G_k \otimes G_k \in \mathbb{R}^{d^2\times d^2}$, and $\mathrm{vec}(M)$ denotes the vector by stacking the columns of matrix M vertically.

By definition, $P(t) \in \mathbb{R}^{d\times d}$ is symmetric, thus $\mathrm{vec}(P(t))$ for all $0 \leqslant t < \infty$ is confined to a proper subspace of $\mathbb{R}^{d^2}$, let us denote this proper subspace W, simple calculation shows that

$$\dim(W) = (d^2 + d)/2\,,$$

where $\dim(W)$ denotes the dimension of the subspace W, that is the number of vectors in any basis for W. In particular, one can find a basis of W denoting as

$$\{\mathrm{vec}(E_{11}), \mathrm{vec}(E_{21}), \mathrm{vec}(E_{22}), \ldots, \mathrm{vec}(E_{dd})\}\,,$$

where E_{ij} stands for a $d \times d$ matrix, with the ij-th and ji-th elements are 1 and all other elements are 0, for all $i, j = 1, \ldots, d$ and $i \geqslant j$.

Suppose there exists t_i's, for $i = 1, \ldots, (d^2+d)/2$, such that $\mathrm{vec}(P(t_1)), \ldots, \mathrm{vec}(P(t_{(d^2+d)/2}))$ are linearly independent, then for all $\boldsymbol{x} \in \mathbb{R}^d$,

$$\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^\top) = l_1\mathrm{vec}(P(t_1)) + \ldots + l_{(d^2+d)/2}\mathrm{vec}(P(t_{(d^2+d)/2}))\,,$$

that is

$$\boldsymbol{x}\boldsymbol{x}^\top = l_1P(t_1) + \ldots + l_{(d^2+d)/2}P(t_{(d^2+d)/2})\,,$$

where $\boldsymbol{l} := \{l_1, \ldots, l_{(d^2+d)/2}\} \in \mathbb{R}^{(d^2+d)/2}$. According to Equation (C.20), it is readily checked that for all $\boldsymbol{x} \in \mathbb{R}^d$,

$$\sum_{k=1}^m G_{1,k}\boldsymbol{x}\boldsymbol{x}^\top G_{1,k}^\top = \sum_{k=1}^m G_{2,k}\boldsymbol{x}\boldsymbol{x}^\top G_{2,k}^\top\,.$$

By [28, Lemma 6.1], there exists $(d^2+d)/2$ t_i's such that $\mathrm{vec}(P(t_1)), \ldots, \mathrm{vec}(P(t_{(d^2+d)/2}))$ are linearly independent, if and only if the orbit of $\mathrm{vec}(P(t))$ (i.e., the trajectory of ODE (C.21) started from initial state $\boldsymbol{v}$), denoting as $\gamma(\mathcal{A}, \boldsymbol{v})$ with $\boldsymbol{v} = \mathrm{vec}(\boldsymbol{x}_0\boldsymbol{x}_0^\top)$, is not confined to a proper subspace of W.

Next, we show that under condition A2, orbit $\gamma(\mathcal{A}, \boldsymbol{v})$ is not confined to a proper subspace of W.

Assume orbit $\gamma(\mathcal{A}, \boldsymbol{v})$ is confined to a proper subspace of W. Then there exists $\boldsymbol{w} \neq \boldsymbol{0} \in W$ such that

$$\boldsymbol{w}^\top e^{\mathcal{A}t} \boldsymbol{v} = 0\,, \quad \forall 0 \leqslant t < \infty\,.$$

By taking j-th derivative with respect to t, we have

$$\boldsymbol{w}^T \mathcal{A}^j e^{\mathcal{A}t} \boldsymbol{v} = 0\,, \quad \forall 0 \leqslant t < \infty, j = 0, \dots, (d^2 + d - 2)/2\,.$$

In particular, for $t = 0$,

$$\boldsymbol{w}^T \mathcal{A}^j \boldsymbol{v} = 0\,, \quad \text{for } j = 0, \dots, (d^2 + d - 2)/2\,.$$

Therefore,

$$\boldsymbol{w}^T [\boldsymbol{v} | \mathcal{A}\boldsymbol{v} | \dots | \mathcal{A}^{(d^2+d-2)/2} \boldsymbol{v}] = 0\,. \tag{C.22}$$

Since $\boldsymbol{w} \in W$, $\boldsymbol{w} \in \operatorname{span}\{\operatorname{vec}(E_{11}), \operatorname{vec}(E_{21}), \dots, \operatorname{vec}(E_{dd})\}$, set $\operatorname{vec}(\overline{\boldsymbol{w}}) := \boldsymbol{w}$, then $\overline{\boldsymbol{w}}$ is a $d \times d$ symmetric matrix. Since $P(t)$ is symmetric for all $0 \leqslant t < \infty$, according to Equation (C.19), simple calculation shows that the j-th derivative of $P(t)$ is also symmetric for all $0 \leqslant t < \infty$, for $j = 0, 1, \dots$. Recall that

$$\operatorname{vec}(P(t)) = e^{\mathcal{A}t} \boldsymbol{v}\,,$$
$$\operatorname{vec}(P^{(j)}(t)) = \mathcal{A}^j e^{\mathcal{A}t} \boldsymbol{v}\,,$$

where $P^{(j)}(t)$ denotes the j-th derivative of $P(t)$ with respect to t. In particular, when $t = 0$, one has

$$\operatorname{vec}(P(0)) = \boldsymbol{v}\,,$$
$$\operatorname{vec}(P^{(j)}(0)) = \mathcal{A}^j \boldsymbol{v}\,,$$

then if we denote matrix $\overline{\mathcal{A}^j \boldsymbol{v}}$ by setting $\operatorname{vec}(\overline{\mathcal{A}^j \boldsymbol{v}}) := \mathcal{A}^j \boldsymbol{v}$, matrices $\overline{\mathcal{A}^j \boldsymbol{v}}$ are symmetric for all $j = 0, 1, \dots$.

Therefore, we can say that there are only $(d^2 + d)/2$ distinct elements in each of the vectors: $\boldsymbol{w}, \boldsymbol{v}, \mathcal{A}\boldsymbol{v}, \dots, \mathcal{A}^{(d^2+d-2)/2} \boldsymbol{v}$ in Equation (C.22). Moreover, since these vectors all correspond to $d \times d$ symmetric matrices, the repetitive elements in each vector appear in the same positions in each vector. Hence, we can focus on checking those distinct elements in each vector, that is Equation (C.22) can be expressed as

$$\underline{\boldsymbol{w}}^T[\underline{\boldsymbol{v}}|\underline{\mathcal{A}\boldsymbol{v}}|\ldots|\underline{\mathcal{A}^{(d^2+d-2)/2}\boldsymbol{v}}] = 0\,, \tag{C.23}$$

where $\underline{\boldsymbol{w}} \in \mathbb{R}^{(d^2+d)/2}$ denotes as

$$\underline{\boldsymbol{w}} := [\overline{w}_{11}, \sqrt{2}\overline{w}_{21}, \ldots, \sqrt{2}\overline{w}_{d1}, \overline{w}_{22}, \ldots, \sqrt{2}\overline{w}_{d2}, \ldots, \overline{w}_{dd}]^\top\,,$$

where $\overline{w}_{ij}$ denotes the ij-th element of $\overline{\boldsymbol{w}}$, with $i, j = 1, \ldots, d$ and $i \geqslant j$. When $i \neq j$, the element is multiplied by a $\sqrt{2}$, that is, $\underline{\boldsymbol{w}}$ only keeps the distinctive elements in $\boldsymbol{w}$, and for each of the repetitive element, we multiply $\sqrt{2}$. We define $\underline{\boldsymbol{v}}, \underline{\mathcal{A}\boldsymbol{v}}, \ldots$ using the same way.

Under condition A2, matrix

$$[\underline{\boldsymbol{v}}|\underline{\mathcal{A}\boldsymbol{v}}|\ldots|\underline{\mathcal{A}^{(d^2+d-2)/2}\boldsymbol{v}}] \in \mathbb{R}^{\frac{(d^2+d)}{2}\times\frac{(d^2+d)}{2}}$$

is easily to be checked to be invertible, then $\underline{\boldsymbol{w}} = 0$, thus $\boldsymbol{w} = \mathbf{0}$. This contradicts to $\boldsymbol{w} \neq \mathbf{0}$, therefore, under condition A2, orbit $\gamma(\mathcal{A}, \boldsymbol{v})$ is not confined to a proper subspace of W. Hence, the theorem is proved. □

C.2 Genericity of the derived identifiability conditions

C.2.1 The identifiability condition stated in Theorem 5.11 is generic

We will show that the identifiability condition stated in Theorem 5.11 is generic. Specifically, we will show that for the set of $(\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m) \in \mathbb{R}^{d+(m+1)d^2}$ such that either condition A1 or A2 stated in Theorem 5.11 is violated, has Lebesgue measure zero in $\mathbb{R}^{d+(m+1)d^2}$. In the following, we first present a lemma we will use to prove our main proposition.

Lemma C.2. *Let $p : \mathbb{R}^n \to \mathbb{R}$ be a non-zero polynomial function. Let $Z := \{\boldsymbol{x} \in \mathbb{R}^n : p(\boldsymbol{x}) = 0\}$. Then Z has Lebesgue measure zero in $\mathbb{R}^n$.*

Proof. When $n = 1$, suppose the degree of $\boldsymbol{x}$ is $k \geqslant 1$, then by the fundamental theorem of algebra, there are at most k $\boldsymbol{x}$'s such that $\boldsymbol{x} \in Z$. Therefore, Z has Lebesgue measure zero, since a finite set has measure zero in $\mathbb{R}$.

