Estimating Treatment Effects from Irregular Time Series Observations with Hidden Confounders

Defu Cao[*1], **James Enouen** [*1], **Yujing Wang** [2], **Xiangchen Song** [3],
Chuizheng Meng [1], **Hao Niu** [4], **Yan Liu** [1]
[1]University of Southern California [2]Peking University
[3]Carnegie Mellon University [4]KDDI Research, Inc.
{defucao, enouen, chuizhem, yanliu.cs}@usc.edu yujwang@pku.edu.cn
xiangchensong@cmu.edu ha-niu@kddi.com

Abstract

Causal analysis for time series data, in particular estimating individualized treatment effect (ITE), is a key task in many real-world applications, such as finance, retail, healthcare, etc. Real-world time series, i.e., large-scale irregular or sparse and intermittent time series, raise significant challenges to existing work attempting to estimate treatment effects. Specifically, the existence of hidden confounders can lead to biased treatment estimates and complicate the causal inference process. In particular, anomaly hidden confounders which exceed the typical range can lead to high variance estimates. Moreover, in continuous time settings with irregular samples, it is challenging to directly handle the dynamics of causality. In this paper, we leverage recent advances in Lipschitz regularization and neural controlled differential equations (CDE) to develop an effective and scalable solution, namely LipCDE, to address the above challenges. LipCDE can directly model the dynamic causal relationships between historical data and outcomes with irregular samples by considering the boundary of hidden confounders given by Lipschitz constrained neural networks. Furthermore, we conduct extensive experiments on both synthetic and real-world datasets to demonstrate the effectiveness and scalability of LipCDE.

1 Introduction

Estimating individualized treatment effects (ITE) for time series data, which makes predictions about causal effects of actions [75], is one key task in many domains, including marketing [10, 1], education [46], healthcare [41], etc. However, the existence of confounders can introduce bias into the estimation [64, 53]. For example, in finance applications, multi-factor investing strategies can give investors a deeper understanding of the risk drivers underlying a portfolio. The unobserved factors (i.e., hidden confounders), which typically happen at irregular time stamps and are not reflected in finance system records or are difficult to observe, could bring bias by influencing both interventions and stock returns. The reason is that even a small number of existing factors (such as Small Minus Big and High Minus Low) could significantly explain the cross-section of stock returns [16]. If we can simulate such hidden confounders within a reasonable range, we are able to obtain treatment estimates with reduced bias and variance by making appropriate impact assumptions on the relationship between treatments and outcomes [71].

Estimating ITE is an extremely challenging task in continuous time settings with hidden confounders. First, estimating treatment effects in large-scale irregular and sparse time series still has considerable room for improvement as previous works fail to consider the continuous time setting, where it is difficult to handle the dynamic behavior and complex interactions of covariates and treatments [25]. Second, hidden confounders' values generated by randomness and noise can introduce high variance and undesirable explanations. For example, in healthcare applications, according to domain knowledge of drug resistance, the response to single-agent immune-checkpoint inhibitors (ICI) in uremic patients ranged from 15% to 31% [77]. Consequently, when left unconsidered, drug resistance will introduce biased estimates of treatment effects. Furthermore, any substitute confounders generated by data-driven methods with an impact on outcomes over 31% can lead to high variance.

Recently, there have been several attempts to address these challenges. To model hidden confounders over time, [4] introduce a new causal prior graph for the confounding information and concept completeness to improve the interpretability of prediction models; [47] study the identification of direct and indirect causes for causal feature selection in time series solely based on observational data. Deconfounding-based models [29, 8] use latent variables given by their factor model as substitutes for the hidden confounders to render the assigned treatments conditionally independent. However, existing works either cannot handle irregular time series [4, 47], or have strong assumptions [29, 8]. Furthermore, the range of hidden confounders generated by previous data-driven works is possibly unjustifiable, which will distort (obscure or augment) the true causal relationship between treatments and outcomes.

In this work, we consider the task of estimating treatment effects under continuous time settings with multi-cause hidden confounders (which affect multiple treatments and the outcome). To tackle the above two challenges, we propose a novel Lipschitz regularized neural controlled differential equation (LipCDE) model for estimation by obtaining the constrained time-varying hidden confounders. Specif-

[*] These authors contributed equally.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

ically, LipCDE first infers the interrelationship of hidden confounders on treatment by estimating the boundary of hidden confounders: we decompose the historical covariates into low-frequency components and high-frequency components in the spectral domain. Then we use Lipschitz regularization [2] on the decomposition to get the latent representation. Afterward, we model the historical trajectories with neural CDE using sparse numerical solvers, which is one of the most suitable methods for large-scale problems under the continuous time setting [24]. In this way, we can explicitly model the observed irregular sequential data as a process evolving continuously in time with a dynamic causal relationship to equip the LipCDE with interpretability. In the outcome model, we re-weight the population of all participating patients and balance the representation via applying the inverse probability of treatment weighting (IPTW) strategy [43].

In this paper, we conduct extensive experiments on both simulated and real-world datasets. Experimental results show that LipCDE outperforms other state-of-the-art estimating treatment effect approaches. From a qualitative perspective, experiments show that LipCDE is in agreement with the true underlying hidden confounders in simulated environments, which can effectively eliminate bias in causal models [53]. In addition, the average RMSE of TSD [8] and SeqConf [29] on MIMIC-III's blood pressure outcome and COVID-19 datasets decreases by 28.7% and 32.3%, respectively. To the best of our knowledge, this is the first complete estimating treatment effects model that considers both the boundary of hidden confounders and the continuous time setting.

We summarize the main **contributions** as follows:

- LipCDE utilizes a convolutional operation with Lipschitz regularization on the spectral domain and neural controlled differential equation from observed data to obtain hidden confounders, which are bounded to reduce the high variance of treatment effect estimation.
- LipCDE can fully use information of observed data and dynamic time intervals, allowing the continuous inclusion of input interventions and supporting irregularly sampled time series.
- Sufficient experiments demonstrate the effectiveness of LipCDE in estimating treatment effect on both synthetic and real-world datasets. Particularly, experiments on MIMIC-III and COVID-19 demonstrate the potential of LipCDE for health care applications in personalized medical recommendation.

2 Related Work

Treatment effects learning in the static setting. In recent years, there has been a significant increase in interest in the study of causal inference accomplished through representational learning [37, 15]. [34] propose to take advantage of the multiple processing methods assigned in a static environment. [63] show that balancing the representational distributions of the treatment and control groups can help upper limits of error for counterfactual outcome estimates. However, these approaches rely on the strong ignorability assumption, which ignores the influence of implicit hidden confounders. Many works focus on relaxing such assumptions with the consideration of hidden confounders including domain adversarial training [6, 15]. [27] and [28] propose to unravel the patterns of hidden confounders from the network structure and observed features by learning the representations of hidden confounders and using the representations for potential outcome prediction. [71] propose to estimate confounding factors in a static setting using a latent factor model and then infer potential outcomes using bias adjustment. Nevertheless, such works fail to take advantage of the dynamic evolution of the observed variables and the inter-individual relationships which are present in the time-dynamic setting.

Treatment effects learning in the dynamic setting without hidden confounders. There are many related previous works estimating treatment effects in dynamic settings including g-computation formula, g-estimation of structural nested mean models [30], IPTW in marginal structural models (MSMs) [56], and recurrent marginal structural networks (RMSNs) [43], CRN [9] etc. In addition, Gaussian processes [61] and bayesian nonparametrics [58] have been tailored to estimate treatment response in a continuous time setting in order to incorporate non-deterministic quantification. Besides, [65] relies on regularization to decompose the observed data into shared and signal-specific components in treatment response curves from multivariate longitudinal data. However, those models still need constraint methods to guarantee the posterior consistency of the sub-component modules and cannot directly model the dynamic causal relationship between different time intervals. While [62, 19] directly model the dynamic causal relationship, they make a strong assumption with no hidden confounders, which does not have the flexibility to be applied to all real-world scenarios.

Treatment effect learning in the dynamic setting with hidden confounders. Rather than making strong ignorability assumptions, [52] and [40] theoretically prove that observed proxy variables can be used to capture hidden confounders and estimate treatment effects. [70] use network information as a proxy variable to mitigate confounding bias without utilizing the characteristics of the instances. TSD [8] introduces recurrent neural networks in the factor model to estimate the dynamics of confounders. In a similar vein, [29] propose a sequential deconfounder to infer hidden confounders by using Gaussian process latent variable model and DTA [41] estimates treatment effects under dynamic setting using observed data as noisy proxies. Besides, DSW [44] infers the hidden confounders by using a deep recursive weighted neural network that combines current treatment assignment and historical information. DNDC [45] aims to learn how hidden confounders behave over time by using current network observation data and historical information. However, previous works have not bounded confounders leading to high variance estimates when the data-driven approach produces anomaly confounders which have exceeded the impact constraint over treatments and outcomes.

Please refer to Appendix. D to see comprehensive related works.

3 Problem Setup

3.1 Estimating treatment effects task

Here we define the problem of estimating treatment effects from irregular time series observations formally: observational data for each patient i at irregular time steps $t_0^i < \cdots < t_{m_i}^i$ for some $m_i \in \mathbb{N}$. We have observed covariates $X^i = [X_{t_0}^i, X_{t_1}^i, \ldots, X_{t_{m_i}}^i] \in \mathcal{X}_t$ and corresponding treatments $A^i = [A_{t_0}^i, A_{t_1}^i, \ldots, A_{t_{m_i}}^i] \in \mathcal{A}_t$, and a_{t_k} is the set of all j possible assigned treatments at timestep t_k. Additionally, we have hidden confounder variables $Z^i = [Z_{t_0}^i, Z_{t_1}^i, \ldots, Z_{t_{m_i}}^i] \in \mathcal{Z}_t$. We omit the patient id i on timestamps unless they are explicitly needed. Combining all hidden confounders, observed covariates, and observed treatments, we define the history before time t_k as $H_{t_k}^i = \{X_{<t_k}^i, A_{<t_k}^i, Z_{<t_k}^i\}$ as the collection of all historical information.

