Towards Reverse Causal Inference on Panel Data: Precise Formulation and Challenges

Jiayao Zhang* **Youngsuk Park†** **Danielle C. Maddix†** **Dan Roth†** **Yuyang Wang†**

Abstract

Seeking causal explanations in panel (or longitudinal/multivariate time-series) data is a difficult problem of both academic and industrial importance. Although there exists a large amount of literature on forward causal inference, where the treatment/outcome/covariates variables are well-defined, it is unclear how to answer the reverse question: *which covariates have effects on the outcome?* In this paper, we set forth our expedition on this reverse question from the first principles. We formulate the precise problem definition in terms of causal patterns and causal paths, and propose a linear-time greedy meta algorithm that makes use of forward causal inference estimators. We further identify a set of optimality conditions under which the proposed algorithm is able to find the optimal causal path. To substantiate our greedy algorithm, we propose a generalized version of the synthetic control estimator by fitting both synthetic treatments and controls by *conditioning* on the partial causal paths. Promising results on on synthetic datasets demonstrate the potential of our method.

1 Introduction

Time-series data are ubiquitous and important in various scientific and business domains spanning longitudinal data analysis, econometrics, epidemiology, cloud computing, supply chain management, labor planning, to name a few (see e.g., Petropoulos et al. [2022] and Benidis et al. [2022] for a comprehensive overview and relevant applications). There is a growing interest on drawing causal inference on time-series data, apart from being an interesting and important research problem on its own, a causal understanding of the underlying mechanisms will facilitate the construction of more robust and generalizable solutions. Causal inference itself has a long history and has become the working horse in many applied research fields such as econometrics, political science, clinical trials, to name a few (see, e.g., Neyman [1923], Fisher [1958], Rubin [1974], Robins [1986], Abadie and Gardeazabal [2003], Imbens and Rubin [2015]), and much effort has been devoted to time-series/panel data (see e.g., Robins [1986], Robins et al. [2000], Bojinov et al. [2021]). There has been an interest in moving from reasoning about effects-of-cause to causes-of-effect, see, e.g., Gelman [2011], Gelman and Imbens [2013]. The term "reverse causal inference" coined by Gelman and Imbens, refer to the problems of this flavor that can be viewed as the inverse of the conventional (forward) causal queries. Although being of great theoretical and practical importance, literature on reverse causal inference usually focuses on the philosophical aspect of the question (see e.g., Dawid et al. [2016]). It has been an intriguing open question of how to formalize and investigate this problem in a practical manner.

In this work, we provide our answer to this question by formulating the reverse causal inference problem through the notions of causal patterns and causal paths (Section 2). Under this formulation, reverse causal inference is equivalent to identifying and estimating the causal patterns. As exact identification requires examining exponentially many combinations of patterns (in both time and the

*University of Pennsylvania. Work done while interning at AWS AI Labs. Email: zjiayao@upenn.edu.
†AWS AI Labs. Email: {pyoungsu,dmmaddix,drot,yuyawang}@amazon.com.

A causal view on dynamical systems workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

number of covariates), we propose to use a greedy linear-time algorithm whose solution becomes exact under a set of "monotonicity" conditions. Given that the core of our procedure solves a subproblem of forward causal inference, our method can be flexibly combined with any good causal estimator, such as the inverse-propensity weighted (IPW) estimator and synthetic control (SC) estimator (see e.g., Abadie and Gardeazabal [2003], Abadie et al. [2010, 2015]). We propose to use a generalized synthetic control estimator, and tested it on synthetic datasets.

Contributions. (1) We give a novel precise formulation of the reversal causal inference problem through causal patterns and causal paths. (2) We propose a linear-time meta algorithm for finding causal paths and derive optimality conditions. (3) We propose a generalized version of the synthetic control estimator, and demonstrate its benefit on synthetic experiments.

1.1 Related Work

Reverse Causal Inference and Causal Evidence. Although huge literature exists for "forward" causal queries as discussed above, it is only recently when practitioners started to contemplate reverse causal inference in applied fields, see, e.g., Gelman [2011], Gelman and Imbens [2013], Dawid et al. [2016], Imbens [2020], where one is interested in identifying the causes of effects instead of evaluating effects of causes. On top of several subtle differences, this problem is also relevant to graphical causal models pioneered by Robins [1986], Pearl [1995], Pearl and Mackenzie [2018]. Nonetheless, much of the recent focus is on distinguishing the "directions of causality" as discussed in Peters et al. [2012], see, e.g., Imbens [2020] for a thorough comparison between these two different approaches for causal inference.

Causal inference in panel data and time-varying treatments. Propensity score-based methods and marginal models are used in various scientific domain Rubin [1974], Rosenbaum and Rubin [1983], Robins [1986], Cole and Hernan [2008], Vansteelandt and Joffe [2014]. Inverse probability weighted methods are sensitive to specification, and even if ground truth is known, the estimator itself may suffer instability issues, and does not cope well with high-dimensional covariates Viviano and Bradic [2021], Khan and Nekipelov [2022]. Bojinov and Shephard [2019], Bojinov et al. [2021] adopted a design-based approach by assuming the treatment assignment mechanisms are known and studied the properties of inverse probability weighted estimators when the lag parameter is known (especially when it is either zero or one). Similar in spirit, Viviano and Bradic [2021] proposed a dynamic covariate balancing by considering a linear model. In the econometrics communities, there is much effort in drawing causal inference on policy evaluations in panel data. Commonly used methods include difference-in-difference (DiD), synthetic control (SC) Abadie and Gardeazabal [2003], Abadie et al. [2010, 2015], two-way fixed-effect (TWFE) modeling, to name a few. See, e.g., Athey et al. [2021] for a unified framework formulated as matrix completion problems and Imbens [2022] for a introduction. Along these lines of research, SC methods are most relevant to our problem in spirit where the treated group is usually few and one constructs a "synthetic control group" from the untreated group that best mimics the treated unit when the intervention had not taken place. See e.g., Doudchenko and Imbens [2017], Abadie [2021] for a recent introduction and progresses.

Inference. Inference is usually performed to test Fisher's sharp null of no-effect Fisher [1958] or to test Neymann's weak null of no average effect Neyman [1923]. Exact inference under randomization can be performed for design-based experiments Bojinov et al. [2021], Bottmer et al. [2021] or under outcome modelling. Inference is more subtle for SC on the effect on the single treated unit, as such usually the average effect over time horizon or over units are studied. The recent work Athey et al. [2021] referrs to the former as the horizontal regression and the latter vertical regression and studied their properties under the matrix completion literature.

2 Preliminaries and Setup

2.1 Background and Notations

We use capital letters to denote potentially multivariate time-series. For each study unit $i \in [N]$, we assume there exists a covariate time-series $X_{i,t} \in \mathcal{X}$ and an outcome time-series $Y_{i,t} \in \mathcal{Y}$ for $t \in [T]$. We write $Z_{i,t}$ for unobserved covariates. For example, each unit may correspond to a cloud

computing instance, $Y_{i,t}$ is the network latency of the i-th instance at time t, $X_{i,t}$ can be one kind of telemetry of the instance such as CPU usage or memory usage, and $Z_{i,t}$ is some unobserved variable of the instance such as internal state of the applications that are running on this particular instance. We focus on the case where all time-series are dichotomized, i.e., binary-valued $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, though our method and analysis can be extended to the more general case where $X_{i,t}$ may be multivariate and takes more values. We omit the unit index i when making general statements for all units. We use the colon in the indices to select a slice (both ends inclusive) along that index, for example, $Y_{i,1:t} = (Y_{i,1}, \ldots, Y_{i,t})$, and use the shorthand notation $\bar{Y}_{i,t}$ for $Y_{i,1:t}$. We also put a superscript on observations such as Y^{obs} to emphasize its non-random nature. Without loss of generality, we assume $Y_{i,t}$ takes place instantaneously after $X_{i,t}$ and $Z_{i,t}$ but before $X_{i,t+1}$ and $Z_{i,t+1}$ for all i and t. We view the potential-outcomes $Y_{i,t}$ as functions of both observed covariates $\bar{X}_{i,t}$ and unobserved covariates $\bar{Z}_{i,t}$, as well as its past history $\bar{Y}_{i,t-1}$. When only partial information is available, $Y_{i,t}$ may contain randomness arisen from other variables. For example, we view $Y_{1,2}(\bar{X}_{1,2})$ as a fixed random function dependent on $Y_{1,1}$ and $\bar{Z}_{1,2}$.