Suppose the lemma is established for polynomials in $n-1$ variables. Let p be a non-zero polynomial in n variables, say of degree $k \geqslant 1$ in x_n, then we can write

$$p(\boldsymbol{x}', x_n) = \sum_{j=0}^{k} p_j(\boldsymbol{x}') x_n^j \,,$$

where $\boldsymbol{x}' = \{x_1, \ldots, x_{n-1}\}$ and $p_0, \ldots, p_k$ are polynomials in the $n-1$ variables $\{x_1, \ldots, x_{n-1}\}$, and there exists $j \in \{0, \ldots, k\}$ such that p_j is a non-zero polynomial since p is a non-zero polynomial. Then we can denote Z as

$$Z = \{(\boldsymbol{x}', x_n) : p(\boldsymbol{x}', x_n) = 0\} \,.$$

Suppose $(\boldsymbol{x}', x_n) \in Z$, then there are two possibilities:

case 1 $p_0(\boldsymbol{x}') = \ldots = p_k(\boldsymbol{x}') = 0$.

case 2 there exists $i \in \{0, \ldots, k\}$ such that $p_i(\boldsymbol{x}') \neq 0$.

Let

$$A := \{(\boldsymbol{x}', x_n) \in Z : \text{ case 1 is satisfied}\} \,,$$
$$B := \{(\boldsymbol{x}', x_n) \in Z : \text{ case 2 is satisfied}\} \,,$$

then $Z = A \cup B$.

For case 1, recall that there exists $j \in \{0, \ldots, k\}$ such that p_j is a non-zero polynomial, let

$$A_j := \{\boldsymbol{x}' \in \mathbb{R}^{n-1} : p_j(\boldsymbol{x}') = 0\} \,,$$

then by the induction hypothesis, A_j has Lebesgue measure zero in $\mathbb{R}^{n-1}$. Therefore, $A_j \times \mathbb{R}$ has Lebesgue measure zero in $\mathbb{R}^n$. Since $A \subseteq A_j \times \mathbb{R}$, A has Lebesgue measure zero in $\mathbb{R}^n$.

For case 2, let λ^n be Lebesgue measure on $\mathbb{R}^n$, then

$$\begin{aligned}\lambda^n(B) &= \int_{\mathbb{R}^n} \mathbb{1}_B(\boldsymbol{x}', x_n) d\lambda^n \\ &= \int_{\mathbb{R}^n} \mathbb{1}_B(\boldsymbol{x}', x_n) d\boldsymbol{x}' dx_n \\ &= \int_{\mathbb{R}^{n-1}} \Big(\int_{\mathbb{R}} \mathbb{1}_B(\boldsymbol{x}', x_n) dx_n \Big) d\boldsymbol{x}' ,\end{aligned} \tag{C.24}$$

where

$$\mathbb{1}_B(\boldsymbol{x}', x_n) = \begin{cases} 1, & \text{if } (\boldsymbol{x}', x_n) \in B \,, \\ 0, & \text{if } (\boldsymbol{x}', x_n) \notin B \,. \end{cases}$$

The inner integral in Equation (C.24) is equal to zero, since for a fixed $\boldsymbol{x}'$, there are finitely many (indeed, at most k) x_n's such that $p(\boldsymbol{x}', x_n) = 0$ under the condition of case 2. Thus, $\lambda^n(B) = 0$, that is, B has Lebesgue measure zero in $\mathbb{R}^n$. Then it is readily checked that Z has Lebesgue measure zero. □

Now we are ready to present the main proposition.

Proposition C.2.1. Let

$$S := \{(\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m) \in \mathbb{R}^{d+(m+1)d^2} : \text{ either condition A1 or A2 in Theorem 5.11 is violated}\} ,$$

then S has Lebesgue measure zero in $\mathbb{R}^{d+(m+1)d^2}$.

Proof. Let

$$S_A := \{(\boldsymbol{x}_0, A) \in \mathbb{R}^{d+d^2} : \text{ condition A1 is violated}\} ,$$

we first show that S_A has Lebesgue measure zero in $\mathbb{R}^{d+d^2}$. Suppose $(\boldsymbol{x}_0, A) \in S_A$, then $(\boldsymbol{x}_0, A)$ satisfies

$$\operatorname{rank}([\boldsymbol{x}_0|A\boldsymbol{x}_0|\ldots|A^{d-1}\boldsymbol{x}_0]) < d \,,$$

that is the set of vectors $\{\boldsymbol{x}_0, A\boldsymbol{x}_0, \ldots, A^{d-1}\boldsymbol{x}_0\}$ are linearly dependent, this means that

$$\det([\boldsymbol{x}_0|A\boldsymbol{x}_0|\ldots|A^{d-1}\boldsymbol{x}_0]) = 0 \,. \tag{C.25}$$

It is a simple matter of algebra that the left side of (C.25) can be expressed as some universal polynomial of the entries of $\boldsymbol{x}_0$ and entries of A, denotes

$p(\boldsymbol{x}_0, A) = p(x_{01}, \ldots, x_{0d}, a_{11}, a_{12}, \ldots, a_{dd})$, where x_{0j} denotes the j-th entry of $\boldsymbol{x}_0$ and

a_{ij} denotes the ij-th entry of A. Therefore, one concludes that

$$p(\boldsymbol{x}_0, A) = p(x_{01}, \ldots, x_{0d}, a_{11}, a_{12}, \ldots, a_{dd}) = 0\,.$$

Thus, S_A can be expressed as

$$S_A = \{(\boldsymbol{x}_0, A) \in \mathbb{R}^{d+d^2} : p(\boldsymbol{x}_0, A) = 0\}\,.$$

Some calculation shows that

$$p(\boldsymbol{x}_0, A) = \sum_{i_1,\ldots,i_d=1}^{d} x_{0i_1} \ldots x_{0i_d} \det([(A^0)_{\cdot i_1}|(A^1)_{\cdot i_2}| \ldots |(A^{d-1})_{\cdot i_d}])\,, \tag{C.26}$$

where $(M)_{\cdot j}$ denotes the j-th column vector of matrix M. Obviously, $p(\boldsymbol{x}_0, A)$ is a non-zero polynomial function of entries of $\boldsymbol{x}_0$ and entries of A, therefore, by Lemma C.2, S_A has Lebesgue measure zero in $\mathbb{R}^{d+d^2}$. Let

$$S_1 := \{(\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m) \in \mathbb{R}^{d+(m+1)d^2} : \text{ condition A1 is violated}\}\,,$$

then it is readily checked that S_1 has Lebesgue measure zero in $\mathbb{R}^{d+(m+1)d^2}$.

Let

$$S_2 := \{(\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m) \in \mathbb{R}^{d+(m+1)d^2} : \text{ condition A2 is violated}\}\,,$$

we then show that S_2 has Lebesgue measure zero in $\mathbb{R}^{d+(m+1)d^2}$.
Suppose $(\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m) \in S_2$, then $(\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m)$ satisfies

$$\operatorname{rank}([\boldsymbol{v}|\mathcal{A}\boldsymbol{v}| \ldots |\mathcal{A}^{(d^2+d-2)/2}\boldsymbol{v}]) < (d^2 + d)/2\,,$$

recall that $\mathcal{A} = A \oplus A + \sum_{k=1}^m G_k \otimes G_k \in \mathbb{R}^{d^2 \times d^2}$ and $\boldsymbol{v} = \operatorname{vec}(\boldsymbol{x}_0 \boldsymbol{x}_0^\top) \in \mathbb{R}^{d^2}$. According to the proof C.1.7 of Theorem 5.11, we obtain that the set of vectors $\{\boldsymbol{v}, \mathcal{A}\boldsymbol{v}, \ldots, \mathcal{A}^{(d^2+d-2)/2}\boldsymbol{v}\}$ are linearly dependent. Because all of these vectors are transferred from vectorizing $d \times d$ symmetric matrices, thus each of these vectors has only $(d^2 + d)/2$ distinct elements and the repetitive elements appear in the same positions in all vectors. Hence, abuse notation a little bit, we can focus on checking those distinct elements in each vector, that is

$$\{\underline{\boldsymbol{v}}|\underline{\mathcal{A}\boldsymbol{v}}| \ldots |\underline{\mathcal{A}^{(d^2+d-2)/2}\boldsymbol{v}}\}\,, \tag{C.27}$$

where $\underline{\boldsymbol{v}} \in \mathbb{R}^{(d^2+d)/2}$ denotes the vector of deleting the repetitive elements of $\boldsymbol{v}$. Since the set of vectors $\{\boldsymbol{v}, \mathcal{A}\boldsymbol{v}, \dots, \mathcal{A}^{(d^2+d-2)/2}\boldsymbol{v}\}$ are linearly dependent, the set of vectors $\{\underline{\boldsymbol{v}}|\underline{\mathcal{A}\boldsymbol{v}}|\dots|\underline{\mathcal{A}^{(d^2+d-2)/2}\boldsymbol{v}}\}$ are linearly dependent, that is

$$\det([\underline{\boldsymbol{v}}|\underline{\mathcal{A}\boldsymbol{v}}|\dots|\underline{\mathcal{A}^{(d^2+d-2)/2}\boldsymbol{v}}]) = 0\,. \tag{C.28}$$

Each entry of $\boldsymbol{v}$ can be written as a non-zero polynomial function of entries of $\boldsymbol{x}_0$ since $\boldsymbol{v} = \text{vec}(\boldsymbol{x}_0\boldsymbol{x}_0^\top)$. Each entry of $\mathcal{A}$ can be written as a non-zero polynomial function of entries of A and G_k with $k = 1, \dots, m$, since $\mathcal{A} = A \oplus A + \sum_{k=1}^m G_k \otimes G_k \in \mathbb{R}^{d^2 \times d^2}$. Hence, the left side of Equation (C.28) can be expressed as some universal polynomial of the entries of $\boldsymbol{x}_0$, A and G_k's, denotes

$$p(\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m) = p(x_{01}, \dots, x_{0d}, a_{11}, \dots, a_{dd}, G_{1,11}, \dots, G_{1,dd}, \dots, G_{m,11}, \dots, G_{m,dd})\,,$$

where $G_{k,ij}$ denotes the ij-th entry of matrix G_k. Therefore, one concludes that

$$p(\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m) = 0\,.$$

Thus, S_2 can be expressed as

$$S_2 := \{(\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m) \in \mathbb{R}^{d+(m+1)d^2} : p(\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m) = 0\}\,.$$

Similar to the calculation of $p(\boldsymbol{x}_0, A)$ in Equation (C.26), $p(\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m)$ can be expressed as a non-zero polynomial function of entries of v and $\mathcal{A}$, thus it can also be expressed as a non-zero polynomial function of entries of $\boldsymbol{x}_0$, A and G_k's. Therefore, by Lemma C.2, S_2 has Lebesgue measure zero in $\mathbb{R}^{d+(m+1)d^2}$.