We focus on one-dimensional outcomes $Y^i = [y_{t_0}^i, y_{t_1}^i, \ldots, y_{t_m}^i] \in \mathcal{Y}_t$ and we will be interested in the final expected outcome $\mathbb{E}[Y_{a_t,t_m}^i | H_t^i, X_t^i, A_t^i, Z_t^i]$, given a specified treatment plan a. In this way, we can define the individual treatment effect (ITE) with historical data as $\tau_t^i = \mathbb{E}[Y_{b_t,t_m}^i | H_t^i, X_t^i, A_t^i, Z_t^i] - \mathbb{E}[Y_{a_t,t_m}^i | H_t^i, X_t^i, A_t^i, Z_t^i]$ for two specified treatments a and b. In practice, we rely on assumptions to be able to estimate τ_t^i for any possible treatment plan, which begins at time step t until just before the final patient outcome Y is measured:

Assumption 1. Consistency [43]. *If $A_{\geq t} = a_{\geq t}$, then the potential outcomes for following the treatment plan $a_{\geq t}$ is the same as the observed (factual) outcome $Y_{a_{\geq t}} = Y$.*

Assumption 2. Positivity (Overlap) [33]. *For any patient, if the probability $P(a_{<t_m}, z_{<t_m}, x_{\leq t_m}) \neq 0$ then the probability of assigning treatment: $P(A_{t_m} = a_{t_m} | a_{<t_m}, z_{<t_m}, x_{\leq t_m}) > 0$ for all a_{t_m}.*

Assumption 1 and Assumption 2 are relatively standard assumptions of causal inference which assume that artificially assigning a treatment has the same impact as if it were naturally assigned and that each treatment has some nonzero probability. Additionally, most previous works in the time series domian make the sequential strong ignorability assumption [56] that if there are no hidden confounders, for all possible treatments A_t, given the historical observed covariates X_t, we have: $Y_{a_{\geq t_m}} \perp\!\!\!\perp A_{t_m} | A_{<t_m}, X_{<t_m}$. However, this assumption is often untestable due to the presence of hidden confounders in the real-world. Inspired by [71] and [8], we assume sequential single strong ignorability in the continuous time setting:

Assumption 3. Sequential single strong ignorability in continuous time setting. *If there exists multi-cause confounders, we have $Y_{a_{\geq t_m}} \perp\!\!\!\perp A_{t_m} | X_{t_m}, H_{<t_m}$, for all $a_{\geq t_m}$ and all j possible assigned treatments.*

Assumption 3 expands the sequential single strong ignorability assumption from [8] to the continuous time setting. Thus, only multi cause hidden confounders exist at every time stamp, having a causal effect on the treatment A_t and potential outcome Y_t. One of our goals is to learn representations of hidden confounders under the line of deconfounding works, which aim to eliminate bias, based on the following theorem:

Theorem 1. *If the distribution of the assigned causes $p(a_T)$ can be written as $p(\theta, x_T, z_T, a_T)$, we can obtain sequential ignorable treatment assignment:*

$$Y_{a_{\geq t_m}} \perp\!\!\!\perp A_{t_m} | X_{t_m}, H_{<t_m}, \tag{1}$$

for all $a_{\geq t_m}$ with possible assigned treatments, where θ are the parameters of the causal model.

Thm. 1 is proved by [8] and [29] in the discrete case. Here, we extend Thm. 1 to the continuous-time setting. Nevertheless, there are still existing challenges in applying the deconfounder framework to longitudinal data in the continuous time setting. After its original publication, [71] has been met with concerns of difficulty in reconstructing confounders in practical applications and the deconfounder assumption itself has been challenged. We leave a more detailed discussion to the Appendix. Towards the necessity of further constraints on the latent confounding, we introduce a frequency-based Lipschitz assumption on the structure of the hidden confounders in Assumption 4.

Assumption 4. Decomposition of time-varying hidden confounders. *The hidden confounders Z_t can be decomposed into high-frequency components Z_t^h and low-frequency Z_t^l with distinguishable frequency gap ω, i.e., $Z_{t_m} = (Z_{t_m}^h, Z_{t_m}^l)$ such that low-frequency confounders have Lipschitz bounded influence and high-frequency confounders are sufficiently covered by proxy variables in X_t.*

In this sense, we combine two existing extensions of TSD under a unifying assumption. Z_t^l contains smooth information (the trend of the confounding data) bounded by its maximal frequency ω_l. The functional outcomes are then Lipschitz bounded by constant L. Further, its distance and influence from its original value Z_0 will be bounded, reflecting its bounded variation from a static confounder U, as explored in [29]. Further, the high-frequency components are assumed to have corresponding noisy proxy variables available in the measured covariates X. Consequently, sufficient information about these high-frequency confounders can be derived from the observed proxy variables, as explored in [41]. Unified together, our assumption explores a semiparametric assumption enhancing the practicality of applying the deconfounder setup to longitudinal data. We provide a further discussion in the Appendix.

3.2 Neural Controlled Differential Equations

Starting from an initial state $u(t_0)$, neural ordinary differential equations (ODE) evolve following a neural network based differential equations. The state at any time t_i is given by integrating an ODE forward in time:

$$\frac{du(t)}{dt} = F(u(t), t; \theta), u(t_i) = u(t_0) + \int_{t_0}^{t_i} \frac{du(t)}{dt} dt, \tag{2}$$

where $F \in \mathcal{F}$, parametrized by θ with $(\mathcal{F}, || \cdot ||)$ a normed vector space and $u(t_0)$ is the initial state. Neural CDEs are a family of continuous time models that explicitly define the latent vector field f_θ by a neural network parameterized by

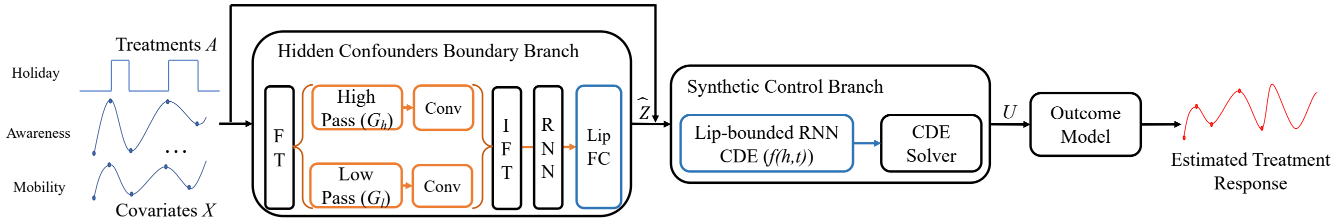


Figure 1: Architecture of LipCDE.

θ, and allow for the dynamics to be modulated by the values of an auxiliary path over time. To constrain the ODE into CDE format, let $\mathbf{H}_t = (H_t^1, H_t^2, \cdots, H_t^n) : t \in [t_0, t_m] \to \mathbb{R}^{n \times m}$ be the m dimensional representation of historical data with all n observed history control paths, the integral be a Riemann-Stieltjes integral and F be a continuous function acting on all control path [38]. For continuous time synthetic control, we estimate the latent representation of treatment effect H_t through: $H_t = H_{t_0} + \int_{t_0}^{t} f_\theta(H_s) d\mathbf{H}_s, t \in (t_0, t_m]$.

4 Lipschitz Bounded Neural Controlled Differential Equations (LipCDE)

To address the treatment effect estimation task from irregular time series observation, we must avoid inference bias caused by hidden confounders. Thus, we propose an approach called Lipschitz bounded neural controlled differential equations (LipCDE). As shown in Figure 1, LipCDE first infers the interrelationship of hidden confounders on treatment by bounding the boundary of hidden confounders via the hidden confounders boundary branch. After that, LipCDE feeds the history trajectories into the synthetic control branch, which utilizes both observed data and hidden confounders to generate the latent representation of each patient. Besides, we re-weight the population of all participating patients and balance the representation via applying a time-varying inverse probability of treatment weighting (IPTW) strategy. Combined with the LSTM layer, the outcome model can get the final estimate of the treatment effect.

4.1 Hidden confounders Boundary Branch

In this section, we focus on how to use Lipschitz regularized convolutional operation to infer the hidden confounders from both high-frequency signals and low-frequency signals of observed data. As shown in Fig 1, the Fourier transform $\mathcal{F}$ on observed data first converts the time-domain signals of history trajectories h_t [12, 11], including covariates and treatments with length N, into the corresponding amplitude and phase at different frequencies. Then, we sort the spectrum so that the spectrum corresponding to low-frequency information is concentrated at the origin after Fourier transform, and high-frequency information is far from the origin and contains rich boundary and detail information. After that, we use Gaussian high-pass filter G_h and Gaussian low-pass filter G_l to get high-frequency components and low-frequency components, respectively:

$$\begin{cases} G_h(h_t) = G_h(\mathcal{F}(h_t)) = 1 - e^{\frac{-D^2(\mathcal{F}(h_t))}{2D_0^2}} \\ G_l(h_t) = G_l(\mathcal{F}(h_t)) = e^{\frac{-D^2(\mathcal{F}(h_t))}{2D_0^2}} \end{cases} \tag{3}$$

The use of spectral-domain analysis enables change detection in certain frequency bands where the influence of trends (low frequency) or daily and seasonal cycles can be considered as time-invariant hidden confounders. The high-frequency components are easily perturbed, which can be treated as noisy proxies. We extract the influence of hidden confounders on the covariates by analyzing the presence of the covariates we extract. After that, both components are fed into convolutional operation:

$$F_c(h_t) = Conv(G_h(h_t)) + Conv(G_l(h_t)) \tag{4}$$

Next, we use the inverse Fourier transform $\mathcal{F}^{-1}$ converts the spectrum information of latent representation back to the time-domain signals. Then, the RNN layer takes the representation $\mathcal{F}^{-1}(F_c(h_t))$ as input and outputs the hidden states h_{hc} of hidden confounders. Note that, after the Fourier transform, time series no longer consider specific timesteps in the spectral domain. In addition, in contrast to directly handling irregular time series as [73], we use the processing of the Fourier transform as a mathematical component without considering time intervals, and irregular sampling is enabled in the next component.