Causal inference revolves around making comparisons between potential-outcomes at different treatment levels, which is in essence a missing data problem [Neyman, 1923, Fisher, 1958]. For example, when the treatment is known and when there is only contemptuous effect (i.e., $Y_{i,t}$ is only affected by $X_{i,t}$, a common assumption made to simplify theoretical analysis e.g., in Arkhangelsky et al. [2021]), the potential-outcomes $Y_{i,t}$ is a function of $X_{i,t}$ and we write $Y_{i,t}(1)$ and $Y_{i,t}(0)$ for the potential-outcomes under $X_{i,t} = 1$ and 0, respectively. Here one is usually interested in studying the treatment effect for unit i at time t defined as

$$\tau_{i,t} = Y_{i,t}(1) - Y_{i,t}(0) \tag{1}$$

and perform hypothesis test on no effect across all units (i.e., Fisher's sharp null) or estimating an average treatment effect (e.g., average over the temporal or unit axis, or both). Specifically in the SC literature, usually one focuses on the case when the treated units are few, e.g., a single treated unit, and wishes to estimate

$$\tau_{1,t} = Y_{1,t}(1) - Y_{1,t}(0) = Y_{1,t}^{\text{obs}} - Y_{1,t}(0) \tag{2}$$

where we assume the first unit is the single treated unit whose observed outcome is by assumption the potential-outcome under treatment. This setup is particularly appealing for comparative case studies where potential-outcomes under treatment for other units may not always be well-defined, see e.g., west Germany reunificaiton study Abadie and Gardeazabal [2003].

This canonical setup in SC literature closely mirrors our problem of causal explanation: the units that exhibit abnormal behaviors is few (compared with few treated units); estimation of $\tau_{1,t}$ is also of interest on top of inferences on average effects; units may not be comparable in the sense that exchangeability assumption may fail. Furthermore, the most significant complication is one does not know *a priori* whether there exists "multiple levels" of treatments when dynamic treatment effects are possible. To this ends, we first set up definitions that fasciliate our exposition.

2.2 Partial and Complete Covariate Paths

The first complication for drawing reverse causal inference lies in the fact that compared with forward causal queries, the plausible answers may not be unique or even exist. As such we first define the history up to time t as $H_t \coloneqq \mathcal{X}^t$, and write $\mathcal{H}_t = \bigcup_{s=1}^{t} H_t$ as the set of (partial or complete) *covariate paths*. Note that not all paths $\boldsymbol{x} \in \mathcal{H}_t \coloneqq \mathcal{X}^t$ is realizable (has non-zero probability of being observed in the population) as we allow arbitrary dependency of $Y_{i,t'}$ on $X_{i,1:t'}$ for any $t' \leq t$. For example, in Figure 1, when $T = 3$, $(1, 1)$ is a partial covariate path and $(1, 1, 1)$ is a complete covariate path. We further define for each index t, the intervention $A(s, a; \bar{X}_{i,t}) = (\bar{X}_{i,s-1}, a, X_{i,s+1:t})$ for any $s \leq t$ and use the shorthand notation $A(s, \bar{X}_{i,t}) = A(s, 1 - X_{i,s}; \bar{X}_{i,t})$. In other words, these interventions alter the s-th component of the covariate series at time t, and leave everything else unchanged. We use the notation $\cup$ to define path concatenation, i.e., if for some $t \leq T$, $\omega \in H_t$ then $\omega' = (\omega, a) = \omega \cup \{a\} \in H_{t+1}$ with $\omega'_s = \omega_s$ for $s \leq t$ and $\omega_{t+1} = a$. With these definitions, at any time t, the potential-outcomes can be viewed as a random function taking an element in $\mathcal{H}_t$ (or more generally, any ordered subset of $\mathcal{X}^T$) as an argument where the randomness arises from the other (observed or unobserved) covariates that may affect it.

With these definitions, at any time t, the potential outcome can be viewed as a function taking an element in $\mathcal{H}_t$ as an argument. We denote the observed outcome by $Y_{i,t}^{\text{obs}}$, which connects with the

potential outcome via the *consistency assumption* $Y_{i,t}^{\text{obs}} = \sum_{\boldsymbol{x} \in \mathcal{H}_t} \mathbb{1}_{\{\boldsymbol{x}=X_{i,1:t}^{\text{obs}}\}} Y_{i,t}(\boldsymbol{x})$, where $X_{i,t}^{\text{obs}}$ is the observed outcome at time t. We overload the notation $\boldsymbol{x} = \boldsymbol{y}$ to return false whenever $\boldsymbol{x}$ and $\boldsymbol{y}$ are of different dimensions. A concrete example in our problem for the outcome series Y is the indicator of presence of anomalies at each time step and X the metrics and other time-varying auxiliary information we collect. We summarize our assumptions below.

Assumption 2.1 (Assumptions)**.** *We assume for all $i \in [N]$ and $t \in [T]$ the following assumptions.* ***1. Unconfoundedness:*** $X_{i,t} \perp \bar{Y}_{i,T} | \{\bar{X}_{i,t-1}, \bar{Z}_{i,t-1}\}$. ***2. Non-anticipation:*** $Y_{i,t}(\cdot) \in \mathcal{F}_{i,t} := \sigma\left(\{X_{i,1:t}, Z_{i,1:t}, Y_{i,1:t-1}\}\right)$. ***3. Consistency:*** $Y_{i,t}^{obs} \sim \sum_{\boldsymbol{x} \in H_t} \mathbb{1}_{\{\boldsymbol{x}=\bar{X}_{i,t}^{obs}\}} Y_{i,t}(\boldsymbol{x})$. ***4. Individualistic/No-interference:*** *$\mathcal{F}_{i,t} \perp \mathcal{F}_{j,s}$ for all $i \neq j$ and $t, s \in [T]$.*

Remark. Note that the set of assumptions are similar in spirit to those usually made in the literature Robins [1986], Bojinov and Shephard [2019], Viviano and Bradic [2021]. Comparing with the work studying effect estimation in time-varying/dynamic treatments, our formulation have the several most distinctive features: (1) we do not explicitly specify which component of $X_{i,t}(\cdot)$ are the "treatment" and which are "covariates;" instead, there might be a non-trivial pattern/dependency among them that jointly serve as the most causally relevant explanation for the anomaly. (2) we treat the outcome as a function of (partial and complete) covariate paths, which allows us to *specify* covariates after the intervention time or averaging over all paths therein (which requires an exponentially many paths and manually selected weights), as are usually done in e.g., Bojinov and Shephard [2019].