We know that $S \subseteq S_1 \cup S_2$, let λ be Lebesgue measure on $\mathbb{R}^{d+(m+1)d^2}$, then one has

$$\lambda(S) \leqslant \lambda(S_1) + \lambda(S_2) = 0.$$

Thus S has Lebesgue measure zero in $\mathbb{R}^{d+(m+1)d^2}$. □

C.2.2 The identifiability condition stated in Theorem 5.8 is generic

We will show that the identifiability condition stated in Theorem 5.8 is generic.

Proposition C.2.2. Let

$$S := \{(\boldsymbol{x}_0, A, G) \in \mathbb{R}^{d+d^2+dm} : \text{condition (5.14) in Theorem 5.8 is violated}\},$$

then S has Lebesgue measure zero in $\mathbb{R}^{d+d^2+dm}$.

Proof. Suppose $(\boldsymbol{x}_0, A, G) \in S$, then $(\boldsymbol{x}_0, A, G)$ satisfies

$$\operatorname{rank}([\boldsymbol{x}_0|A\boldsymbol{x}_0|\ldots|A^{d-1}\boldsymbol{x}_0|H_{\cdot 1}|AH_{\cdot 1}|\ldots|A^{d-1}H_{\cdot 1}|\ldots|H_{\cdot d}|AH_{\cdot d}|\ldots|A^{d-1}H_{\cdot d}]) < d\,,$$

recall that $H := GG^T$, and $H_{\cdot j}$ stands for the j-th column vector of matrix H, for all $j = 1, \cdots, d$.

Let

$$S' := \{(\boldsymbol{x}_0, A, G) \in \mathbb{R}^{d+d^2+dm} : \operatorname{rank}([\boldsymbol{x}_0|A\boldsymbol{x}_0|\ldots|A^{d-1}\boldsymbol{x}_0]) < d\},$$

one observes that $S \subseteq S'$. According to the proof of Proposition C.2.1, it is readily checked that S' has Lebesgue measure zero in $\mathbb{R}^{d+d^2+dm}$. Thus, S has Lebesgue measure zero in $\mathbb{R}^{d+d^2+dm}$. □

C.3 Simulation settings

We present the true underlying system parameters along with the initial states of the SDEs employed in the simulation experiments. We randomly generate the true system parameters that satisfy or violate the corresponding identifiability conditions.

For the SDE (5.1):

1. identifiable case: satisfy condition (5.14) stated in Theorem 5.8:

$$\boldsymbol{x}_0^{\text{id}} = \begin{bmatrix} 1.87 \\ -0.98 \end{bmatrix}, \quad A^{\text{id}} = \begin{bmatrix} 1.76 & -0.1 \\ 0.98 & 0 \end{bmatrix}, \quad G^{\text{id}} = \begin{bmatrix} -0.11 & -0.14 \\ -0.29 & -0.22 \end{bmatrix};$$

2. unidentifiable case: violate condition (5.14):

$$\boldsymbol{x}_0^{\text{un}} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad A^{\text{un}} = \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}, \quad G^{\text{un}} = \begin{bmatrix} 0.11 & 0.22 \\ -0.11 & -0.22 \end{bmatrix}.$$

For the SDE (5.2):

1. identifiable case: satisfy both A1 and A2 stated in Theorem 5.11:

$$\boldsymbol{x}_0^{\text{id}} = \begin{bmatrix} 1.87 \\ -0.98 \end{bmatrix}, \quad A^{\text{id}} = \begin{bmatrix} 1.76 & -0.1 \\ 0.98 & 0 \end{bmatrix}, \quad G_1^{\text{id}} = \begin{bmatrix} -0.11 & -0.14 \\ -0.29 & -0.22 \end{bmatrix}, \quad G_2^{\text{id}} = \begin{bmatrix} -0.17 & 0.59 \\ 0.81 & 0.18 \end{bmatrix};$$

2. unidentifiable case1: violate A1 satisfy A2:

$$\boldsymbol{x}_0^{\text{un-A1}} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad A^{\text{un-A1}} = \begin{bmatrix} 2 & 1 \\ 3 & 0 \end{bmatrix}, \quad G_1^{\text{un-A1}} = \begin{bmatrix} -0.11 & -0.14 \\ -0.29 & -0.22 \end{bmatrix}, \quad G_2^{\text{un-A1}} = \begin{bmatrix} -0.17 & 0.59 \\ 0.81 & 0.18 \end{bmatrix};$$

3. unidentifiable case2: satisfy A1 violate A2:

$$\boldsymbol{x}_0^{\text{un-A2}} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad A^{\text{un-A2}} = \begin{bmatrix} 1 & -2 \\ -1 & 0 \end{bmatrix}, \quad G_1^{\text{un-A2}} = \begin{bmatrix} -0.3 & 0.4 \\ -0.7 & 0.2 \end{bmatrix}, \quad G_2^{\text{un-A2}} = \begin{bmatrix} 0.8 & 0.2 \\ -0.2 & -0.4 \end{bmatrix}.$$

We have discussed in Section 5.4 that we use MLE method to estimate the system parameters from discrete observations sampled from the corresponding SDEs. Specifically, the negative log-likelihood function was minimized using the 'scipy.optimize.minimize' library in Python.

For all of our experiments, we initialized the parameter values as the true parameters plus 2. In the case of the SDE (5.1), we utilized the 'trust-constr' method with the hyper-parameter 'gtol'= 1e-3 and 'xtol'= 1e-3. On the other hand, for the SDE (5.2), we applied the 'BFGS' method and set the hyper-parameter 'gtol'= 1e-2. The selection of the optimization method and the corresponding hyper-parameters was determined through a series of experiments aimed at identifying the most suitable configuration.

C.4 Examples for reliable causal inference for linear SDEs

C.4.1 Example for reliable causal inference for the SDE (5.1)

This example is inspired by [21, Example 5.4]. Recall that the SDE (5.1) is defined as

$$dX_t = AX_t dt + GdW_t, \quad X_0 = \boldsymbol{x}_0,$$

where $0 \leqslant t < \infty$, $A \in \mathbb{R}^{d\times d}$ and $G \in \mathbb{R}^{d\times m}$ are constant matrices, W is an m-dimensional standard Brownian motion. Let $X(t; \boldsymbol{x}_0, A, G)$ denote the solution to the SDE (5.1). Let $\tilde{A} \in \mathbb{R}^{d\times d}$ and $\tilde{G} \in \mathbb{R}^{d\times m}$ define the following SDE:

$$d\tilde{X}_t = \tilde{A}\tilde{X}_t dt + \tilde{G} dW_t, \quad \tilde{X}_0 = \boldsymbol{x}_0,$$

such that

$$X(\cdot; \boldsymbol{x}_0, A, G) \overset{\mathrm{d}}{=} \tilde{X}(\cdot; \boldsymbol{x}_0, \tilde{A}, \tilde{G}).$$

Then under our proposed identifiability condition stated in Theorem 5.8, we have shown that the generator of the SDE (5.1) is identifiable, i.e., $(A, GG^\top) = (\tilde{A}, \tilde{G}\tilde{G}^\top)$. Till now, we have shown that under our proposed identifiability conditions, the observational distribution $\xrightarrow{\text{identity}}$ the generator of the observational SDE. Then we will show that the post-intervention distribution is also identifiable. For notational simplicity, we consider intervention on the first coordinate, making the intervention $X_t^1 = \xi$ and $\tilde{X}_t^1 = \xi$ for $0 \leqslant t < \infty$. It will suffice to show equality of the distributions of the non-intervened coordinates (i.e., $X_t^{(-1)}$ and $\tilde{X}_t^{(-1)}$, note the superscripts do not denote reciprocals, but denote the $(d-1)$-coordinates without the first coordinate). Express the matrices of A and G in blocks

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad G = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix},$$

where $A_{11} \in \mathbb{R}^{1\times 1}$, $A_{22} \in \mathbb{R}^{(d-1)\times(d-1)}$, $G_1 \in \mathbb{R}^{1\times m}$ and $G_2 \in \mathbb{R}^{(d-1)\times m}$. Also, consider corresponding expressions of matrices $\tilde{A}$ and $\tilde{G}$. By making intervention $X_t^1 = \xi$, one obtains the post-intervention process of the first SDE satisfies:

$$dX_t^{(-1)} = (A_{21}\xi + A_{22}X_t^{(-1)})dt + G_2 dW_t\,, \quad X_0^{(-1)} = \boldsymbol{x}_0^{(-1)}\,,$$

which is a multivariate Ornstein-Uhlenbeck process, according to [162, Corollary 1], this process is a Gaussian process, assuming A_{22} is invertible, then the mean vector can be described as

$$E[X_t^{(-1)}] = e^{A_{22}t}\boldsymbol{x}_0^{(-1)} - (I - e^{A_{22}t})A_{22}^{-1}A_{21}\xi,$$

and based on [162, Theorem 2], the cross-covariance can be described as

$$\begin{aligned} V(X_{t+h}^{(-1)}, X_t^{(-1)}) :&= \mathbb{E}\{(X_{t+h}^{(-1)} - \mathbb{E}[X_{t+h}^{(-1)}])(X_t^{(-1)} - \mathbb{E}[X_t^{(-1)}])^\top\} \\ &= \int_0^t e^{A_{22}(t+h-s)} G_2 G_2^\top e^{A_{22}^\top (t-s)} ds\,. \end{aligned}$$

Similarly, one can obtain that the mean vector and cross-covariance of the distribution of the post-intervention process of the second SDE by making intervention $\tilde{X}_t^1 = \xi$ satisfy:

$$E[\tilde{X}_t^{(-1)}] = e^{\tilde{A}_{22}t}\boldsymbol{x}_0^{(-1)} - (I - e^{\tilde{A}_{22}t})\tilde{A}_{22}^{-1}\tilde{A}_{21}\xi,$$

and

$$\begin{aligned} V(\tilde{X}_{t+h}^{(-1)}, \tilde{X}_t^{(-1)}) :&= \mathbb{E}\{(\tilde{X}_{t+h}^{(-1)} - \mathbb{E}[\tilde{X}_{t+h}^{(-1)}])(\tilde{X}_t^{(-1)} - \mathbb{E}[\tilde{X}_t^{(-1)}])^\top\} \\ &= \int_0^t e^{\tilde{A}_{22}(t+h-s)} \tilde{G}_2 \tilde{G}_2^\top e^{\tilde{A}_{22}^\top (t-s)} ds \end{aligned}.$$

Then we will show that $E[X_t^{(-1)}] = E[\tilde{X}_t^{(-1)}]$, and $V(X_{t+h}^{(-1)}, X_t^{(-1)}) = V(\tilde{X}_{t+h}^{(-1)}, \tilde{X}_t^{(-1)})$ for all $0 \leqslant t, h < \infty$. Recall that we have shown $(A, GG^\top) = (\tilde{A}, \tilde{G}\tilde{G}^\top)$, thus, $A_{22} = \tilde{A}_{22}$ and $A_{21} = \tilde{A}_{21}$, then it is readily checked that $E[X_t^{(-1)}] = E[\tilde{X}_t^{(-1)}]$ for all $0 \leqslant t < \infty$.