To define the boundary of hidden confounders' value interval, following the RNN layer, the confounders encoder uses a Lipschitz bounded linear fully-connected (FC) layer with Lipschitz regularization [54] to map the output of RNN layer into a hidden embedding, i.e. $z = g(h_{hc}) = W_g h_{hc} + b_g$. The function $g : \mathbb{R}^n \to \mathbb{R}^K$ can be said as L-Lipschitz if there exists an L such that for all $x, y \in \mathbb{R}^n$, we have $||f(x) - f(y)|| \leq L||x - y||$ [7]. In this work, we enforce the function g to satisfy the 1-Lipschitz constraint, where g is the linear FC layer. Following spectral normalization of [26]:

$$\text{Lip}(g) \leq 1, \text{if } ||W_g||_2 \leq 1, \tag{5}$$

where $|| \cdot ||_2$ is the spectral matrix norm, we enforce the linear weights W_g be at most 1-Lipschitz by having a spectral norm less than one. This constraint ensures that when the observed data is within the normal interval, the inferred hidden confounders satisfy the corresponding bound interval with constant L.

4.2 Synthetic Control Branch

Since the neural ordinary differential equations ODE family is effective in continuous time problems, we use neural

CDE to estimate latent factors and treatment effects. Inspired by [5], let $u_t := g_\eta(H_t) = g_\eta([x_t, a_t, \hat{z}_t, H_{t-1}])$, where $g_\eta : \mathbb{R}^{n\times m} \rightarrow \mathbb{R}^{l\times m}$ is a set of functions that embeds the historical data into a l-dimensional latent state. Let f be a neural network parameterizing the latent vector field. To apply Lipschitz constraint on f, following [22], we define f as a continuous time Lipschitz RNN:

$$f(h,t) = A_R h + \sigma\left(W_R h + U u(s) + b\right), \tag{6}$$

where hidden-to-hidden matrices A_R and W_R are trainable matrices and nonlinearity $\sigma(\cdot)$ is an 1-Lipschitz function. Now $\dot{f} = \frac{\partial f(t)}{\partial t}$ is the time derivative and f considers both controlling the history path of observed data and the hidden state of RNN. A latent path can be expressed as the solution to a controlled differential equation of the form:

$$u_t = u_{t_0} + \int_{t_0}^{t} f\left(u_s, s\right) d\mathbf{H}_s^0, \quad t \in (t_0, t_m] \tag{7}$$

In that way, we can directly utilize adjoint methods [13] of CDEs to enable computing the gradient with a dynamic causal relationship between historical information controlled by $\mathbf{H}$ and outcomes. For each estimate of f_θ and g_η the forward latent trajectory in time that these functions defined through (7) can be computed using any numerical ODE solver as those equations continuously incorporate incoming interventions, without interrupting the differential equation:

$$\hat{u}_{t_1}, \ldots, \hat{u}_{t_k} = \mathrm{ODESolve}\left(f_\theta, u_{t_0}, \mathbf{H}_{t_1}, \ldots, \mathbf{H}_{t_k}\right) \tag{8}$$

4.3 Outcome Model

After sampling the latent representation $U_t = (\hat{u}_{t_1}, \ldots, \hat{u}_{t_k})$ of historical trajectories on each patient, we use the outcome model to estimate the treatment effect. To adjust the treatment assignment and get the final estimates, we first re-weight the population via an RNN model, which can handle time-varying treatment assignment [43], to estimate the propensity scores and IPTW of each dynamic time steps. After that, we use two stacked LSTM layers to decode the padded hidden sequence of irregular inputs. Then we use a linear fully-connected layer mapping the output of the LSTM layer into an unbiased estimated treatment response over time. For the loss function part, we weight each patient via the generated score of IPTW, w^i, and use the mean squared error (MSE) function as our target loss function: $L = \frac{1}{N}\sum_{i=1}^{N} w^i(\hat{y}^i_{t_{m+1}} - y^i_{t_{m+1}})^2$.

Empirically, the identifiability can be assessed on the synthetic data via sample hidden confounders Z_t repeatedly to evaluate the uncertainty of the outcome model estimates. However, identifiability might not be guaranteed under the framework of deconfounding in the completely general case [18, 51]. Previous works find that the estimates may have a high variance when the treatment effects are non-identifiable [8, 29, 41]. To achieve the goal of identifiability and obtain unbiased ITE estimates, [29] introduces the assumption of *Time-Invariant Unobserved Confounding*, which requires the hidden confounders are invariant for different timestamps, and [41] claim that we can learn the hidden embedding to make *Sequential Strong Ignorability* assumption hold via the observed noised proxies. Thus, the greater identifiability of our work follows both [29] and [41] as it utilizes both time-invariant hidden confounders from low-frequency components and dynamic noisy proxies from the high-frequency component of the observed data simultaneously in practice.

5 Experiments

5.1 Experiments Setting

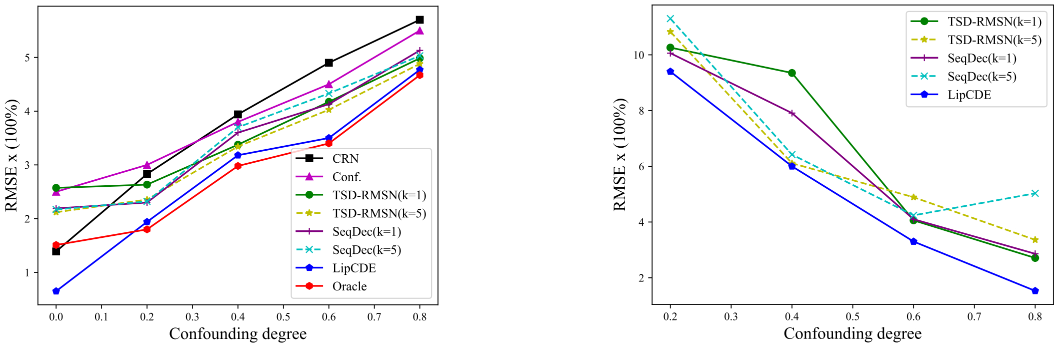
In this section, we estimate the treatment effects for each time step by one-step ahead predictions on both synthetic dataset and real-world datasets including MIMIC-III [35] dataset and COVID-19 [67] dataset. Hidden confounders in such real-world datasets is present as variables not included in the records. However, for real-world data, it is untestable to estimate the oracle treatment responses and we only evaluate the factual treatment effects. Refer to Appendix A for more detail on experiment settings.

Baselines. LipCDE is evaluated by examining the degree of control it has over hidden confounders. The baselines used in these experiments are: **Oracle**, which estimates ITE with simulated (oracle) confounders; **Conf. (No-hidden)**, which assumes no hidden confounders and can make it clear how hidden confounders here impact the performance of treatment effect prediction models; **CRN** [9], which introduces a sequence-to-sequence counterfactual recurrent network to estimate treatment effects and utilizes domain adversarial training to handle the bias from time-varying confounders; **TSD** [8], which leverages the assignment of multiple treatments over time to enable the estimation of treatment effects in the presence of multi-cause hidden confounders; **DTA** [41], which combines a LSTM autoencoder with a causal regularization penalty to learn dynamic noisy proxies and render the potential outcomes and treatment assignment conditionally independent; **SeqDec** [29] , which utilizes a Gaussian process latent variable model to infer substitutes for the hidden confounders; **OriCDE** [5], which can estimate ITE explicitly using the formalism of linear controlled differential equations.

Outcome model. Except OriCDE, all baselines share the same design of the outcome model, i.e. *MSM* [57], which uses inverse probability of treatment weighting (IPTW) to adjust for the time-dependent confounding bias by linear regression and then constructs a pseudo-population to compute final outcome, and *RMSN* [43], which estimates IPTW using RNNs instead of logistic regressions. OriCDE and LipCDE use the outcome model introduced in previous section.

5.2 Estimating Treatment Effects Experiments

Synthetic experiments. For the synthetic dataset, we can simulate data in which we are able to control the amount of hidden confounders and decide the treatment plan. Therefore in addition to estimating factual treatment responses, we will also perturb the inputs to quantify how accurate counterfactual relationships are captured by LipCDE. Following [8], we have $T = 30$ max time steps and $N = 5000$ patient trajectories, where each patient has $p = 5$ observed covariates and different treatments. We vary the confounding degree parameter γ to produce a varying amount of hidden



(a) RMSE results on treatment effects

(b) RMSE results on counterfactual treatment effects

Figure 2: Results on synthetic data

Outcome Model	-	MSM (RMSE%)			RMSN (RMSE%)				Ours (RMSE%)	
Methods	CRN	Conf.	DTA	TSD	Conf.	DTA	TSD	SeqDec	OriCDE	LipCDE
Blo. pre.	12.43	14.54	13.31	13.57	14.46	18.33	12.11	13.74	10.55	**9.19**
Oxy. sat.	4.17	4.72	4.65	4.33	4.22	4.21	4.25	4.19	4.24	**4.15**
COVID-19	-	15.10	13.93	13.07	11.48	13.52	11.08	11.43	11.36	**7.56**

Table 1: Results for real-world data (MIMIC-III and COVID-19) experiments. Lower is better.

confounders. Factual results use the outcome results corresponding to the real-world treatment we simulate. For the counterfactual estimations, we set all the treatments to 0 at the timestamp interval of $[\frac{l_i}{2}, l_i]$, where l_i is the sequential length of patient i, and get the outcome of the counterfactual world. For more details of the synthetic dataset, we refer the reader to Appendix. As shown on Figure 2, for the factual treatment effects results, methods considering hidden confounders are generally better than the models without the hidden confounders (CRN, Conf.). Note that, LipCDE achieves better results on all different levels of confounders and its outcome is closest to the estimates obtained using simulated (oracle) confounders, which means LipCDE can yield less biased estimates compared with other baselines. In addition, LipCDE remains stable and becomes closer to the simulated (oracle) confounders baseline when we increase the degree of confounders influence, which indicates that our model can effectively constrain the influence boundary of hidden confounders based on observed data. For the counterfactual path results, we interestingly observe that the RMSE decreases as the confounding degree increases. The reason is that when the degree increase, Z_t gets easier to handle with fixed treatment plans referring to the data generation method. Besides, LipCDE still performs better than the current baselines in the counterfactual world, indicating the stability of LipCDE for hidden confounder's reasoning and the validity of the estimation.