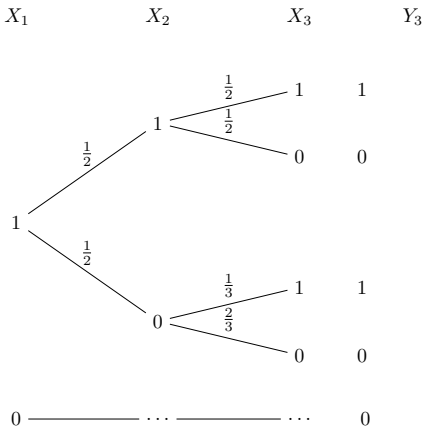


Figure 1: Example of causal paths (the subtree starting with $X_1 = 0$ is omitted). Here $\mathbb{P}(X_1 = 1) = 1/2$ and all conditional probabilities that are not $\frac{1}{2}$ are marked on the edge. The potential outcomes are such that $Y_1 = Y_2 = 0$ and $Y_3 = \mathbb{1}_{\{X_{1:3} \in \{(1,1,1),(1,0,1)\}\}}$.

3 The Reverse Causal Inference Problem

3.1 Problem Formulation via Causal Paths

We have used "causal" several times in the colloquial sense so far, yet it avoids confusion and is more illuminating when we formally define what do we mean by "causation:" we say X "causes" Y or X "has an effect on" Y or X "provides causally relevant explanations" for Y if (1) with every pre-treatment (covariates that occur prior to X) controlled, X and Y are correlated; and (2) X precedes Y in time. This perspective dates back to the very first works on causation Neyman [1923], Fisher [1958] and was inherited by the potential-outcomes framework and adopted by causal practitioners in various fields. We formulate our problem of reverse causal query as below.

Problem 3.1 (Reverse causal inference)**.** Given covariate time-series X_t and (observed) outcome series Y_t, obtain the set $\Omega \subset [T]$ such that $X_\Omega := \{X_t : t \in \Omega\}$ provides causal explanations for Y_T in the sense that *the change of any covariate X_Ω indexed by Ω from 1 to 0 leads to the change in the potential-outcome.*

The nature of this problem that multiple causes are plausible precludes putting assumption on the necessity between the change in X_Ω and the change in the potential outcome. We use the following example to illustrate the rationale behind.

Example 3.2. Suppose $T = 3$, $Y_1 = Y_2 = 0$ and $Y_3 = \mathbb{1}_{\{X_{1:3} \in \{(1,1,1),(1,0,1)\}\}}$, then $\Omega = \{1, 3\}$, and changing either X_1 or X_3 results a change in Y_3. However, when there is a change in Y_3, it is not necessary that the change was due to X_1 (or X_3).

Note that in this definition, we are essentially concerned with finding combinations of 1's in the covariate series that jointly have an effect on the outcome while treating 0's as "reference levels." The

condition for the set Ω in the problem statement is made precise in the following definition of causal patterns.

Definition 3.3 (Causal patterns)**.** *Given a set $\Omega \subset [T]$, we write $X_\Omega \in \mathcal{X}^{|\Omega|}$ to be a subset of $\bar{X}_T$ such that $X_s = 1$ if $s \in \Omega$ and X_s is left undetermined otherwise. We say Ω is a set of causal patterns if the potential outcomes*

$$Y_T(X_\Omega) \neq Y_T(A(s, X_\Omega)), \quad \forall s \in \Omega \tag{3}$$

differ when applying intervention A_s on any index $s \leq t$ in the set.

In Example 3.2, the set $\Omega = \{1, 3\}$ (which selects the first and the third covariates, X_1 and X_3) satisfies the above definition. Note that depending on different modelling choices, the expression $Y_T(X_\Omega)$ can be viewed as either a fixed quantity or a random variable. In latter case the non-equality should be interpreted in the sense of random variables (e.g., in the almost sure sense). Note that in Definition 3.3, X_Ω is a subset of $X_{1:T}$ that may contain "holes" in the middle. It is more convenient to work with an alternative definition in terms of covariate paths.

Definition 3.4 (Causal paths)**.** *Fix $s \leq T$, we say $\omega \in H_s = \mathcal{X}^s$ is an s-causal path for Y_T if*

$$\begin{cases} \mathbb{P}(Y_T = 1 | X_{1:s} = \omega) > 0 \text{ and} \\ \mathbb{P}(Y_T = 1 | X_{1:s} = \omega) > \mathbb{P}(Y_T = 1 | X_{1:s} = A(s, \omega)). \end{cases} \tag{4}$$

In other words, ω defines a causal path from $t = 0$ to $t = s$ if the potential outcome along this path has a higher probability (over the randomness in the future covariates and unobserved covariates) of being positive and intervening the end point of this path reduces this probability. For example, in Example 3.2, with $s = 2$, neither $(1, 0)$ nor $(1, 1)$ is a causal path but when $s = 3$, both $(1, 0, 1)$ and $(1, 1, 1)$ are causal paths.

As causal paths need not to be unique, we study their qualities through the following notion of globally and locally maximal probable causal paths, and we will study a weaker version of the reverse causal inference that attempts to find the globally maximal probable causal path.

Definition 3.5. *Given $s \leq T$, an s-causal path ω per Definition 3.4 is maximal probable if*

$$\begin{cases} \omega \in \operatorname{argmax}_{\omega \in \mathcal{C}_s} \mathbb{P}(Y_T = 1 | X_{1:s} = \omega), \\ \mathcal{C}_s := \{\omega \in \mathcal{H}_s : \omega \text{ is an } s\text{-causal path}\}. \end{cases} \tag{5}$$

We say a path ω^* is globally maximal probable *if ω^* satisfies Definition 3.5 with $s = T$; we say a path $\widehat{\omega}$ is* locally maximal probable *if for any $s \in [T-1]$, $\bar{\omega}_{s+1}$ maximizes over the same conditional probability conditioning that $\bar{X}_s = \bar{\omega}_s$.*

3.2 Optimality Condition for Local Causal Paths

Algorithm 1

Input: Sample set $\{(X_{i,t}, Y_{i,t}) : i \in [N], t \in [T]\}$.
Output: Causal path ω.

```
ω ← ∅
for t ∈ [1..T] do
    if Switch-Condition then
        ω ← ω ∪ {1}
    else
        ω ← ω ∪ {0}
    end if
end for
```

Algorithm 1: Here the `Switch-Condition` can be (1) $\mathbb{P}(Y_T = 1 | \bar{X}_t = (\omega, 1)) \geq \mathbb{P}(Y_T = 1 | \bar{X}_t = (\omega, 0))$ for the deterministic case; (2) The test of $\bar{\tau}_{i,t} = 0$ is rejected for a prescribed level α.

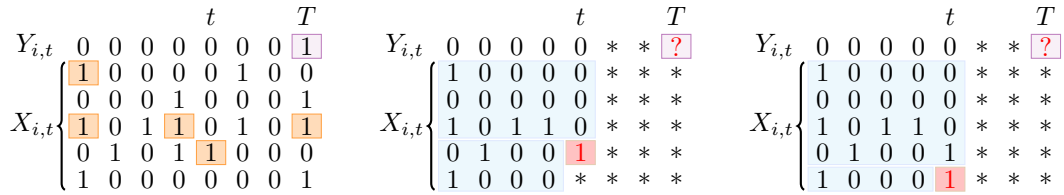
Definition 3.5 provides a way of quantifying the "quality" of different causal paths, and studying global maximal probable causal path provides a reasonable way towards the reverse causal inference problem. Nonetheless, exact identification of the global maximal probable causal path requires examination of exponentially many paths. In practice, the local maximal probable path can be found efficiently (to be discussed shortly), but it could be arbitrarily bad (measured by the conditional probability $\mathbb{P}(Y_T = 1 | \bar{X}_T = \widehat{\omega})$), as shown in the following theorem.