Since

$$GG^\top = \begin{bmatrix} G_1G_1^\top & G_1G_2^\top \\ G_2G_1^\top & G_2G_2^\top \end{bmatrix} = \tilde{G}\tilde{G}^\top,$$

thus, $G_2G_2^\top = \tilde{G}_2\tilde{G}_2^\top$, then it is readily checked that $V(X_{t+h}^{(-1)}, X_t^{(-1)}) = V(\tilde{X}_{t+h}^{(-1)}, \tilde{X}_t^{(-1)})$ for all $0 \leqslant t, h < \infty$. Since both of these two post-intervention processes are Gaussian processes, according to Lemma 5.6, the distributions of these two post-intervention processes are the same. That is, the post-intervention distribution is identifiable.

C.4.2 Example for reliable causal inference for the SDE (5.2)

Recall that the SDE (5.2) is defined as

$$dX_t = AX_t dt + \textstyle\sum_{k=1}^m G_k X_t dW_{k,t}, \quad X_0 = \boldsymbol{x}_0,$$

where $0 \leqslant t < \infty$, $A, G_k \in \mathbb{R}^{d\times d}$ for $k = 1, \ldots, m$ are some constant matrices, $W := \{W_t = [W_{1,t}, \ldots, W_{m,t}]^\top : 0 \leqslant t < \infty\}$ is an m-dimensional standard Brownian motion.

Let $X(t;\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m)$ denote the solution to the SDE (5.2). Let $\tilde{A}, \tilde{G}_k \in \mathbb{R}^{d\times d}$ for $k = 1, \ldots, m$ define the following SDE:

$$d\tilde{X}_t = \tilde{A}\tilde{X}_t dt + \textstyle\sum_{k=1}^m \tilde{G}_k \tilde{X}_t dW_{k,t}, \quad \tilde{X}_0 = \boldsymbol{x}_0,$$

such that

$$X(\cdot;\boldsymbol{x}_0, A, \{G_k\}_{k=1}^m) \overset{\mathrm{d}}{=} \tilde{X}(\cdot;\boldsymbol{x}_0, \tilde{A}, \{\tilde{G}_k\}_{k=1}^m)\,.$$

Then under our proposed identifiability condition stated in Theorem 5.11, we have shown that the generator of the SDE (5.2) is identifiable,
i.e., $(A, \sum_{k=1}^m G_k \boldsymbol{x}\boldsymbol{x}^\top G_k^\top) = (\tilde{A}, \sum_{k=1}^m \tilde{G}_k \boldsymbol{x}\boldsymbol{x}^\top \tilde{G}_k^\top)$ for all $\boldsymbol{x} \in \mathbb{R}^d$. Till now, we have shown that under our proposed identifiability conditions, the observational distribution $\xrightarrow{\text{identity}}$ the generator of the observational SDE. Then we aim to show that the post-intervention distribution is also identifiable. For notational simplicity, we consider intervention on the first coordinate, making the intervention $X_t^1 = \xi$ and $\tilde{X}_t^1 = \xi$ for $0 \leqslant t < \infty$. It will suffice to show equality of the distributions of the non-intervened coordinates (i.e., $X_t^{(-1)}$ and $\tilde{X}_t^{(-1)}$). Express the matrices of A and G_k for $k = 1, \ldots, m$ in blocks

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad G_k = \begin{bmatrix} G_{k,11} & G_{k,12} \\ G_{k,21} & G_{k,22} \end{bmatrix},$$

where $A_{11}, G_{k,11} \in \mathbb{R}^{1\times 1}$, $A_{22}, G_{k,22} \in \mathbb{R}^{(d-1)\times(d-1)}$. Also consider corresponding expressions of matrices $\tilde{A}$ and $\tilde{G}_k$ for $k = 1, \ldots, m$. By making intervention $X_t^1 = \xi$, one obtains the post-intervention process of the first SDE satisfies:

$$dX_t^{(-1)} = (A_{21}\xi + A_{22}X_t^{(-1)})dt + \textstyle\sum_{k=1}^m (G_{k,21}\xi + G_{k,22}X_t^{(-1)})dW_{k,t}, \quad X_0^{(-1)} = \boldsymbol{x}_0^{(-1)}.$$

Since this post-intervention process is not a Gaussian process, one cannot explicitly show that the post-intervention distribution is identifiable. Instead, we check the surrogate of the post-intervention distribution, that is the first- and second-order moments of the post-intervention process $X_t^{(-1)}$. Which denote as $\boldsymbol{m}(t)^{(-1)} = \mathbb{E}[X_t^{(-1)}]$ and $P(t)^{(-1)} = \mathbb{E}[X_t^{(-1)}(X_t^{(-1)})^\top]$ respectively. Then $\boldsymbol{m}(t)^{(-1)}$ and $P(t)^{(-1)}$ satisfy the following ODE systems:

$$\frac{d\boldsymbol{m}(t)^{(-1)}}{dt} = A_{21}\xi + A_{22}\boldsymbol{m}(t)^{(-1)}, \quad \boldsymbol{m}(0)^{-1} = \boldsymbol{x}_0^{(-1)},$$

and

$$\begin{aligned}\frac{dP(t)^{(-1)}}{dt} &= \boldsymbol{m}(t)^{(-1)}\xi^\top A_{21}^\top + A_{21}\xi(\boldsymbol{m}(t)^{(-1)})^\top + P(t)^{(-1)}A_{22}^\top + A_{22}P(t)^{(-1)} \\ &+ \textstyle\sum_{k=1}^m (G_{k,21}\xi\xi^\top G_{k,21}^\top + G_{k,22}\boldsymbol{m}(t)^{(-1)}\xi^\top G_{k,21}^\top + G_{k,21}\xi(\boldsymbol{m}(t)^{(-1)})^\top G_{k,22}^\top \\ &+ G_{k,22}P(t)^{(-1)}G_{k,22}^\top), \quad P(0)^{(-1)} = \boldsymbol{x}_0^{(-1)}(\boldsymbol{x}_0^{(-1)})^\top.\end{aligned} \tag{C.29}$$

Similarly, one can obtain the ODE systems describing the $\tilde{\boldsymbol{m}}(t)^{(-1)}$ and $\tilde{P}(t)^{(-1)}$. Then we will show that $\boldsymbol{m}(t)^{(-1)} = \tilde{\boldsymbol{m}}(t)^{(-1)}$ and $P(t)^{(-1)} = \tilde{P}(t)^{(-1)}$ for all $0 \leqslant t < \infty$. Recall that we have shown $A = \tilde{A}$, thus $A_{21} = \tilde{A}_{21}$ and $A_{22} = \tilde{A}_{22}$, then it is readily checked that $\boldsymbol{m}(t)^{(-1)} = \tilde{\boldsymbol{m}}(t)^{(-1)}$ for all $0 \leqslant t < \infty$.

In the proof of Theorem 5.11, we have shown that

$$\textstyle\sum_{k=1}^m G_k P(t) G_k^\top = \sum_{k=1}^m \tilde{G}_k P(t) \tilde{G}_k^\top$$

for all $0 \leqslant t < \infty$, where $P(t) = \mathbb{E}[X_t X_t^\top]$. Simple calculation shows that

$$\sum_{k=1}^m G_k P(t) G_k^\top = \sum_{k=1}^m \left(\begin{bmatrix} G_{k,11} & G_{k,12} \\ G_{k,21} & G_{k,22} \end{bmatrix} \begin{bmatrix} \xi\xi^\top & \xi(\boldsymbol{m}(t)^{(-1)})^\top \\ \boldsymbol{m}(t)^{(-1)}\xi^\top & P(t)^{(-1)} \end{bmatrix} \begin{bmatrix} G_{k,11}^\top & G_{k,21}^\top \\ G_{k,12}^\top & G_{k,22}^\top \end{bmatrix} \right),$$

Then one can get that the $(2,2)$-th block entry of the matrix $\sum_{k=1}^m G_k P(t) G_k^\top$ is the same as the $\sum_{k=1}^m(\ldots)$ part in the ODE corresponds to $P(t)^{(-1)}$ (i.e., Equation (C.29)), since $\sum_{k=1}^m G_k P(t) G_k^\top = \sum_{k=1}^m \tilde{G}_k P(t) \tilde{G}_k^\top$, then the $\sum_{k=1}^m(\ldots)$ part in the ODEs correspond to both $P(t)^{(-1)}$ and $\tilde{P}(t)^{(-1)}$ are the same. Thus, it is readily checked that $P(t)^{(-1)} = \tilde{P}(t)^{(-1)}$ for all $0 \leqslant t < \infty$.

Though we cannot explicitly show that the post-intervention distribution is identifiable, showing that the first- and second-order moments of the post-intervention process is identifiable can indicate the identification of the post-intervention distribution to a considerable extent.

