Applied Artificial Intelligence Institute (A2I2)
Deakin University
April 30, 2025

Problem Overview

Long Van Tran

s224930257@deakin.edu.au

1 Problem Setting

In a conventional setting, we have a time-series dataset $\{\mathbf{X}^{(n)}\}_{n=1}^{N}$,

$$\mathbf{X}^{(n)} = \left[\mathbf{X}_{t_1}^{(n)}, \mathbf{X}_{t_2}^{(n)}, \ldots, \mathbf{X}_{t_k}^{(n)}\right],$$

where $\mathbf{X}_{t_i}^{(n)}$ can be vectors ($\mathbf{X}_{t_i}^{(n)} \in \mathbf{R}^d$) or matrices ($\mathbf{X}_{t_i}^{(n)} \in \mathbf{R}^{t\times d}$), depending on the unit of time being considered here is a single time step or a segment of t steps. Here we consider the latter matrix case. Each of these trajectories describe patient records across k time points/segments (e.g., patient measurements throughout k hours).

We assume that the data is governed by a time-homogeneous linear additive noise stochastic differential equation (SDE), which has the form:

$$\mathrm{d}\mathbf{X}_t = \mathbf{A}\mathbf{X}_t\mathrm{d}t \; + \; \mathbf{G}\mathrm{d}\mathbf{W}_t,$$

where $\mathbf{A} \in \mathbf{R}^{d\times d}$ and $\mathbf{G} \in \mathbf{R}^{d\times m}$ are the unknown drift-diffusion parameters. $\mathbf{W}$ is an m-dimensional Brownian motion.

However, due to various practical reasons (such as data anonymization, incomplete data logs, etc.), the temporal ordering (i.e. the time dimension) might be lost. Usually, in a time-series dataset we would know the process states through a sequence of time steps. But when this "hidden" time dimension is lost, we face another challenge of sorting the data in the correct temporal order before using it to infer the underlying SDE's parameters.

Our goal is to correctly estimate the parameters $\mathbf{A}$ (drift), and $\mathbf{H} = \mathbf{G}\mathbf{G}^{\mathrm{T}}$ (observational diffusion) when given only the observational data. Furthermore, building on the recent identifiability theory of Wang et al., 2023 [2], who show that for linear SDEs driven by additive Brownian

noise the infinitesimal generator are generically identifiable from i.i.d. trajectories, provided a simple full-rank moment condition:

$$\operatorname{rank}\left[x_0, \mathbf{A}x_0, ..., \mathbf{A}^{d-1}x_0, \mathbf{G}\mathbf{G}^T, \mathbf{A}\mathbf{G}\mathbf{G}^T, ..., \mathbf{A}^{d-1}\mathbf{G}\mathbf{G}^T\right] = d \tag{1}$$

is satisfied.

Their result guarantees that, once the condition is met, every post-intervention distribution in the sense of stochastic structural causal models is determined by the observational data.

2 Proposed Method

We propose our base method with an extension based on Wang et al., [2].

2.1 Iterative Two-Step Scheme

First, we randomly initialize the parameters $\mathbf{A}^{(0)}$ and $\mathbf{H}^{(0)}(= \mathbf{G}\mathbf{G}^{\mathrm{T}})$. Then, we follow a two-step iterative scheme as follows:

(a) **Update Previously-Sorted Segments**
Keeping $\mathbf{A}^{(n)}$, and $\mathbf{H}^{(n)}$ at iteration n fixed, rearrange the previously sorted segments for each trajectory to maximize the log-likelihood:

$$\{\widetilde{\mathbf{X}}^{(m)}\} = \arg\max_{\{\widetilde{\mathbf{X}}^{(m)}\}} \sum_{m=1}^{M} \ln p\big(\widetilde{\mathbf{X}}^{(m)} \mid \mathbf{A}^{(n)}, \mathbf{H}^{(n)}\big).$$

(b) **Update SDE Parameters**
With the newly completed trajectories fixed, we can update the SDE parameters also by maximum likelihood estimation:

$$\mathbf{A}^{(k)} = \frac{1}{\Delta t}\Big(\sum_{i,j}(\Delta \mathbf{X}_i^{(j)})\mathbf{X}_i^{(j)\mathrm{T}}\Big)\Big(\sum_{i,j}\mathbf{X}_i^{(j)}\mathbf{X}_i^{(j)\mathrm{T}}\Big)^{-1},$$

$$\mathbf{H}^{(k)} = \frac{1}{T}\sum_{i,j}\big(\Delta \mathbf{X}_i^{(j)} - \mathbf{A}^{(k)}\mathbf{X}_i^{(j)}\Delta t\big)\big(\Delta \mathbf{X}_i^{(j)} - \mathbf{A}^{(k)}\mathbf{X}_i^{(j)}\Delta t\big)^{\mathrm{T}},$$

where $T = (N-1)\Delta t$ is the total time and $\Delta \mathbf{X}_i^{(j)} = \mathbf{X}_{i+1}^{(j)} - \mathbf{X}_i^{(j)}$.

2.2 Extension to Generator Identification

After we arrive at our estimated parameters, we can use the test (1) to check for generator identifiability. Namely, we test if:

$$\operatorname{rank}\big[\hat{x}_0, \hat{\mathbf{A}}\hat{x}_0, ..., \hat{\mathbf{A}}^{d-1}\hat{x}_0, \widehat{\mathbf{G}\mathbf{G}^T}, \hat{\mathbf{A}}\widehat{\mathbf{G}\mathbf{G}^T}, ..., \hat{\mathbf{A}}^{d-1}\widehat{\mathbf{G}\mathbf{G}^T}\big] = d$$

If our estimated params and re-ordered data pass this test, we can leverage the results in [2] to move forward with the SDE's generator ddentification.

3 Notes

Some notes to keep in mind:

- Proving convergence of the iterative scheme.
- Leveraging the identifiability conditions presented in [1].

References

[1] Vincent Guan, Joseph Janssen, Hossein Rahmani, Andrew Warren, Stephen Zhang, Elina Robeva, Geoffrey Schiebinger, Identifying drift, diffusion, and causal structure from temporal snapshots, 2024, available from: `https://arxiv.org/abs/2410.22729`.

[2] Yuanyuan Wang, Xi Geng, Wei Huang, Biwei Huang, Mingming Gong, Generator identification for linear sdes with additive and multiplicative noise, in A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), *Advances in Neural Information Processing Systems*, Volume 36. year, Curran Associates, Inc., 2023, available from: `https://proceedings.neurips.cc/paper_files/paper/2023/file/ca642f8e1174012d67c05c1c9f969644-Paper-Conference.pdf`.