Theorem 3.6 (Hardness of the problem)**.** *Fix T, suppose*

$$\varepsilon := \min_{2 \leq t \leq T, a, \bar{x}} \mathbb{P}(X_t = a | \bar{X}_{t-1} = \bar{x}) > 0. \tag{6}$$

Let ω^* and $\widehat{\omega}$ be a global and local maximal probable causal path, then we have

$$\frac{\mathbb{P}(Y_T = 1 | \bar{X}_T = \widehat{\omega})}{\mathbb{P}(Y_T = 1 | \bar{X}_T = \omega^*)} \geq \left(\frac{\varepsilon}{1 - \varepsilon} \right)^{1 + \frac{T^2}{2}}, \tag{7}$$



(a) Ground truth X_Ω (shaded in orange).

(b) Condition on the partial path (shaded in blue) and decide whether to add the next covariate (shaded in red).

(c) Repeat this procedure while iterating according to the temporal order until a complete path is constructed.

Figure 2: Illustration of the linear-time meta algorithm where the covariate series is multivariate. (a) Suppose the ground truth X_Ω for the effect on $Y_{i,T}$ consists of the covariates shaded in orange (hence any path containing it is a causal path). (b) The meta algorithm starts from the very beginning and iteratively decides whether to add covariate X_s (marked in red) given the partial path already identified (an example of such path is shaded in blue). The complexity of this procedure is linear in the total number of covariates. (c) Repeat this process while iterating according to the temporal order until all covariates are decided (thus a complete path is constructed).

where the inequality becomes equality in the worse-case scenario.

The proof is done by analyzing the structure of the covariate paths and is deffered to the Appendix. This theorem suggests that a greedy algorithm could result to an arbitrarily bad path if we could choose the path weights adversarial. We now investigate conditions on on the assignment mechanism $\mathbb{P}(X_{t+1} = 1 | \bar{X}_t = \boldsymbol{x})$ that would make locally maximal probable causal path more desirable. To that end, we consider the scenario where all causal paths satisfy a form of a "monotonicity condition" under which the greedy algorithm is optimal.

Definition 3.7 (Monotonic paths)**.** *We say the causal paths are monotonic if*

$$\mathbb{P}(Y_T = 1 | \bar{X}_T = \widehat{\omega}) \geq \mathbb{P}(Y_T = 1 | \bar{X}_T = \widehat{\omega'}) \tag{8}$$

implies that

$$\mathbb{P}(Y_T = 1 | \bar{X}_s = \widehat{\omega}_{1:s}) \geq \mathbb{P}(Y_T = 1 | \bar{X}_s = \widehat{\omega'}_{1:s}) \tag{9}$$

for all $s \in [T]$.

Theorem 3.8 (Optimality condition)**.** *When all causal paths are monotonic, the greedy procedure returns the globally maximally probable causal path.*

Proof. Suppose the greedy path $\widehat{\omega}$ differ from the global maximal path ω^*, inspecting the child node of their lowest common ancestor implies a contradiction with the monotonicity assumption.

Note that the monotonic paths condition include the important case of Bernoulli random trials (where X_t' s are iid Bernoulli-$\frac{1}{2}$ random variables), which is itself an important model in the literature (e.g., Bojinov et al. [2021]).

4 Estimation via Generalized Synthetic Control

In this section, we substantiate Algorithm 1 by providing estimators for causal estimands that can be used to implement the `Switch-Condition` in the meta algorithm.

4.1 The Impulse Effect Estimand

To substantiate the `Switch-Condition` in the meta algorithm, based on our discussion in Section 2, we consider the following estimand evaluating the impulse effect at each covariate and timestamp t on the outcome at time T conditioning on the history seen so far, while being *oblivious* to the covariates between $t+1$ to T:

$$\tau_{i,t} = Y_{i,T}(A(t, 1; \bar{X}_{i,t})) - Y_{i,T}(A(t, 0; \bar{X}_{i,t})). \tag{10}$$

We define the average effect as $\bar{\tau}_t = \frac{1}{N} \sum_{i=1}^{N} \tau_{i,t}$. Now given a partial causal path $\omega \in \mathcal{H}_s$ for $s < T$, Algorithm 1 decides whether X_{s+1} should be 1 or 0 in the complete causal path by examining

$$\tau_{i,t}(\omega) = Y_{i,T}(\omega \cup \{1\}) - Y_{i,T}(\omega \cup \{0\}), \tag{11}$$

where, recall that we use $\cup$ notation to denote path concatenation.

Remark. The rationale behind choosing this estimand is to ensure no post-treatments/mediators are adjusted Rosenbaum [1984]. Through enumerating all combinations of covariate-time indices, we are evaluating the *direct* effect from each covariate. Compared with causal estimands on time-series proposed by Bojinov and Shephard [2019], Bojinov et al. [2021], our formulation allows for identifying specific patterns in multivariate series that serve as a cause for the outcome, and can be extended into an online algorithm.

4.2 A Generalized Synthetic Control Estimator

Outcome modeling is done by directly modelling the outcome on the treatment and pre-treatment variables. In our setup, it amounts to modelling the outcome on the past history:

$$\mu_{i,t} = \mathbb{E}[Y_{i,t}|\bar{X}_{i,t}] \tag{12}$$

for treatment D_i and covariates X_i. For any $s \leq t$, given partial path $\omega \in H_s$, we use the shorthand notation $\mu_{i,t}(\omega)$ to denote $\mathbb{E}[Y_{i,t}|\bar{X}_{i,s} = \omega]$. Synthetic control type estimators is one good choice for this problem as the problem setups are similar: the units that exhibit abnormal behaviors is few (compared with few treated units); and estimation of $\tau_{1,t}$ is also of interest on top of inferences on average effects; units may not be comparable in the sense that the exchangeability assumption may fail. The SC method then aims at finding the weights $\boldsymbol{w} = (w_1, \ldots, w_N)$ such that

$$\tau_{1,t} = Y_{1,t}(X_{1,t}^{\text{obs}}) - Y_{1,t}(0) \approx w_1 Y_{1,t}(X_{1,t}^{\text{obs}}) - \sum_{i=2}^{n} w_i Y_{i,t}^{\text{obs}}. \tag{13}$$

The weights are usually selected according to pre-treatments and are constrained as

$$\boldsymbol{w} = \operatorname{argmin} \left\| \bar{X}_{1,t} - \bar{\boldsymbol{X}}_{2:N,t} \boldsymbol{w}_{2:N} \right\|_2^2, \quad \boldsymbol{w} \in \left\{ w_1 = 1, \sum_{j=2}^{N} w_j = 1 \right\}, \tag{14}$$

where we wrote $\bar{\boldsymbol{X}}_{2:N,t} = (\bar{X}_{i,t})_{i=2}^{N} \in \mathcal{X}^{t \times (N-1)}$ for the matrix with $\bar{X}_{i,t}$ for $2 \leq i \leq N$ as its columns. We will consider a generalization of the SC method to estimate Equation (11): given a partial path $\omega \in H_s$, we fit both "synthetic treatment" $\widehat{\mu}_{s+1}(\omega \cup \{1\})$ and "synthetic control" $\widehat{\mu}_{s+1}(\omega \cup \{0\})$ on a "hypothetical unit" (say with unit index $i = 0$) with $\bar{X}_{0,s} = \omega$ using two donor pools $\mathcal{D}_1 = \{j : X_{j,s+1}^{\text{obs}} = 1\}$ and $\mathcal{D}_0 = \{j : X_{j,s+1}^{\text{obs}} = 0\}$ respectively. For example, we fit $\widehat{\mu}_{s+1}(\omega \cup \{1\}) = \sum_{j \in \mathcal{D}_0} w_j Y_{j,s+1}^{\text{obs}}$ via