C.5 Conditions for identifying the generator of a linear SDE system with multiplicative noise when its explicit solution is available

Proposition C.5.1. Let $\boldsymbol{x}_0 \in \mathbb{R}^d$ be fixed. The generator of the SDE (5.2) is identifiable from $\boldsymbol{x}_0$ if the following conditions are satisfied:

C1 $\mathrm{rank}([\boldsymbol{x}_0|A\boldsymbol{x}_0|\ldots|A^{d-1}\boldsymbol{x}_0|H_{\cdot 1}|AH_{\cdot 1}|\ldots|A^{d-1}H_{\cdot 1}|\ldots|H_{\cdot d}|AH_{\cdot d}|\ldots|A^{d-1}H_{\cdot d}]) = d\,,$

C2 $\mathrm{rank}([\boldsymbol{v}|\mathcal{A}\boldsymbol{v}|\ldots|\mathcal{A}^{(d^2+d-2)/2}\boldsymbol{v}]) = (d^2+d)/2\,,$

C3 $AG_k = G_kA$ and $G_kG_l = G_lG_k$ for all $k, l = 1, \ldots, m\,.$

where $H := \sum_{k=1}^m G_k\boldsymbol{x}_0\boldsymbol{x}_0^\top G_k^\top$, and $H_{\cdot j}$ stands for the j-th column vector of matrix H, for all $j = 1, \cdots, d$. And $\mathcal{A} = A \oplus A + \sum_{k=1}^m G_k \otimes G_k \in \mathbb{R}^{d^2\times d^2}$, $\oplus$ denotes Kronecker sum and $\otimes$ denotes Kronecker product, $\boldsymbol{v}$ is a d^2-dimensional vector defined by $\boldsymbol{v} := \mathrm{vec}(\boldsymbol{x}_0\boldsymbol{x}_0^\top)$, where $\mathrm{vec}(M)$ denotes the vectorization of matrix M.

Proof. Let $\tilde{A}, \tilde{G}_k \in \mathbb{R}^{d\times d}$ and $\tilde{A}\tilde{G}_k = \tilde{G}_k\tilde{A}$, $\tilde{G}_k\tilde{G}_l = \tilde{G}_l\tilde{G}_k$ for all $k, l = 1, \ldots, m$, such that $X(\cdot; \boldsymbol{x}_0, A, \{G_k\}_{k=1}^m) \overset{\mathrm{d}}{=} X(\cdot; \boldsymbol{x}_0, \tilde{A}, \{\tilde{G}_k\}_{k=1}^m)$, we denote as $X \overset{\mathrm{d}}{=} \tilde{X}$, we will show that under our identifiability condition, for all $\boldsymbol{x} \in \mathbb{R}^d$, $(A, \sum_{k=1}^m G_k\boldsymbol{x}\boldsymbol{x}^\top G_k^\top) = (\tilde{A}, \sum_{k=1}^m \tilde{G}_k\boldsymbol{x}\boldsymbol{x}^\top \tilde{G}_k^\top)$. By applying the same notations used in the proof of Theorem 5.11, in the following, we denote $A_1 := A$, $A_2 := \tilde{A}$, $G_{1,k} := G_k$ and $G_{2,k} := \tilde{G}_k$, and denote $X \overset{\mathrm{d}}{=} \tilde{X}$ as $X^1 \overset{\mathrm{d}}{=} X^2$.

We first show that $H_1 = H_2$ ($H_i := \sum_{k=1}^m G_{i,k}\boldsymbol{x}_0\boldsymbol{x}_0^\top G_{i,k}^\top$). Indeed, since X^1, X^2 have the same distribution, one has

$$\mathbb{E}[f(X_t^1)] = \mathbb{E}[f(X_t^2)] \tag{C.30}$$

for all $0 \leqslant t < \infty$ and $f \in C^\infty(\mathbb{R}^d)$. By differentiating (C.30) at $t = 0$, one finds that

$$(\mathcal{L}_1 f)(\boldsymbol{x}_0) = (\mathcal{L}_2 f)(\boldsymbol{x}_0)\,, \tag{C.31}$$

where $\mathcal{L}_i$ is the generator of X^i $(i=1,2)$. Based on the Proposition 5.2.1,

$$(\mathcal{L}_i f)(\boldsymbol{x}_0)=\sum_{k=1}^{d}\sum_{l=1}^{d}(A_i)_{kl}x_{0l}\frac{\partial f}{\partial x_k}(\boldsymbol{x}_0)+\frac{1}{2}\sum_{k,l=1}^{d}(H_i)_{kl}\frac{\partial^2 f}{\partial x_k \partial x_l}(\boldsymbol{x}_0)\,,$$

where $(M)_{kl}$ denotes the kl-entry of matrix M, and x_{0l} is the l-th component of $\boldsymbol{x}_0$. Since (C.31) is true for all f, by taking

$$f(\boldsymbol{x})=(x_p-x_{0p})(x_q-x_{0q})\,,$$

it is readily checked that

$$(H_1)_{pq}=(H_2)_{pq}\,,$$

for all $p,q=1,\ldots,d$. As a result, $H_1=H_2$. Let us call this matrix H. That is

$$H:=H_1=\sum_{k=1}^{m}G_{1,k}\boldsymbol{x}_0\boldsymbol{x}_0^\top G_{1,k}^\top=\sum_{k=1}^{m}G_{2,k}\boldsymbol{x}_0\boldsymbol{x}_0^\top G_{2,k}^\top=H_2\,.$$

In the proof of Theorem 5.11, we have shown that

$$A_1A^{j-1}\boldsymbol{x}_0=A_2A^{j-1}\boldsymbol{x}_0 \quad \text{for all } j=1,2,\ldots\,, \tag{C.32}$$

next, we will derive the relationship between A_i and H. Under condition C3, the SDE system (5.2) has an explicit solution (cf. [51]):

$$X_t:=X(t;\boldsymbol{x}_0,A,\{G_k\}_{k=1}^m)=\exp\left\{\left(A-\frac{1}{2}\sum_{k=1}^{m}G_k^2\right)t+\sum_{k=1}^{m}G_kW_{k,t}\right\}\boldsymbol{x}_0\,, \tag{C.33}$$

then the covariance of X_t, $P(t,t+h)=\mathbb{E}[X_tX_{t+h}^\top]$ can be calculated as

$$\begin{aligned}&\mathbb{E}[X_tX_{t+h}^\top]\\ =&\,\mathbb{E}\Bigg[\exp\left\{\left(A-\frac{1}{2}\sum_{k=1}^{m}G_k^2\right)t+\sum_{k=1}^{m}G_kW_{k,t}\right\}\boldsymbol{x}_0\boldsymbol{x}_0^\top\exp\left\{\left(A^\top-\frac{1}{2}\sum_{k=1}^{m}(G_k^2)^\top\right)(t+h)\right.\\ &\left.+\sum_{k=1}^{m}G_k^\top W_{k,t+h}\right\}\Bigg]\\ =&\,e^{At}e^{-\frac{1}{2}\sum_{k=1}^m G_k^2 t}\mathbb{E}\left[e^{\sum_{k=1}^m G_kW_{k,t}}\boldsymbol{x}_0\boldsymbol{x}_0^T e^{\sum_{k=1}^m G_k^\top W_{k,t+h}}\right]e^{-\frac{1}{2}\sum_{k=1}^m(G_k^2)^\top(t+h)}e^{A^\top(t+h)}\,,\end{aligned} \tag{C.34}$$

where

$$\begin{aligned}
&\mathbb{E}\big[e^{\sum_{k=1}^m G_k W_{k,t}} \boldsymbol{x}_0 \boldsymbol{x}_0^T e^{\sum_{k=1}^m G_k^\top W_{k,t+h}}\big] \\
&= \mathbb{E}\big[e^{\sum_{k=1}^m G_k W_{k,t}} \boldsymbol{x}_0 \boldsymbol{x}_0^T e^{\sum_{k=1}^m G_k^\top W_{k,t}} e^{-\sum_{k=1}^m G_k^\top W_{k,t}} e^{\sum_{k=1}^m G_k^\top W_{k,t+h}}\big] \\
&= \mathbb{E}\big[e^{\sum_{k=1}^m G_k W_{k,t}} \boldsymbol{x}_0 \boldsymbol{x}_0^T e^{\sum_{k=1}^m G_k^\top W_{k,t}} e^{\sum_{k=1}^m G_k^\top (W_{k,t+h}-W_{k,t})}\big] \\
&= \mathbb{E}\big[e^{\sum_{k=1}^m G_k W_{k,t}} \boldsymbol{x}_0 \boldsymbol{x}_0^T e^{\sum_{k=1}^m G_k^\top W_{k,t}}\big] \mathbb{E}\big[e^{\sum_{k=1}^m G_k^\top (W_{k,t+h}-W_{k,t})}\big] ,
\end{aligned} \tag{C.35}$$

because the Brownian motion $W_{k,t}$ has independent increments.

It is known that, for $Z \sim \mathcal{N}(0,1)$, we have that the jth moment is

$$\mathbb{E}(Z^j) = \begin{cases} 0\,, & j \text{ is odd}\,, \\ 2^{-j/2} \frac{j!}{(j/2)!}\,, & j \text{ is even}\,. \end{cases}$$

Since $W_{k,t} \sim \mathcal{N}(0,t)$, we have

$$\begin{aligned}
\mathbb{E}[e^{G_k W_{k,t}}] &= \mathbb{E}\left[\sum_{j=0}^{\infty} \frac{(G_k)^j (W_{k,t})^j}{j!}\right] \\
&= \sum_{j=0}^{\infty} \frac{(G_k)^j \mathbb{E}[(W_{k,t})^j]}{j!} \\
&= \sum_{j=0,2,4\ldots}^{\infty} \frac{(G_k)^j (t/2)^{j/2}}{(j/2)!} \\
&= \sum_{i=0}^{\infty} \frac{(G_k^2 t/2)^i}{i!} \\
&= e^{G_k^2 t/2}\,.
\end{aligned}$$

Similarly, we have

$$\mathbb{E}[e^{G_k^\top (W_{k,t+h}-W_{k,t})}] = e^{(G_k^\top)^2 h/2}\,.$$

Simple calculation shows that

$$\mathbb{E}\big[e^{\sum_{k=1}^m G_k^\top (W_{k,t+h}-W_{k,t})}\big] = e^{\sum_{k=1}^m (G_k^\top)^2 h/2}\,. \tag{C.36}$$