real-world experiments on MIMIC-III & COVID-19. real-world data allow us to demonstrate LipCDE has strong scalability and interpretability in real-world applications. MIMIC-III dataset contains 5000 patient records with 3 treatments, 20 covariates of patients and 2 outcomes including blood pressure (Blo. pre.), and oxygen saturation (Oxy. sat.). The COVID-19 dataset contains 401 German districts over the period of 15 February to 8 July 2020. We extract 10 time-varying covariates and 2 treatments with 2 outcomes, 'active cases', in each district. Please refer to Appendix for more details. The results in Table 1 show that LipCDE outperforms existing baselines in all cases. By modeling the dependence of the assigned treatment for each patient, LipCDE is able to infer latent variables and make orderly use of the causal relationship between latent variables and observed data. This result is consistent with what we have seen in the simulated dataset. Specifically, the average RMSE on MIMIC-III's blood pressure outcome and COVID-19 datasets is decreased by 28.7% and 32.3% over TSD and SeqConf respectively. Besides, the small increase in oxygen saturation is thought to be due to the fact that oxygen saturation itself is not dependent on current covariates and is less influenced by treatment. Although these results on real data require further validation by physicians, they demonstrate the potential of the method to be applied in real medical scenarios.

6 Analysis

Irregular time series with missing values. We emphasize that our model is suitable for irregular time series sampling. Therefore, we remove randomly 15% and 30% of the aligned synthetic data with different confounding degrees, independently for each unit. Except for CDE-based methods, all the baselines require some form of prior interpolation. Results shown in Table 2 demonstrate that our model achieves a comparable performance with irregularly aligned data. Note that, comparing with SeqDec which only models irregular samples via an indirect simple multivariate Gaussian distribution, LipCDE shows the ability of handling continuous time setting by utilizing the CDE module.

Ablation study. We conduct ablation experiments to verify the effectiveness of the proposed components on the synthetic data (with confounding degrees 0.2 and 0.8) and real-world datasets, which can clearly prove that our proposed

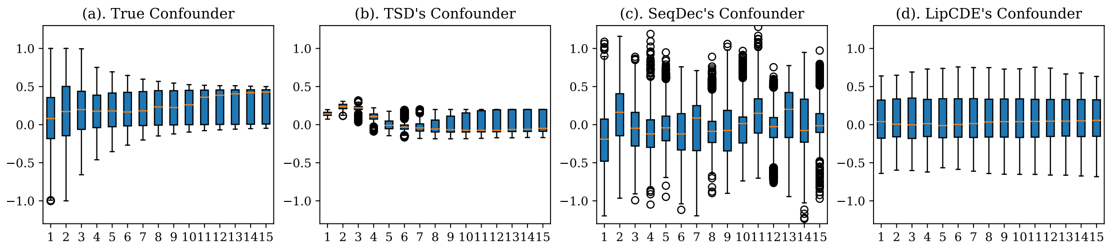


Figure 3: Analysis of the hidden confounders' boundary on synthetic data with first 15 timestamps. Here, the closer the shape of the box plot is to the true confounder, the less discrete the value is, the more accurate we consider the hidden confounder.

Degree	MR	Conf.	TSD	SeqDec	LipCDE	MR	Conf.	TSD	SeqDec	LipCDE
0	15%	3.43	2.83	2.43	**1.19**	30%	3.32	2.84	3.19	**2.29**
0.2		3.47	2.84	2.69	**2.6**		4.66	3.65	2.95	**2.62**
0.4		3.45	3.67	3.7	**3.39**		4.19	4.06	3.89	**3.61**

Table 2: Irregular data with missing value

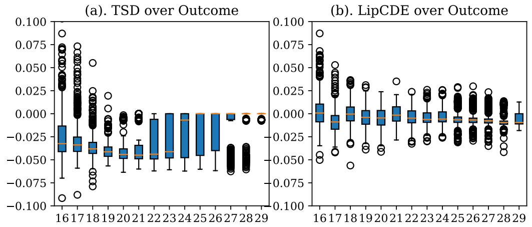


Figure 4: Analysis of the outcome on synthetic data's counterfactual path. Comparing with baseline models, LipCDE can estimate treatment effects with lower variance

	RMSE (%)				
	Synthetic		real-world		
	0.2	0.8	Blo. pre.	Oxy. sat.	COVID-19
LipCDE	**1.94**	**4.77**	**9.19**	**4.15**	**7.56**
-w/o-hc	2.13	4.93	10.54	4.27	10.69
-w/o-lip	2.11	5.00	9.51	4.24	10.5
-w/o-high	1.97	4.98	10.10	4.21	10.69
-w/o-low	2.05	4.85	9.73	4.29	10.5

Table 3: Ablation study on both synthetic dataset and real-world dataset.

method is an effective model for estimating treatment effect. Here, we discuss 4 variants of LipCDE: *w/o-hc*, which estimates ITE without considering hidden confounders in our model structure; *w/o-lip*, which estimates ITE without Lipschitz constraint; *w/o-high* and *w/o-low*, which reduce the high-frequency components and low-frequency component in LipCDE, respectively. As shown in Table 3, the estimation error of *w/o-hc* is larger than that with hidden confounders, demonstrating that our proposed method can take advantage of the hidden information to better estimate the treatment effects. Besides, the gap between *w/o-lip* and LipCDE shows that the Lipschitz regularization effectively avoids the negative impact of the presumed anomaly hidden confounders on the outcomes. Furthermore, we examine that both high-frequency information and low-frequency information can contribute to reducing the variance of estimating treatment effects, which is aligned with the separate statement on dynamic noisy proxies paper [41] and time static confounders paper [29].

Analysis on bounded hidden confounders. We perform the analysis using simulated datasets and evaluate the hidden confounder's quality on LipCDE with TSD and SeqDec. As shown on Figure 3, TSD cannot accurately predict the boundaries of the hidden confounders, which leads to the inability to model the impact of the hidden confounders accurately. Further, we find that TSD can induce highly confident posterior distributions with lower bounds of the hidden confounders, which can yield highly confident biased predictions [76]. The seqDec model has more discrete points and no obvious boundary, which also leads to the degradation of the model performance. LipCDE controls the data distribution of hidden confounders more accurately by filtering the convolutional neural network and Lipschitz regularization, which has higher similarity to the originally hidden confounder compared with other baselines.

In addition, we have a detailed analysis of the outcome's performance over the counterfactual path, referring to Figure 4, LipCDE can achieve better estimate results with lower variance compared with the previous strong baseline. Please refer to the Appendix B for further analysis.

7 Conclusion

In this paper, we proposed the Lipschitz-bounded neural controlled differential equation (LipCDE), a novel neural network that utilizes hidden confounders for estimating treatment effect in the case of irregular time series observations. For one thing, it uses the performance of time-varying observations in the frequency domain to infer the hidden confounders under Lipschitz regularization. For another thing, a well-designed CDE explicitly models the combinational latent path of observed time series, which can effectively capture underlying temporal dynamics and intervention effects. With experimental results on synthetic and real datasets, we demonstrate the effectiveness of LipCDE in reducing bias in the task of estimating treatment effects.

Acknowledgements This project is partially supported by KDDI Research, Inc. Defu Cao, James Enouen, and Chuizheng Meng are partially supported by the Annenberg Fellowship of the University of Southern California.

References

[1] Abadie, A.; Diamond, A.; and Hainmueller, J. 2010. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American statistical Association*, 105(490): 493–505.

[2] Araujo, A.; Negrevergne, B.; Chevaleyre, Y.; and Atif, J. 2021. On Lipschitz Regularization of Convolutional Layers using Toeplitz Matrix Theory. *Thirty-Fifth AAAI Conference on Artificial Intelligence*.

[3] Athey, S.; Bayati, M.; Doudchenko, N.; Imbens, G.; and Khosravi, K. 2021. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 1–15.

[4] Bahadori, M. T.; and Heckerman, D. 2021. Debiasing Concept-based Explanations with Causal Analysis. In *International Conference on Learning Representations*.

[5] Bellot, A.; and van der Schaar, M. 2021. Policy Analysis using Synthetic Controls in Continuous-Time. In *ICML*.

[6] Berrevoets, J.; Jordon, J.; Bica, I.; Gimson, A.; and van der Schaar, M. 2020. OrganITE: Optimal transplant donor organ offering using an individual treatment effect. *https://proceedings. neurips. cc/paper/2020*, 33.

[7] B'ethune, L.; Gonz'alez-Sanz, A.; Mamalet, F.; and Serrurier, M. 2021. The Many Faces of 1-Lipschitz Neural Networks. *ArXiv*, abs/2104.05097.

[8] Bica, I.; Alaa, A.; and Van Der Schaar, M. 2020. Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. In *International Conference on Machine Learning*, 884–895. PMLR.

[9] Bica, I.; Alaa, A. M.; Jordon, J.; and van der Schaar, M. 2020. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*.

[10] Brodersen, K. H.; Gallusser, F.; Koehler, J.; Remy, N.; and Scott, S. L. 2015. Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1): 247–274.

[11] Cao, D.; Li, J.; Ma, H.; and Tomizuka, M. 2021. Spectral temporal graph neural network for trajectory prediction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 1839–1845. IEEE.

[12] Cao, D.; Wang, Y.; Duan, J.; Zhang, C.; Zhu, X.; Huang, C.; Tong, Y.; Xu, B.; Bai, J.; Tong, J.; et al. 2020. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in neural information processing systems*, 33: 17766–17778.