$$\boldsymbol{w} = \operatorname{argmin} \left\| \omega - \bar{\boldsymbol{X}}_{\mathcal{D}_0,s} \boldsymbol{w} \right\|_2^2, \quad \boldsymbol{w} \in \left\{ \boldsymbol{w} \in \mathbb{R}^{|\mathcal{D}_0|} : \sum_{j \in \mathcal{D}_0} w_j = 1 \right\}, \tag{15}$$

where $\bar{\boldsymbol{X}}_{\mathcal{D}_0,t}$ similarly denotes the matrix with $\bar{X}_{i,s}$ for $s \in \mathcal{D}_0$ as its columns. We then form the *generalized SC estimator* for Equation (11) given time $t \leq T$ and partial path $\omega \in H_{t-1}$ as

$$\widehat{\tilde{\tau}}_t(\omega) = \widehat{\mu}_t(\omega \cup \{1\}) - \widehat{\mu}_t(\omega \cup \{0\}). \tag{16}$$

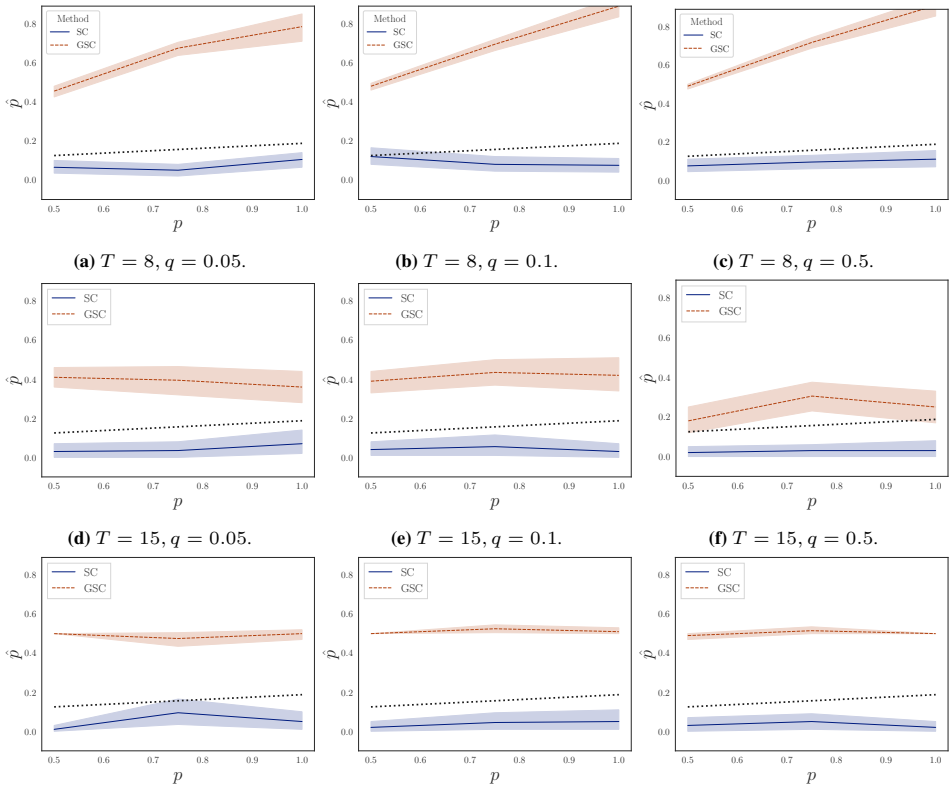
The rationale is that as we refine the causal path, there might be no unit that has the exact partial path.

Theoretical studies of the properties of SC estimators in the literature usually adopts the linear factor model (see e.g., Abadie et al. [2010]). Analogously, we study a generalized version of the linear factor model. The full details and the proof are postponed to the appendix, and the proof is similar to that in Abadie et al. [2010].

Theorem 4.1. *Suppose at time T_0, there are $N = N_0 + N_1$ units among which N_0 are treated (i.e., $X_{i,T_0} = 1$), fix $T_0 \leq T$, assume the potential-outcomes follows a general form of the linear factor model*

$$Y_{j,t} = Y_{j,T}(\bar{X}_{j,t}) = \delta_t + u_t^\top X_{j,1:t} + \lambda_t^\top Z_{j,1:t} + \sum_{k=1}^{K_t} \tau_k^{(t)} \mathbb{1}_{\left\{\bar{X}_{j,t} = \omega_k^{(t)}\right\}} + \varepsilon_{jt}, \tag{17}$$

where δ_t is known, ε is the noise, and u_t and λ_t are known. Under regularity conditions on λ_t, u_t, ε and assume SC fits are perfectly well, which are given in the Appendix, the SC estimators are consistent and asymptotically normal.



(a) $T = 8, q = 0.05$. **(b)** $T = 8, q = 0.1$. **(c)** $T = 8, q = 0.5$.

(d) $T = 15, q = 0.05$. **(e)** $T = 15, q = 0.1$. **(f)** $T = 15, q = 0.5$.

(g) $T = 100, q = 0.05$. **(h)** $T = 100, q = 0.1$. **(i)** $T = 100, q = 0.5$.

Figure 3: Causal path probability ($\widehat{p}$) found by various methods versus the true value p on the synthetic dataset with various time period T and background noise q. Lines of different color and style correspond to different background noise level q; black dotted lines correspond to results correspond to a random classifier $((p + 0.5)/8)$. Note that any causal path has a conditional probability no less than 0.5 by construction. We observe that the baseline SC (solid blue lines) is not able to identify causal paths while the proposed generalized SC (GSC) improves the performance significantly. However, as time horizon T becomes larger, it is harder to GSC to identiy global causal paths.

5 Empirical Studies

We now test our method on a synthetic dataset. Although the dataset is fairly simple, it helps to illustrate the challenges of the problem and fascilates the comparison of different methods.

5.1 Setup and Synthetic Dataset

Pattern $\boldsymbol{x}$	(1,0,1)	(1,0,0)	Other
Assignment Prob	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{8}$

Table 1: Assignment mechanism (pattern and the associated assignment probability) in the synthetic dataset.

We consider a generalization of the model in Figure 1. Given T, we consider a set of patterns of length $l = 3$ that starts at $T - d$ ($d \geq l$) with $Y_t = 0$ for all $t \neq T$. That is, there is an anomaly at Y_T only if certain patter emerges at time $T - d$. Consider the assignment probability mechanism in Table 1 and the outcome model as $Y_T = Z_1 \mathbb{1}_{\{X_{T-d-l-1:T-d}=(1,1,1)\}} + Z_2 \mathbb{1}_{\{X_{T-d-l-1:T-d}=(1,0,1)\}}$ for independent Bernoulli random variables $Z_1 \sim \text{Ber}(1/2)$ and $Z_2 \sim \text{Ber}(p)$ for some parameter p. Finally, we sample $X_j \overset{\text{iid}}{\sim} \text{Ber}(q)$ for a fixed parameter q, which models the background noise due of the covariate series. We experiment with both the case where the time horizon is short ($T = 8$ or $T = 15$) and is moderately long ($T = 100$), we set $d = 3$ in the first case and $d = 30$ in the latter. In both cases, we experiment over $p \in \{0.5, 0.75, 1\}$ and $q \in \{0.01, 0.1, 0.5\}$. Note that by construction, the globally maximal probable causal path has the conditional probability $\mathbb{P}(Y_{i,T} = 1 | X_{i,T-d-l-1:T-d}) = p$. For each parameter configuration, we generate $N = 500$ units on which we compute the optimal causal path ω and record the mean outcome probability $\mathbb{E}_Z[Y_T|\omega]$. We repeat this procedure for $m = 50$ independent runs.