By combining Equations (C.34), (C.35) and (C.36), one readily obtains that

$$P(t,t+h) = P(t,t) e^{A^\top h}\,, \tag{C.37}$$

we denote $P(t) := P(t,t)$. Set $P_i(t,t+h) = \mathbb{E}[X_t^i(X_{t+h}^i)^\top]$, since $X^1 \overset{\mathrm{d}}{=} X^2$, it follows that

$$P_1(t,t+h) = \mathbb{E}[X_t^1(X_{t+h}^1)^\top] = \mathbb{E}[X_t^2(X_{t+h}^2)^\top] = P_2(t,t+h) \quad \forall t,h \geqslant 0\,.$$

To obtain information about A, let us fix t for now and take j-th derivative of (C.37) with respect to h. One finds that

$$\frac{d^j}{dh^j}\bigg|_{h=0} P(t,t+h) = P(t)(A^\top)^j\,, \tag{C.38}$$

for all $j = 1,2,\ldots$. It is readily checked that

$$P_1(t)(A_1^\top)^j = P_2(t)(A_2^\top)^j \quad \forall 0 \leqslant t < \infty\,. \tag{C.39}$$

We know the function $P_i(t)$ satisfies the ODE

$$\begin{aligned}
&\dot{P}_i(t) = A_i P_i(t) + P_i(t)A_i^\top + \sum_{k=1}^m G_{i,k}P_i(t)G_{i,k}^\top\,, \quad \forall 0 \leqslant t < \infty\,,\\
&P_i(0) = \boldsymbol{x}_0\boldsymbol{x}_0^\top\,.
\end{aligned} \tag{C.40}$$

In particular,

$$\dot{P}_i(0) = A_i\boldsymbol{x}_0\boldsymbol{x}_0^\top + \boldsymbol{x}_0\boldsymbol{x}_0^\top A_i^\top + \sum_{k=1}^m G_{i,k}\boldsymbol{x}_0\boldsymbol{x}_0^\top G_{i,k}^\top\,.$$

By differentiating (C.39) with respect to t at $t=0$, it follows that

$$\begin{aligned}
&A_1\boldsymbol{x}_0\boldsymbol{x}_0^\top(A_1^\top)^j + \boldsymbol{x}_0\boldsymbol{x}_0^\top(A_1^\top)^{j+1} + \left(\sum_{k=1}^m G_{1,k}\boldsymbol{x}_0\boldsymbol{x}_0^\top G_{1,k}^\top\right)(A_1^\top)^j\\
=&A_2\boldsymbol{x}_0\boldsymbol{x}_0^\top(A_2^\top)^j + \boldsymbol{x}_0\boldsymbol{x}_0^\top(A_2^\top)^{j+1} + \left(\sum_{k=1}^m G_{2,k}\boldsymbol{x}_0\boldsymbol{x}_0^\top G_{2,k}^\top\right)(A_2^\top)^j\,.
\end{aligned}$$

Since we have known that $A_1^j\boldsymbol{x}_0 = A_2^j\boldsymbol{x}_0$ for all $j = 1,2,\ldots$, it is readily checked that

$$\left(\sum_{k=1}^m G_{1,k}\boldsymbol{x}_0\boldsymbol{x}_0^\top G_{1,k}^\top\right)(A_1^\top)^j = \left(\sum_{k=1}^m G_{2,k}\boldsymbol{x}_0\boldsymbol{x}_0^\top G_{2,k}^\top\right)(A_2^\top)^j\,,$$

that is $A_1^jH = A_2^jH$ for all $j = 1,2,\ldots$. Let us denote this matrix A^jH. Obviously, by rearranging this matrix, one gets

$$A_1A^{j-1}H = A_2A^{j-1}H \quad \text{ for all } j = 1,2,\ldots$$

Therefore, under condition C1, that is $\mathrm{rank}(M) = d$ with

$$M := [\boldsymbol{x}_0 | A\boldsymbol{x}_0 | \ldots | A^{d-1}\boldsymbol{x}_0 | H_{\cdot 1} | AH_{\cdot 1} | \ldots | A^{d-1}H_{\cdot 1} | \ldots | H_{\cdot d} | AH_{\cdot d} | \ldots | A^{d-1}H_{\cdot d}] \,. \quad \text{(C.41)}$$

if we denote the j-th column in M as $M_{\cdot j}$, one gets $A_1 M_{\cdot j} = A_2 M_{\cdot j}$ for all $j = 1, \ldots, d+d^2$ by equations (C.32) and (C.41).

This means one can find a full-rank matrix $B \in \mathbb{R}^{d\times d}$ by horizontally stacking d linearly independent columns from matrix M, such that $A_1 B = A_2 B$. Since B is invertible, one thus concludes that $A_1 = A_2$.

In the proof of Theorem 5.11, we have shown that when $A_1 = A_2$, under condition C2, for all $\boldsymbol{x} \in \mathbb{R}^d$,

$$\sum_{k=1}^{m} G_{1,k}\boldsymbol{x}\boldsymbol{x}^\top G_{1,k}^\top = \sum_{k=1}^{m} G_{2,k}\boldsymbol{x}\boldsymbol{x}^\top G_{2,k}^\top \,.$$

Thus the proposition is proved. □

It is noteworthy that Proposition C.5.1 is established on the explicit solution assumption of the SDE (5.2), which requires both sets of vectors $\{A, \{G_k\}_{k=1}^m\}$ and $\{\tilde{A}, \{\tilde{G}_k\}_{k=1}^m\}$ to satisfy condition C3. As aforementioned, condition C3 is very restrictive and impractical, rendering the identifiability condition derived in this proposition unsatisfactory. Nonetheless, this condition is presented to illustrate that condition C1 is more relaxed compared to condition A1 stated in Theorem 5.11 when identifying A with the incorporation of G_k's information.

Bibliography

[1] Yuanyuan Wang, Wei Huang, Mingming Gong, Xi Geng, Tongliang Liu, Kun Zhang, and Dacheng Tao. Identifiability and asymptotics in learning homogeneous linear ode systems from discrete observations. *Journal of Machine Learning Research*, 25 (154):1–50, 2024.

[2] Yuanyuan Wang, Biwei Huang, Wei Huang, Xi Geng, and Mingming Gong. Identifiability analysis of linear ode systems with hidden confounders. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[3] Yuanyuan Wang, Xi Geng, Wei Huang, Biwei Huang, and Mingming Gong. Generator identification for linear sdes with additive and multiplicative noise. *Advances in Neural Information Processing Systems*, 36, 2024.

[4] Erina Sakamoto and Hitoshi Iba. Inferring a system of differential equations for a gene regulatory network by using genetic programming. In *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No. 01TH8546)*, volume 1, pages 720–726. IEEE, 2001.

[5] Hulin Wu, Tao Lu, Hongqi Xue, and Hua Liang. Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *Journal of the American Statistical Association*, 109(506):700–716, 2014.

[6] Jutta Gebert, Nicole Radde, and G-W Weber. Modeling gene regulatory networks with piecewise linear differential equations. *European Journal of Operational Research*, 181(3):1148–1165, 2007.

[7] Wayne Materi and David S Wishart. Computational systems biology in drug discovery and development: methods and applications. *Drug discovery today*, 12(7-8): 295–303, 2007.

[8] Carla Rezende Barbosa Bonin, Guilherme Cortes Fernandes, Rodrigo Weber Dos Santos, and Marcelo Lobosco. Mathematical modeling based on ordinary differential equations: A promising approach to vaccinology. *Human vaccines & immunotherapeutics*, 13(2):484–489, 2017.

[9] Brian Ewald and Cécile Penland. Numerical generation of stochastic differential equations in climate models. In *Handbook of numerical analysis*, volume 14, pages 279–306. Elsevier, 2009.

[10] Jeehyun Hwang, Jeongwhan Choi, Hwangyong Choi, Kookjin Lee, Dongeun Lee, and Noseong Park. Climate modeling with neural diffusion equations. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 230–239. IEEE, 2021.

[11] Andrew J Majda, Ilya Timofeyev, and Eric Vanden Eijnden. A mathematical framework for stochastic climate models. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 54(8): 891–974, 2001.

[12] Bernt Oksendal. *Stochastic differential equations: an introduction with applications.* Springer Science & Business Media, 2013.

[13] Clifford R Wymer. Econometric estimation of stochastic differential equation systems. *Econometrica: Journal of the Econometric Society*, pages 565–577, 1972.

[14] Abodolsadeh Neisy and Moslem Peymany. Financial modeling by ordinary and stochastic differential equations. *World Applied Sciences Journal*, 13(11):2288–2295, 2011.

[15] Antti Saarinen, Marja-Leena Linne, and Olli Yli-Harja. Stochastic differential equation model for cerebellar granule cell excitability. *PLoS computational biology*, 4(2): e1000004, 2008.

[16] Mahnoosh Tajmirriahi and Zahra Amini. Modeling of seizure and seizure-free eeg signals based on stochastic differential equations. *Chaos, Solitons & Fractals*, 150: 111104, 2021.

[17] James A Roberts, Karl J Friston, and Michael Breakspear. Clinical applications of stochastic dynamic models of the brain, part i: a primer. *Biological psychiatry: cognitive neuroscience and neuroimaging*, 2(3):216–224, 2017.

[18] JM Mooij, D Janzing, and B Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. In *29th Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, pages 440–448. AUAI Press, 2013.

[19] P Rubenstein, S Bongers, B Schölkopf, and JM Mooij. From deterministic odes to dynamic structural causal models. In *34th Conference on Uncertainty in Artificial Intelligence (UAI 2018)*, pages 114–123. Curran Associates, Inc., 2018.

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

[21] Niels Hansen and Alexander Sokol. Causal interpretation of stochastic differential equations. *Electronic Journal of Probability*, 19:1–24, 2014.

[22] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[23] Ror Bellman and Karl Johan Åström. On structural identifiability. *Mathematical biosciences*, 7(3-4):329–339, 1970.

[24] Keith Glover and Jan Willems. Parametrizations of linear dynamical systems: Canonical forms and identifiability. *IEEE Transactions on Automatic Control*, 19 (6):640–646, 1974.

[25] MS Grewal and Keith Glover. Identifiability of linear and nonlinear dynamical systems. *IEEE Transactions on automatic control*, 21(6):833–837, 1976.

[26] BC Gargash and DP Mital. A necessary and sufficient condition of global structural identifiability of compartmental models. *Computers in biology and medicine*, 10(4): 237–242, 1980.