[13] Chen, R. T.; Rubanova, Y.; Bettencourt, J.; and Duvenaud, D. 2018. Neural ordinary differential equations. *arXiv preprint arXiv:1806.07366*.

[14] Chernozhukov, V.; Wüthrich, K.; and Zhu, Y. 2021. An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 1–16.

[15] Curth, A.; and van der Schaar, M. 2021. On Inductive Biases for Heterogeneous Treatment Effect Estimation. *arXiv preprint arXiv:2106.03765*.

[16] D'Acunto, G.; Bajardi, P.; Bonchi, F.; and De Francisci Morales, G. 2021. The evolving causal structure of equity risk factors. In *Proceedings of the Second ACM International Conference on AI in Finance*, 1–8.

[17] D'Amour, A. 2019. Comment: Reflections on the Deconfounder. *Journal of the American Statistical Association*, 114(528): 1597–1601.

[18] D'Amour, A. 2019. On Multi-Cause Causal Inference with Unobserved Confounding: Counterexamples, Impossibility, and Alternatives. *ArXiv*, abs/1902.10286.

[19] De Brouwer, E.; Gonzalez, J.; and Hyland, S. 2022. Predicting the impact of treatments over time with uncertainty aware neural differential equations. In *International Conference on Artificial Intelligence and Statistics*, 4705–4722. PMLR.

[20] Ding, Y.; and Toulis, P. 2020. Dynamical systems theory for causal inference with application to synthetic control methods. In *International Conference on Artificial Intelligence and Statistics*, 1888–1898. PMLR.

[21] Doudchenko, N.; and Imbens, G. W. 2016. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research.

[22] Erichson, N. B.; Azencot, O.; Queiruga, A.; Hodgkinson, L.; and Mahoney, M. W. 2020. Lipschitz recurrent neural networks. *arXiv preprint arXiv:2006.12070*.

[23] Erichson, N. B.; Azencot, O.; Queiruga, A.; Hodgkinson, L.; and Mahoney, M. W. 2020. Lipschitz recurrent neural networks. *arXiv preprint arXiv:2006.12070*.

[24] Fröhlich, F.; Loos, C.; and Hasenauer, J. 2019. Scalable inference of ordinary differential equation models of biochemical processes. *Gene Regulatory Networks*, 385–422.

[25] Gao, T.; Subramanian, D.; Bhattacharjya, D.; Shou, X.; Mattei, N.; and Bennett, K. P. 2021. Causal inference for event pairs in multivariate point processes. *Advances in Neural Information Processing Systems*, 34: 17311–17324.

[26] Gouk, H.; Frank, E.; Pfahringer, B.; and Cree, M. J. 2021. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110(2): 393–416.

[27] Guo, R.; Li, J.; and Liu, H. 2020. Counterfactual evaluation of treatment assignment functions with networked observational data. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, 271–279. SIAM.

[28] Guo, R.; Li, J.; and Liu, H. 2020. Learning individual causal effects from networked observational data. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, 232–240.

[29] Hatt, T.; and Feuerriegel, S. 2021. Sequential Deconfounding for Causal Inference with Unobserved Confounders. *arXiv preprint arXiv:2104.09323*.

[30] Hernán, M. A.; and Robins, J. M. 2010. Causal inference.

[31] Huang, B.; Zhang, K.; Gong, M.; and Glymour, C. 2019. Causal discovery and forecasting in nonstationary environments with state-space models. In *International Conference on Machine Learning*, 2901–2910. PMLR.

[32] Imai, K.; and Jiang, Z. 2019. Comment: The Challenges of Multiple Causes. *Journal of the American Statistical Association*, 114(528): 1605–1610.

[33] Imai, K.; and Van Dyk, D. A. 2004. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467): 854–866.

[34] Johansson, F.; Shalit, U.; and Sontag, D. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*, 3020–3029. PMLR.

[35] Johnson, A. E.; Pollard, T. J.; Shen, L.; Li-Wei, H. L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L. A.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.

[36] Kallenberg, O.; and Kallenberg, O. 1997. *Foundations of modern probability*, volume 2. Springer.

[37] Kallus, N.; Mao, X.; and Udell, M. 2018. Causal inference with noisy and missing covariates via matrix factorization. *arXiv preprint arXiv:1806.00811*.

[38] Kidger, P.; Morrill, J.; Foster, J.; and Lyons, T. 2020. Neural controlled differential equations for irregular time series. *arXiv preprint arXiv:2005.08926*.

[39] Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[40] Kuroki, M.; and Pearl, J. 2014. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2): 423–437.

[41] Kuzmanovic, M.; Hatt, T.; and Feuerriegel, S. 2021. Deconfounding Temporal Autoencoder: estimating treatment effects over time using noisy proxies. In *Machine Learning for Health*, 143–155. PMLR.

[42] Langley, P. 2000. Crafting Papers on Machine Learning. In Langley, P., ed., *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, 1207–1216. Stanford, CA: Morgan Kaufmann.

[43] Lim, B.; Alaa, A.; and van der Schaar, M. 2018. Forecasting Treatment Responses Over Time Using Recurrent Marginal Structural Networks. *NeurIPS*, 18: 7483–7493.

[44] Liu, R.; Yin, C.; and Zhang, P. Nov 2020. Estimating Individual Treatment Effects with Time-Varying Confounders. 382–391. IEEE.

[45] Ma, J.; Guo, R.; Chen, C.; Zhang, A.; and Li, J. 2021. Deconfounding with Networked Observational Data in a Dynamic Environment. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 166–174.

[46] Mandel, T.; Liu, Y.-E.; Levine, S.; Brunskill, E.; and Popovic, Z. 2014. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, volume 1077.

[47] Mastakouri, A. A.; Schölkopf, B.; and Janzing, D. 2021. Necessary and sufficient conditions for causal feature selection in time series with latent common causes. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 7502–7511. PMLR.

[48] Miao, W.; Geng, Z.; and Tchetgen, E. T. ???? Identifying Causal Effects With Proxy Variables of an Unmeasured Confounder.

[49] Morrill, J.; Kidger, P.; Yang, L.; and Lyons, T. 2021. Neural Controlled Differential Equations for Online Prediction Tasks. *arXiv preprint arXiv:2106.11028*.

[50] Ogburn, E. L.; Shpitser, I.; and Tchetgen, E. J. T. 2019. Comment on “Blessings of Multiple Causes”. *Journal of the American Statistical Association*, 114(528): 1611–1615.

[51] Ogburn, E. L.; Shpitser, I.; and Tchetgen, E. J. T. 2020. Counterexamples to "The Blessings of Multiple Causes" by Wang and Blei.

[52] Pearl, J. 2012. On measurement bias in causal inference. *arXiv preprint arXiv:1203.3504*.

[53] Pearl, J.; et al. 2000. Causality: Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19.

[54] Perugachi-Diaz, Y.; Tomczak, J. M.; and Bhulai, S. 2021. Invertible DenseNets with Concatenated LipSwish. .

[55] Peters, J.; Mooij, J. M.; Janzing, D.; and Schölkopf, B. 2014. Causal discovery with continuous additive noise models.

[56] Robins, J. M.; and Hernán, M. A. 2009. Estimation of the causal effects of time-varying exposures. *Longitudinal data analysis*, 553: 599.

[57] Robins, J. M.; Hernan, M. A.; and Brumback, B. 2000. Marginal structural models and causal inference in epidemiology.

[58] Roy, J.; Lum, K. J.; and Daniels, M. J. 2017. A Bayesian nonparametric approach to marginal structural models for point treatments and a continuous or survival outcome. *Biostatistics*, 18(1): 32–47.

[59] Rubanova, Y.; Chen, R. T.; and Duvenaud, D. 2019. Latent odes for irregularly-sampled time series. *arXiv preprint arXiv:1907.03907*.

[60] Rubenstein, P. K.; Bongers, S.; Schölkopf, B.; and Mooij, J. M. 2016. From deterministic ODEs to dynamic structural causal models. *arXiv preprint arXiv:1608.08028*.

[61] Schulam, P.; and Saria, S. 2017. Reliable decision support using counterfactual models. *Advances in Neural Information Processing Systems*, 30: 1697–1708.

[62] Seedat, N.; Imrie, F.; Bellot, A.; Qian, Z.; and van der Schaar, M. 2022. Continuous-Time Modeling of Counterfactual Outcomes Using Neural Controlled Differential Equations. *arXiv preprint arXiv:2206.08311*.

[63] Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, 3076–3085. PMLR.

[64] Simpson, E. H. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2): 238–241.

[65] Soleimani, H.; Subbaswamy, A.; and Saria, S. 2017. Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions. *arXiv preprint arXiv:1704.02038*.

[66] Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, prediction, and search*. MIT press.

[67] Steiger, E.; Mussgnug, T.; and Kroll, L. E. 2020. Causal analysis of COVID-19 observational data in German districts reveals effects of mobility, awareness, and temperature. *medRxiv*.

[68] Tchetgen, E. J. T.; Ying, A.; Cui, Y.; Shi, X.; and Miao, W. 2020. An Introduction to Proximal Causal Learning.

[69] Tian, J.; and Pearl, J. 2013. Causal discovery from changes. *arXiv preprint arXiv:1301.2312*.

[70] Veitch, V.; Sridhar, D.; and Blei, D. 2020. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, 919–928. PMLR.

[71] Wang, Y.; and Blei, D. M. 2019. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528): 1574–1596.

[72] Wang, Y.; and Blei, D. M. 2020. Towards Clarifying the Theory of the Deconfounder.

[73] Ware, A. F. 1998. Fast approximate Fourier transforms for irregularly spaced data. *SIAM review*, 40(4): 838–856.

[74] Wu, S.; Zhang, H. R.; and Ré, C. 2019. Understanding and Improving Information Transfer in Multi-Task Learning. In *ICLR*.

[75] Zhang, Y.; Cao, D.; and Liu, Y. 2022. Counterfactual Neural Temporal Point Process for Estimating Causal Influence of Misinformation on Social Media. In *Advances in Neural Information Processing Systems*.