5.2 Methods and Evaluation

We compare the following methods. **Point causal estimand using SC (SC).** Here we directly apply synthetic control to estimate the effect of each covariate X_t to Y_T for $t \leq T$ by fitting SC of a randomly chosen unit with $X_{i,t} = 1$ and estimate $\bar{\tau}_t$. In this way, the baseline has the same computational complexity with the generalized SC estimator. and 0 otherwise.

Generalized SC (GSC). Here we apply Algorithm 1 with the generalized SC estimator introduced in Section 4.2 and append 1 to the partial path if the estimate is positive.

Evaluation. We use the true conditional probability $\mathbb{P}(Y_T = 1 | \bar{X}_T = \widehat{\omega})$, which is known from dataset construction, to evaluate different methods by computing the average conditional probabilities $\widehat{p}$ of causal paths found over independent runs. In our synthetic dataset, the perfect algorithm should have $\widehat{p}^{\text{OPT}} = \max\{0.5, p\}$ while a random algorithm will have $\widehat{p}^{\text{random}} = (0.5 + p)/8$ since we are focusing on the pattern of length 3.

5.3 Simulation Results and Discussions

Baseline method performs poorly. We observe that in Figure 3, the baseline method, where we fit SC on point effect without adjusting for dependencies among covariates, performs poorly, though the synthetic dataset seems simple. This suggests that unlike forward causal inference problems, reverse causal inference is harder and much care needs to be taken when performing identification.

GSC identifies the *a* maximal probable causal path reasonably well. We observe that the Greedy-SC method is able to identify the globally probable causal path reasonably well as the GSC estimates is consistently above 0.5, indicating a causal path is returned.

GSC fails to identify global causal paths when the time horizon T becomes larger. As we increase the time horizon T from 8 to 15 and to 100, we note the it is gradually harder for GSC to find a *global* causal path (while still outputting a local causal path, in contrast to the vanilla SC estimator). **This issue can be alleviated through constructing a doubly-robust version of the GSC estimator, which we discuss in a following-up work.**

Effects of p and q on the difficulty of the problem. We observe all estimators have comparable performances as we vary the background noise level q, indicating q kis not a decisive factor that governs the difficulty of the problem.

6 Concluding Remarks and Broader Impacts

In this paper, we give a precise formulation of the reverse causal inference problem through causal patterns and causal paths. We propose to use a linear-time approximate procedure that can be flexibly combined with any causal estimator for this purpose and analyze its optimality conditions. We propose to use a generalized synthetic control estimator for this problem. There are several avenues for future work, for example, we only attempted to find causal paths but not causal patterns, which is more difficult. As reverse causal inference is a challenging problem, we hope our expositions could cast new lights into the community and inspire the practitioners.

References

A. Abadie. Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, 59(2):391–425, June 2021. ISSN 0022-0515. doi: 10.1257/jel.20191450. URL https://pubs.aeaweb.org/doi/10.1257/jel.20191450.

A. Abadie and J. Gardeazabal. The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, 93(1):113–132, Feb. 2003. ISSN 0002-8282. doi: 10.1257/000282803321455188. URL https://pubs.aeaweb.org/doi/10.1257/000282803321455188.

A. Abadie, A. Diamond, and J. Hainmueller. 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, June 2010. ISSN 0162-1459, 1537-274X. doi: 10.1198/jasa.2009.ap08746. URL http://www.tandfonline.com/doi/abs/10.1198/jasa.2009.ap08746.

A. Abadie, A. Diamond, and J. Hainmueller. Comparative Politics and the Synthetic Control Method: COMPARATIVE POLITICS AND THE SYNTHETIC CONTROL METHOD. *American Journal of Po-*

litical Science, 59(2):495–510, Feb. 2015. ISSN 00925853. doi: 10.1111/ajps.12116. URL https://onlinelibrary.wiley.com/doi/10.1111/ajps.12116.

D. Arkhangelsky, G. W. Imbens, L. Lei, and X. Luo. Double-Robust Two-Way-Fixed-Effects Regression For Panel Data, July 2021. URL http://arxiv.org/abs/2107.13737. arXiv:2107.13737 [econ, q-fin, stat].

S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi. Matrix Completion Methods for Causal Panel Data Models. *Journal of the American Statistical Association*, 116(536):1716–1730, Oct. 2021. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2021.1891924. URL http://arxiv.org/abs/1710.10251. arXiv:1710.10251 [econ, math, stat].

K. Benidis, S. S. Rangapuram, V. Flunkert, Y. Wang, D. Maddix, C. Turkmen, J. Gasthaus, M. Bohlke-Schneider, D. Salinas, L. Stella, et al. Deep learning for time series forecasting: Tutorial and literature survey. *ACM Computing Surveys (CSUR)*, 2022.

I. Bojinov and N. Shephard. Time series experiments and causal estimands: Exact randomization tests and trading. *Journal of the American Statistical Association*, 114(528):1665–1682, 2019. doi: 10.1080/01621459.2018.1527225. URL https://doi.org/10.1080/01621459.2018.1527225.

I. Bojinov, A. Rambachan, and N. Shephard. Panel experiments and dynamic causal effects: A finite population perspective. *Quantitative Economics*, 12(4):1171–1196, 2021. ISSN 1759-7323. doi: 10.3982/QE1744. URL http://qeconomics.org/ojs/index.php/qe/article/view/QE1744.

L. Bottmer, G. Imbens, J. Spiess, and M. Warnick. A Design-Based Perspective on Synthetic Control Methods, Oct. 2021. URL http://arxiv.org/abs/2101.09398. arXiv:2101.09398 [econ, stat].

S. R. Cole and M. A. Hernan. Constructing Inverse Probability Weights for Marginal Structural Models. *American Journal of Epidemiology*, 168(6):656–664, July 2008. ISSN 0002-9262, 1476-6256. doi: 10.1093/aje/kwn164. URL https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwn164.

P. Dawid, M. Musio, and S. E. Fienberg. From Statistical Evidence to Evidence of Causality. *Bayesian Analysis*, 11(3), Sept. 2016. ISSN 1936-0975. doi: 10.1214/15-BA968. URL http://arxiv.org/abs/1311.7513. arXiv:1311.7513 [math, stat].

N. Doudchenko and G. W. Imbens. Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis, Sept. 2017. URL http://arxiv.org/abs/1610.07748. arXiv:1610.07748 [stat].

R. A. Fisher. Cancer and smoking. *Nature*, 182(4635):596–596, 1958.

A. Gelman. Causality and Statistical Learning. *American Journal of Sociology*, 117(3):955–966, 2011. doi: 10.1086/662659. URL https://doi.org/10.1086/662659. _eprint: https://doi.org/10.1086/662659.

A. Gelman and G. Imbens. Why ask Why? Forward Causal Inference and Reverse Causal Questions. Technical Report w19614, National Bureau of Economic Research, Cambridge, MA, Nov. 2013. URL http://www.nber.org/papers/w19614.pdf.