[27] Howard Harry ROSENBROCK. Structural properties of linear dynamical systems. *International Journal of Control*, 20(2):191–202, 1974.

[28] S Stanhope, Jonathan E Rubin, and David Swigon. Identifiability of linear and linear-in-parameters dynamical systems from a single trajectory. *SIAM Journal on Applied Dynamical Systems*, 13(4):1792–1815, 2014.

[29] Xing Qiu, Tao Xu, Babak Soltanalizadeh, and Hulin Wu. Identifiability analysis of linear ordinary differential equation systems with a single trajectory. *Applied Mathematics and Computation*, 430:127260, 2022.

[30] Hans Georg Bock. Recent advances in parameteridentification techniques for ode. *Numerical treatment of inverse problems in differential and integral equations*, pages 95–121, 1983.

[31] LT Biegler, JJ Damiano, and GE Blau. Nonlinear parameter estimation: a case study comparison. *AIChE Journal*, 32(1):29–45, 1986.

[32] Hongqi Xue, Hongyu Miao, and Hulin Wu. Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. *Annals of statistics*, 38(4):2351, 2010.

[33] James M Varah. A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing*, 3 (1):28–46, 1982.

[34] Jianwei Chen and Hulin Wu. Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to hiv-1 dynamics. *Journal of the American Statistical Association*, 103(481):369–384, 2008.

[35] Jim O Ramsay. Principal differential analysis: Data reduction by differential operators. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(3): 495–508, 1996.

[36] Nancy E Heckman and James O Ramsay. Penalized regression with model-based penalties. *Canadian Journal of Statistics*, 28(2):241–258, 2000.

[37] Sanmitra Ghosh, Paul Birrell, and Daniela De Angelis. Variational inference for nonlinear ordinary differential equations. In *International Conference on Artificial Intelligence and Statistics*, pages 2719–2727. PMLR, 2021.

[38] Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

[39] Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

[40] Yu A Kutoyants. *Statistical inference for ergodic diffusion processes*. Springer Science & Business Media, 2004.

[41] A Le-Breton and M Musiela. Some parameter estimation problems for hypoelliptic homogeneous gaussian diffusions. *Banach Center Publications*, 1(16):337–356, 1985.

[42] Ana Prior, Marina Kleptsyna, and Paula Milheiro-Oliveira. On maximum likelihood estimation of the drift matrix of a degenerated o–u process. *Statistical Inference for Stochastic Processes*, 20:57–78, 2017.

[43] Jason R Blevins. Identifying restrictions for finite parameter continuous time models with discrete time data. *Econometric Theory*, 33(3):739–754, 2017.

[44] Lars Peter Hansen and Thomas J Sargent. The dimensionality of the aliasing problem in models with rational spectral densities. *Econometrica: Journal of the Econometric Society*, pages 377–387, 1983.

[45] Mathieu Kessler and Anders Rahbek. Identification and inference for multivariate cointegrated and ergodic gaussian diffusions. *Statistical inference for stochastic processes*, 7:137–151, 2004.

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

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

[48] Alexis Bellot, Kim Branson, and Mihaela van der Schaar. Neural graphical modelling in continuous-time: consistency guarantees and algorithms. In *International Conference on Learning Representations*, 2021.

[49] Simo Särkkä and Arno Solin. *Applied stochastic differential equations*. Cambridge University Press, 2019.

[50] Ioannis Karatzas, Steven Shreve, and Steven E Shreve. *Brownian motion and stochastic calculus*. Springer Science & Business Media, 1991.

[51] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer Berlin, Heigelberg, 1992.

[52] James E Ferrell Jr, Tony Yu-Chen Tsai, and Qiong Yang. Modeling the cell cycle: why do certain circuits oscillate? *Cell*, 144(6):874–885, 2011.

[53] Wu Xiu-Li, Chen Si-Yang, and Hong Ji. Oscillation of a class of second-order nonlinear ode with impulses. *Applied mathematics and computation*, 138(2-3):181–188, 2003.

[54] Paul Massatt. Limiting behavior for strongly damped nonlinear wave equations. *Journal of Differential Equations*, 48(3):334–349, 1983.

[55] PW Hemker. Numerical methods for differential equations in system simulation and in parameter estimation. *Analysis and Simulation of biochemical systems*, 28:59–80, 1972.

[56] Yangxin Huang, Dacheng Liu, and Hulin Wu. Hierarchical bayesian methods for estimation of parameters in a longitudinal hiv dynamic system. *Biometrics*, 62(2): 413–423, 2006.

[57] Zheng Li, Ping Li, Arun Krishnan, and Jingdong Liu. Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic bayesian network analysis. *Bioinformatics*, 27(19):2686–2691, 2011.

[58] Tao Lu, Hua Liang, Hongzhe Li, and Hulin Wu. High-dimensional odes coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *Journal of the American Statistical Association*, 106(496):1242–1258, 2011.

[59] Pierre NV Tu. *Dynamical systems: an introduction with applications in economics and biology*. Springer Science & Business Media, 2012.

[60] Thomas A Weber and AV Kryazhimskiy. *Optimal control theory with applications in economics*, volume 10. MIT press Cambridge, 2011.

[61] Roger D Norton and Pasquale L Scandizzo. Market equilibrium computations in activity analysis models. *Operations Research*, 29(2):243–262, 1981.

[62] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[63] Hanshu Yan, Jiawei Du, Vincent YF Tan, and Jiashi Feng. On robustness of neural ordinary differential equations. *arXiv preprint arXiv:1910.05513*, 2019.

[64] Han Zhang, Xi Gao, Jacob Unterman, and Tom Arodz. Approximation capabilities of neural ordinary differential equations. *arXiv preprint arXiv:1907.12998*, 2(4):3–1, 2019.

[65] Zhengfeng Li, Michael R Osborne, and Tania Prvan. Parameter estimation of ordinary differential equations. *IMA Journal of Numerical Analysis*, 25(2):264–285, 2005.

[66] Hua Liang and Hulin Wu. Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association*, 103(484):1570–1583, 2008.

[67] Jim O Ramsay, Giles Hooker, David Campbell, and Jiguo Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796, 2007.

[68] AA Poyton, M Saeed Varziri, Kim B McAuley, P James McLellan, and Jim O Ramsay. Parameter estimation in continuous-time dynamic models using principal differential analysis. *Computers & chemical engineering*, 30(4):698–708, 2006.

[69] Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2): 431–441, 1963.

[70] Michael L Johnson, John J Correia, David A Yphantis, and Herbert R HalvorSON. Analysis of data from the analytical ultracentrifuge by nonlinear least-squares techniques. *Biophysical journal*, 36(3):575–588, 1981.

[71] Chien-Fu Wu. Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics*, 9(3):501–513, 1981.

[72] Michael L Johnson and Susan G Frasier. [16] nonlinear least-squares analysis. In *Methods in enzymology*, volume 117, pages 301–342. Elsevier, 1985.

[73] Marcel Maeder and Andreas D Zuberbuehler. Nonlinear least-squares fitting of multivariate absorption data. *Analytical Chemistry*, 62(20):2220–2224, 1990.

[74] Andrea Silvestrini and David Veredas. Temporal aggregation of univariate and multivariate time series models: a survey. *Journal of Economic Surveys*, 22(3): 458–497, 2008.

[75] Rajkumar Sharma, DK Pathak, and VK Dwivedi. Modeling & simulation of spring mass damper system in simulink environment. *Proceedings of SOM*, 2014, 2014.

[76] Mira-Cristiana Anisiu. Lotka, volterra and their model. *Didáctica mathematica*, 32 (01), 2014.

[77] Hans Petter Langtangen and Geir K Pedersen. *Scaling of differential equations*. Springer Nature, 2016.

[78] Nils Lehmann and Hans-Jürgen Sommers. Eigenvalue statistics of random real matrices. *Physical review letters*, 67(8):941, 1991.

[79] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[80] M Gong, K Zhang, B Schölkopf, C Glymour, and D Tao. Causal discovery from temporally aggregated time series. In *Uncertainty in Artificial Intelligence-Proceedings of the 33rd Conference, UAI 2017*, 2017.

[81] Robert I Jennrich. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643, 1969.

[82] Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

[83] Debasis Kundu. Asymptotic theory of least squares estimator of a particular nonlinear regression model. *Statistics & probability letters*, 18(1):13–17, 1993.

[84] Santiago Mira and Alvaro Escribano. Nonlinear time series models: consistency and asymptotic normality of nls under new conditions. 1995.

[85] A.A. Borokov. *Mathematical Statistics.* Gordon and Breach Science Publishers, 2018. ISBN 9780203749326.

[86] Jun Shao. *Mathematical statistics.* Springer Science & Business Media, 2003.

[87] Mary Ann Branch, Thomas F Coleman, and Yuying Li. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal on Scientific Computing*, 21(1):1–23, 1999.

[88] John R Dormand and Peter J Prince. A family of embedded runge-kutta formulae. *Journal of computational and applied mathematics*, 6(1):19–26, 1980.

[89] Jianwei Chen and Hulin Wu. Estimation of time-varying parameters in deterministic dynamic models. *Statistica Sinica*, pages 987–1006, 2008.

[90] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

[91] Xin Qi and Hongyu Zhao. Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *The Annals of Statistics*, 38(1):435–481, 2010.

[92] Tong Qin, Kailiang Wu, and Dongbin Xiu. Data driven governing equations approximation using deep neural networks. *Journal of Computational Physics*, 395: 620–635, 2019.

[93] Odd O Aalen, Kjetil Røysland, Jon Michael Gran, and Bruno Ledergerber. Causality, mediation and time: a dynamic viewpoint. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(4):831–861, 2012.

[94] Manuel C Voelkle and Johan HL Oud. Continuous time modelling with individually varying time intervals for oscillating and non-oscillating processes. *British Journal of Mathematical and Statistical Psychology*, 66(1):103–126, 2013.

[95] I Bzhikhatlov and S Perepelkina. Research of robot model behaviour depending on model parameters using physic engines bullet physics and ode. In *2017 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, pages 1–4. IEEE, 2017.

[96] Miglena N Koleva and Lubin G Vulkov. Two-grid quasilinearization approach to odes with applications to model problems in physics and mechanics. *Computer Physics Communications*, 181(3):663–670, 2010.