[76] Zheng, J.; D'Amour, A.; and Franks, A. 2021. Bayesian Inference and Partial Identification in Multi-Treatment Causal Inference with Unobserved Confounding. *arXiv preprint arXiv:2111.07973*.

[77] Zibelman, M.; Ramamurthy, C.; and Plimack, E. R. 2016. Emerging role of immunotherapy in urothelial carcinoma—advanced disease. In *Urologic Oncology: Seminars and Original Investigations*, volume 34, 538–547. Elsevier.

A Details for Experiments

A.1 Experiments Setting

LipCDE described in Experiments Section was implemented in PyTroch and trained on 4 NVIDIA GeForce RTX 2080 TI GPUs. We adopt the Adam [39] optimizer with learning rate 0.01. The training epoch is set as 10 with 10 iterations on each batch. The batch size is 16 and each dataset follows a 80%/10%/10% split for training/validation/testing respectively. LipCDE is trained by an end-to-end way with the same loss function with [5]. In addition, we report the average RMSE over 10 model runs on each experiment and report the mean results of LipCDE.

Specifically, in hidden confounders boundary branch, we first adjust the *fft-conv* open source code: https://github.com/fkodom/fft-conv-pytorch to fit the irregular time setting with the convolutional kernel with 3 × 3. After that, we apply RNN on that branch with 32 hidden units which are decided by grid search from [16, 32, 64, 128] on the validation dataset and then apply Lipschitz constraint on FC layer without tunable hyperparameters. For the synthetic control branch, we adapt Lipschitz bounded RNN from [23] with:

$$\begin{cases} A_{\beta_A,\gamma_A} = (1-\beta_A)\left(M_A + M_A^T\right) + \beta_A\left(M_A - M_A^T\right) - \gamma_A I \\ W_{\beta_W,\gamma_W} = (1-\beta_W)\left(M_W + M_W^T\right) + \beta_W\left(M_W - M_W^T\right) - \gamma_W I \end{cases} \tag{9}$$

where $\beta_A, \beta_W \in [0,1], \gamma_A, \gamma_W > 0$ are tunable parameters and hidden-to-hidden matrices $A_{\beta,\gamma} \in \mathbb{R}^{N\times N}$ and $W_{\beta,\gamma} \in \mathbb{R}^{N\times N}$ are trainable matrices. The LSTM layers in our outcome model are with 64,32 hidden states, respectively, which are also decided by grid search.

A.2 Simulated Dataset

Synthetic data allows us to simulate data in which we can control the amount of hidden confounders. Following [8], the observed data of patients is $H = (\{x_t^i, a_t^i, y_{t+1}^i\}_{t=1}^T)_{i=1}^N$ and the hidden confounders is $(\{z_t^i\}_{t=1}^T)_{i=1}^N$, where we have $T = 30$ max time steps and $N = 5000$ patient trajectories, where each patient has $p = 5$ observed covariates and different treatments. We vary the confounding degree parameter $\gamma \in \{0, 0.2, 0.4, 0.6, 0.8\}$ to produce a varying amount of hidden confounders. Factual results use the outcome results corresponding to the real-world treatment we simulate. For the counterfactual path, we set all the treatments to 0 at the timestamp interval of $[\frac{l_i}{2}, l_i]$, where l_i is the sequential length of patient i, and get the outcome of the counterfactual world. Then, we calculate the RMSE of model simulated outcome and counterfactual world outcome.

We build a dataset using p-order autoregressive processes. At each timestep t, we simulate k time-varying covariates $X_{t,k}$ representing single cause confounders and a multi-cause hidden confounders Z_t as follows:

$$X_{t,j} = \frac{1}{p}\sum_{i=1}^{p}\left(\alpha_{i,j}X_{t-i,j} + \omega_{i,j}A_{t-i,j}\right) + \eta_t \tag{10}$$

$$Z_t = \frac{1}{p}\sum_{i=1}^{p}\left(\beta_i Z_{t-i} + \sum_{j=1}^{k}\lambda_{i,j}A_{t-i,j}\right) + \epsilon_t \tag{11}$$

for $j = 1, \ldots, k, \alpha_{i,k}, \lambda_{i,j} \sim \mathcal{N}\left(0, 0.5^2\right), \omega_{i,k}, \beta_i \sim \mathcal{N}\left(1-(i/p), (1/p)^2\right)$, and $\eta_t, \epsilon_t \sim \mathcal{N}\left(0, 0.01^2\right)$. The value of Z_t changes over time and is affected by the treatment assignments.

Each treatment assignment $A_{t,j}$ depends on the single-cause confounders $X_{t,j}$ and multi-cause hidden confounders Z_t :

$$\pi_{tj} = \gamma_A \hat{Z}_t + (1-\gamma_A)\hat{X}_{tj} \tag{12}$$

$$A_{tj} \mid \pi_{tj} \sim \text{Bernoulli}\left(\sigma\left(\lambda\pi_{tj}\right)\right) \tag{13}$$

where $\hat{X}_{tj}$ and $\hat{Z}_t$ are the sum of the covariates and confounders respectively over the last p timestamps, $\lambda = 15$, $\sigma(\cdot)$ is the sigmoid function and γ_A controls the amount of hidden confounding applied to the treatment assignments. The outcomes are also obtained as a function of covariates and hidden confounders.

$$\mathbf{Y}_{t+1} = \gamma_Y Z_{t+1} + (1-\gamma_Y)\left(\frac{1}{k}\sum_{j=1}^{k}X_{t+1,j}\right), \tag{14}$$

where γ_Y controls the amount of hidden confounding applied to the outcome. We simulate datasets consisting of 5000 patients, with trajectories between 20 and 30 timestamps, and $k = 3$ covariates and treatments. To induce time dependencies we set $p = 5$.

A.3 real-world dataset

MIMIC-III dataset contains three treatment options: antibiotics, vasopressors, and mechanical ventilators. We extract 20 covariates of patients, including laboratory tests and vital signs for each patient and 2 outcomes including blood pressure (Blo. pre.), and oxygen saturation (Oxy. sat.). We extract up to 30 days of 5000 patient records from the dataset for training and testing following the same setting with [8], and infer treatment response within 24 hours.

Then, we apply experiments on the COVID-19 dataset, which contains 401 German districts over the period of 15 February to 8 July 2020. We extract 10 time-varying covariates which focus on mobility including parks mobility, workplaces mobility, etc; weather including rainfall and temperature; awareness such as searches corona, etc. The task is to infer the effects of multiple treatments including 'Holiday' and 'Weekday' over the 10 covariates for the outcome 'activate cases' in each district.

B Additional Analysis

In general, models that do not take into account hidden confounders cannot obtain correct causality because they assume that treatment assignment only depends on the observed history, which means that any unobserved probability confounder can lead to a biased estimate of outcome. LipCDE obtains the final representation used to infer the outcome by analysing all possible factors, thus reducing the bias caused by the presence of hidden confounders. In this section, we show additional LipCDE analysis for Analysis Section.

Miss Rate	Degree	Conf.	TSD-RMSN(K=5)	SeqDec(K=5)	LipCDE
0%	0	2.5	2.11	2.17	**0.65**
	0.2	3.01	2.633	2.32	**1.94**
	0.4	3.83	3.37	3.70	**3.18**
	0.6	4.51	4.17	4.33	**3.50**
	0.8	5.55	4.98	5.03	**4.77**
15%	0	3.43	2.83	2.43	**1.19**
	0.2	3.47	2.84	2.69	**2.60**
	0.4	3.45	3.67	3.70	**3.39**
	0.6	6.28	4.94	5.28	**4.45**
	0.8	7.07	**5.48**	6.17	5.62
30%	0	3.32	2.84	3.19	**2.29**
	0.2	4.66	3.65	2.95	**2.62**
	0.4	4.19	4.06	3.89	**3.61**
	0.6	6.90	6.87	6.33	**4.66**
	0.8	10.06	7.70	8.01	**5.91**

Table 4: Irregular data with missing data rate in {0%, 15%, 30%}.

B.1 Irregular time series with missing values

For **irregular time series with missing values** analysis part, we add the results when there is no data missing. It can be seen from Table 4 that our model outperforms the baseline in all cases.

B.2 Analysis on hidden confounders

For **analysis on bounded hidden confounders** part, we further use "covariance similarity score CovSim" [74] to show the similarity between true confounders and hideen confounders inferred by proposed methods. Given true confounders i and hideen confounders j, the "covariance similarity score" between them can be calculated by:

$$\text{CovSim}_{i,j} = \frac{||(U_{i,r_i} D_{i,r_i}^{1/2})^\top U_{j,r_j} D_{j,r_j}^{1/2}||_F}{||U_{i,r_j} D_{i,r_i}^{1/2}||_F \cdot ||U_{j,r_j} D_{i,r_j}^{1/2}||_F}, \tag{15}$$

where U_{i,r_i} and D_{i,r_i} denote matrices consisting of the top r_i eigenvalues and eigenvectors respectively, calculated by SVD over the instance-level representation matrix for each task i; r_i is chosen to contain 99% of the eigenvalues. As a result, the mean CovSim on different confounding degrees from 0.0 to 0.8 between true confounders and hidden confounders inferred by LipCDE is 0.82, which is **15%** and **30%** higher than the CovSim between true confounders and hidden confounders inferred by TSD-RMSN (0.71) and by SeqConf (0.63), respectively. This shows that the hidden confounders extracted by LipCDE matches the real information to a greater extent than other baselines, proving that our proposed model can effectively constrain the hidden confounders.

For **analysis on outcomes of counterfactual path** part, we plot all the degrees' results on Figure5, 6, 7, 8, 9. We plot the difference between the estimated outcome and the synthetic outcome on the counterfactual path. In addition to using the RMSE to illustrate our model's performance directly, when estimating the treatment outcome, LipCDE is able to make more accurate estimates above different time steps.