G. Imbens. Causal Panel Data Models, July 2022.

G. W. Imbens. Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics, Mar. 2020. URL http://arxiv.org/abs/1907.07271. Number: arXiv:1907.07271 arXiv:1907.07271 [stat].

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

S. Khan and D. Nekipelov. On uniform inference in nonlinear models with endogeneity. *Journal of Econometrics*, page S0304407622000409, Apr. 2022. ISSN 03044076. doi: 10.1016/j.jeconom.2021.07.016. URL https://linkinghub.elsevier.com/retrieve/pii/S0304407622000409.

J. S. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Annals of Agricultural Sciences*, 10(4):1–51, 1923.

J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. ISSN 00063444. URL http://www.jstor.org/stable/2337329.

J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

J. Peters, D. Janzing, and B. Schölkopf. Causal Inference on Time Series using Structural Equation Models. *arXiv:1207.5136 [cs, stat]*, July 2012. URL http://arxiv.org/abs/1207.5136. arXiv: 1207.5136.

F. Petropoulos, D. Apiletti, V. Assimakopoulos, M. Z. Babai, D. K. Barrow, S. B. Taieb, C. Bergmeir, R. J. Bessa, J. Bijak, J. E. Boylan, J. Browell, C. Carnevale, J. L. Castle, P. Cirillo, M. P. Clements, C. Cordeiro, F. L. C. Oliveira, S. D. Baets, A. Dokumentov, J. Ellison, P. Fiszeder, P. H. Franses, D. T. Frazier, M. Gilliland, M. S. Gönül, P. Goodwin, L. Grossi, Y. Grushka-Cockayne, M. Guidolin, M. Guidolin, U. Gunter, X. Guo, R. Guseo, N. Harvey, D. F. Hendry, R. Hollyman, T. Januschowski, J. Jeon, V. R. R. Jose, Y. Kang, A. B. Koehler, S. Kolassa, N. Kourentzes, S. Leva, F. Li, K. Litsiou, S. Makridakis, G. M. Martin, A. B. Martinez, S. Meeran, T. Modis, K. Nikolopoulos, D. Önkal, A. Paccagnini, A. Panagiotelis, I. Panapakidis, J. M. Pavía, M. Pedio, D. J. Pedregal, P. Pinson, P. Ramos, D. E. Rapach, J. J. Reade, B. Rostami-Tabar, M. Rubaszek, G. Sermpinis, H. L. Shang, E. Spiliotis, A. A. Syntetos, P. D. Talagala, T. S. Talagala, L. Tashman, D. Thomakos, T. Thorarinsdottir, E. Todini, J. R. T. Arenas, X. Wang, R. L. Winkler, A. Yusupova, and F. Ziel. Forecasting: theory and practice. *International Journal of Forecasting*, 38(3):705–871, jul 2022. doi: 10.1016/j.ijforecast.2021.11.001. URL https://doi.org/10.1016%2Fj.ijforecast.2021.11.001.

J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12):1393–1512, 1986. ISSN

02700255. doi: 10.1016/0270-0255(86)90088-6. URL https://linkinghub.elsevier.com/retrieve/pii/0270025586900886.

J. M. Robins, M. A. Hernan, and B. Brumback. Marginal Structural Models and Causal Inference in Epidemiology:. *Epidemiology*, 11(5):550–560, Sept. 2000. ISSN 1044-3983. doi: 10.1097/00001648-200009000-00011. URL http://journals.lww.com/00001648-200009000-00011.

P. R. Rosenbaum. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A (General)*, 147(5):656–666, 1984.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

S. Vansteelandt and M. Joffe. Structural Nested Models and G-estimation: The Partially Realized Promise. *Statistical Science*, 29(4), Nov. 2014. ISSN 0883-4237. doi: 10.1214/14-STS493. URL http://arxiv.org/abs/1503.01589. arXiv:1503.01589 [stat].

D. Viviano and J. Bradic. Dynamic covariate balancing: estimating treatment effects over time, June 2021. URL http://arxiv.org/abs/2103.01280. Number: arXiv:2103.01280 arXiv:2103.01280 [econ, math, stat].

A Miscellaneous Proofs

A.1 Proof of Theorem 3.6

Proof. For ease of notation, write $f(\omega_{1:s}) = \mathbb{P}(Y_T = 1 | \bar{X}_{1:s} = \omega_{1:s})$ and $f(\omega) = f(\omega_{1:T})$. First consider the case where there are only two causal paths and let $T - K - 1 \geq 0$ be the time index of the lowest common ancestor of ω^* and $\widehat{\omega}$. Note that when $K = 1$, $\widehat{\omega}$ cannot be worse than ω^*, hence it suffices to consider the cases for $2 \leq K < T$. We have

$$\begin{aligned} f(\widehat{\omega}_{1:T-K}) &= \sum_{\omega' \in \mathcal{X}^K} f((\widehat{\omega}_{1:T-K}, \omega')) \mathbb{P}((\widehat{\omega}_{T-K}, \omega') | \widehat{\omega}_{1:T-K-1}) \\ &\leq (2(1-\varepsilon))^{K-1} f(\widehat{\omega}), \end{aligned} \tag{18}$$

where the factor 2 arises as in the worst case there are $\left|\mathcal{X}^{K-1}\right| = 2^{K-1}$ choices for ω'. Similarly,

$$f(\omega^*_{1:T-K}) \geq \varepsilon^{K-1} f(\omega^*) \tag{19}$$

by leaving out every path that is not ω^*. But by construction

$$f(\widehat{\omega}_{1:T-K}) \geq f(\omega^*_{1:T-K}), \tag{20}$$

hence

$$f(\widehat{\omega}) \geq \left(\frac{\varepsilon}{2(1-\varepsilon)}\right)^{K-1} f(\omega^*), \tag{21}$$

for any $K \in [T-1]$. In general when there are multiple causal paths, let ω^* be the path that has higher probability than $\widehat{\omega}$ which has the lowest common ancestor, and iteratively applying the above reasoning, we see that

$$\frac{f(\widehat{\omega})}{f(\omega^*)} \geq \prod_{K=2}^{T-1} \left(\frac{\varepsilon}{1-\varepsilon}\right)^K \geq \left(\frac{\varepsilon}{1-\varepsilon}\right)^{1+\frac{T^2}{2}}, \tag{22}$$

thus completing the proof. □

A.2 Proof of Theorem 4.1

Proposition A.1 (Formal). *Given $N = N_0 + N_1$ units among which N_0 are treated, fix $T_0 \leq T$, assume $\{u_t\}_{t=1}^{T_0}$ and $\{\lambda_t\}_{t=1}^{T_0}$ are known, writing*

$$\Lambda \coloneqq (\lambda_t)_{t=1}^{T_0} \in \mathbb{R}^{M_z \times T_0}, \quad U \coloneqq (u_t)_{t=1}^{T_0} \in \mathbb{R}^{M_x \times T_0}, \tag{23}$$

if $\Lambda\Lambda^\top$ is non-singular, $\varepsilon_{i,t}$ is independent with $\{X_{i,t}, Z_{i,t}\}$ for all i. Suppose we fit SC (or synthetic treatment) for all units whence there exists weights $\{w_{ij}\}_{i,j}^N$ such that

$$\begin{cases} \sum_{j=1}^N w_{ij} = 0, \quad w_{ii} = 1, \quad \forall i \in [N], \\ \sum_{j=1}^N w_{ij} Y_{j,s}(\bar{X}_{j,s}) = \sum_{j=1}^N w_{ij} Y_{j,T}(\bar{X}_{j,s}) = 0 \quad \forall j \in [N], s \in [T_0 - 1], \\ \sum_{j=1}^N w_{ij} \bar{X}_{j,T_0-1} = 0 \quad \forall j \in [N], \end{cases} \tag{24}$$