[97] VB Mandelzweig and F Tabakin. Quasilinearization approach to nonlinear problems in physics with application to nonlinear odes. *Computer Physics Communications*, 141(2):268–281, 2001.

[98] Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Symplectic ode-net: Learning hamiltonian dynamics with control. *arXiv preprint arXiv:1909.12077*, 2019.

[99] Diana-Elena Gratie, Bogdan Iancu, and Ion Petre. Ode analysis of biological systems. *Formal Methods for Dynamical Systems: 13th International School on Formal Methods for the Design of Computer, Communication, and Software Systems, SFM 2013, Bertinoro, Italy, June 17-22, 2013. Advanced Lectures*, pages 29–62, 2013.

[100] Athanasios Polynikis, SJ Hogan, and Mario Di Bernardo. Comparing different ode modelling approaches for gene regulatory networks. *Journal of theoretical biology*, 261(4):511–530, 2009.

[101] Minh Quach, Nicolas Brunel, and Florence d'Alché Buc. Estimating parameters and hidden variables in non-linear state-space models based on odes for biological networks inference. *Bioinformatics*, 23(23):3209–3216, 2007.

[102] Philipp Städter, Yannik Schälte, Leonard Schmiester, Jan Hasenauer, and Paul L Stapor. Benchmarking of numerical integration methods for ode models of biological systems. *Scientific reports*, 11(1):2696, 2021.

[103] Wei-Hung Su, Ching-Shan Chou, and Dongbin Xiu. Deep learning of biological models from data: applications to ode models. *Bulletin of Mathematical Biology*, 83:1–19, 2021.

[104] Engelbert Dockner. *Differential games in economics and management science*. Cambridge University Press, 2000.

[105] Anastasios Tsoularis. On some important ordinary differential equations of dynamic economics. *Recent developments in the solution of nonlinear differential equations*, pages 147–153, 2021.

[106] Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. *Advances in neural information processing systems*, 32, 2019.

[107] Wei Chen, Kun Zhang, Ruichu Cai, Biwei Huang, Joseph Ramsey, Zhifeng Hao, and Clark Glymour. Fritl: A hybrid method for causal discovery in the presence of latent confounders. *arXiv preprint arXiv:2103.14238*, 2021.

[108] Biwei Huang, Charles Jia Han Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. *Advances in Neural Information Processing Systems*, 35:5549–5561, 2022.

[109] David Kaltenpoth and Jilles Vreeken. Causal discovery with hidden confounders using the algorithmic markov condition. In *Uncertainty in Artificial Intelligence*, pages 1016–1026. PMLR, 2023.

[110] Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. *Advances in neural information processing systems*, 33:14891–14902, 2020.

[111] Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. Identification of linear non-gaussian latent hierarchical structure. In *International Conference on Machine Learning*, pages 24370–24387. PMLR, 2022.

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

[113] Thomas Verma and Judea Pearl. Causal networks: Semantics and expressiveness. In *Machine intelligence and pattern recognition*, volume 9, pages 69–76. Elsevier, 1990.

[114] Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

[115] Howard Straubing. A combinatorial proof of the cayley-hamilton theorem. *Discrete Mathematics*, 43(2-3):273–279, 1983.

[116] Gerald Teschl. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Soc., 2012.

[117] Samantha C Burnham, Dominic P Searson, Mark J Willis, and Allen R Wright. Inference of chemical reaction networks. *Chemical Engineering Science*, 63(4):862–873, 2008.

[118] Gheorghe Craciun and Martin Feinberg. Multiple equilibria in complex chemical reaction networks: I. the injectivity property. *SIAM Journal on Applied Mathematics*, 65(5):1526–1546, 2005.

[119] Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, pages 32540–32560. PMLR, 2023.

[120] Gregory C Chow. Optimum control of stochastic differential equation systems. *Journal of Economic Dynamics and Control*, 1(2):143–175, 1979.

[121] LS Lima and LLB Miranda. Price dynamics of the financial markets using the stochastic differential equation for a potential double well. *Physica A: Statistical Mechanics and its Applications*, 490:828–833, 2018.

[122] PCB Phillips. The structural estimation of a stochastic differential equation system. *Econometrica: Journal of the Econometric Society*, pages 1021–1041, 1972.

[123] Zeev Schuss. Singular perturbation methods in stochastic differential equations of mathematical physics. *SIAM Review*, 22(2):119–155, 1980.

[124] Kazimierz Sobczyk. *Stochastic differential equations: with applications to physics and engineering*. Springer Science & Business Media, 2001.

[125] Nicolaas G Van Kampen. Stochastic differential equations. *Physics reports*, 24(3): 171–228, 1976.

[126] Linda JS Allen. *An introduction to stochastic processes with applications to biology*. CRC press, 2010.

[127] Carlos A Braumann. *Introduction to stochastic differential equations with applications to modelling in biology and finance*. John Wiley & Sons, 2019.

[128] Darren J Wilkinson. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, 10(2):122–133, 2009.

[129] Douglas Henderson and Peter Plaschko. *Stochastic Differential Equations In Science And Engineering (With Cd-rom)*. World Scientific, 2006.

[130] SITU Rong. *Theory of stochastic differential equations with jumps and applications: mathematical and analytical techniques with applications to engineering*. Springer Science & Business Media, 2006.

[131] Stefano M Iacus et al. *Simulation and inference for stochastic differential equations: with R examples*. Springer, 2008.

[132] Junteng Jia and Austin R Benson. Neural jump stochastic differential equations. *Advances in Neural Information Processing Systems*, 32, 2019.

[133] Maria C Mariani and Osei K Tweneboah. Stochastic differential equations applied to the study of geophysical and financial time series. *Physica A: Statistical Mechanics and its Applications*, 443:170–178, 2016.

[134] Søren Wengel Mogensen, Daniel Malinsky, and Niels Richard Hansen. Causal learning for partially observed stochastic dynamical systems. In *UAI*, pages 350–360, 2018.

[135] Pedro Sanchez and Sotirios A Tsaftaris. Diffusion causal models for counterfactual estimation. In *Conference on Causal Learning and Reasoning*, pages 647–668. PMLR, 2022.

[136] Imanol Perez Arribas, Cristopher Salvi, and Lukasz Szpruch. Sig-sdes model for quantitative finance. In *Proceedings of the First ACM International Conference on AI in Finance*, pages 1–8, 2020.

[137] Marek Capiński and Ekkehard Kopp. *The Black–Scholes Model*. Cambridge University Press, 2012.

[138] Peter Klein. Pricing black-scholes options with correlated credit risk. *Journal of Banking & Finance*, 20(7):1211–1229, 1996.

[139] J Roderick Mccrorie. The problem of aliasing in identifying finite parameter continuous time stochastic models. *Acta Applicandae Mathematicae*, 79(1-2):9, 2003.

[140] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and

H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[141] Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 Computing in Cardiology*, pages 245–248. IEEE, 2012.

[142] Alexander P Browning, David J Warne, Kevin Burrage, Ruth E Baker, and Matthew J Simpson. Identifiability analysis for stochastic differential equation models in systems biology. *Journal of the Royal Society Interface*, 17(173):20200652, 2020.

[143] Jack Collins, David Howard, and Jurgen Leitner. Quantifying the reality gap in robotic manipulation tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6706–6712. IEEE, 2019.

[144] Philipp Kügler. Moment fitting for parameter inference in repeatedly and partially observed stochastic biological models. *PLOS ONE*, 7:1–15, 2012.

[145] RB Levien and SM Tan. Double pendulum: An experiment in chaos. *American Journal of Physics*, 61(11):1038–1044, 1993.

[146] KK Sankaran, R Perez, and KV Jata. Effects of pitting corrosion on the fatigue behavior of aluminum alloy 7075-t6: modeling and experimental studies. *Materials Science and Engineering: A*, 297(1-2):223–229, 2001.

[147] Leslaw Socha. *Linearization methods for stochastic dynamic systems.* Springer Science & Business Media, 2007.

[148] Gisiro Maruyama. Continuous markov processes and stochastic equations. *Rendiconti del Circolo Matematico di Palermo*, 4:48–90, 1955.

[149] Jan Nygaard Nielsen, Henrik Madsen, and Peter C Young. Parameter estimation in stochastic differential equations: an overview. *Annual Reviews in Control*, 24: 83–94, 2000.

[150] Patrik Axelsson and Fredrik Gustafsson. Discrete-time solutions to the continuous-time differential lyapunov equation with applications to kalman filtering. *IEEE Transactions on Automatic Control*, 60(3):632–643, 2014.

[151] Simo Särkkä et al. *Recursive Bayesian inference on stochastic differential equations.* Helsinki University of Technology, 2006.

[152] Xuerong Mao. The truncated euler–maruyama method for stochastic differential equations. *Journal of Computational and Applied Mathematics*, 290:370–384, 2015.

[153] Peter CB Phillips. The problem of identification in finite parameter continuous time models. *Journal of Econometrics*, 1(4):351–362, 1973.

[154] Michael Eichler. Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137(2):334–353, 2007.

[155] Michael Eichler and Vanessa Didelez. On granger causality and the effect of interventions in time series. *Lifetime data analysis*, 16:3–32, 2010.

[156] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[157] JD Kalbfleisch and Jerald Franklin Lawless. The analysis of panel data under a markov assumption. *Journal of the american statistical association*, 80(392):863–871, 1985.

[158] Henghsiu Tsai and KS Chan. A note on parameter differentiation of matrix exponentials, with applications to continuous-time modelling. *Bernoulli*, 9(5):895–919, 2003.

[159] Ken-iti Sato. Stochastic integrals with respect to lévy processes and infinitely divisible distributions. *Sugaku Expositions*, 27(1):19–42, 2014.

[160] Moshe Zakai and Jakov Snyders. Stationary probability measures for linear differential equations driven by white noise. *Journal of Differential Equations*, 8(1):27–33, 1970.

[161] Martin Jacobsen. *Homgeneous Gaussian Diffusions in Finite Dimensions.* Preprint, Institute of Mathematical Statistics University of Copenhagen, 1991.

[162] Pat Vatiwutipong and Nattakorn Phewchean. Alternative way to derive the distribution of the multivariate ornstein–uhlenbeck process. *Advances in Difference Equations*, 2019(1):1–7, 2019.