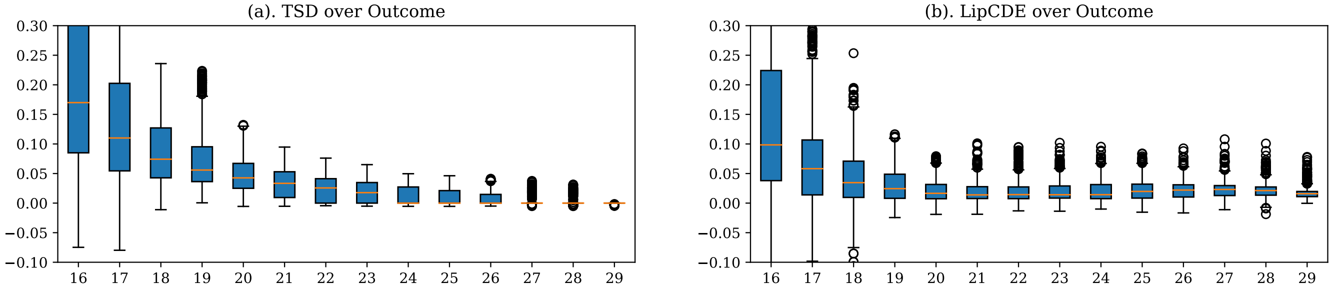


Figure 5: Analysis of the outcome on synthetic data's counterfactual path with degree 0.0.

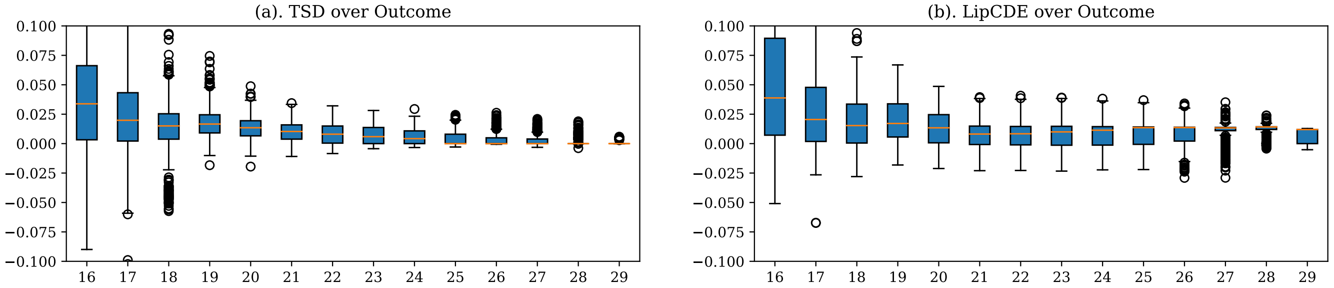


Figure 6: Analysis of the outcome on synthetic data's counterfactual path with degree 0.2.

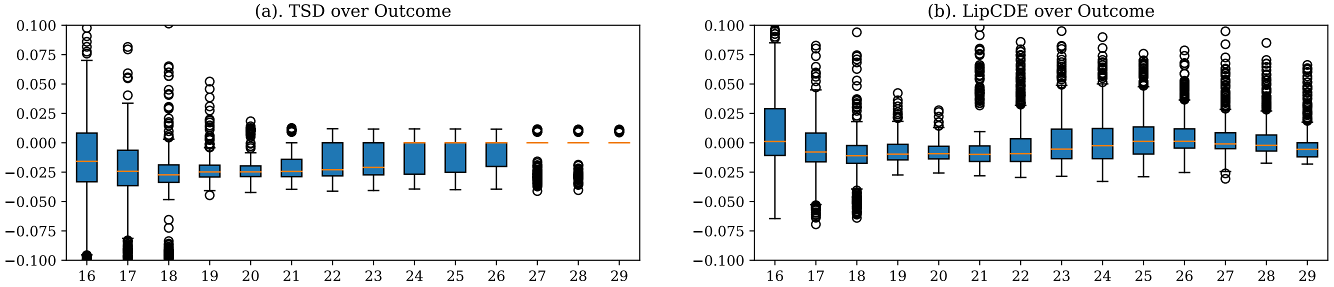


Figure 7: Analysis of the outcome on synthetic data's counterfactual path with degree 0.4.

C Theoretical Discussion

C.1 The Deconfounding Assumption

Issues with Deconfounding Despite one of the fundamental assumptions of the Time Series Deconfounder [8] being the Deconfounder Assumption from [71], there has been increasing concern with the validity of causal inference under the deconfounder assumption, including [32, 17, 50, 72, 51]. In light of these concerns, which indicate that the deconfounder assumption alone is insufficient to do full causal inference in the presence of hidden confounders, there are a number of pathways suggested for still utilizing the deconfounder assumption effectively: assuming uniqueness and exact reconstruction of confounders from treatments alone [32]; further parametric identification assumptions [17]; and identifiable or sufficiently rich proxy variables [48, 68, 40]. Consequently, the view that applying the Deconfounder Assumption to confounded longitudinal studies directly seemingly faces the majority of these existing issues. Following up work in the domain of deconfounding

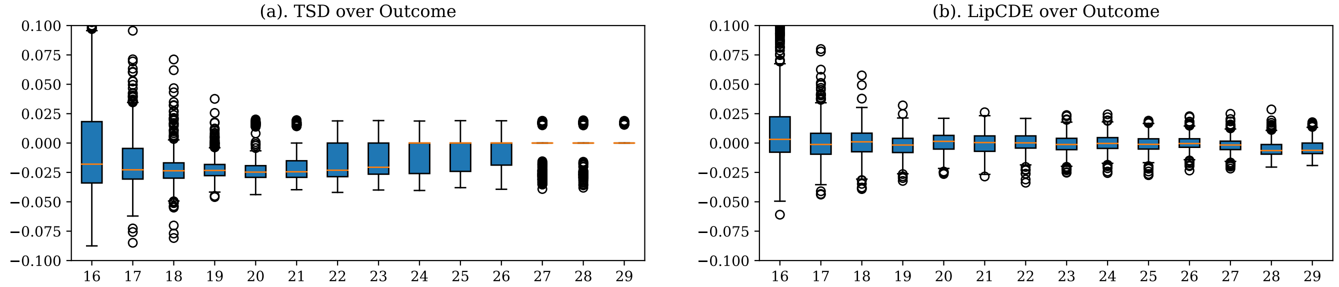


Figure 8: Analysis of the outcome on synthetic data's counterfactual path with degree 0.6.

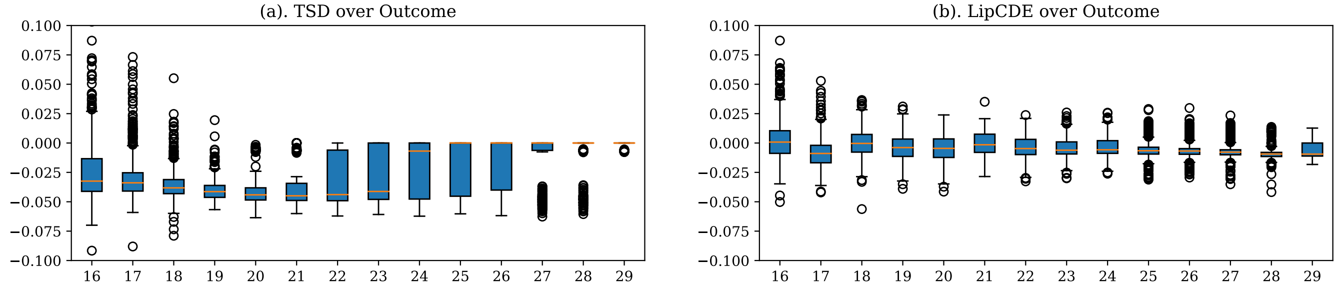


Figure 9: Analysis of the outcome on synthetic data's counterfactual path with degree 0.8.

observational data in the longitudinal domain has considered different sets of assumptions to recover unbiased identification of effects. [29] considers the assumption that there is a confounder that is static over time. Without requiring the single sequential strong ignorability assumption, they then show strong ignorability under the deconfounding assumption. Although the deconfounding assumption seems better warranted in the case of static confounders, it is possible that necessitating the static confounders assumption will exclude many potential cases of interest for causal inference on longitudinal data. [41] instead considers that the observed covariates might not directly measure unobserved confounders but instead can be considered as noisy measurements of the true confounders. Unfortunately, no theoretical results are provided and although an appealing notion in the presence of abundant data, often greater care is required in the domain of causal inference. Both negative controls and proximal causal learning [48, 68] further delineate between proxy covariates which influence the outcome and proxy covariates which influence the treatment. In the absence of such distinctions, causal inference may fail to achieve unbiased estimates. Consequently, it is likely that nonparametric treatment of time-dynamic data will require leveraging emerging work such as negative controls and proximal causal learning.

Discussion on LipCDE In this work, we attempt to make practical assumptions toward achieving unbiased estimates for causal inference in the longitudinal setting. As a consequence, we further existing work on time series deconfounding with an extended assumption combining both the regimes of static confounders and noisy proxy variables. In this way, we subdivide hidden confounders into two independent regimes: the low-frequency components and the high-frequency components. After this division, we can treat each uniquely. The low-frequency confounders represent the static confounders under which reconstruction of the confounding variable using the deconfounder assumption seems plausible after sufficient observation of treatments. With sufficiently bounded frequency alongside the Lipschitz influence assumption, we can achieve a bounded range of plausible confounders.

Although unlikely to completely debias an estimation procedure, it is possible to significantly reduce the bias and variance in a way dependent on the nuisance parameter of the low-frequency bound ω_ℓ. Such a balance should be considered in the context of the high-frequency confounders, which are then assumed to have sufficient proxy variables available in the observed data. Afterward, additional care should be taken for these proxy variables in terms of sufficient assumptions for identifiability. Possible routes include making assumptions directly on the noise distribution or identifying confounders represented by each covariate. Both of these paths often require greater care, inspection, or prior knowledge of the data which is being studied than is warranted by many practitioners of machine learning. Consequently, it is likely that methods developing the work of proximal causal learning in extension to longitudinal data will be required to suit such needs flexibly. We delay such a study detailing the application of proximal causal learning to dynamic time for future work.