then

$$\widehat{\tau}^{SC}_{i,t} = \sum_{j=1}^N w_j Y_{j,t} \tag{25}$$

is a consistent estimator for $\tau_{i,t}$, and consequently,

$$\widehat{\bar{\tau}}^{SC}_t \coloneqq \frac{1}{N} \sum_{i=1}^N \widehat{\tau}^{SC}_{i,t} \tag{26}$$

is a consistent estimator for the ATE $\bar{\tau}_t$. In fact, we have

$$\mathrm{Bias}(\widehat{\bar{\tau}}^{SC}_t) \leq C_p^{1/p} \left(\frac{c^2 M_x}{\xi}\right) \rho^{1/p} \max\left\{\frac{\bar{m}_p^{1/p}}{T_0^{1-1/p}}, \frac{\bar{m}_2}{T_0^{1/2}}\right\}, \tag{27}$$

where C_p is a universal constant that depends only on p, $\rho_N \coloneqq \max\{N_0, N_1\}/N$, and $\bar{m}_p \coloneqq \max_j T_0^{-1} \sum_{j=1}^N |\varepsilon_{jt}|^p$.

Proof. First consider the case where there is only a single unit under treatment at time $T_0 \leq T$, which we assume without loss of generality to be the first unit:

$$Y^{\text{obs}}_{1,T} = Y_{1,T}(1), \tag{28}$$

then the error r_t at time $t \le T_0$ satisfies

$$\begin{aligned} r_t &\coloneqq \sum_{j=1}^{N} w_j Y_{j,t} = Y_{1,t} - \sum_{j=2}^{N} w_j Y_{j,t} \\ &= u_t^\top \sum_j w_j \bar{X}_{j,t} + \lambda_t^\top \sum_j w_j \bar{Z}_{j,t} + \sum_j w_j \varepsilon_{j,t}. \end{aligned} \tag{29}$$

Write $\boldsymbol{r} = (r_t)_{t=1}^{T_0} \in \mathbb{R}^{T_0}$, $\boldsymbol{Y}_i = (Y_{i,t})_{t=1}^{T_0} \in \mathbb{R}^{T_0}$, then

$$\bar{r}_{T_0} = \sum_{j=1}^{N} w_j \bar{Y}_{j,T_0} = U^\top \sum_j w_j \bar{X}_{j,T_0} + \Lambda^\top \sum_j w_j \bar{Z}_{j,T_0} + \sum_j w_j \bar{\varepsilon}_{j,T_0}. \tag{30}$$

Thus

$$r_t = \lambda_t^\top \Lambda' \sum_j w_j \bar{Y}_{j,T_0} + (u_t^\top - \lambda_t \Lambda' U^\top) \sum_j w_j \bar{X}_{j,T_0} + \sum_j w_j \varepsilon_{j,t} - \lambda_t^\top \Lambda' \sum_j w_j \bar{\varepsilon}_{j,T_0}, \tag{31}$$

where we write $\Lambda' \coloneqq (\Lambda\Lambda^\top)^{-1}\Lambda$. Now taking expectation with respect to the randomness in $\varepsilon_{j,T}$ and those in $\boldsymbol{Y}_j$, recall that $w_1 = 1$ is fixed, hence the only non-zero term is

$$\lambda_t \Lambda' \sum_{j \ge 2} w_j \bar{\varepsilon}_{j,T_0} = \sum_{j=1}^{N} w_j \sum_{s=1}^{T_0} \lambda_t^\top \left(\sum_{i=1}^{T_0} \lambda_i \lambda_i^\top \right)^{-1} \lambda_s \varepsilon_{j,s} \tag{32}$$

since the condition Equation (24) may introduces correlation between $\bar{\varepsilon}_{j,t}$. Write

$$\mu_s \coloneqq \lambda_t^\top \left(\sum_{i=1}^{T_0} \lambda_i \lambda_i^\top \right)^{-1} \lambda_s, \quad \varepsilon_j' \coloneqq \sum_{s=1}^{T_0} \mu_s \varepsilon_{j,s}, \tag{33}$$

we have by Cauchy-Schwarz inequality

$$|\mu_s|^2 \le \left(\lambda_t^\top \left(\sum_{i=1}^{T_0} \lambda_i \lambda_i^\top \right)^{-1} \lambda_t \right) \left(\lambda_s^\top \left(\sum_{i=1}^{T_0} \lambda_i \lambda_i^\top \right)^{-1} \lambda_s \right) \tag{34}$$

since we have assumed $\Lambda\Lambda^\top$ to be non-singular, hence so is $(\Lambda\Lambda^\top)^{-1}$. Now assume further that

$$\lambda_{\min} \left(\frac{1}{T_0} \sum_{s=1}^{T_0} \lambda_s \lambda_s^\top \right) \ge \xi > 0, \quad \forall T_0 \le T, \tag{35}$$

and $|\lambda_{sk}| \le c$ for all $s \in [T_0]$ and $k \in [M_X]$, then

$$|\mu_s|^2 \le \left(\frac{c^2 M_x}{T_0 \xi} \right)^2. \tag{36}$$

By Hölder's inequality,

$$\sum_{j=2}^{N} w_j \left| \varepsilon_j' \right| \le \left(\sum_{j=2}^{N} \left| \varepsilon_j' \right|^p \right)^{1/p}, \tag{37}$$

and

$$\mathbb{E} \sum_{j=2}^{N} w_j \left| \varepsilon_j' \right| \le \left(\mathbb{E} \sum_{j=2}^{N} \left| \varepsilon_j' \right|^p \right)^{1/p}, \tag{38}$$

where we assume the p-moment of ε_{js} exists. Applying Rosenthal's inequality yields

$$\mathbb{E} \left| \varepsilon_j' \right|^p \le C_p \left(\frac{c^2 M_x}{\xi} \right)^p \max \left\{ \frac{1}{T_0^{p-1}} \bar{m}_{p,j}, \left(\frac{1}{T_0} \bar{m}_{2,j} \right)^{p/2} \right\} \tag{39}$$

where we write

$$\bar{m}_{p,j} = \frac{1}{T_0} \sum_{s=1}^{T_0} \mathbb{E} \left| \varepsilon_{j,s} \right|^p, \tag{40}$$

and C_p is an absolute constant that only depends on p. Thus

$$\mathbb{E} \left| r_t \right| \le C_p^{1/p} \left(\frac{c^2 M_x}{\xi} \right) N^{1/p} \max \left\{ \frac{\bar{m}_p^{1/p}}{T_0^{1-1/p}}, \frac{\bar{m}_2}{T_0^{1/2}} \right\}, \tag{41}$$

where $\bar{m}_p = \max_j \bar{m}_{p,j}$.

In the general case, assume there are N_0 treated units and N_1 untreated such that $N = N_0 + N_1$. We then fit SC for each treated units and similarly synthetic treatments for each untreated units. Write $\rho_N \coloneqq \max\{N_0, N_1\}/N$, which we assume $0 < \rho_N < 1$, then we have by applying the above result,

$$\mathrm{Bias}(\widehat{\overline{\tau}}_t^{\mathrm{SC}}) \leq C_p^{1/p} \left(\frac{c^2 M_x}{\xi} \right) \rho^{1/p} \max \left\{ \frac{\bar{m}_p^{1/p}}{T_0^{1-1/p}}, \frac{\bar{m}_2}{T_0^{1/2}} \right\}. \tag{42}$$

□