C.2 Extension of Theorem 1 to continuous time setting

Following [8, 71], we introduce several definitions and lemmas that will help us relocate Theorem 1 for discrete-time setting to adapt it to continuous time setting. As a reminder, at each timestep t_k under continues-time setting, the random variable $z_{t_k} \in Z_{t_k}$ is constructed as a function of the history path until timestep $t_k : z_{t_k} = g\left(H_{t_{(k-1)}}\right)$, where $H_{t_{(k-1)}} = \left(\mathbf{Z}_{t_{(k-1)}}, \mathbf{X}_{t_{(k-1)}}, \mathbf{A}_{t_{(k-1)}}\right)$ takes values in $\mathcal{H}_{t_{(k-1)}} = \mathcal{Z}_{t_{(k-1)}} \times \mathcal{X}_{t_{(k-1)}} \times \mathcal{A}_{t_{(k-1)}}$ and $g : \mathcal{H}_{t_{(k-1)}} \to \mathcal{Z}$. In order to obtain unbiased estimation using hidden confounders Z_{t_k}, the following property needs to hold:

$$Y(a_{\geq t_k}) \perp\!\!\!\perp (A_{t_{k1}}, \cdots, A_{t_{kj}}) | X_{t_k}, A_{t_{(k-1)}}, Z_{t_k}, \tag{16}$$

$\forall a_{\geq t_k}$ and for all t_k in irregular samples.

Definition 1. ***Continuous sequential Kallenberg construction***
At timestep t_k, we say that the distribution of assigned treatment $(A_{t_k 1}, \ldots A_{t_k j})$ admits a continuous sequential Kallenberg construction from the random variables $Z_{t_k} = g\left(H_{t_{(k-1)}}\right)$ and X_{t_k} if there exist measurable functions $f_{t_k j} : \mathcal{Z}_{t_k} \times \mathcal{X}_{t_k} \times [0,1] \to \mathcal{A}_{t_k j}$ and random variables $U_{j t_k} \in [0,1]$ for j assigned treatments, such that:

$$A_{t_k j} = f_{t_k j}\left(Z_{t_k}, X_{t_k}, U_{t_k j}\right), \tag{17}$$

where $U_{t_k j}$ marginally follow Uniform $[0,1]$ and jointly satisfy:

$$U_{t_k j} \perp\!\!\!\perp Y(a_{\geq t_k}) | X_{t_k}, H_{t_{(k-1)}}, Z_{t_k} \tag{18}$$

for all $a_{\geq t_k}$ with j assigned treatments, where $f_{t_k j}$ are measurable.

(Continuous sequential Kallenberg construction $t \Rightarrow$ Sequential strong ignorability.) If at every timestep t_k, the distribution of assigned treatments $A_{t_k j}$ admits a continuous sequential Kallenberg construction from Z_{t_k} and X_{t_k} then we obtain sequential strong ignorability.

Proof. Without loss of generality, we assume that $\mathcal{A}_j$ are Borel spaces. For any irregular timestamps t_k we assume $\mathcal{Z}_{t_k}$ and $\mathcal{X}_{t_k}$ are measurable spaces. As $A_{t_k j}$ admits continuous sequential Kallenberg construction, we have

$$U_{t_k j} \perp\!\!\!\perp Y(a_{\geq t_k}) | X_{t_k}, H_{t_{(k-1)}}, Z_{t_k} \tag{19}$$

for all $a_{\geq t_k}$ with j assigned treatments. This implies that:

$$(Z_{t_k}, X_{t_k}, U_{t_k j}) \perp\!\!\!\perp Y(a_{\geq t_k}) | X_{t_k}, H_{t_{(k-1)}}, Z_{t_k} \tag{20}$$

Since the $A_{t_k j}$ are measurable functions according to Eq. 17 and $H_{t_{(k-1)}} = \left(X_{t_{(k-1)}}, A_{t_{(k-1)}}, Z_{t_{(k-1)}}\right)$, we have that sequential strong ignorability holds:

$$A_{t_k j} \perp\!\!\!\perp Y(a_{\geq t_k}) | X_{t_k}, A_{t_{(k-1)}}, Z_{t_k} \tag{21}$$

for all $a_{\geq t_k}$ with j assigned treatments of irregular samples. □

(Factor models for the assigned treatments $\Rightarrow$ Sequential continuous sequential Kallenberg construction.) Under weak regularity conditions[1], if the distribution of assigned causes $p\left(\mathbf{a}_T\right)$ can be written as the factor model $p\left(\theta, \mathbf{x}_T, \mathbf{z}_T, \mathbf{a}_T\right)$ then we obtain a continuous sequential Kallenberg construction for irregular timestamps.

The proof for Lemma C.2 uses Lemma 2.22 in [36] (kernels and randomization): Let μ be a probability kernel from a measurable space S to a Borel space T. Then there exists some measurable function $f : S \times [0,1] \to T$ such that if ϑ is $U(0,1)$, then $f(s, \vartheta)$ has distribution $\mu(s,)$, for every $s \in S$.

Proof. For timestep t_k, consider the random variables $A_{t_k j} \in \mathcal{A}_{t_k j}, X_{t_k} \in \mathcal{X}_{t_k}, Z_{t_k} = g\left(H_{t_{(k-1)}}\right) \in \mathcal{Z}_{t_k}$ and $\theta_j \in \Theta$. We assume sequential single strong ignorability holds. Without loss of generality, we assume $\mathcal{A}_{t_k j} = [0,1]$ for jth treatment.

From Lemma 2.22 in [36] , there exists some measurable function $f_{t_k j} : \mathcal{Z}_{t_k} \times \mathcal{X}_{t_k} \times [0,1] \to [0,1]$ such that $U_{t_k j} \sim$ Uniform $[0,1]$ and:

$$A_{t_k j} = f_{t_k j}\left(Z_{k_t}, X_{k_t}, U_{t_k j}\right) \tag{22}$$

with $U_{t_k} \perp\!\!\!\perp A_{t_1}$ for all irregular samples. It remains to show that

$$U_{t_k j} \perp\!\!\!\perp Y(a_{\geq t_k}) | X_{t_k}, H_{t_{(k-1)}}, Z_{t_k} \tag{23}$$

This can be seen by a distinction of cases. For any j treatment: if there exists a random variable V_{t_k} (not equal to Z_{t_k} or X_{t_k} almost surely) that confounds U_{t_k} and $Y\left(a_{\geq t_k}\right)$, it is either (i) time-invariant or (ii) time-varying. (i) If U_{t_k} is time-invariant, then it would also confound U_s for $s \neq t_k$, which introduces depedences between the random variables U_{t_k} for all irregular samples. However, since U_{t_k} are drawn iid from Uniform $[0,1]$, this cannot be the case. Otherwise, U_{t_k} and U_s for $s \neq t_k$ would not be jointly independent. (ii) If V_{t_k} is time-varying, then V_{t_k} would confound A_{t_k} through U_{t_k}. As a consequence, V_{t_k} would be also a confounders for A_{t_k}, which means $V_{t_k} \subseteq U_{t_k}$. As a result, there cannot be another random variable that confounds U_{t_k}, and therefore $U_{t_k j} \perp\!\!\!\perp Y(a_{\geq t_k}) | X_{t_k}, H_{t_{(k-1)}}, Z_{t_k}$ holds true.

□

[1] Regularity condition: The domains of the causes $\mathcal{A}_j$ are Borel subsets of compact intervals. Without loss of generality, we assume $\mathcal{A}_j = [0,1]$ for jth treatment.

D Related Work

Another different line of research involves the use of synthetic control measures for counterfactual estimation [1, 14]. [21] use negative weights and intercept terms to estimate weights depending on data structure. [20] propose time-varying weights to model the changing correlation between variables, and [3] interpret counterfactual estimation as a matrix completion problem using matrix normalization. However, although the goodness of fit can be improved, matching in discrete time is still difficult with irregularly aligned data, i.e., unit observations that are not aligned in time. To this end, differential equations have been introduced into causal and counterfactual inference. [60] propose that the equilibrium state of a first-order ODE system [13] can be described by a deterministic structural causal model, even with non-constant interventions. ODE-RNN [59] moderates the trajectory of interest using irregularly-sampled data. Additionally, [5] estimate counterfactual path explicitly using the formalism of controlled differential equations (CDE). The synthetic control literature is usually discussed in contrast to structural time series model. Besides, some structural methods rely on the regularity of the treated group trajectories over time to infer counterfactual estimates to balance the distribution between treated and control groups, whereas synthetic controls rely on the regularity of each group to infer counterfactual estimates to matche treated units to control units [5, 49]. Note that our work is different from another research line of identifying causal structure, i.e. causal discovery, where proposed approaches include: [66, 69, 55, 31], etc.

E Broader Impact

Causal analysis is one of the most fundamental problems in time series analysis and easily finds many applications in finance, retail, healthcare, transportation, manufacturing, etc. Among them, Estimating treatment effect for time series data is one key task in many industries and research scenarios which are extremely challenging due to the existence of hidden confounders and dynamic causal relationships of irregular samples or missing observations.

In this paper, we leverage recent advances in Lipschitz regularization and neural controlled differential equation (CDE) to tackle the above challenges, leading to effective and scalable solutions to causal analysis for time series applications in the wild. LipCDE can directly model the dynamic causal relationships between historical data and outcomes with irregular samples by considering the boundary of hidden confounders given by Lipschitz constraint neural networks.

Although current models including LipCDE are still far away from using causality to figure out the counterfactual world and to estimate treatment effect absolutely correctly, we do believe that the margin is decreasing rapidly. We would like to highlight that researchers are able to understand and mitigate the potential risks in estimating treatment effects. Especially in the healthcare domain, illogical or unreasonable treatments would be a disaster for an individual or humanity. In addition, we suggest researchers take a people-centered approach to build the responsible AI system with the features of fairness, interpretability, privacy, security, and accountability.